

Contents

Part I: Invited Papers

M.J. BAYARRI ET AL.: Statistical Issues in the Utilization of Computer Models	3
G. CASELLA ET AL.: Consistent Variable Selection	8
N. KEIDING: Design and analysis of time-to-pregnancy	16
G. MOLENBERGHS ET AL.: The Meta-analytic Framework for the Evaluation of Surrogate Endpoints in Clinical Trials	21
D. PFEFFERMANN: Modelling of complex survey data, Why is it different and what can be done about it?	28
S. WOOD: Generalized additive smooth modelling	35

Part II: Contributed Papers

T. ADAMSKI ET AL.: A multivariate analysis of DH lines experiments repeated over a period of years	39
M. AERTS ET AL.: Direct Models for Multiple Infection Measurements of Antibody Levels	43
M. ALFÓ AND A. MARUOTTI: Semiparametric models for longitudinal binary responses with attrition	48
J. ALMANSA ET AL.: Analyzing Mental Comorbidity through LCA. Results of the ESEMeD project	52
A.M. ALONSO ET AL.: Time Series Classification based on Functional Depth	56
A.M. ALONSO ET AL.: Forecasting the Spanish mortality rates	60
A. AREIRA ET AL.: Modelling of local elections in Portugal	65
I. AROSTEGUI AND V. NÚÑEZ-ANTÓN: Alternative modelling approaches for the SF-36 health questionnaire	69

A. ASSAF AND K.M. MATAWIE: A Bayesian Approach to the Estimation of Technical Efficiency in Health Care Foodservice Operations.....	73
A. BARBER ET AL.: A Bayesian Hierarchical spatial model for the bioclimatic classification of Cyprus island	77
C. BARCELÓ-VIDAL ET AL.: Compositional Time Series: A First Approach..	81
S. BARRY AND A. BOWMAN: Modelling longitudinal spatial curve data.....	87
F. BARTOLUCCI AND M. LUPPARELLI: The multilevel latent Markov model .	93
A. BATCHELOR ET AL.: Nonlinear discrete-time hazard models for the rate of first marriage	99
M. BÉCUE ET AL.: Mixed Text and Data Mining through a principal axes method. Application to legal documents.....	103
M. BLAGOJEVIC: CTDL-Positive Stable Frailty Model.....	107
A. BLANCE ET AL.: Beyond Kappa: Use of multifaceted RASCH analysis and multilevel modelling to investigate observer effects.....	111
H. BOLFARINE: Asymmetric distributions generated by the normal distribution function.....	114
F. BOTELLA ET AL.: Spatio-Temporal Bayesian Model for studying waterbird biodiversity in artificial ponds.....	117
M.J. BREWER ET AL.: Temporal Smoothing of Compositional Data on Water Quality	121
S. BRONER AND P. DELICADO: Explaining electoral participation by an economic capacity index in Barcelona.....	126
J.A. BROWN ET AL.: Modeling long memory time series: the Shihua Cave speleothems.....	130
A. BUIL ET AL.: Mixed-Models for Genetic Linkage Analysis of Quantitative Traits: Analysis of APTT in the GAIT Project	136
R. CABALLERO-ÁGUILA ET AL.: Estimation of signals transmitted by different randomly delayed sensors using covariance information	140
R. CABALLERO-ÁGUILA ET AL.: Estimation from observations with randomly missing signals using an innovation approach.....	144
C.G. CAMARDA ET AL.: Modelling General Patterns of Digit Preference.....	148
A.I. CARITA ET AL.: Following brake reaction time in total knee arthroplasty: analysis of variance for repeated measures	154
M. CAZZARO ET AL.: Testing Markov Chain Lumpability	158

B. CERANKA AND M. GRACZYK: Note on A-optimal chemical balance weighing design.....	164
B. CERANKA AND M. GRACZYK: Optimum chemical balance weighing design for p+1 objects	168
S.J. CHUA ET AL.: Small Sample Properties of Maximum Likelihood Estimators for Type II Censored Data.....	172
S. CONDE AND G. MACKENZIE: Modelling High Dimensional Sets of Binary Co-morbidities	177
D. CONESA ET AL.: Bayesian Markov switching models for epidemiologic surveillance	181
F. CONSENTINO AND G. CLAESKENS: Model Selection With Missing Covariates Under Ignorable Missingness.....	185
C. CORDEIRO AND M.M. NEVES: Bootstrap prediction intervals: a case-study	191
M. CORREAL: A model for a system of flow rivers with non-linear behavior..	195
M.J. COSTA AND J. E. H. SHAW: Parameterization and Penalties in Spline Models.....	199
M. CRUYFF ET AL.: A ZIP model accounting for response bias in randomized response.	205
A.H.M.A. CYSNEIROS ET AL.: Modified Profile Likelihood for the Birnbaum-Saunders Distribution	211
R. DITTRICH ET AL.: On the Treatment of Missing Observations in Paired Comparisons Experiments.....	215
I.L. DRYDEN ET AL.: Factored principal components analysis and likelihood ratio based face recognition.....	221
T. ECONOMOU ET AL.: Bayesian modelling of time aggregated water pipe bursts with a zero-inflated, non-homogeneous Poisson process.....	227
P.H.C. EILERS: The Smooth Complex Logarithm Model for Quasi-Periodic Signals.....	233
P.H.C. EILERS ET AL.: Modulation Models for Seasonal Incidence Tables	239
J. EINBECK ET AL.: Smoothing, sampling, and Basu's elephants.....	245
A. ESTEVE ET AL.: Adaptive Distance-Based Classification.....	249
C. FAES ET AL.: A High-Dimensional Joint Model for Longitudinal Endpoints of Different Type.....	253
A. DE FALGUEROLLES: From Dunkirk to Barcelona with GLIM [®] . A tribute to least-squares	259

J.S. FENLON AND M.J. FADDY: Modelling and Analysis of Superparasitism Data.....	265
M. FRIENDLY AND J. FOX: Visualizing hypothesis tests in multivariate linear models.....	269
M.J. GARCÍA-LIGERO ET AL.: Image estimation from signal-dependent noise observations.....	273
R. GIRALDO ET AL.: Ordinary kriging for functional data.....	277
G. GÓMEZ AND O. JULIÀ: Inverse weighted estimators when there is double censoring.....	283
E. GONZÁLEZ-DÁVILA ET AL.: Small Area Estimation using Spanish Labour Force Survey in Canary Islands.....	287
A. GRANÉ AND H. VEIGA: Conditional Heteroscedasticity or Stochastic Volatility in Financial Risk Management?.....	291
M. GREENACRE: Diagnosing Models from Maps based on Weighted Logratio Analysis.....	295
S. GREVEN ET AL.: Likelihood ratio testing for zero variance components in linear mixed models.....	300
L. GRILLI AND C. RAMPICHINI: Endogeneity issues in mixed models.....	306
A. GUOLO: A flexible approach to measurement error correction in case-control studies.....	310
I.D. HA AND Y. LEE: On Likelihood Estimation in Semiparametric Frailty Models.....	314
K. HEINER AND J. HINDE: Generalized Linear Models for Assessing Performance.....	319
G. HELLER ET AL.: Randomly Stopped Models.....	323
A. HERMOSO-CARAZO AND J. LINARES-PÉREZ: Recursive estimation of the uncertainty probability in nonlinear systems with uncertain observations..	329
J. HOFRICHTER AND H. FRIEDL: Change Point Detection for Panel Data Models.....	333
E. HOLIAN ET AL.: Mixture-Regression Cluster Model applied to Longitudinal Microarray Experiments.....	339
A. VAN DEN HOUT AND F. E. MATTHEWS: A hidden illness-death model to estimate life expectancies.....	344
C.-H. HSU ET AL.: A Weighted Kaplan-Meier Approach for Estimation of Recurrence of Colorectal Adenomas.....	350

V. JOWAHEER AND B. SUTRADHAR: Stationary versus Non-stationary Correlation Models for Familial Longitudinal Count Data	354
Z. KACZMAREK ET AL.: Some regression methods in evaluation of genotypes in series of experiments.....	360
D. KARLIS ET AL.: Discrete valued time series models for examining weather effects in daily accident counts	364
J. KIRKBY AND I. CURRIE: Smooth models of mortality with period shocks .	370
A. KOMÁREK AND E. LESAFFRE: Generalized linear mixed model with a flexible random-effects distribution.....	376
I.KOSMIDIS: Penalized likelihood for a three-parameter Rasch Model.....	382
E. KULINSKAYA ET AL.: Cochran’s Q -test for variance stabilized effect size estimates and a random effect size model	386
P. LAMBERT AND P.H.C. EILERS: Bayesian Density Estimation from Grouped Observations	390
D.-J. LEE AND M. DURBÁN: Smoothing mixed models for overdispersed spatial count data	396
Y. LI ET AL.: Assessing Surrogacy in the Counterfactual Framework Using Bayesian Models	400
J. LYNCH AND G. MACKENZIE: Analysis of Breast Cancer Survival in Local Health Authorities	404
C. MACHADO ET AL.: An analysis of deprivation in Portugal based on Bayesian latent class models	408
G. MACKENZIE AND I.D. HA: Modelling Survival Data with Crossing Hazards	412
Y.C. MACNAB: Bayesian multivariate disease mapping and ecological models with errors-in-covariates: Mapping disability adjusted life years	417
P. MAIR AND A. ZEILEIS: Out-of-Sample Bootstrap Tests for Non-Nested Models.....	423
J.A. MARTÍN-FERNÁNDEZ ET AL.: Compositional modelling of sediment formation at the surface of Mars	427
G. MATEU-FIGUERAS ET AL.: Balances versus amalgamations in compositional data with an application in welfare research	431
I. MEJZA AND S. MEJZA: On split plot type experiments with subsamples...	437
S. MEJZA ET AL.: On a modelling environmental indexes.....	441
A. MERCATANTI: Identifiability of causal models with ignorable assignments and non-ignorable treatments	445

J. NEWELL AND J. EINBECK: A comparative study of nonparametric derivative estimators	449
J. ORBE AND V. NÚÑEZ-ANTÓN: Censored partial regression models and the study of the determinants of survival of Russian commercial banks	453
M.I. ORTEGO, AND J.J. EGOZCUE: Copulas and their extremal transformations.....	459
J. PALAREA-ALBALADEJO ET AL.: A convenient device for replacing rounded zeros in compositional data: <i>aln</i> model	463
A.L. PAPOILA AND C.S. ROCHA: Modelling Survival Data using Generalized Additive Models with Flexible Link	467
G.A. PAULA AND F.J.A. CYSNEIROS: Local Influence under Parameter Constraints	473
D. PENG AND G. MACKENZIE: On the analysis of censored reliability data .	477
R. PENMAN ET AL.: Modelling IVF Data using an Extended Continuation Ratio Random Effects Model.	481
D. PEREIRA AND J.T. MEXIA: Overview of Joint Regression Analysis.....	486
N. PEREZ-ALVAREZ ET AL.: Study of the 1 st , 2 nd and 3 rd Guided Interruption Periods in an HIV Clinical Trial	490
L.C. PÉREZ-RUÍZ AND G. ESCARELA: A Discretised-Copula-Based Transition Model for Binary Longitudinal Data	494
C. PFEIFER ET AL.: Damage detection of structures by analyzing embedded time series of vibration signals.....	500
F.Z. POLETO ET AL.: A product-multinomial framework for categorical data analysis with missing responses.....	504
N. PORTA ET AL.: Regression Modelling of Competing Risks in a Bladder Cancer Study	508
C. RIVERO AND T. VALDES: Robust estimation of linear models with grouped data and arbitrary errors with unknown scale parameter.....	514
D. RIZOPOULOS ET AL.: Joint modelling of time-to-event and longitudinal binary data with excess zeros.....	518
P.C. RODRIGUES AND J.A. BRANCO: Principal Component Analysis of Electoral Data.....	524
M.X. RODRÍGUEZ-ÁLVAREZ ET AL.: Comparing different approaches to regression analysis of Receiver Operating Characteristic curves. An application to Endocrinology data	528

J.A. SANTOS AND M. MANUELA NEVES: A Local Maximum Likelihood Estimator for Logistic Regression..... 532

C. SERRAT ET AL.: Joint Modelling of a Longitudinal Variable and a Time to Event Data: Methodological and Computational Issues 536

G.L. SILVA AND M.A. AMARAL-TURKMAN: Additive Survival Models with Shared Frailty..... 540

I. SOLIS-TRAPALA ET AL.: Statistical modelling of development of executive function in early childhood 544

K. STEFANOVA ET AL.: Spatial Modelling of Field Experiments: Sample Variogram and Enhanced Diagnostics..... 548

G. STREFTARIS ET AL.: Hierarchical and empirical Bayes estimators in the analysis of insurance claims 552

J. TELEX ET AL.: Bayesian model selection criteria: a comparative study through simulation 556

R. TOLOSANA-DELGADO ET AL.: A Bayesian alternative to Indicator Kriging 560

R. TSONAKA ET AL.: Marginalized Semi-Parametric Shared Parameter Models for Incomplete Ordinal Responses 564

M.D. UGARTE ET AL.: MSE of the log-risk predictor in a mixed Poisson model with spatial dependence..... 568

F. VAIDA AND L. LIU: Fast Implementation For Mixed Effects Models with Censored Response..... 574

O. VALERO ET AL.: Study of ewe’s milk composition using a combination of multivariate techniques and linear mixed models with random effects 580

C. VARIN AND C. CZADO: Pairwise likelihood inference in dynamic models for longitudinal ordinal outcomes 584

H. WAGNER ET AL.: Auxiliary Mixture Sampling for Non-normal data 587

R.M. WEST AND M.S. GILTHORPE: Use of functional data analysis and longitudinal latent class analysis to investigate the developmental origins of disease..... 593

P. WILSON: A Hybrid Test for Non-Nested Models 597

Author Index 603

Part I

Invited Papers

Statistical Issues in the Utilization of Computer Models

M.J. Bayarri¹, J. Berger², F. Liu², R. Paulo³ and J. Sacks⁴

¹ Department of Statistics and O.R., University of Valencia, Av. Dr. Moliner 50, Burjassot 46100 Spain

² ISDS, Duke University, Box 90251, Durham, NC, 27708-0251, USA

³ ISEG, Technical University of Lisbon, Rua do Quelhas, 6, Lisbon, 1200-781 Portugal

⁴ National Institute of Statistical Sciences, T.W. Alexander Drive, 27709, Research Triangle Park, USA

Abstract: Computer models are numerical implementations of mathematical models intended as ‘surrogates’ of reality. With the increase use of this ‘simulators’ comes a need to ‘validate’ them. Our approach to statistical validation of computer models aims to answer the question does the computer model adequately represent reality? It is based on comparison of computer model runs with field data of the process being modelled; of crucial importance is to explicitly account for simulator deficiencies, which also avoids ‘overtuning’. A Bayesian analysis is particularly suited to treating the major issues associated with the validation process: quantifying multiple sources of error and uncertainty in computer models; combining several sources of information; being able to adapt to different – but related – scenarios, and dealing with the unavoidable confounding present in the process. The methods and analyses are illustrated with a test bed dynamic stress analysis for a particular engineering system,

Keywords: Bayesian Analysis; Confounding; Model Inadequacy; Over-tuning; Validation of Computer Models.

1 Computer Models

Computer models/simulators are numerical implementations of mathematical models intending to approximately reproduce and predict real processes. The most popular ones are perhaps the computer models for weather forecast, but they are becoming ubiquitous. Simulators have been developed in all areas of science, engineering, economics, social sciences, medicine, biology, This models are increasingly used as aids in crucial decision and policy making: models for water quality in bays are used to dictate the policies of waste disposal along rivers discharging in that bay, economic models might dictate important economic policies adopted, weather predictions are used to evacuate areas menaced by hurricanes, models for volcano avalanches might determinate big areas declared out of limits for human settlements, and so on.

With the increase use of computer models, comes a need to check their adequacy or ‘validate’ them. Statistical evaluation of computer models is performed primarily by comparing model output to field data from the real process being modelled. The rationale for this *predictive approach* is simple: the only way to see if a model actually works is to see if its predictions are correct.

We regard the computer model as a function $y^M(\mathbf{x}, \mathbf{u})$, of (high dimensional) inputs $\mathbf{z} = (\mathbf{x}, \mathbf{u})$ where \mathbf{x} is a vector of controllable (known) inputs, and \mathbf{u} denotes calibration/tuning (unknown) parameters. Components of \mathbf{u} that have physical meaning, are often called ‘calibration’ parameters, whereas those that are somewhat artificial, required by the mathematical model, are referred to as ‘tuning’ parameters.

The simulators $y^M(\mathbf{x}, \mathbf{u})$ are usually numerical implementations to complex math models (defined by systems of ODE’s or PDE’s); we do not distinguish between the math model and its numerical implementation. We denote by $y^R(\mathbf{x})$ the real process that $y^M(\mathbf{x}, \mathbf{u})$ tries to simulate. As remarked earlier, \mathbf{u} is an input only to $y^M(\cdot)$ (reality occurs at the “true” unknown value). We concentrate in this exposition on *deterministic* models, that is, computer models that produce the same outputs when repeatedly run at a given set of inputs. However, the methodology can be also applied to stochastic simulators.

We concentrate here on Calibration, Validation and Prediction using computer models. Calibration of computers models is basically learning about \mathbf{u} (solving the ‘inverse problem’, estimating), but it is very important to account and transmit uncertainties. The crucial question for computer models validation is does the model adequately represent reality? A common approach is to ask – is the model correct? This is rarely useful (since we know it is not). Instead we aim to answering the question does the model provide predictions that are accurate enough for the intended use? by providing predictions *and* tolerance bounds, which accounts for uncertainties and model inadequacies. In this problems, it is *crucial* to simultaneously learn about model adequacy and uncertainty in input parameters to avoid over fitting

2 Acknowledging Model Deficiencies and Uncertainties

‘Standard’ statistical analyses are not adequate when combining computer model and field data. Since uncertainties are crucial, merely feeding the simulator with an estimate of \mathbf{u} does not work. Moreover, since the hypothesized model is incorrect, ‘over-fitting’ will typically have occurred; the fit tries to ‘make up’ for the model inadequacy by over-shifting \mathbf{u} to compensate. This over-fitting makes it problematical to believe any structure found in the residuals. Hence, it is very important to introduce (and account for) model inadequacy while calibrating the model. An extended discussion (with a pedagogical example) can be found in Bayarri et. all (2007a).

2.1 Bayesian analysis with a fast simulator

We follow work by Keneddy and O’Hagan (2001) and others and explicitly recognize that the computer model differs from reality by a bias term and not simply by random error. Specifically, we assume that the computer model is related to reality via

$$y^R(\mathbf{x}) = y^M(\mathbf{x}, \mathbf{u}^*) + b_{\mathbf{u}^*}(\mathbf{x}),$$

where \mathbf{u}^* is the true (unknown) value of \mathbf{u} , and $b_{\mathbf{u}^*}(\mathbf{x})$ is *model bias*, an unknown function of the inputs \mathbf{x} . As usual, we assume that reality is observed with error, so field data $y^F(\mathbf{x}_i)$ at inputs $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are obtained, and modelled as

$$y^F(\mathbf{x}_i) = y^R(\mathbf{x}_i) + \epsilon_i^F,$$

where ϵ_i^F are i.i.d. $\text{Normal}(0, 1/\lambda^F)$ random errors. We denote the vector of observed field data by $\mathbf{y}^F = (y^F(\mathbf{x}_1), \dots, y^F(\mathbf{x}_n))$.

We take a Bayesian approach so that uncertainties are acknowledged and transmitted. Most importantly, Bayesian analysis can also handle the lack of identifiability of \mathbf{u} and $b(\cdot)$ (and also λ^F if there were no replicates), through prior assessments. Indeed, in the computer model scenario, \mathbf{u} may have physical meaning or, at least, physical limits, so that experts may be able to construct a fairly tight prior distribution for \mathbf{u} . Moreover, we adopt a (nonparametric) prior distribution for the bias function which ‘encourages’ $b(\cdot)$ to be zero, allowing a correct computer model to emerge with little bias if supported by the data.

The Bayesian analysis puts a prior on all unknowns $\pi(\mathbf{u}, b, \lambda^F)$ and derives the posterior via Bayes Theorem as

$$\pi(\mathbf{u}, b, \lambda^F | \mathbf{y}^F) \propto f(\mathbf{y}^F | \mathbf{u}, b, \lambda^F) \times \pi(\mathbf{u}, b, \lambda^F)$$

where the likelihood function, $f(\mathbf{y}^F | \mathbf{u}, b, \lambda^F)$ is a multivariate normal with mean $([y^M(\mathbf{x}_1, \mathbf{u}) + b(\mathbf{x}_1)], \dots, [y^M(\mathbf{x}_n, \mathbf{u}) + b(\mathbf{x}_n)])$ and covariance matrix $\frac{1}{\lambda^F} \mathbf{I}$.

Metropolis-Hastings MCMC then produces simulations from the joint posterior. Notice, however, that this straight analysis requires running the simulator $y^M(\mathbf{x}, \mathbf{u})$ hundreds of thousands of times, and hence it is only feasible for extremely fast simulators. But in practice, complex computer models are very slow.

2.2 Emulating a slow simulator

We use an *objective Bayesian* ‘spatial’ response surface approach (Sacks et al., 1989) which provides a smooth interpolation/extrapolation at new inputs while estimating the accuracy of the model approximation. Statistical approximations to the simulators are called emulators and have several uses. We can also interpret them as a distributions for the simulator; indeed, since $y^M(\mathbf{z})$ is an *unknown* function (known only at few input values) we assign $y^M(\cdot)$ a prior distribution which gets revised by data. A very convenient interpolator/prior is a GASP (Gaussian Separable Process). A function has a GASP distribution

$$y(\cdot) \sim \text{GASP}(\mu(\cdot), \frac{1}{\lambda} c(\cdot, \cdot))$$

if, for any $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, $(y(\mathbf{z}_1), \dots, y(\mathbf{z}_n))$ has a multivariate normal distribution with mean vector $(\mu(\mathbf{z}_1), \dots, \mu(\mathbf{z}_n))$, and covariance matrix $[\frac{1}{\lambda} c(\mathbf{z}_i, \mathbf{z}_j)]$, where λ is the precision and $c(\cdot, \cdot)$ is the correlation function. For the computer model, we consider a *mean function* $\mu^M(\mathbf{z}) = \mathbf{\Psi}(\mathbf{z}) \boldsymbol{\theta}^L$, with specified *basis functions* $\mathbf{\Psi} = (\Psi_1, \dots, \Psi_k)$ and unknown $\boldsymbol{\theta}^L = (\theta_1^L, \dots, \theta_k^L)'$. (A constant mean θ is often satisfactory, but more general means are useful for extrapolation.) As a *correlation function* for the d -dimensional \mathbf{z} , we utilize the *separable* power exponential family.

$$c^M(\mathbf{z}, \mathbf{z}^*) = \prod_{j=1}^d \exp\left(-\beta_j^M |z_j - z_j^*|^{\alpha_j^M}\right)$$

where $\beta_j^M > 0$ determines how fast the correlation decays to 0, and $\alpha_j^M \in (0, 2]$ determines continuity, differentiability, \dots , etc. The product form greatly speeds computation and allows stochastic inputs to be handled easily. Although we were not explicit in the previous section, we also use a GASP prior for the unknown bias function, $b(\cdot) \sim \text{GASP}(\theta^b, c^b(\cdot, \cdot)/\lambda^b)$; often we assume $\theta^b = 0$ and $\alpha^b = 2$.

The posterior analysis proceeds basically as before, but whenever the analysis requires running the (slow) simulator, $y^M(\cdot)$, we treat it as an unknown (latent) variable with the GASP prior distribution.

In this case model data \mathbf{y}^M is also needed, and the likelihood $f(\mathbf{y}^F, \mathbf{y}^M | y^M(\cdot), u, b, \lambda^F)$ is still multivariate normal but with a more complicated covariance matrix.

The analysis is conceptually straightforward, but the special characteristic of these problems, with high dimensional, highly correlated parameter vector, and little data, makes the numerical analysis tricky. We have found that a very useful simplification is the *Modular Approach* which estimates the GASP parameters in $y^M(\cdot)$ based only on computer model data \mathbf{y}^M . For very complex problems, with little uncertainty in the GASP model approximation compared to uncertainties on (b, \mathbf{u}) we often use still a further simplification: we simply use MLE for GASP parameters in $y^M(\cdot)$ which are then fixed for the rest of the analysis. Other uses of modularization and its intuitive justification will be provided in the talk.

3 Calibration, Validation, Prediction.

The Metropolis-Hastings MCMC of the previous section produces N (very large) simulations $(\mathbf{u}_i, y^M(\mathbf{x}, \mathbf{u}_i), b_i(\mathbf{x}), \lambda_i^F)$ from the joint posterior. A variety of analyses are then possible, including all our previous goals.

For instance, $\mathbf{u}_1, \dots, \mathbf{u}_N$ is a sample from the marginal posterior distribution of \mathbf{u} , so, for the calibration problem, $\hat{\mathbf{u}} = \frac{1}{N} \sum_{i=1}^N \mathbf{u}_i$ provides an estimate of \mathbf{u} and $\widehat{\text{var}}(\mathbf{u}) = \frac{1}{N} \sum_{i=1}^N (\mathbf{u}_i - \hat{\mathbf{u}})^2$ is a measure of its accuracy. Indeed, the histogram of the \mathbf{u}_i represents the posterior density of \mathbf{u} .

Similarly, for the validation task, $\hat{b}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N b_i(\mathbf{x})$ is an estimate of the bias function, and the 5% and 95% percentiles yield 90% tolerance bounds.

To predict the real process at some (new) input \mathbf{x} , there are several possibilities (see Bayarri et. al, 2007b). The one we ended up recommending (if feasible) is to first run the computer model, obtaining $y^M(\mathbf{x}, \hat{\mathbf{u}})$; sometimes modelers wish to use $y^M(\mathbf{x}, \hat{\mathbf{u}})$ directly as the prediction of reality, and then the variance $V_{\hat{\mathbf{u}}}(\mathbf{x})$ is available. Better still is to estimate the bias $\hat{b}_{\hat{\mathbf{u}}}(\mathbf{x})$, and use the bias-corrected prediction $\hat{y}^R(\mathbf{x}) = y^M(\mathbf{x}, \hat{\mathbf{u}}) + \hat{b}_{\hat{\mathbf{u}}}(\mathbf{x})$, with considerable gains in precision if the bias is not negligible. The distribution of $y^R(\cdot)$ produces predictive confidence bands.

4 Functional outputs

When $y^M(\mathbf{x}, \mathbf{u})$ is a function of time t , say, we can approach the problem in several ways:

If the function is smooth, we can discretize it and include t as another model input in \mathbf{z} . This makes the analysis relatively simple, but usually produces covariance matrices

of huge dimensions, which can make this procedure unfeasible. However, considerable simplifications occur if all correlation structures involving t can be assumed to be of the same form. In this case (with common design spaces) the variance matrices take a kronecker product form, considerably reducing the dimensions of the matrices to be inverted. An application can be found in Bayarri et al. 2005.

Often, the functional outputs are complicated, rough functions of t , and alternative methods are then required. The most popular approaches use some basis expansion of the function (principal components, wavelets, ... etc.) and then apply the previous methodology to the coefficients of the expansion. This usually requires more complex, hierarchical priors for the coefficients of the bias functions, and careful modelling of the coefficients of the field error. (See Bayarri et al., 2007a for an application.)

Bayesian analysis is conceptually simple, incorporates all uncertainties, calibrates and validates simultaneously, avoids over tuning, and improves prediction, but the devil is in the details.

Acknowledgments: Special thanks to SAMSI (Statistical and Applied Mathematical Sciences Institute), and ISDS, Duke University, for their generous hosting during the 2006-2007 academic year. This work is supported in part by the Spanish Ministry of Science and Technology under Grant MTM2004-03290.

References

- Bayarri, M.J., Berger, J.O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R.J., Paulo, R., Sacks, J. and Walsh, D. (2007a). Computer Model Validation with Functional Output. *Annals of Statistics*. To appear.
- Bayarri, M.J. , Berger, J.O., Kennedy, M.C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Lin, C. H. and Tu, J. (2005). Bayesian Validation of a Computer Model for Vehicle Crashworthiness. *Tech. Rep. 163*, National Institute of Statistical Sciences.
- Bayarri, M.J. , Berger, J.O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H. and Tu, J. (2007b). A Framework for Validation of Computer Models. *Technometrics* **49**, 2 (in press).
- Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society B* **63**, 425-464.
- Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments *Statistical Science* **4**, 409-423.

Consistent Variable Selection

George Casella¹, F. Javier Girón² and Elías Moreno³

¹ Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grant DMS-04-05543, casella@stat.ufl.edu

² Professor, Department of Statistics, University of Málaga, fj_giron@uma.es

³ Professor, Department of Statistics, University of Granada, 18071, Granada, Spain. Supported by Ministerio de Ciencia y Tecnología, Grant BEC2001-2982, emoreno@ugr.es

Abstract: In choosing between two models, it is well known that the Bayes factor produces a consistent model selector (in the frequentist sense). Here we show that for intrinsic priors, the corresponding Bayesian procedure for variable selection in normal regression is consistent in the entire class of normal linear models. We also find that the asymptotics of the Bayes factors for intrinsic priors are equivalent to those of the Schwarz (BIC) criterion.

Keywords: Bayes factors, intrinsic priors, linear models, consistency.

1 Introduction

Bayesian estimation of the parameters of a given sampling model is, under wide conditions, consistent. That is, the posterior probability of the parameter is concentrated around the true value as the sample size increases, assuming that the true value belongs to the parameter space being considered. The case where the dimension of the parameter space is infinite can be an exception (see Diaconis and Friedman 1986 for examples of inconsistency of Bayesian methods).

For nested models and proper priors for the model parameters, the consistency of the Bayesian pairwise model comparison is a well established result (see O'Hagan and Forster 2004, and references therein). Assuming that we are sampling from one of the models, say M_1 , which is nested in M_2 , consistency is understood in the sense that the posterior probability of the true model tends to 1 as the sample size tends to infinity. We observe that the posterior probability is defined on the space of models $\{M_1, M_2\}$. An equivalent result is that the Bayes factor $BF_{21} = m_2(\mathbf{X}_n)/m_1(\mathbf{X}_n)$ tends in probability $[P_1]$ to zero, where $\mathbf{X}_n = (X_1, \dots, X_n)$.

The extension of this result to the case of a collection of models $\{P_i : i = 1, 2, \dots\}$ for which the condition $\lim_{n \rightarrow \infty} m_i(\mathbf{X}_n)/m_1(\mathbf{X}_n) = 0, [P_1]$, holds for any $i \geq 2$ has been established by Dawid (1992). We note that this condition is satisfied when the model P_1 is nested into any other. For nonnested models the condition does not necessarily hold.

For pairwise comparison between nested linear models the consistency of the intrinsic Bayesian procedure has already been established (Moreno and Girón 2005). The present paper is an extension of this result, and we prove here consistency of the in-

trinsic model posterior probabilities in the class of all linear models, where many of the models involved are nonnested.

2 Intrinsic Bayes Variable Selection

Suppose that Y represents an observable random variable and X_1, X_2, \dots, X_k a set of k potential explanatory covariates related through the normal linear model

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_k X_k + \varepsilon, \quad \varepsilon \sim N(\cdot | 0, \sigma^2).$$

The variable selection problem consists of reducing the complexity of this model by identifying a subset of the α_i coefficients that have a zero value based on an available dataset (\mathbf{y}, \mathbf{X}) , where \mathbf{y} is a vector of observations of size n and \mathbf{X} an $n \times k$ design matrix of full rank.

This is by nature a model selection problem where we have to choose a model among the 2^k possible submodels of the above full one. It is common to set $X_1 = 1$ and $\alpha_1 \neq 0$ to include the intercept in any model. In this case the number of possible submodels is 2^{k-1} . The class of models with i regressors will be denoted as \mathfrak{M}_i and hence the class of all possible submodels can be written as $\mathfrak{M} = \cup_i \mathfrak{M}_i$.

Consider the pairwise model comparison between a generic submodel M_j and the model

$$Y = \alpha_1 + \varepsilon, \quad \varepsilon \sim N(\cdot | 0, \sigma^2),$$

that contains the intercept only, which is denoted as M_1 . Formally, this comparison is made through the hypothesis test

$$H_0 : \text{Model } M_1 \text{ vs. } H_A : \text{Model } M_j. \quad (1)$$

Notice that M_1 is nested in M_j , for any j , so that the corresponding intrinsic priors can be derived. In the space of models $\{M_1, M_j\}$ the intrinsic posterior probability

$$P(M_j | \mathbf{y}, \mathbf{X}) = \frac{BF_{j1}}{1 + BF_{j1}}$$

is computed and it gives a new ordering of the models $\{M_j, M_j \in \mathfrak{M}\}$.

Although this procedure is based on multiple pairwise comparisons it is easy to see that it is equivalent to ordering the models according to the intrinsic model posterior probabilities computed in the space of all models \mathfrak{M} as

$$P(M_j | \mathbf{y}, \mathbf{X}) = \frac{BF_{j1}}{1 + \sum_{j' \neq 1} BF_{j'1}}, \quad M_j \in \mathfrak{M}. \quad (2)$$

This intrinsic Bayesian procedure has previously been considered by Girón *et al.* (2006a).

The intrinsic priors utilized in the variable selection methods are defined from the comparison of two nested linear models, and we now give a general expression of the intrinsic priors and the Bayes factor associated with them.

Suppose we want to choose between the following two linear models

$$M_i : \mathbf{y} = \mathbf{X}_i \alpha_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2 \mathbf{I}_n),$$

and

$$M_j : \mathbf{y} = \mathbf{X}_j \beta_j + \varepsilon_j, \varepsilon_j \sim N_n(0, \sigma_j^2 \mathbf{I}_n).$$

We again can do this formally through the hypothesis test

$$H_0 : \text{Model } M_i \text{ vs. } H_A : \text{Model } M_j, \quad (3)$$

where M_i is nested in M_j . Since the models are nested, this implies that the $n \times i$ design matrix \mathbf{X}_i is a submatrix of the $n \times j$ design matrix \mathbf{X}_j , so that $\mathbf{X}_j = (\mathbf{X}_i | \mathbf{Z}_{ij})$. Then, model M_j can be written as

$$M_j : \mathbf{y} = \mathbf{X}_i \beta_i + \mathbf{Z}_{ij} \beta_0 + \varepsilon_j, \varepsilon_j \sim N_n(0, \sigma_j^2 \mathbf{I}_n).$$

Comparing model M_i versus M_j is equivalent to testing the hypothesis $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$. A Bayesian setup for this problem is that of choosing between the Bayesian models

$$\begin{aligned} M_i & : N_n(\mathbf{y} | \mathbf{X}_i \alpha_i, \sigma_i^2 \mathbf{I}_n), \pi^N(\alpha_i, \sigma_i) = \frac{c_i}{\sigma_i}, \\ \text{and} & \\ M_j & : N_n(\mathbf{y} | \mathbf{X}_j \beta_j, \sigma_j^2 \mathbf{I}_n), \pi^N(\beta_j, \sigma_j) = \frac{c_j}{\sigma_j}, \end{aligned} \quad (4)$$

where π^N denotes the improper reference prior and c_i, c_j are arbitrary constants (Berger and Bernardo, 1992).

The direct use of improper priors for computing model posterior probabilities is not possible since they depend on the arbitrary constant c_i/c_j ; however, they can be converted into suitable intrinsic priors (Berger and Pericchi 1996). Intrinsic priors for the parameters of the above nested linear models provide a Bayes factor (Moreno *et al.* 1998), and, more importantly, posterior probabilities for the models M_i and M_j , assuming that prior probabilities are assigned to them. Here we will use an objective assessment for this model prior probability, $P(M_i) = P(M_j) = 1/2$.

Application of the standard intrinsic prior methodology yields the intrinsic prior distribution for the parameters β_j, σ_j of model M_j , conditional on a fixed parameter point α_i, σ_i of the reduced model M_i ,

$$\pi^I(\beta_j, \sigma_j | \alpha_i, \sigma_i) = \frac{2}{\pi \sigma_i (1 + \frac{\sigma_j^2}{\sigma_i^2})} N_j(\beta_j | \tilde{\alpha}_j, (\sigma_j^2 + \sigma_i^2) \mathbf{W}_j^{-1})$$

where $\tilde{\alpha}'_j = (\mathbf{0}', \alpha'_i)$ with $\mathbf{0}$ being the null vector of $j - i$ components and

$$\mathbf{W}_j^{-1} = \frac{n}{j+1} (\mathbf{X}'_j \mathbf{X}_j)^{-1}.$$

The unconditional intrinsic prior for (β_j, σ_j) is obtained from $\pi^I(\beta_j, \sigma_j) = \int \pi^I(\beta_j, \sigma_j | \alpha_i, \sigma_i) \pi^N(\alpha_i, \sigma_i) d\alpha_i d\sigma_i$, yielding the intrinsic priors for comparing models M_i and M_j as $\{\pi^N(\alpha_i, \sigma_i), \pi^I(\beta_j, \sigma_j)\}$. The computation of the Bayes factor

to compare these models using the intrinsic priors is a straightforward calculation (see Casella *et al.* 2006) and turns out to be

$$BF_{ij}^n = \left(\frac{2}{\pi} (j+1)^{(j-i)/2} \int_0^{\pi/2} \frac{\sin^{j-i} \varphi (n + (j+1) \sin^2 \varphi)^{(n-j)/2}}{(n \mathcal{B}_{ij}^n + (j+1) \sin^2 \varphi)^{(n-i)/2}} d\varphi \right)^{-1}, \quad (5)$$

where the statistics \mathcal{B}_{ij}^n is the ratio of the residual sum of squares

$$\mathcal{B}_{ij}^n = \frac{RSS_j}{RSS_i} = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{H}_j)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{H}_i)\mathbf{y}}.$$

Note that as M_i is nested in M_j the values of the statistic \mathcal{B}_{ij}^n lie in the interval $[0, 1]$.

3 Sampling distribution of \mathcal{B}_{ij}^n

If we denote the true model by M_T , so that the distribution of the vector of observations \mathbf{y} follows $N_n(\mathbf{y}|\mathbf{X}_T\alpha_T, \sigma_T^2\mathbf{I}_n)$, the sampling distribution of the statistic \mathcal{B}_{ij}^n is characterized in the following results, whose proofs are omitted.

Theorem 1 *If M_i is nested in M_j and M_T is the true model, then the sampling distribution of \mathcal{B}_{ij}^n is the doubly noncentral beta distribution*

$$\mathcal{B}_{ij}^n | M_T \sim Be'' \left(\frac{n-j}{2}, \frac{j-i}{2}; \lambda_1, \lambda_2 \right)$$

where the noncentrality parameters are

$$\lambda_1 = \frac{1}{2\sigma_T^2} \alpha_T' \mathbf{X}_T' (\mathbf{I} - \mathbf{H}_j) \mathbf{X}_T \alpha_T,$$

and

$$\lambda_2 = \frac{1}{2\sigma_T^2} \alpha_T' \mathbf{X}_T' (\mathbf{H}_j - \mathbf{H}_i) \mathbf{X}_T \alpha_T.$$

Note that the models M_i and M_j need not be nested in the true model M_T , and the true model is not necessarily nested in M_i or M_j . However, the distribution of \mathcal{B}_{ij}^n simplifies whenever M_i or M_j is the true model. Thus we have the following corollary.

Corollary 1 ,

(i) *If the smallest model M_i is the true one, then*

$$\mathcal{B}_{ij}^n | M_i \sim Be \left(\frac{n-j}{2}, \frac{j-i}{2} \right).$$

(ii) If the largest model M_j is the true one, then

$$\mathcal{B}_{ij}^n | M_j \sim Be' \left(\frac{n-j}{2}, \frac{j-i}{2}; 0, \lambda \right).$$

where

$$\lambda = \frac{1}{2\sigma_j^2} \alpha_j' \mathbf{X}_j' (\mathbf{H}_j - \mathbf{H}_i) \mathbf{X}_j \alpha_j.$$

The limiting value of \mathcal{B}_{ij}^n is important because it bears directly on the evaluation of the consistency of the Bayes factors. That value is given in the following theorem.

Theorem 2 Let $\{X_n, n \geq 1\}$ be a sequence of random variables with distribution $Be''((n - \alpha_0)/2, \beta_0/2; n\delta_1, n\delta_2)$, where $\alpha_0, \beta_0, \delta_1, \delta_2$ are positive constants. Then

(i) the sequence X_n converges in probability to the constant

$$\frac{1 + \delta_1}{1 + \delta_1 + \delta_2}.$$

(ii) If $\delta_1 = \delta_2 = 0$ then X_n degenerates in probability to 1. However, the random variable $-n/2 \log X_n$ does not degenerate and has an asymptotic Gamma distribution, $Ga(\beta_0, 1)$.

4 Consistency

The steps in proving consistency of the intrinsic Bayesian procedures are

1. Derive an asymptotic approximation for the Bayes factor for nested models given in expression (5).
2. From this approximation derive another which is valid for any arbitrary pair of models.
3. Use Theorems 1 and 2 to prove consistency.

For large n , we can get an approximation of BF_{ij}^n of (5) that is valid whenever model M_i is nested in M_j . The approximation turns out to be equivalent to the Schwarz (1978) Bayes factor approximation.

Theorem 3 When M_i is nested in M_j , for large values of n the Bayes factor given in (5) can be approximated by

$$BF_{ij}^n \approx \frac{\pi}{2} (j+1)^{(i-j)/2} I(\mathcal{B}_{ij}^n)^{-1} \exp \left(\frac{j-i}{2} \log n + \frac{n-i}{2} \log \mathcal{B}_{ij}^n \right) \quad (6)$$

where

$$\begin{aligned} I(\mathcal{B}_{ij}^n) &= \int_0^{\pi/2} \sin^{j-i}(\varphi) \exp \left[\frac{j+1}{2} \sin^2(\varphi) \left(1 - \frac{1}{\mathcal{B}_{ij}^n} \right) \right] d\varphi \\ &= \frac{1}{2} Be \left(\frac{1}{2}, \frac{j-i+1}{2} \right) {}_1F_1 \left(\frac{j-i+1}{2}; \frac{j-i+2}{2}; \frac{j+1}{2} \left(1 - \frac{1}{\mathcal{B}_{ij}^n} \right) \right), \end{aligned}$$

and ${}_1F_1(a; b; z)$ denotes the Kummer confluent hypergeometric function.

Proof: The proof follows by approximating the integrand in (5) and passing to the limit.

We note that $I(\mathcal{B}_{ij}^n)^{-1}$ has a finite value for all values of the statistic \mathcal{B}_{ij}^n except when it goes to zero. For this unrealistic case the approximation is not needed.

Therefore, BF_{ij}^n can be approximated, up to a multiplicative constant, by the exponential function in (6). This exponential function turns out to be the Schwarz approximation S_{ij}^n to the Bayes factor for comparing linear models (Schwarz 1978). Of course, the normal linear models are regular so that the Laplace approximation can be applied to obtain the Schwarz approximation although for intrinsic priors the ratio BF_{ij}^n/S_{ij}^n does not go to 1 (only for particular priors this holds; see, Kass and Wasserman 1995). However, for proving consistency we can ignore terms of constant order and the Bayes factor for intrinsic priors can be approximated by the Schwarz approximation

$$BF_{ij}^n \approx S_{ij}^n = \exp\left(\frac{j-i}{2} \log n + \frac{n}{2} \log \mathcal{B}_{ij}^n\right). \quad (7)$$

Given an arbitrary model M_j and the true model M_T in the class \mathfrak{M}_T , we will assume that the design matrix of the linear models satisfy the following condition (D): the matrix

$$\mathbf{S}_{jT} = \lim_{n \rightarrow \infty} \frac{\mathbf{X}'_T(\mathbf{I} - \mathbf{H}_j)\mathbf{X}_T}{n} \quad (8)$$

is a positive semidefinite matrix. This is not a too demanding condition.

To characterize the asymptotic behavior of the model posterior probabilities, we can work with BF_{ij}^n of (6) ignoring the positive terms that do not depend on n (as we are only interested in limiting values of 0 or ∞ .)

To test the hypothesis (1) with data (\mathbf{y}, \mathbf{X}) , we note that the intrinsic model posterior probability of model M_j , defined in the class of all models \mathfrak{M} given by (2), is an increasing function of BF_{j1} , where BF_{j1} denotes the Bayes factor for intrinsic priors for comparing the nested models M_1 versus M_j . Hence, from the asymptotic approximation (7) we can write

$$P(M_j|\mathbf{y}, \mathbf{X}) \propto BF_{j1} \approx \exp\left(-\frac{j-1}{2} \log n - \frac{n}{2} \log \mathcal{B}_{1j}^n\right). \quad (9)$$

Similarly, for the true model M_T we can write

$$P(M_T|\mathbf{y}, \mathbf{X}) \propto BF_{T1} \approx \exp\left(-\frac{T-1}{2} \log n - \frac{n}{2} \log \mathcal{B}_{1T}^n\right),$$

and consequently the ratio is approximated by

$$\frac{P(M_j|\mathbf{y}, \mathbf{X})}{P(M_T|\mathbf{y}, \mathbf{X})} \approx \exp\left(\frac{T-j}{2} \log n + \frac{n}{2} \log \frac{\mathcal{B}_{1T}^n}{\mathcal{B}_{1j}^n}\right). \quad (10)$$

(As a curiosity note that this formula provides an exact approximation to the ratio for the case when $M_j = M_T$, when its value is exactly equal to one.)

We now have the following theorem.

Theorem 4 *In the class of linear models \mathfrak{M} with design matrices satisfying condition (D), the intrinsic Bayesian variable selection procedure is consistent. That is, when sampling from M_T we have that*

$$\frac{P(M_j|\mathbf{y}, \mathbf{X})}{P(M_T|\mathbf{y}, \mathbf{X})} \rightarrow 0, [P_t],$$

whenever the model $M_j \neq M_T$.

Proof: See Casella *et al.* 2006.

5 Discussion

It has long been known that when choosing between two models, when one of which is true, selecting according to Bayes factors provides a consistent decision function in the sense that the *frequentist* probability of selecting the true model approaches 1 as $n \rightarrow \infty$. In this paper, for the case of variable selection, we have extended this result to selection among an entire class of linear models and a wide class of priors, and shown that selecting according to Bayes factors yields a decision rule with the property that the frequentist probability of selecting the true model approaches 1 as $n \rightarrow \infty$, and the frequentist probability of selecting any other model approaches 0 as $n \rightarrow \infty$. Intrinsic priors have been used successfully in both variable selection and changepoint problems (Casella and Moreno 2006, Girón *et al.* 2006ab), where excellent small sample properties were exhibited.

Lastly, we note that implementation of the model selection procedure is best done with a stochastic search algorithm. As there are 2^{k-1} possible models, enumeration quickly becomes infeasible. We have implemented Metropolis-Hastings driven stochastic searches for both variable selection (Casella and Moreno 2006) and changepoint problems (Girón *et al.* 2006b) with good results.

References

- Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method. In *Bayesian Statistics 4*, J.M. Bernardo *et al.* (eds), 35-60, London: Oxford University Press.
- Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109-122.
- Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association* **101**, 157 - 167.
- Casella, G., Girón, F. J., Martínez, M. L., Moreno, E. (2006). Consistency of Bayesian Procedures for Variable Selection. Technical Report, University of Florida. Available at <http://www.stat.ufl.edu/casella/Papers>.
- Dawid, A.P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In: *Bayesian Statistics 4*, J.M. Bernardo *et al.* (eds), 109-125, London: Oxford University Press.

- Diaconis, P. and Friedman, D. (1986). On the consistency of Bayes estimates (with discussion). *The Annals of Statistics* **14**, 1-67.
- Girón, F. J., Moreno, E. and Martínez, M. L. (2006a). An objective Bayesian procedure for variable selection in regression. In *Advances on Distribution Theory, Order Statistics and Inference*, 393–408. N. Balakrishnan *et al.* (eds), Birkhauser: Boston.
- Girón, F. J., Moreno, E. and Casella, G. (2006b). Objective Bayesian analysis of multiple changepoint models (with discussion). To appear in *Bayesian Statistics 9*, Oxford Press.
- Kass, R.E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90**, 928-934.
- Moreno, E., Bertolino, F. and Racugno, W. (1998). An Intrinsic Limiting Procedure for Model Selection and Hypothesis Testing. *Journal of the American Statistical Association* **93**, 1451-1460.
- Moreno, E. and Girón, F.J. (2005). Consistency of Bayes factors for linear models. *C.R. Acad. Sci. Paris, Ser I* **340**, 911-914.
- O'Hagan, A. and Forster, J. (2004). *Bayesian Inference*. Kendall's Advanced Theory of Statistics (Vol. 2B). London: Arnold.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

Design and analysis of time-to-pregnancy

Niels Keiding¹

¹ Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, PO Box 2099, DK-1014 Copenhagen K, Denmark

Abstract: The time from initiating attempts to become pregnant until conception occurs (*time-to-pregnancy* or TTP) is gaining importance as a measure of natural fecundity. This talk highlights some special features in modeling and design for this special application of survival analysis methods.

Keywords: survival analysis; truncation in survival analysis; cross-sectional sampling; backward recurrence time; accelerated failure time.

1 Introduction

Time-to-pregnancy (TTP), the duration that a couple waits from initiating attempts to conceive until conception occurs, is regarded as one of the direct measures of natural fecundity. TTP data can be used to study effects of environmental and occupational exposures on human reproduction. Statistical tools for designing and analysing TTP studies belong to the general area of survival analysis, with focus on intricate sampling patterns and random heterogeneity between couples. This talk shows how to perform valid analyses under various prospective, retrospective and cross-sectional sampling frames. For a recent survey see Scheike & Keiding (2006).

2 Sampling Designs

The two most common and obvious designs are a cohort (follow-up) study where couples are followed forward in time from when they start attempting to become pregnant, or a retrospective study of pregnant women where couples are interviewed about when they started their attempt to become pregnant. Variants of the cohort study are the *historically prospective* design where a general sample (usually of women) from the population is asked to recall their reproductive history, and the *prevalent cohort study*, see below, where we discuss these designs as well as the possibilities of using cross-sectional samples.

2.1 Prospective Sampling

In principle, the cohort approach leads to standard right-censored survival data, where the couples who have not conceived at the end of follow-up are counted as *right-censored*. For a careful such study see Bonde et al. (1998). In the prevalent cohort

study couples are recruited at a known time t after initiation and will have to be counted with delayed entry (left truncation) at t . In practice, prospective studies are not very common, usually rather small, and often marred by considerable self-selection problems. When assessing the effect of calendar time it is important to score it at initiation, rather than at conception, or as current calendar time along the way. The historically prospective study suffers from recall bias and also mixes experience over a long calendar time period.

2.2 Retrospective Sampling

Large TTP surveys are often retrospective, data being gathered from pregnant women. There are obvious weaknesses with these popular studies, primarily the biased sampling based on fecundity, particularly, the nonpresence of the sterile or nonfecund couples, but also under-representation of the subfecund. Juul et al. (2000) demonstrated how a true age-decreasing fecundity in a heterogeneous population can be made to look age-increasing by naïve analysis of a retrospective sample.

However, even beyond these unavoidable difficulties, the correct analysis of retrospective TTP data is more intricate than often realized, particularly when the focus is on revealing the trends in initiation intensity, which must be behind the observed secular trends in birth rates. As an example, in a common design, the data are gathered from interviews in a fixed time window. It is then clear that if calendar time is related to initiation, long TTPs will be over represented in the early phase, short TTPs in the late phase, with intricate patterns of left and right truncations (Scheike & Jensen, 1997). As pointed out by Jensen et al. (2000) dramatic artificial temporal trends in fecundity may be generated by disregarding the effects of these truncations. Incorporation of several TTPs per couple in a retrospective design is possible through careful likelihood constructions; see Scheike et al. (1999) for details.

2.3 Current duration designs

The practical difficulties in establishing representative prospective TTP cohorts and the shortcomings of the information retrievable from retrospectively obtained TTPs led Weinberg and Gladen (1986) and Keiding et al. (2002) to study a *current duration* design where a cross-sectional sample of women are asked whether they are currently attempting to get pregnant, and if so, for how long they have attempted. (As pointed out by Slama, the relevant time to pregnancy is now the time to *discovery* of the pregnancy.)

Let us summarize the distributions involved in the main sampling design as follows. For each attempt at becoming pregnant, let T be the time to pregnancy, U the time to discontinuation without pregnancy (for reasons such as death of the woman, disappearance of partner, couple give up trying; in some cases start of fertility treatment should perhaps be included) and V the time to discontinuation of follow-up since the start of attempt. We are interested in the distribution of T . In a *prospective design*, the problem reduces to standard survival analysis with T as the time to endpoint and $\min(U, V) = U \wedge V$ the time of censoring. In the *retrospective* design (based on pregnant women) we have a truncated sample from the conditional distribution of

$T|T < U$. (Note that this situation is different from right-truncation of T by U , which corresponds to observing the conditional distribution of $(T, U)|T < U$.)

In the *current-duration* design let $X = T \wedge U$ be the waiting time until termination for whatever reason, successful or not, with probability density $f(x)$, survival function $S(x) = \int_x^\infty f(a)da$ and expectation $\mu_X = \int_0^\infty xf(x)dx = \int_0^\infty S(x)dx$ which we shall assume finite. Cross-sectional sampling takes place at some fixed time t_0 , and assume that initiations happen according to a Poisson process in calendar time t with intensity $\beta(t)$. In the time-homogeneous situation $\beta(t) = \beta$, which should suffice in most situations where only short calendar intervals are considered for each “cross-section”, the observed experienced waiting time at t_0 (“current duration”), $Y = X \wedge V = T \wedge U \wedge V$ will be distributed as a backward recurrence time in a renewal process in equilibrium with renewal distribution $f(x)$, that is, the *density* of Y is

$$g(y) = S(y)/\mu_X$$

Note in particular that $0 < g(0) < \infty$. Thus, Y has a *decreasing density* proportional to the *survival function* of X .

Estimation of the density $f(x)$ of interest may be based on estimating g , either postulating some parametric distribution using beta-binomial (a beta-mixture of geometric distributions) (Weinberg & Gladen, 1986); a Pareto (a gamma mixture of exponentials) class of distributions or non-parametrically (Keiding et al., 2002). In the latter case, a technical difficulty is that the non-parametric maximum likelihood estimator $\hat{g}(0)$ of $g(0)$, necessary to obtain $\hat{S}(y) = \hat{g}(y)/\hat{g}(0)$, is inconsistent.

The current duration approach cannot distinguish between attempts ending in a pregnancy ($X = T$) and attempts terminated for other reasons ($X = U$), and neither can it catch spontaneous pregnancies (TTP=0). Furthermore, it is not yet empirically clarified how practical it is to distinguish long current durations from sterile or behaviourally non-fecund couples.

Example. The observatory of fecundity in France. A large-scale study is currently field-testing this approach (Slama et al., 2006). The statistical analysis of these data is partly based on the observation (Yamaguchi, 2003, Keiding et al., 2005, Mokveld, 2007) that accelerated failure time models for the observed current durations carry over to accelerated failure time models for the underlying times to pregnancy: if

$$P(Y > y|z) = S_0(ye^{\beta z})$$

is an accelerated failure time model for Y with underlying survival function S_0 then $g(y|z) = g_0(ye^{\beta z})e^{\beta z}$ with g_0 the density of S_0 so that

$$S(x|z) = P(X > x|z) = \frac{g(x|z)}{g(0|z)} = \frac{g_0(xe^{\beta z})}{g_0(0)}$$

which is an accelerated failure time model for X with the same β and baseline survival function $g_0(\cdot)/g_0(0)$. In particular, this requires g_0 to be decreasing with $0 < g_0(0) < \infty$.

3 Conclusion

The study of TTP is an important component of reproductive epidemiology. We have discussed here the main designs for obtaining and analysing such TTP data. All designs have both advantages and drawbacks and it is important to realize these and to understand their consequences. The prospective design is the easiest to understand conceptually and to analyse, but it is expensive to carry out a prospective study and very difficult to obtain a representative sample of couples from the population.

In the retrospective design couples are asked, at some point during pregnancy, to recall their TTP. Such data are cheap and easy to obtain in large quantities but technically difficult to analyse, because of the truncation biases that are an intrinsic consequence of the design. Another important limitation, that is not always realized sufficiently clearly, is that the data do not form a direct sample of the TTP distribution but in reality only those times to pregnancy that are not censored before being observed.

The current duration designs are new developments that aim at remedying some of the drawbacks of the prospective and retrospective designs. Notably, they should be relatively cheap to carry out, and should avoid some of the sampling biases that are feared in, for example, the prospective design. The possibility of re-using the currently trying couples in a prevalent cohort study is a promising further possibility.

References

- Bonde, J.P.E., Ernst, E., Jensen, T.K., Hjollund, N.H.I., Kolstad, H., Henriksen, T.B., Scheike, T., Givercman, A., Olsen, J. and Skakkebaek, N.E. (1998). Relation between semen quality and fertility: a population-based study of 430 first-pregnancy planners. *The Lancet* **352**, 1172-1177.
- Jensen, T.K., Keiding, N., Scheike, T., Slama, R. and Spira, A. (2000). Declining human fertility? *Fertility and Sterility* **73**, 421-422.
- Juul, S., Keiding, N. and Tvede, M. (2000). Retrospectively sampled time-to-pregnancy data may make age-decreasing fecundity look increasing. *Epidemiology* **11**, 717-719.
- Keiding, N., Kvist, K., Hartvig, H., Tvede, M. and Juul, S. (2002). Estimating time to pregnancy from current durations in a cross-sectional sample. *Biostatistics* **3**, 565-578.
- Keiding, N., Fine, J.P., Carstensen, L. and Slama, R. (2005). Accelerated failure time regression for backward recurrence times and current durations. Preprint 2005-86 Institute of Mathematical Sciences, National University of Singapore.
- Mokveld, P.J. (2007). The accelerated failure time model under cross sectional sampling schemes. Ph.D. dissertation, University of Amsterdam.
- Scheike, T.H. and Jensen, T.K. (1997). A discrete survival model with random effects: An application to time to pregnancy. *Biometrics* **53**, 318-329.

- Scheike, T.H. and Keiding, N. (2006). Design and analysis of time-to-pregnancy. *Statistical Methods in Medical Research* **15**, 127-140.
- Scheike, T.H., Petersen, J.H. and Martinussen, T. (1999). Retrospective ascertainment of recurrent events: An application to time to pregnancy. *Journal of the American Statistical Association* **94**, 713-725.
- Slama, R., Ducot, B., Carstensen, L., Lorente, C., de La Rochebrochard, E., Leridon, H., Keiding, N. and Bouyer, J. (2006). Feasibility of the current duration approach to study human fecundity. *Epidemiology* **17**, 440-449.
- Weinberg, C.S. and Gladen, B.C. (1986). The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* **42**, 547-560.
- Yamaguchi, K. (2003). Accelerated failure-time mover-stayer regression models for the analysis of last episode data. *Sociological Methodology* **33**, 81-110.

The Meta-analytic Framework for the Evaluation of Surrogate Endpoints in Clinical Trials

Geert Molenberghs¹, Tomasz Burzykowski¹, Ariel Alonso¹, Pryseley Assam¹, Abel Tilahun¹, and Marc Buyse²

¹ Hasselt University, Center for Statistics, Diepenbeek, Belgium

² International Drug Development Institute, Ottignies Louvain-la-Neuve, Belgium

Abstract: Frequently, surrogate endpoints are considered instead of the true endpoint in clinical studies. Building on the seminal work of Prentice (1989) and Freedman *et al* (1992), Buyse *et al* (2000) framed the evaluation exercise within a meta-analytic setting, in an effort to overcome difficulties that necessarily surround evaluation efforts based on a single trial. In this paper, we review the meta-analytic approach for continuous outcomes, discuss extensions to non-normal and longitudinal settings, as well as proposals to unify the somewhat disparate collection of validation measures currently on the market. Implications for design and for predicting the effect of treatment in a new trial, based on the surrogate, are discussed.

Keywords: Hierarchical model; Meta-analysis; Random-effects model; Surrogate endpoint; Surrogate threshold effect.

1 Introduction

The use of surrogate endpoints in the development of new therapies has always been very controversial, partly owing to a number of unfortunate historical instances where treatments showing a highly positive effect on a surrogate endpoints were ultimately shown to be detrimental to the subjects' clinical outcome, and conversely, some instances of treatments conferring clinical benefit without measurable impact on presumed surrogates. In spite of this, there presently is a lot of interest in surrogate endpoints, owing to the advent of a large number of biomarkers that closely reflect the disease process.

It is crucial to use *validated* surrogates, based on a process where statistics plays a role next to a variety of substantive considerations. The ICH Guidelines on Statistical Principles for Clinical Trials state that "In practice, the strength of the evidence for surrogacy depends upon (i) the biological plausibility of the relationship, (ii) the demonstration in epidemiological studies of the prognostic value of the surrogate for the clinical outcome and (iii) evidence from clinical trials that treatment effects on the surrogate correspond to effects on the clinical outcome".

2 A Meta-analytic Framework

The first formal single trial approach to validate markers is due to Prentice (1989), who gave a definition of the concept of a surrogate endpoint, followed by a series of

operational criteria. Freedman *et al* (1992) augmented Prentice’s hypothesis-testing based approach, with the estimation paradigm, through the so-called *proportion of treatment effect explained*. In turn, Buyse and Molenberghs (1998) added two further measures: the *relative effect* and the *adjusted association*. All of these proposals are hampered by their single-trial basis, lacking trial-level replication. Daniels and Hughes (1997), Buyse *et al* (2000), and Gail *et al* (2000) have introduced the meta-analytic approach, based on a hierarchical two-level model.

Let T_{ij} and S_{ij} be the random variables denoting the true and surrogate endpoints for the j th subject in the i th trial, respectively, and let Z_{ij} be the indicator variable for treatment. Consider the set of models:

$$S_{ij} = \mu_S + m_{Si} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{Sij}, \quad (1)$$

$$T_{ij} = \mu_T + m_{Ti} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{Tij}. \quad (2)$$

Here, μ_S and μ_T are fixed intercepts, α and β are fixed treatment effects, m_{Si} and m_{Ti} are random intercepts, and a_i and b_i are random treatment effects in trial i for the surrogate and true endpoints, respectively. The random effects $(m_{Si}, m_{Ti}, a_i, b_i)$ are assumed to be mean-zero normally distributed with general covariance matrix D . The error terms ε_{Sij} and ε_{Tij} follow a zero-mean normal with covariance matrix Σ .

After fitting the above models, surrogacy is captured by means of two quantities: trial-level and individual-level coefficients of determination. The former is given by: $R_{\text{trial}}^2 = R_{b_i|m_{Si}, a_i}^2$. The above quantity is unitless and, at the condition that the corresponding variance-covariance matrix is positive definite, lies within the unit interval.

Apart from estimating the strength of surrogacy, the above model can also be used for prediction purposes. To this end, observe that $(\beta + b_0|m_{S0}, a_0)$ follows a normal distribution with conditional mean $E(\beta + b_0|m_{S0}, a_0)$ and conditional variance $\text{Var}(\beta + b_0|m_{S0}, a_0)$ following from the usual expressions. A surrogate could be adopted when R_{trial}^2 is sufficiently large. Arguably, rather than using a fixed cutoff above which a surrogate would be adopted, there always will be clinical and other judgment involved in the decision process. The $R_{\text{indiv}}^2 = R_{\varepsilon_{Tij}|\varepsilon_{Si}}^2$ is based on Σ .

2.1 Ramifications

Though the above hierarchical modelling is elegant, it often poses a considerable computational challenge (Burzykowski, Molenberghs, and Buyse, 2005). To address this problem, Tibaldi *et al* (2003) suggested and studied several simplifications. They are based on a combination of the following: (1) replacing the random-effects approach with a simpler two-stage fixed-effects approach, (2) analyzing each of the endpoints separately, in a couple of univariate, rather than a single bivariate meta-analysis, and then (3) acknowledging for the differing amounts of information apported by the various trials through simpler or more sophisticated methods. Further issues, reported in Cortiñas *et al* (2004) and Tilahun *et al* (2007) are (a) the impact of the choice of the unit of analysis to define the meta-analysis, (b) the coding of treatment in the random-effects model (e.g., 0/1 versus $-1/+1$), and the occurrence of ill-conditioned D matrices, leading to unreliable estimates of the R^2 measures.

Renard *et al* (2002) have shown that extension to binary outcomes is easily done using a latent variable formulation. Using a bivariate copula approach, Burzykowski *et al* (2001) were able to deal with the situation where the outcomes are of a time-to-event type. Burzykowski *et al* (2004) discuss one of several mixed-outcomes type: an ordinal surrogate and a survival true endpoint, which is relevant, in, for example, oncology.

3 Longitudinal Endpoints

Alonso *et al* (2003) showed that going from a univariate setting to a multivariate or longitudinal framework represents new challenges. They assume that information from $i = 1, \dots, N$ trials is available, in the i th of which, $j = 1, \dots, n_i$ subjects are enrolled and they denoted the time at which subject j in trial i is measured as t_{ijk} . If T_{ijk} and S_{ijk} denote the associated true and surrogate endpoints, respectively, and Z_{ij} is a binary indicator variable for treatment then along the ideas of Galecki (1994), they proposed the following joint model, at the first stage, for both responses

$$\begin{cases} T_{ijk} = \mu_{Ti} + \beta_i Z_{ij} + g_{Tij}(t_{ijk}) + \varepsilon_{Tijk}, \\ S_{ijk} = \mu_{Si} + \alpha_i Z_{ij} + g_{Sij}(t_{ijk}) + \varepsilon_{Sijk}, \end{cases} \quad (3)$$

where μ_{Ti} and μ_{Si} are trial-specific intercepts, β_i and α_i are trial-specific effects of treatment Z_{ij} on the two endpoints and g_{Tij} and g_{Sij} are trial-subject-specific time functions that can include treatment-by-time interactions. They also assume that the vectors, collecting all information over time for patient j in trial i , $\tilde{\varepsilon}_{Tij}$ and $\tilde{\varepsilon}_{Sij}$ are correlated error terms, following a mean-zero multivariate normal distribution with covariance matrix

$$\Sigma_i = \begin{pmatrix} \Sigma_{TTi} & \Sigma_{TSi} \\ \Sigma'_{TSi} & \Sigma_{SSi} \end{pmatrix} = \begin{pmatrix} \sigma_{TTi} & \sigma_{TSi} \\ \sigma_{TSi} & \sigma_{SSi} \end{pmatrix} \otimes R_i. \quad (4)$$

Here, R_i is a correlation matrix for the repeated measurements.

If treatment effect can be assumed constant over time, then R_{trial}^2 can still be useful to evaluate surrogacy at the trial level. However, at the individual level the situation is totally different, the R_{ind}^2 no longer being applicable, and new concepts are needed.

Using multivariate ideas, Alonso *et al* (2003, 2005) proposed, to capture individual-level surrogacy: (1) the *variance reduction factor* (VRF):

$$VRF_{\text{ind}} = \frac{\sum_i \{\text{tr}(\Sigma_{TTi}) - \text{tr}(\Sigma_{(T|S)i})\}}{\sum_i \text{tr}(\Sigma_{TTi})}, \quad (5)$$

with obvious notation; (2) the quantity

$$\theta_p = \sum_i \frac{1}{Np_i} \text{tr} \{ (\Sigma_{TTi} - \Sigma_{(T|S)i}) \Sigma_{TTi}^{-1} \}, \quad (6)$$

and (3) the so-called R_{Λ}^2 :

$$R_{\Lambda}^2 = \frac{1}{N} \sum_i (1 - \Lambda_i), \quad (7)$$

where: $\Lambda_i = |\Sigma_i|/|\Sigma_{TTi}||\Sigma_{SSi}|$. All three range in the unit interval and reduce to the R_{ind}^2 in the cross-sectional case. The later proposals enjoy progressively more desirable properties than the earlier ones, and all can be embedded in uncountable families for meta-analyses. Nevertheless, they all hinge upon normality, which can be relaxed, too. To achieve generality, Alonso *et al* (2004) considered the following generalized linear models in the i th trial

$$g_T(T_{ij}) = \mu_{Ti} + \beta_i Z_{ij}, \quad (8)$$

$$g_T(T_{ij}) = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}. \quad (9)$$

The longitudinal case would be covered by considering particular functions of time in (8) and (9). Consider G_i^2 as the log-likelihood ratio test statistics to compare (8) with (9) in trial i , and quantify the association between both endpoints at the individual level using a scaled likelihood reduction factor (LRF)

$$\text{LRF} = 1 - \frac{1}{N} \sum_i \exp\left(-\frac{G_i^2}{n_i}\right). \quad (10)$$

Alonso *et al* (2004) established a number of properties for LRF, in particular its ranging in the unit interval, and its reduction to R_Λ^2 in the longitudinal and to R_{ind}^2 in the cross-sectional case. However, further generality can be achieved using an information-theoretic approach, based on the so-called entropy. The proposal avoids the needs for a joint, hierarchical model, and allows for unification across different types of endpoints. The entropy of a random variable (Shannon 1948) for a discrete r.v. with probability function $P(Y = k_i) = p_i$ is defined as:

$$H(Y) = \sum_i p_i \log\left(\frac{1}{p_i}\right). \quad (11)$$

The differential entropy $h_d(X)$ of a continuous variable X with density $f_X(x)$ and support S_{f_X} equals

$$h_d(Y) = -E[\log f_X(X)] = - \int_{S_{f_X}} f_X(x) \log f_X(x) dx. \quad (12)$$

The joint and conditional (differential) entropies are defined in an analogous fashion. Defining the information of a single event as $I(A) = \log p_A$, the entropy is $H(A) = -I(A)$. No information is gained from a totally certain event, $p_A \approx 1$, so $I(A) \approx 0$, while an improbable event is informative. $H(Y)$ is the average uncertainty associated with P .

We can now quantify the amount of uncertainty in Y , expected to be removed if the value of X were known, by $I(X, Y) = h_d(Y) - h_d(Y|X)$, the so-called *mutual information*. It is always non-negative, zero if and only if X and Y are independent, symmetric, invariant under bijective transformations of X and Y , and $I(X, X) = h_d(X)$.

Let X be a continuous n -dimensional random vector. Its entropy-power is

$$\text{EP}(X) = \frac{1}{(2\pi e)^n} e^{2h(X)}. \quad (13)$$

The differential entropy of a continuous normal random variable is $h(X) = \frac{1}{2} \log(2\pi\sigma^2)$, a simple function of the variance and, on the natural logarithmic scale: $\text{EP}(X) = \sigma^2$. We can now define an information-theoretic measure of association: $R_h^2 = \text{EP}(Y) - \text{EP}(Y|X)/\text{EP}(Y)$, which ranges in the unit interval, equals zero if and only if (X, Y) are independent, is symmetric, is invariant under bijective transformation of X and Y , and, when $R_h^2 \rightarrow 1$ for continuous models, there is usually some degeneracy appearing in the distribution of (X, Y) . In the meta-analytic case, the family of weighted combinations of the trial-specific measures produces an uncountably large family of measures.

An important ramification is Fano's inequality, showing the relationship between entropy and prediction:

$$\text{E}[(T - g(S))^2] \geq \text{EP}(T)(1 - R_h^2) \quad (14)$$

where $\text{EP}(T) = 1/(2\pi e)e^{2h(T)}$. Note that nothing has been assumed about the distribution of our responses and no specific form has been considered for the prediction function g . Also, (14) shows that the predictive quality strongly depends on the characteristics of the endpoint, specifically on its power-entropy. Fano's inequality states that the prediction error increases with $\text{EP}(T)$ and therefore, if our endpoint has a large power-entropy then a surrogate should produce a large R_h^2 to have some predictive value. This means that, for some endpoints, the search for a good surrogate can be a dead end street: the larger the entropy of T the more difficult it is to predict.

4 Prediction and Design Aspects

An important application of surrogacy evaluation is the prediction of treatment effect on the true endpoint *without measuring the latter*, supplemented with appropriate quantification of uncertainty. We will review the work done in this respect by Burzykowski and Buyse (2006).

Two components contribute to such a prediction: (a) information obtained in the validation process based on trials $i = 1, \dots, N$, and (b) the estimate of the effect of Z on S in a new trial $i = 0$ providing data on the surrogate endpoint but not on the true endpoint. We can then fit the following linear model to the surrogate outcomes S_{0j} : $S_{0j} = \mu_{s0} + \alpha_0 Z_{0j} + \varepsilon_{s0j}$. Based on this, we observe that the treatment effect on the true endpoint, $(\beta + b_0 | m_s, a_0)$, follows a normal distribution with mean linear in μ_{s0} , μ_s , α_0 , and α , and variance

$$\text{Var}(\beta + b_0 | m_{s0}, a_0) = (1 - R_{\text{trial}}^2) \text{Var}(b_0), \quad (15)$$

where m_{s0} and a_0 are the surrogate-specific random intercept and treatment effect in the new trial, respectively, and $\text{Var}(b_0)$ denotes the unconditional variance of the trial-specific random effect. Group the fixed-effects parameters and variance components into ϑ , with $\hat{\vartheta}$ the corresponding estimates. The prediction variance can then be written as:

$$\text{Var}(\beta + b_0 | \mu_{s0}, \alpha_0, \vartheta) \approx f\{\text{Var}(\hat{\mu}_{s0}, \hat{\alpha}_0)\} + f\{\text{Var}(\hat{\vartheta})\} + (1 - R_{\text{trial}}^2) \text{Var}(b_0),$$

where $f\{\text{Var}(\widehat{\mu}_{s0}, \widehat{\alpha}_0)\}$ and $f\{\text{Var}(\widehat{\vartheta})\}$ are functions of the asymptotic covariance matrices of $(\widehat{\mu}_{s0}, \widehat{\alpha}_0)^T$ and $\widehat{\vartheta}$, respectively. The third term on the right of the prediction variance, describes the prediction's variability if μ_{s0} , α_0 , and ϑ were known. The first two terms describe the contribution to the variability due to the need for estimation. It is useful to consider three scenarios. It is useful to think of three scenarios. In the first scenario, both the meta-analysis and the new trial are considered of finite size. Under scenario true, the new trial is considered infinitely large, while the meta-analysis remains finite. Under scenario 3, both are considered of infinite size. While scenarios 2 and 3 have no practical relevance, they are important to reflect about information limits.

Burzykowski and Buyse (2006) defined the minimal difference, needed to establish significance of the treatment effect on the true endpoint, using the prediction variance. Precisely, the lower bound of the corresponding confidence interval is called the *surrogate threshold effect* (STE). The larger the prediction variance, the larger the absolute value of STE. From a clinical point of view, a large STE points to the need for observing a large treatment effect on the surrogate endpoint, which may cast doubts on a surrogate's usefulness, even when its $R_{\text{trial}}^2 \simeq 1$.

Acknowledgments: We gratefully acknowledge support from Belgian Science Policy IUAP/PAI network "Statistical Techniques and Modelling for Complex Substantive Questions with Complex Data".

References

- Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal* **45**, 931–945.
- Alonso, A., Molenberghs, G., Geys, H., and Buyse, M. (2005). A unifying approach for surrogate marker validation based on Prentice's criteria. *Statistics in Medicine* **25**, 205–211.
- Alonso, A. and Molenberghs, G. (2006). Surrogate marker evaluation from an information theoretic perspective. *Biometrics* (to appear).
- Alonso, A. Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J., and Buyse, M. (2004). Prentice's approach and the meta analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics* **60**, 724–728.
- Burzykowski, T. and Buyse, M. (2006). Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharmaceutical Statistics* **5**, 173–186.
- Burzykowski, T., Molenberghs, G., and Buyse, M. (2005). *The Evaluation of Surrogate Endpoints*. New York: Springer.
- Burzykowski, T., Molenberghs, G., and Buyse, M. (2004). The validation of surrogate endpoints using data from randomized clinical trials: a case-study in advanced colorectal cancer. *Journal of the Royal Statistical Society, Series A* **167**, 103–124.

- Burzykowski, T., Molenberghs, G., Buyse, M., Renard, D., and Geys, H. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics* **50**, 405–422.
- Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029.
- Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–67.
- Cortiñas Abrahantes, J., Molenberghs, G., Burzykowski, T., Shkedy, Z., and Renard, D. (2004). Choice of units of analysis and modelling strategies in multilevel hierarchical models. *Computational Statistics and Data Analysis* **47**, 537–563.
- Daniels, M.J. and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine* **16**, 1515–1527.
- Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine* **11**, 167–178.
- Gail, M.H., Pfeiffer, R., van Houwelingen, H.C., Carroll, R.J. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246.
- Galecki, A. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics: theory and methods* **23**, 3105–3119.
- Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine* **8**, 431–440.
- Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal* **44**, 1–15.
- Shannon, C. (1948). A mathematical theory of communication, *Bell System Technical Journal* **27** 379–423 and 623–656.
- Tibaldi, F.S, Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statistical Computation and Simulation* **73**, 643–658.
- Tilahun, A., Assam, P., Alonso, A., and Molenberghs, G. (2007). Flexible surrogate marker evaluation from several randomized clinical trials with continuous endpoints, using R and SAS. *Computational Statistics and Data Analysis* (to appear).
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Modelling of complex survey data, Why is it different and what can be done about it?

Danny Pfeffermann¹

¹ Hebrew University and University of Southampton

Statistical models are increasingly being fitted to survey data. There are three main reasons for the use of models for inference from sample surveys:

1- Models are used in cases where the randomization distribution over repeated sampling from a fixed population is not operational or is inefficient. Examples are prediction problems, such as the prediction of a time series or of a small area mean, for an area with no sample units. Another example is the modelling of measurement errors in values of the target outcome variable, where again no randomization-based theory exists. A direct survey estimator of an area mean, which is based on a very small sample in the area, is an example where a randomization-based theory exists, but is very inefficient, requiring instead, the use of a model that allows borrowing information across the areas or over time.

2- Models are often used to aid in the specification of the sampling design and the choice of the corresponding estimators of the finite population quantities of interest. However, the inference process (estimation of population quantities of interest) is, in many cases, based on the randomization distribution, rather than on the model assumptions. This is known in the sampling literature as “model assisted inference” (to distinguish from “model based” inference that is based entirely on the model). Stratified sampling and the classical stratified estimator of the population mean are essentially driven by the one-way analysis of variance model with unequal variances. Poststratification into weighting cells in order to adjust for nonresponse is another example of the implicit use of a model before or after the sample is drawn. The ratio estimator, the regression estimator, and the generalized regression estimator (Sarndal, 1980) of finite population means or totals are motivated by appropriate regression models, but their bias, variance and consistency are evaluated under the randomization distribution.

3- The third reason for the use of models in sample survey inference is the same as in other statistical applications, i.e., the identification and estimation of structural relationships between variables, prediction of unobserved values etc., with no direct reference to the prediction of finite population quantities like means or totals. Examples include the estimation of elasticities of demand from household expenditure surveys, assessing the effects of variables explaining the prevalence of a disease using data from health surveys, assessment and comparison of students’ proficiencies from educational surveys, etc.

What makes survey data different from other data sets? There are four important features characterizing survey data. The first and most unique feature is that the sample is often drawn with unequal probabilities, at least at some stage of the sampling process. The selection probabilities are generally known and accessible, at least for the sampled units, in the form of sampling weights (inverse of the sampling probabilities), and they are used in the modelling process, either for randomization-based inference, by weighting the sample data, or for model-based inference (see below). Sampling with probabilities proportional to a size variable that is related to the target outcome variable is an example of this kind of sampling. Another simple example is two-stage cluster sampling, under which the clusters are sampled with probabilities proportional to the cluster sizes and the units within the selected clusters are sampled with equal probabilities. When the selection probabilities are correlated with the outcome values, even after conditioning on the model covariates, the sample model is different from the model holding in the population and the sampling process needs to be accounted for during the modelling process - see below.

The second feature of survey data is that the individual outcomes are usually correlated because of the use of clustered samples. There is nothing unique in modelling correlated continuous data, but survey data are often discrete and presented in large contingency tables of high dimension. Testing goodness of fit, or testing independence or conditional independence of correlated categorical data is not trivial, and the standard Chi-square statistics, for example, no longer have the nominal asymptotic Chi-square distribution. The third feature is that survey data are often, or almost always subject to nonresponse, with nonresponse rates of 20%-30% not being abnormal. When the nonresponse is “missing at random” (MAR), it can be ignored in the inference process. However, when modelling variables like income, the MAR assumption usually does not hold, requiring instead the use of models that account for the nonresponse. These models cannot be easily validated from the available data - see also below.

The last feature of survey data, which is also quite unique, is that because of confidentiality restrictions, the data available to the modeler is often masked in some clever way, and may actually not be the correct data, because of data swapping or other methods that are used to protect the anonymity of the respondents. This problem features, in particular, when analyzing high dimensional contingency tables with very small sample sizes in some of the cells. In what follows, I distinguish between the model that holds for the population data, hereafter the *population model*, and the *sample model*, which holds for the sample data. As mentioned before, the two models can be quite different because of the sample selection process and nonresponse. Theoretically, the sampling and response effects can be accounted for by including among the model covariates all the variables and interactions that are related to the outcome values and might affect the sample selection and response probabilities. However, this paradigm is often not practical because there may be too many variables to include in the model and some, or all of them, may not be known or accessible to the modeler. The theoretical and empirical tasks of fitting and validating such models seem formidable for many surveys. Notice also that by including these variables among the model covariates, the resulting model may no longer be of scientific interest. This may require integrating them out of the model at a later stage, which again can be very complicated and not always feasible.

Next I consider an alternative approach of modelling survey data. This approach is quite general and it accounts for the possible bias resulting from the sample selection (known in the sampling literature as informative sampling). The idea is to fit a model to the sample data and base the inference on the sample model. Denote the population model by $f_p(y|x)$, where y is the outcome variable and x is a set of covariates. Following Pfeffermann et al. (1998), the sample model is defined as,

$$\begin{aligned} f_s(y_i|x_i) &\stackrel{\text{def}}{=} f(y_i|x_i, i \in s) = \frac{\Pr(i \in s|y_i, x_i)f_p(y_i|x_i)}{\Pr(i \in s|x_i)} = \\ &= \frac{E_p(\pi_i|y_i, x_i)f_p(y_i|x_i)}{E_p(\pi_i|x_i)}, \end{aligned} \quad (1)$$

where $\pi_i = \Pr(i \in s)$ is the sample inclusion probability (probability to be selected to the sample and respond).

Remark 1. By (1), the sample model is the same as the population model if $\Pr(i \in s|y_i, x_i) = \Pr(i \in s|x_i) \forall y_i$, in which case the sampling process is ignorable

Remark 2. $\Pr(i \in s|y_i, x_i)$ is generally not the same as π_i , which may depend on all the population values $\{y_i, x_i\}$, $i \in U$, and possibly also on design variables used for the sample selection and latent variables underlying the response process. However, the use of the sample model only requires modelling $\Pr(i \in s|y_i, x_i)$ or $E_p(\pi_i|y_i, x_i)$. We may distinguish between the sample selection and the response by denoting the original sample before nonresponse by \tilde{s} , and defining a response indicator variable R , taking the value $R_i = 1$ if sample unit i responds;

$$f(y_i|x_i \in \tilde{s}, R_i=1) = \frac{\Pr(R_i = 1|y_i, x_i, i \in \tilde{s})\Pr(i \in \tilde{s}|y_i, x_i)f_p(y_i|x_i)}{\Pr(R_i = 1|x_i, i \in \tilde{s})\Pr(i \in \tilde{s}|x_i)}. \quad (2)$$

The following relationship between the population model and the sample model in (1) is established in Pfeffermann and Sverchkov (1999), where $w_i = 1/\pi_i$ and $E_s(\bullet)$ is the expectation under the sample model.

$$f_p(y_i|x_i) = \frac{E_s(w_i|y_i, x_i)f_s(y_i|x_i)}{E_s(w_i|x_i)}. \quad (3)$$

Thus, one can identify and estimate the population model by fitting the sample model $f_s(y_i|x_i)$ to the sample data and estimating the expectation $E_s(w_i|y_i, x_i)$, again using the sample data. Note, however, that where as the sample selection probabilities $\pi_{\tilde{s}i} = \Pr(i \in \tilde{s})$ are generally known and can be used for modelling $\Pr(i \in \tilde{s}|y_i, x_i)$, (using the relationship, $\Pr(i \in \tilde{s}|y_i, x_i) = E_p(\pi_{\tilde{s}i}|y_i, x_i) = 1/E_s(\pi_{\tilde{s}i}^{-1}|y_i, x_i)$ established in Pfeffermann and Sverchkov 1999), the probability $\Pr(R_i = 1|y_i, x_i, i \in \tilde{s})$ is generally unknown and needs to be modelled. However, the goodness of fit of the resulting sample model in (3) can be tested using standard test statistics since it refers to the sample data. In the rest of this paper I assume full response such that $s = \tilde{s}$, $\pi_i = \pi_i$ and $f_s(y_i|x_i)$ in (1) defines the model for units selected to the original sample.

Clearly, both the sample model and the expectations $E_s(w_i|y_i, x_i)$ depend in general on unknown parameters that need to be estimated. Note in this respect that if the outcomes are independent under the population model, they are also ‘‘asymptotically independent’’ under the sample model when increasing the population size but

holding the sample size fixed. See Pfeffermann et al. (1998) for details. Pfeffermann and Sverchkov (2003) discuss alternative approaches of estimating the model parameters. Denote the population by U and the population model by $f_p(y|x;\theta)$, where $\theta = (\theta_0, \theta_1, \dots, \theta_k)'$ represents the model parameters. The vector θ can be defined as the unique solution of the equations,

$$W_p(\hat{a}) = \sum_{i \in U} E_p[d_{pi}|x_i] = 0, \quad (4)$$

where $d_{pi} = (d_{pi,0}, d_{pi,1}, \dots, d_{pi,k})' = \partial \log f_p(y_i|x_i; \theta) / \partial \theta$ is the i^{th} score function. One possible approach of estimating θ is to redefine (4) with respect to the sample distribution. Assuming that the conditional expectations $E_p(\pi_i|y_i, x_i)$ are known or have been estimated and that the expectations $E_p(\pi_i|x_i) = \int_y E_p(\pi_i|y, x_i) f_p(y|x_i; \theta) dy$ are differentiable with respect to θ , the parameter equations are,

$$\begin{aligned} W_{1,s}(\theta) &= \sum_{i \in s} E_s \{ [\partial \log f_s(y_i|x_i; \theta) / \partial \theta] |x_i \} = \\ &= \sum_{i \in s} E_s \{ [d_{pi} - \partial \log E_p(\pi_i|x_i) / \partial \theta] |x_i \} = 0. \end{aligned} \quad (5)$$

The vector parameter θ is estimated under this approach by solving the equations,

$$W_{1s,e}(\theta) = \sum_{i \in s} [d_{pi} - \partial \log E_p(\pi_i|x_i) / \partial \theta] = 0. \quad (6)$$

Note that (6) defines the sample likelihood equations.

A second approach uses a relationship established in Pfeffermann and Sverchkov (1999), that for pairs of random variables (u, v) with population values $\{u_i, v_i; i \in U\}$,

$$E_p(u_i|v_i) = E_s(w_i u_i | v_i) / E_s(w_i | v_i). \quad (7)$$

Assuming a random sample of size n from the sample distribution and applying the relationship (7) to (4) yields the parameter equations,

$$W_{2s}(\theta) = \sum_{i \in s} E_s(q_i d_{pi} | x_i) = 0, \quad (8)$$

where $q_i = w_i / E_s(w_i | x_i)$. The vector θ is estimated under this approach by solving,

$$W_{2s,e}(\theta) = \sum_{i \in s} q_i d_{pi} = 0. \quad (9)$$

A third approach uses the property that if θ solves the equations (4), it solves also the equations $\widetilde{W}_p(\theta) = \sum_{i \in U} E_p(d_{pi}) = E_x \left[\sum_{i \in U} E_p(d_{pi} | x_i) \right] = 0$. Application of (7) to each of the terms $E_p(d_{pi})$ (without conditioning on x_i) yields the following parameter equations for a random sample of size n from the sample distribution,

$$W_{3s}(\beta) = \sum_{i \in s} E_s(w_i d_{pi}) / E_s(w_i) = 0. \quad (10)$$

The corresponding estimating equations are,

$$W_{3s,e}(\beta) = \sum_{i \in s} w_i d_{pi} = 0. \quad (11)$$

See Pfeffermann and Sverchkov (2003) for appropriate variance estimators under each of the three approaches.

Remark 3. The equations (11) coincide with the familiar pseudo-likelihood equations obtained by estimating the “census” likelihood equations, (the equations that would be obtained if all the population values were observed), by the Horvitz-Thompson estimator. For the concept and uses of the pseudo-likelihood approach, see the discussion and references in Pfeffermann, 1993. Comparing (9) with (11) shows that (9) uses the adjusted weights, $q_i = w_i/E_s(w_i|x_i)$ instead of the standard weights w_i used in (11). The weights q_i account for the net sampling effects on the target conditional distribution of $y_i|x_i$, whereas the weights w_i account also unnecessarily for the sampling effects on the marginal distribution of x_i , resulting therefore in higher variability.

Example. Suppose that the population model is,

$$y_i = x_i' \beta + \varepsilon_i, \quad E_p(\varepsilon_i|x_i) = 0, \quad E_p(\varepsilon_i^2|x_i) = \sigma_\varepsilon^2.$$

Application of (9) yields the estimator, $\tilde{\beta}_q = \left[\sum_{i \in s} q_i x_i x_i' \right]^{-1} \sum_{i \in s} q_i x_i y_i$. Application of (11) yields the estimator, $-\tilde{\beta}_q = \left[\sum_{i \in s} w_i x_i x_i' \right]^{-1} \sum_{i \in s} w_i x_i y_i$, which is also the classical probability weighted estimator.

Remark 4. Instead of basing the likelihood on the sample distribution, one could base it on the joint distribution of the sample data and the sample membership indicators,

$$f(s, y_s | x_s, x_{\bar{s}}) = \prod_{i \in s} \Pr(i \in s | y_i, x_i) f_p(y_i | x_i) \prod_{j \notin s} [1 - \Pr(j \in s | x_j)], \quad (12)$$

where $\Pr(i \in s) = \int \Pr(i \in s | y_i, x_i) f_p(y_i | x_i) dy_i$; see, e.g. Gelman et al. (2003), Pfeffermann and Sverchkov (2003) and Little (2004). The use of (12) has the theoretical advantage of employing the information on the sample selection probabilities for units outside the sample, but it requires knowledge of the covariates for every unit in the population, unlike the use of the approaches mentioned above. Modelling the joint distribution of the covariates and integrating them out of the likelihood is complicated and seems formidable with high dimensional covariates.

So far I considered model estimation but the sample distribution enables also to predict the missing population values. For this we need to define the *sample-complement* model,

$$\begin{aligned} f_c(y_i | x_i) &\stackrel{\text{def}}{=} f(y_i | x_i, i \notin s) = \frac{\Pr(i \notin s | y_i, x_i) f_p(y_i | x_i)}{\Pr(i \notin s | x_i)} = \\ &= \dots = \frac{E_s[(w_i - 1) | y_i, x_i] f_s(y_i | x_i)}{E_s[(w_i - 1) | x_i]} \end{aligned} \quad (13)$$

with the last equation established in Sverchkov and Pfeffermann (2004). Note that the sample-complement model is again a function of the sample model $f_s(y_i|x_i)$, and thus can be estimated from the sample data. The optimal predictor of the population total under a quadratic loss function is,

$$\begin{aligned} \hat{Y} &= \sum_{i \in s} y_i + \sum_{j \notin s} E(y_i|j \notin s) = \sum_{i \in s} y_i + \sum_{j \notin s} E_c(y_j|x_j) = \\ &= \sum_{i \in s} y_i + \sum_{j \notin s} \frac{E_s[(w_j - 1)y_j|x_j]}{E_s[(w_j - 1)|x_j]}. \end{aligned} \quad (14)$$

The second equality follows from (13), with the sample expectations in the numerator and the denominator either being modelled based on the sample data or simply estimated by the corresponding sample means (application of the method of moments). As shown in Sverchkov and Pfeffermann (2004), familiar estimators of finite population means such as the estimator $\bar{y}_w = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$ or the generalized regression estimator (Sarndal, 1980) are obtained as special cases of this theory by specifying appropriate population or sample models.

Remark 5. When having to predict the outcome values for a specific nonsampled unit say, a unit with a given set of covariates, or the mean of a given nonsampled area in a small area estimation problem, and the sampling process is informative, there seems to be no alternative but to model and estimate the sample complement distribution. Classical randomization based inference is suited for estimating population quantities of the population from which the sample is drawn, but not for prediction problems.

During my presentation I shall show an empirical example of the use of the sample distribution for fitting multi-level models with application to small area prediction.

References

- Gelman, A., Carling, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.
- Little, R.J. (2004). To model or not to model? competing modes of inference for finite population sampling, *Journal of the American Statistical Association* **99**, 546-556.
- Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. *International statistical review* **61**, 317-337.
- Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistics Sinica* **8**, 1087-1114.
- Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya, Series B* **61**, 166-186.

- Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In, *Analysis of survey Data*, Eds. C. Skinner and R. Chambers, New York: Wiley, 175-195.
- Sarndal, C.E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67**, 639-650.
- Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology* **30**, 79-92.

Generalized additive smooth modelling

Simon N. Wood¹

¹ University of Bath, UK

Abstract: Generalized additive models (GAMs) are GLMs in which the linear predictor is specified in terms of a sum of smooth functions of predictor variables. They provide an appealing mix of flexibility and structure for approaching many regression modelling problems. In recent years the trend to represent the smooth functions in GAMs using penalized regression splines has enhanced this appeal by allowing the development of effective and practical means for smoothness selection with such models. Illustrated with examples from ecology and epidemiology, this talk will review the penalized regression spline approach to GAMs, showing how a wide range of practically useful model structures can be constructed. The major estimation and inference problems that arise when using such flexible models will also be illustrated, and the key ideas for resolving them will be presented. The methods discussed are implemented in R package mgcv.

Part II

Contributed Papers

A multivariate analysis of DH lines experiments repeated over a period of years

Tadeusz Adamski¹, Maria Surma¹, Zygmunt Kaczmarek¹

¹ Institute of Plant Genetics, Polish Academy of Sciences, 60-479 Poznań, ul. Strzeszyńska 34, Poland, tada@igr.poznan.pl

Abstract: A statistical model for the series of trials repeated with a number of genotypes at one location over a period of years was described. The model involves the use of analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) with some multivariate data-analytic methods, in particular such as the canonical variate analysis, associated with relevant graphical techniques (Caliński et al., 1997). The paper contains the applicability of this methodology to real data resulting from a series of trials with barley doubled haploid (*DH*) lines. Thirty genotypes were studied in a series of field experiments repeated in five successive years. Malt characters such as protein content, extract yield and Kolbach index were measured. Simultaneous estimation of malting quality characters and stability of barley *DH* lines was performed. Some classification of genotypes was made with the aim to find which of the doubled haploids were the best taking into account all three studied malt characters.

Keywords: multivariate methods; *GE* interaction; barley; doubled haploids; malting quality.

1 Introduction

In the past years, techniques allowing for haploidization of hybrids and production of doubled haploid (*DH*) lines have been applied in many breeding programs (Pickering, Devaux 1992). These lines are completely homozygous and homogenous. It is interesting whether these properties of *DH* lines affect their response to varied environmental conditions for agronomical traits, especially malting quality. The statistical analysis of data from such repeated experiments involve the use of MANOVA and other multivariate methods for testing the hypotheses interesting with regard to genotype classification. In particular the analysis of doubled haploids derived from parents with dissimilar malt parameters permits to asses stability of individual lines. Therefore, the principal aim of this work was the classification of barley doubled haploid lines with regard to malt characters examined in a series of experiments. Because the malt traits are closely connected, statistical analysis for all traits simultaneously, was also performed.

2 The data

Twenty eight doubled haploid (*DH*) lines of *Hordeum vulgare L.* and two parental genotypes Maresi and Klimek were used in this study. *DH* lines were derived from

F1 hybrids of the 2-rowed malting cultivar Maresi and 6-rowed non-malting cultivar Klimek by the *H. bulbosum* method, using the standard crossing procedure of *H. vulgare* with *bulbosum* followed by in vitro culture of immature embryos. Thirty genotypes (fifteen 6-rowed *DH* lines, S1-S15, thirteen 2-rowed *DH* lines, D1-D13, and two parental forms) were studied in one location over five successive years (2001-2005). Experiments were carried out in a randomized complete block design with three replications. The following malt traits were measured: total protein content, Kolbach index and fine extract yield. The malt analyses were performed using standard methods according to European Brewery Convention.

3 Statistical methods

Taking into account the aim of the study, the data were analyzed in two stages. At the first stage (more general) all malt traits were simultaneously tested by the multivariate analysis of variance. The rejection of the multidimensional hypothesis allowed us to examine the additional information concerning the differences between *DH* lines for the individual malting quality traits.

At the second stage, the model and the methods for the analysis of genotype environment (*GE*) interaction and their structure given by Caliński et al. (1997) were used.

Suppose that in a series of trials each experiment contains the same number of genotypes, and each is laid out in the same block design, which may differ only with regard to the randomizations. Further, assume that the observed value of given trait of genotype i ($i = 1, 2, \dots, I$) in block b ($b = 1, 2, \dots, B$) of the experiment conducted at one locality in year k ($k = 1, 2, \dots, K$) can be described by the model

$$y_{ikb} = m_i(k, b) + e_{ikb},$$

where y_{ikb} is the observed value of the trait, $m_i(k, b)$ is the "true" value of the trait and e_{ikb} is the relevant experimental error. Under the usual assumption of the genotype-block additivity, the "true" value of the trait can be expressed as

$$m_i(k, b) = m_i(k) + m_b(k),$$

where $m_i(k)$ denotes the "true" mean value of the trait of genotype i in the experiment conducted in year k , and $m_b(k)$ denotes corresponding deviation summing up to zero over all B blocks. From the analysis of individual experiments the observed means of the trait are obtainable for all the genotypes. Their model can be written as

$$y_{ik.} = m_i(k) + e_{ik.},$$

where $e_{ik.}$ is the mean random error from B replications. Further considerations (see Kaczmarek, 1986) lead to the model

$$y_{ik.} = \mu_i + \alpha_i^E(k) + e_{ik.}, \quad (1)$$

valid from the point of view of the analysis of the series of experiments, where the environmental conditions of the experiments are considered as random. Due to this

assumption, μ_i represents the average "true" value of the trait of genotype i taken over period represented by years of experiments, while $a_i^E(k)$ represents the potential of genotype i to the particular environmental conditions of the experiment in year k . The former is considered as a fixed parameter, the latter as a random variable. In matrix notation model (1) can be written as

$$\mathbf{y}_k = \boldsymbol{\mu} + \mathbf{a}^E(k) + \mathbf{e}_k, \quad (2)$$

To apply the MANOVA methods to (2), it is assumed that the random vectors \mathbf{y}_k are all independent, each of an I -variate normal distribution, of the form

$$\mathbf{y}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y)$$

with the vector of expected values $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_y$.

4 Analysis of the data

For comparing barley *DH* lines with respect to all the malt traits simultaneously, the multivariate analysis of variance was applied. The general hypothesis about the equality of all genotypes was rejected at $P = 0.1\%$ ($F_{0.01} = 1.41$). It means that doubled haploids differed in malt quality. The hypotheses implied by the general hypothesis from MANOVA enabled also to test the differences between the genotypes in their main effects. The results of testing the hypotheses concerning the main effects of individual *DH* lines for each trait separately allowed to find the genotypes with significant and positive estimates of main effects for all the studied malt characters. The analysis of *GE* interaction for each trait separately enabled to find the estimates of main effects and the results of testing the hypotheses concerning *DH* lines and their interaction with environments (years) for individual traits. On the basis of these results the classification of *DH* lines in respect to their main effects and their stability was done and genotypes which had high or at least average main effect for each malt trait were distinguished (Table 1).

TABLE 1. Genotypes selected on the basis of main effects for each malt trait (h – high main effect, a – average main effect, * – genotype stable)

Genotype	Protein content	Extract yield	Kolbach index
Maresi	[h,	h,	h]
D2	[a,	h*,	h]
D5	[h*,	a,	a]
S12	[a,	h*,	a]
D12	[a*,	h,	a]
S14	[a,	h,	a]
D9	[a,	a,	a]

5 Conclusion

Classification of *DH* lines with regard to malting quality index allowed to find the group of the best genotypes. Apart from malting cultivar Maresi, one 6-rowed (S12) and two 2-rowed (D2 and D5) *DH* lines were distinguished as malting genotypes.

References

- Caliński T., Czajka S. and Kaczmarek Z. (1997). A multivariate approach to analyzing genotype-environment interactions. In: *Advances in Biometrical Genetics*. Proceedings of the Tenth Meeting of the EUCARPIA Section Biometrics in Plant Breeding, Poznań 14–16 May 1997, P. Krajewski and Z. Kaczmarek (eds.), 3–14.
- Caliński T., Czajka S., Kaczmarek Z., Krajewski P. and Siatkowski I. (1998). *SERGEN – 3 Statistical methodology and usage of the program SERGEN* (Version 3 for Windows 95), IGR PAN, Poznań.
- Kaczmarek Z. (1986). The analysis of a series of experiments in incomplete block designs (in Polish) *Roczniki Akademii Rolniczej w Poznaniu, Rozprawy Naukowe, Zeszyt 155*.
- Pickering R. A. and Devaux P. (1992). Haploid production: approaches and use in plant breeding. In: *Barley: Genetics, Biochemistry, Molecular Biology and Biotechnology* (P.R. Shewry, ed.), 519–547. CAB Int., Wallingford.

Direct Models for Multiple Infection Measurements of Antibody Levels

Marc Aerts¹, Kaatje Bollaerts¹, Niel Hens¹, Zip Shkedy¹, Christen Faes¹, Pierre Van Damme² and Philippe Beutels²

¹ Center for Statistics, Hasselt University, Belgium

² Center for the Evaluation of Vaccination, Antwerp University, Belgium, Corresponding author: marc.aerts@uhasselt.be

Abstract: Examining humans or animals for infectious diseases is often done by assessing the level of disease-specific antibodies in serum samples. In this paper we propose a new method to estimate an age-dependent disease prevalence and force of infection, directly from antibody levels. The use of one (or two) threshold values in order to diagnose each individual subject as being infected or as susceptible for a specific disease (or in some cases as equivocal), which is always prone to false positives, false negative or inconclusive classifications, is not needed for this approach. The method is based on an underlying age-dependent mixture model and can be extended to the joint analysis of two or more diseases with similar transmission routes. Such joint analyses lead to new insights in the dynamics of related diseases.

Keywords: Infectious Diseases; Force of Infection; Mixture Model; Prevalence; Penalized Splines.

1 Introduction

Mathematical compartmental models are often used to describe the process of infectious diseases at population level (Anderson and May 1991). We assume that immunity inferred by infection is lifelong, and that mortality caused by infection is negligible. In a steady state, the equation

$$dq(a)/da = -\ell(a)q(a), \quad (1)$$

describes the change in the susceptible fraction $q(a)$ with host age a . Here $\ell(a)$ denotes the force of infection (FOI), i.e. the instantaneous rate at which susceptible individuals become infected.

Typically the FOI can be estimated from serological data. A subject of age a at the time of the test is considered to be infected by (and probably recovered from) the disease of interest (denoted by $y = 1$) if the subject's (log of the) antibody level z exceeds a threshold value ζ , and is considered as still susceptible (denoted by $y = 0$) if z does not exceed that threshold value ζ . The seroprevalence is given by

$$\tilde{\pi}(a) = P(y = 1|a) = P(z > \zeta|a). \quad (2)$$

Noting that $\tilde{q}(a)$ in equation (3) corresponds to $P(y = 0|a) = 1 - \tilde{\pi}(a)$, the force of infection $\ell(a)$ can be estimated using the identity

$$\ell(a) = \frac{\tilde{\pi}'(a)}{1 - \tilde{\pi}(a)}, \quad (3)$$

where $\tilde{\pi}'(a)$ denotes the derivative of $\tilde{\pi}(a)$.

Several flexible methods (parametric and nonparametric) to estimate $\ell(a)$ via $\tilde{\pi}(a)$ have recently been proposed by Shkedy *et al.*, (2003), Shkedy *et al.*, (2006), Namata *et al.*, (2007).

The seroprevalence $\tilde{\pi}(a)$ however is not identical to the true disease prevalence $\pi(a) = 1 - q(a)$. The choice of ζ is crucial, determining the sensitivity and specificity of the serological test. In the next section we introduce a new method, which allows modelling of the disease prevalence and FOI, directly from antibody levels.

For feasibility and financial reasons, sera are often tested for more than one antigen. Studying diseases with similar transmission routes can lead to new insights for disease dynamics. Hens *et al.* (2007) describe the use of flexible marginal and conditional models to model multi-sera prevalence data on the Varicella-Zoster Virus and the Parvo B19-virus in Belgium. These models allows one to study the association among the occurrence and acquisition of both infections. In Section 4 we indicate how the models of Hens *et al.* (2007) can be adopted for “direct” modelling, on the scale of the antibody levels, without the need of any thresholds.

2 A Direct Method

Define $G(a)$ as the distribution of the log antibody levels of all subjects of age a . Some of these subjects have been infected, others are still susceptible. This can be described by the mixture

$$G(a) = (1 - \pi(a))G_s(a) + \pi(a)G_i(a), \quad (4)$$

where G_s (resp. G_i) denotes the distribution of all susceptible (resp. infected/recovered) subjects of age a . The mean log antibody level can then be decomposed as

$$\mu(a) = (1 - \pi(a))\mu_s(a) + \pi(a)\mu_i(a). \quad (5)$$

Assuming that the mean log antibody levels for the infected and susceptible compartments do not depend on age, some basic calculus leads to the identity

$$\ell(a) = \mu'(a)/(\mu_i - \mu(a)). \quad (6)$$

The overall mean level $\mu(a)$ can be estimated directly from the observed log antibody levels. Since $\pi(a)$ is monotone increasing as a function of age a and since $\mu_i \geq \mu_s$, the mean level $\mu(a)$ also has a monotone increasing behaviour. To allow for sufficient model flexibility, a non-parametric approach such as monotone penalized splines (see Bollaerts *et al.*, 2006) can be used.

Theoretically it can be shown that the classical threshold based approach leads to a structural bias in both age-dependent parameters, the true disease prevalence as well as the true force of infection. Finite sample simulations confirm these findings and illustrate that direct estimation performs better.

3 Example: Varicella-Zoster Virus and Parvo B19 in Belgium

Serum samples were tested for Varicella-Zoster Virus (VZV) and Parvo B19-virus (B19) in Belgium (Nardone and Miller, 2004, Ory *et al.*, 2006). In total 2381 serum samples were collected in a period from November 2001 until March 2003. The Varicella-Zoster Virus, also known as human herpes virus 3 (HHV- 3), is one of eight herpes viruses known to affect humans (and other vertebrates). Primary VZV infection results in chickenpox (varicella), which may rarely result in complications including bacterial surinfection, encephalitis, pneumonia and death. It has a two-week incubation period and is highly contagious by air droplets starting two days before symptoms appear. Infectiousness is known to last up to ten days. Therefore, chickenpox spreads quickly through close social contacts. Parvovirus B19 was the first human Parvovirus to be discovered, in 1975. In clinical terms Parvovirus B19 is best known for causing a childhood exanthem called fifth disease or erythema infectiosum. The virus is primarily spread by infected respiratory droplets.

Figure 1 shows the resulting fitted curves (for Parvo B19) for the mean level of antibodies (on the log scale, in the upper panel) and the FOI (in the lower panel), using monotone P-splines. The two dashed horizontal lines show the ELISA-kit supplied thresholds (with the equivocal area in between the two thresholds). In this example, the directly estimated FOI curve shows essentially the same qualitative characteristics as the curve estimated by the classical threshold based “indirect” approach.

4 Extension to Multiple Infection Measurements

The direct method can be extended to two or more diseases. For diseases with similar transmission routes (such as VZV and Parvo B19), this multivariate approach allows estimation of age-dependent association or disease synchrony measures, which provide new insights in the joint dynamics of such diseases. For the situation of two diseases, the mixture model has to be extended to four bivariate distributions, representing the joint compartments $\boxed{S_1 \mid S_2}$, $\boxed{S_1 \mid I_2}$, $\boxed{I_1 \mid S_2}$, and $\boxed{I_1 \mid I_2}$, where S_i (I_i) denotes the compartment “Susceptible for disease i ” (respectively “Infected by disease i ”). A quadrinomial distribution with age-dependent probabilities $\pi_{SS}(a)$, $\pi_{SI}(a)$, $\pi_{IS}(a)$ and $\pi_{II}(a)$ governs the mixture distribution. Using for instance a multivariate spline model to model the mean of (z_1, z_2) , with z_i the (log of the) antibody level of disease i , as a function of the age of the subject, allows estimation of marginal epidemiological parameters such as the marginal force of infection, as well as joint parameters, such as the age-dependent correlation, or other measures of association of disease “synchrony”, and conditional versions of the force of infection. The method has been applied to the joint estimation of Varicella and Parvo B19 in Belgium. In general, for K diseases, the approach leads to a mixture of $K \times K$ distributions of dimension K .

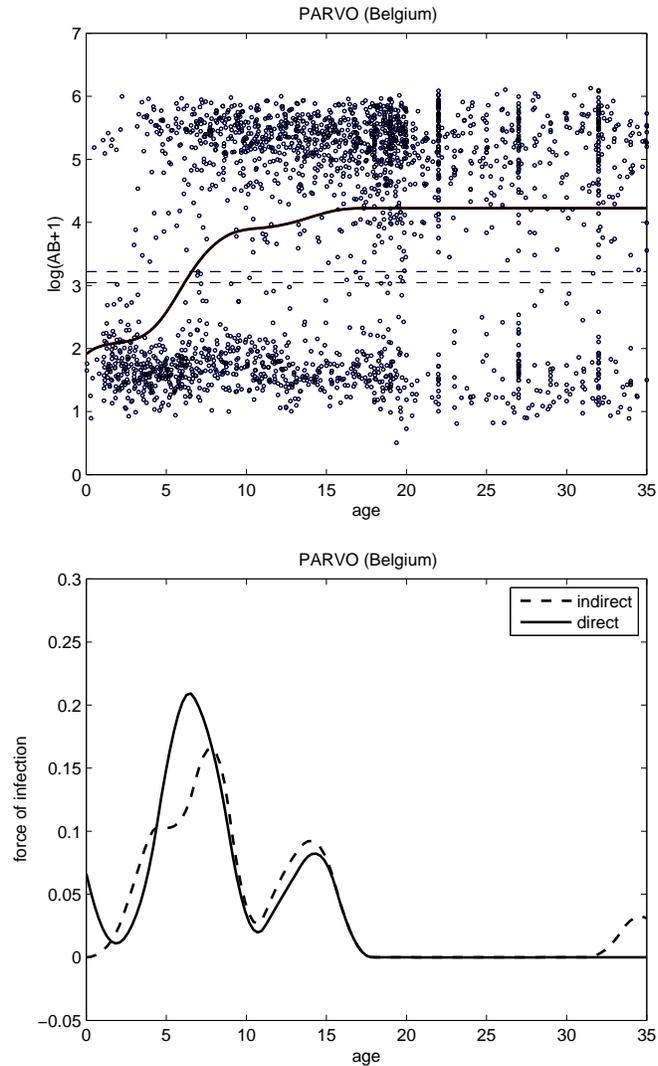


FIGURE 1. Parvo B19 example: Monotone P-spline fit for the mean log antibody level $\mu(a)$ (upper panel) and associated fit for the force of infection $\ell(a)$ (lower panel).

Acknowledgments: This work has been partly funded by POLYMOD, a European Commission project funded within the Sixth Framework Programme, Contract number: SSP22-CT-2004-502084, by the Fund of Scientific Research (FWO, Research Grant # G039304) in Flanders, Belgium and by the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

References

- Anderson, R.M. and May, R.M. (1991). *Infectious diseases of humans, dynamic and control*. Oxford University Press Inc., New York.
- Bollaerts, K., Eilers, P.H.C. and Van Mechelen, I. (2006). Simple and multiple P-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology* **59**, 451–496 .
- Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P. and Beutels, P. (2007). Modelling multi-sera data: The estimation of new joint and conditional epidemiological parameters. Submitted to *Statistics in Medicine*.
- Namata, H., Shkedy, Z., Faes, C., Aerts, M., Molenberghs, G., Theeten, H., Van Damme, P. and Beutels, P. (2007). Estimation of the force of infection from current status data using generalized linear mixed models. *Journal of Applied Statistics*. To appear.
- Nardone, A. and Miller, E. (2004). Serological surveillance of rubella in europe: European sero-epidemiology network (ESEN2). *Euro-surveillance*.
- Ory, F., Echevarria, J. and Kafatos, G. e. a. (2006). European seroepidemiology network 2: Standardisation of assays for seroepidemiology of varicella zoster virus. *Journal of Clinical Virology* **36**, 111–118.
- Shkedy, Z, Aerts, M., Molenberghs, G, Beutels, P. and Van Damme, P. (2003). Modelling forces of infection by using monotone local polynomials. *Applied Statistics* **52**, 469–485.
- Shkedy, Z, Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2006). Estimating the force of infection from serological data using fractional polynomials. *Statistics in Medicine* **25**, 1577–1591.

Semiparametric models for longitudinal binary responses with attrition

Marco Alfó¹ and Antonello Maruotti¹

¹ Dipartimento di Statistica, Probabilità e Statistiche Applicate, “Sapienza” Università di Roma, P.le Aldo Moro, 5 00185 Rome - ITALY.

Corresponding author e-mail: antonello.maruotti@uniroma1.it

Abstract: Longitudinal studies collect information on a sample of individuals which is followed over time, to analyze the effects of individual and time dependent characteristics on the observed response. These studies may, however, suffer from *attrition*: some individuals drop out of the study before its completion time and thus may present incomplete data records. If missing and complete sequences differ in any way relevant to the analysis, the dropout mechanism is non-ignorable and standard analyses are potentially biased. We extend semiparametric variance component models for longitudinal binary responses to allow for unobservable sources of variation influencing the missing-data mechanism as well as the primary response process. Applications to some benchmark datasets are discussed.

Keywords: longitudinal binary responses; non-ignorable dropouts; random coefficient based dropout model

1 Introduction

This paper discusses regression models for the analysis of longitudinal binary responses; attention is focused on those empirical situations where some measurements are missing for some units in the analyzed population, due to irretrievable non compliance or attrition.

We discuss cases where the dropout probability depends on unobserved sources of variation which influence also the primary response. In this case, the dropout process is *non ignorable*: when estimating parameters in the primary model, we need to specify a secondary model for the dropout process. We define a random effect based dropout model (see Little, 1995), which assumes that dependence between (the observed and unobserved) primary response and the dropout mechanism is driven by unobservable sources of random variation. The latter represent a set of omitted (unobservable) covariates influencing the primary response as well as the time spent in the analyzed panel. We propose to model the primary response and the dropout process through a selection model with *shared* or *correlated* random effects. Model parameters are estimated by NPML, avoiding parametric assumptions upon the random effect distribution.

2 Variance component models for complete sequences

Let Y_{it} represent a binary response recorded on n individuals ($i = 1, \dots, n$) at T_i time occasions ($t = 1, \dots, T_i$) together with a set of p explanatory variables, $\mathbf{x}_{it} = (x_{it1}, \dots, x_{itp})^\top$. Since the sequence of the primary response could be partially observed, we will write $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i}) = (\mathbf{y}_i^{obs}, \mathbf{y}_i^{mis})$, where the terms within parentheses denote the observed and the unobserved part of \mathbf{y}_i , respectively. A natural way to deal with complete sequences in a longitudinal setting is via the usual variance components GLM, which can be formulated as follows. Responses Y_{it} , $i = 1, \dots, n$, $t = 1, \dots, T_i$ are assumed to be conditionally independent Bernoulli variates with canonical parameter defined through the following linear function:

$$\theta_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + u_i \quad (1)$$

where u_i , $i = 1, \dots, n$ represent unobserved *individual-specific* sources of heterogeneity common to each lower-level unit (time) within the same i -th upper-level unit (individual), and $\boldsymbol{\beta}$ is the p -dimensional vector of *fixed* regression parameters. In particular let us consider the following mixed effects AR(1) model:

$$\theta_{it} = \mathbf{x}_{it}^\top \boldsymbol{\beta} + u_i + \alpha y_{i,t-1} \quad (2)$$

The log-likelihood function is calculated by integrating out the u_i 's as follows:

$$l(\cdot) = \sum_{i=1}^n \log \left\{ \int \prod_{j=2}^{T_i} f(y_{ij} | y_{i,j-1}, \mathbf{x}_{it}, u_i) f(y_{i1} | \mathbf{x}_{i1}, u_i) dG(u_i) \right\} \quad (3)$$

The term $f(y_{i1} | \mathbf{x}_{i1}, u_i)$ is not specified by adopted modeling assumptions; however, estimation is possible by defining different model structures on the first and subsequent occasions, as detailed in Aitkin and Alfó (2003). We will implicitly adopt the latter approach. The joint likelihood is defined by the following integral:

$$l(\cdot) = \sum_{i=1}^n \log \left\{ \int \prod_{t=1}^{T_i} f(y_{it} | y_{i,t-1}, \mathbf{x}_{it}, u_i) g(u_i) du_i \right\} \quad (4)$$

The distribution of u_i , say $G(u_i)$, may be left unspecified; NPML estimation of locations and corresponding masses can be achieved in a general finite mixture framework (see e.g. Aitkin, 1999).

3 Semiparametric selection model

Longitudinal studies may suffer from the problem of *dropout*, i.e. some individuals may leave the study before its completion time, and thus present incomplete data records. If the designed completion time is denoted by T_i , we will have $\mathcal{T}_i \leq T_i$ measures for each unit. If the decision to dropout is related to the primary response, a selection problem can result, just as in the cross section case. We assume that, once a person drops out, he or she is out forever: attrition is, in this sense, an absorbing state.

We denote with \mathbf{R}_i the missing-data indicator, i.e. a \mathcal{T}_i -dimensional random variable with $R_{it} = 0$ if observations for the i -th unit are available at the t -th occasion, $t = 1, \dots, \mathcal{T}_i$, and $R_{it} = 1$ otherwise. The key question is whether those who dropout may differ (in any way relevant to the analysis) from those who still continue to participate. Little and Rubin (2002) discuss two classes of models for handling non-ignorable missing data: namely, *selection* and *pattern mixture* models. In the former case, a complete-data model is defined for the primary response and joined to a model describing the missing-data mechanism conditional on the complete data. Observed data are considered as selected from the complete data through a process which is dependent on the *complete* primary outcome.

Selection models are based on the bivariate distribution $f(y_{it}, R_{it})$, $i = 1, \dots, n$, $t = 1, \dots, \mathcal{T}_i$ (see Little and Rubin, 2002). The term *selection* is used because the process of obtaining the observed data can be viewed as due to a form of selection from the *complete* data set. To specify a formal model for the pair (y_{it}, R_{it}) , we follow Verzilli and Carpenter (2002) and use a set of shared random effects, say u_i , for the primary response and the missing-data indicator. In particular, the latter is modelled using a discrete-time proportional hazards model. Such model corresponds to a generalized linear model where the missing-data indicator has (conditional on u_i) a Bernoulli distribution with complementary log-log link function. The corresponding model could be therefore written as:

$$\Pr(R_{it} = 1 \mid \mathbf{v}_{it}, u_i) = 1 - \exp \left[- \exp(\mathbf{v}_{it}^\top \boldsymbol{\phi} + u_i + \psi_t) \right] \quad (5)$$

where \mathbf{x}_{it} and $\boldsymbol{\phi}$ represent dropout specific covariates and parameter vectors. The terms ψ_t represent, instead, fixed effects contrasts for time period $t = 1, \dots, \mathcal{T}_i$. The joint marginal distribution is defined as follows:

$$f(\mathbf{y}_i, \mathbf{r}_i) = \int_{\mathcal{U}} f(\mathbf{y}_i \mid u_i) f(\mathbf{r}_i \mid u_i) dG(u_i)$$

where $f(\mathbf{y}_i \mid u_i) = \prod_t f(y_{it} \mid u_i)$ and $h(\mathbf{r}_i \mid u_i) = \prod_{t=1}^{\min\{\mathcal{T}_i+1, \mathcal{T}_i\}} f(r_{it} \mid u_i)$. The first term on right-hand side defines a complete-data model for the primary response, while the second term models the missing-data (i.e. dropout) mechanism. We leave $G(\cdot)$ unspecified, and use the finite mixture approach described above. This can be simply done by using a multilevel structure where the primary response and the missing-data indicator are simply two outcomes registered on the same individual (with time periods nested within individuals as well). In this case, the log-likelihood function is given by

$$\begin{aligned} l(\cdot) &= \sum_{i=1}^n \left\{ \sum_{k=1}^K f(\mathbf{y}_i \mid \mathbf{x}_i, u_k) h(\mathbf{r}_i \mid \mathbf{v}_i, u_k) \pi_k \right\} \\ &= \sum_{i=1}^n \left\{ \sum_{k=1}^K f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{x}_i, \mathbf{v}_i, u_k) \pi_k \right\}, \end{aligned} \quad (6)$$

where the dropout mechanism is explained through a set of explanatory variables, \mathbf{v}_i . An EM algorithm can be straightforwardly adapted. The *shared* random effect approach may, however, lack generality: we implicitly assume the same heterogeneity

sources affect all outcomes, with the effects having the same (or proportional) *size* in both equations. The previous model can be generalized by allowing for *correlated* random effects, as in multivariate models. To explain, let $\mathbf{u}_i = (u_{i1}, u_{i2}) \sim G(\cdot)$ denote a set of subject and outcome-specific random effects. They control for overdispersion and association in the univariate profiles as well as for association between the primary response and the dropout process. The log-likelihood function is thus defined by:

$$\begin{aligned} l(\cdot) &= \sum_{i=1}^n \left\{ \sum_{k=1}^K f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{u}_{k1}) h(\mathbf{r}_i \mid \mathbf{v}_i, u_{k2}) \pi_k \right\} \\ &= \sum_{i=1}^n \left\{ \sum_{k=1}^K f(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{x}_i, \mathbf{v}_i, \mathbf{u}_k) \pi_k \right\}, \end{aligned} \quad (7)$$

where $\pi_k = \Pr(\mathbf{u}_k) = \Pr(u_{k1}, u_{k2})$ denotes the joint probability of locations \mathbf{u}_k . Even in this case the standard EM algorithm for VC models can be simply adapted to the augmented data, by defining a multilevel structure with outcomes (upper level units) nested within individuals (lower level unit).

4 Conclusions

The proposed model has been fitted to data from the schizophrenia longitudinal study (Thara et al., 2004) and from a clinical trial on contracepting women (Machin et al., 1988). The results show that ignoring the influence of the dropout mechanism may lead to biased estimates for both fixed and random effects. Formal tests for independence between the random effects in the two equations are based on penalized likelihood criteria.

References

- Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- Little, R.J.A. (1995) Modelling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.
- Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, 2nd edition. New York: John Wiley.
- Machin, D., et al. (1988) Assessing changes in vaginal bleeding patterns in contracepting women *Contraception* **38**, 165–179.
- Thara, R. (2004) Twenty-Year Course of Schizophrenia: The Madras Longitudinal Study *Canadian Journal of Psychiatry* **49**(8), 564–569.
- Verzilli, C.J. and Carpenter, J.R. (2002) A Monte Carlo EM algorithm for random coefficient-based dropout models *Journal of Applied Statistics* **29**, 1011-1021.

Analyzing Mental Comorbidity through LCA. Results of the ESEMeD project

J. Almansa¹, G. Vilagut¹ and J. Alonso¹

¹ Health Services Research Unit, IMIM-Hospital del mar, Carrer del Doctor Aiguader, 88, Edifici PRBB, 08003 BARCELONA, SPAIN, jalmansa@imim.es

Abstract: This study shows the application of Latent Class Analysis to identify relevant mental comorbidity patterns. Eleven classes fit the data: one no disorder class, eight with mainly one disorder and two with more than one disorder. The latter are the most disabling patterns.

Keywords: Comorbidity; Latent Class Analysis

1 Background and Objectives

Psychiatric comorbidity is defined as the co-occurrence of mental disorders. Mental comorbidity has become a topic of increasing interest in research and clinical practice. Mental comorbidity has been related to increased severity, longer duration of the disorder, greater functional disability, and increased the use of healthcare services. Comorbid disorders have also different risk-factor profiles compared with pure disorders. Latent Class modelization has not been widely used in the analysis of mental health data despite having a great potential in this area. This technique allows for analyzing not directly observable phenomena from a set of manifest variables.

The goals of this analysis are to identify the significant 12 month mental comorbidity patterns in a European mental health survey (ESEMeD), and determine the factors which may be able to predict the presence of some of these patterns.

2 Data Description and Methods

This study is based on the European Study of the Epidemiology of Mental Disorders (ESEMeD), a general population survey carried out in six European countries as part of the WHO World Mental Health (WMH) Survey Initiative (ESEMeD/MHEDEA-2000 Investigators, 2002).

We used the CIDI 3.0, the latest version of the Composite International Diagnostic Interview (CIDI) (Kessler & Ustun, 2004). The CIDI allows for the evaluation of mental health disorders according to the Diagnostic Statistical Manual (4th edition), DSM-IV criteria, through a CAPI interview. Mental health diagnostics were evaluated as dichotomous variables (having or not having the disease). The diagnostics used in this analysis are grouped in 3 main categories: mood disorders (major depression, dysthymia), anxiety disorders (social phobia, specific phobia, generalized anxiety disorder,

agoraphobia, panic disorder, post-traumatic stress disorder) and alcohol disorders (alcohol dependence, alcohol abuse). Impact of comorbidity was assessed with the SF-12 component summary scores (physical: PCS; mental: MCS).

Latent Class Analysis (Goodman, 1974) was conducted to identify the comorbidity patterns in the ESEMeD data. Mental diagnostics are the main variables. Alcohol dependence and alcohol abuse were collapsed into a single variable due to their high association (indeed they both lay on a single latent dimension, see Kahler & Strong, 2006). Latent class analysis postulates a discrete latent variable defining class membership that explains covariance among observed disorders. This model contains 1 parameter for the probability of each disorder conditional to each of k classes of the latent variable, in addition to k parameters for class prevalence. Complex sample design was taken into account (Patterson et al., 2002) and covariables were also included in the model (active: gender and age; inactive: MCS, country and marital status). The number of classes was determined by the best fitting model according to the AIC3 value. All the analyzes were conducted with Mplus 4.2.

3 Results

The main results are shown in Table 1 which contains the conditional probabilities (conditional to a class membership, which is the probability of having each of the diagnoses). Inactive categorical covariates are computed as ratio of the mean membership probability over the population distribution.

Eleven classes were obtained. Class one was forced to be the non-disorder class for those without any 12month disorder (88.05%). The rest of classes were defined by: Specific Phobia (4.56%), Major Depression (2.31%), Social Phobia with some association to Specific Phobia (1.11%), Comorbidity of Depression and Anxiety disorders (1.08%), Dysthymia associated with Major Depression (0.93%), Post-Traumatic Stress (0.65%), Alcohol disorders (0.63%), Panic Disorder (0.29%), General Anxiety Disorder(0.23%) and Agoraphobia (0.16%).

The no-disorder class was distributed equally between males and females. Regarding the disorder classes, males are clearly associated to Alcohol disorders, and females to: specific phobia, agoraphobia, post-traumatic stress, and comorbidity patterns. Alcohol disorder was also more frequent among never married and the younger individuals (<35).

According to the mental component of the SF12 instrument (MCS12) the patterns with worst quality of life are the comorbidity (depression and anxiety disorders) and the depression (Dysthymia with Major Depression) patterns. The latter was also related to older people (>50) and previously married. Post-Traumatic Stress is more frequent in France and The Netherlands.

Most of the diagnoses tend to appear alone, and the existence of comorbidity is related to greater disability. 1% of general population have 12-month mental comorbid disease, and almost another 1% mood comorbidity. Among anxiety disorders, apart from the comorbidity pattern (class 5) social phobia is somewhat related to specific phobia, but no other anxiety comorbidity pattern was found.

TABLE 1. Conditional probability of response, and covariates.

Cluster Num:	1	2	3	4	5	6	7	8	9	10	11
	(88.05)	(4.56)	(2.31)	(1.11)	(1.08)	(0.93)	(0.65)	(0.63)	(0.29)	(0.23)	(0.16)
DYSTHYMIA	0.00	0.00	0.04	0.00	0.18	1.00	0.00	0.03	0.00	0.00	0.00
MAJOR DEPRESSION	0.00	0.05	1.00	0.05	0.74	0.60	0.12	0.09	0.00	0.00	0.16
AGORAPHOBIA	0.00	0.01	0.00	0.06	0.29	0.08	0.00	0.00	0.06	0.00	1.00
GENERAL ANXIETY	0.00	0.00	0.09	0.03	0.35	0.08	0.01	0.00	0.01	1.00	0.00
PANIC DISORDER	0.00	0.01	0.04	0.00	0.30	0.02	0.03	0.01	1.00	0.00	0.00
SOCIAL PHOBIA	0.00	0.00	0.03	1.00	0.32	0.01	0.06	0.09	0.00	0.00	0.00
SPECIFIC PHOBIA	0.00	1.00	0.01	0.24	0.72	0.13	0.00	0.00	0.00	0.05	0.00
POST-TRAUM.STRESS	0.00	0.01	0.04	0.00	0.22	0.07	1.00	0.00	0.00	0.00	0.03
ALCOHOL DISORDER	0.00	0.00	0.01	0.00	0.09	0.00	0.01	1.00	0.00	0.00	0.00
Active Covariates											
AGE											
18-24	0.11	0.13	0.19	0.22	0.30	0.03	0.04	0.32	0.09	0.05	0.03
25-34	0.18	0.15	0.22	0.20	0.16	0.08	0.24	0.36	0.21	0.35	0.02
35-49	0.27	0.35	0.25	0.34	0.32	0.17	0.28	0.20	0.40	0.22	0.82
50-64	0.22	0.21	0.21	0.12	0.17	0.40	0.39	0.11	0.22	0.27	0.13
65+	0.22	0.16	0.14	0.12	0.06	0.32	0.06	0.00	0.08	0.12	0.00
GENDER											
Male	0.50	0.28	0.36	0.33	0.23	0.38	0.15	0.88	0.35	0.64	0.04
Female	0.50	0.72	0.64	0.67	0.77	0.62	0.85	0.12	0.65	0.36	0.96
Inactive Covariates											
NUMBER OF DIAGNOSES											
mean	0.00	1.08	1.26	1.38	3.22	2.00	1.24	1.23	1.08	1.05	1.19
MCS-12											
mean	54.37	50.35	42.42	46.69	38.75	38.18	47.19	52.52	48.05	50.65	51.44
MARITAL STATUS											
Married	1.01	1.05	0.84	0.94	0.76	0.97	1.15	0.59	0.90	0.96	1.42
PreviousMarried	0.99	1.09	1.08	0.43	1.38	2.38	1.20	0.33	1.41	0.35	0.04
NeverMarried	0.99	0.79	1.44	1.46	1.54	0.39	0.44	2.58	1.08	1.44	0.20
COUNTRY											
Belgium	0.99	0.96	1.52	0.83	1.17	1.03	0.57	1.13	2.10	1.31	0.56
France	0.93	1.42	1.60	1.72	1.65	1.34	2.13	0.90	0.61	2.78	2.19
Germany	1.01	1.02	0.60	1.08	1.02	0.71	0.47	1.75	1.29	0.09	0.64
Italy	1.04	0.77	0.83	0.80	0.56	0.99	0.59	0.12	0.74	0.43	0.99
Netherlands	0.99	0.76	1.19	0.59	1.24	0.96	3.08	1.86	1.94	1.21	0.73
Spain	1.03	0.85	1.07	0.39	0.61	1.17	0.46	0.51	0.66	1.17	0.38

Future work will focus in comparing male and female comorbidity patterns and among countries, confirmatory class analysis to test to what extent these results can be generalized, and include some chronic conditions into the analysis of comorbidity patterns.

Acknowledgments: The ESEMeD project was funded by the European Commission (Contracts QLG5-1999-01042; SANCO 2004123), the Piedmont Region (Italy), Fondo de Investigación Sanitaria, Instituto de Salud Carlos III, Spain (FIS 00/0028-02), Ministerio de Ciencia y Tecnología, Spain (SAF 2000-158-CE), Departament de Salut, Generalitat de Catalunya, Spain, and other local agencies and by an unrestricted educational grant from GlaxoSmithKline. Almansa J. is supported by "Fons Social Europeu i del Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya (AGAUR)".

Special acknowledgment to Dave MacFarlane, who helped us to build this document in LaTeX format.

References

- ESEMeD/MHEDEA-2000 Investigators. (2002) The European Study of the Epidemiology of Mental Disorders (ESEMeD/MHEDEA 2000) project: rationale and methods. *Int J Meth Psychiatr Res* **11**, 55–67.
- ESEMeD/MHEDEA 2000 Investigators. (2004) 12-Month comorbidity patterns and

- associated factors in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMED) project. *Acta Psych Scand* **109(Suppl. 420)**, 28–37.
- ESEMED/MHEDEA 2000 Investigators. (2004) Sampling and methods of the European study of epidemiology of mental disorders (ESEMED) project. *Acta Psych Scand* **109(Suppl. 420)**, 8–20.
- Goodman. L.A. (1974) Exploratory Latent Structure Analysis using both identifiable and unidentifiable models. *Biometrika* **61(2)** 215–231.
- Kahler C, Strong D (2006) A Rasch Model Analysis of DSM-IV Alcohol Abuse and Dependence Items in the National Epidemiological Survey on Alcohol and Related Conditions *Alcoholism Clinical and Experimental Research* **30(7)**, 1165–1175(11)
- Kessler RC (1995) Epidemiology of psychiatric comorbidity. In: Tsuang MT, Tohen M, Zahner GEP, eds. *Textbook in psychiatric epidemiology*. New York: Wiley-Liss.
- Kessler RC, Chiu WT, Demler O, Walters EE (2005) Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry* **62**, 617–627.
- Kessler RC, Ustun TB. (2004) The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int J Methods Psychiatr Res* **13(2)**, 93–121.
- Patterson, Blosson H., Dayton CM, Graubard BI (2002) Latent Class Analysis of Complex Sample Survey Data: Application to Dietary Data. *Journal of the American Statistical Association* **97**, 721–728.

Time Series Classification based on Functional Depth

Andrés M. Alonso¹, David Casado¹, Sara López-Pintado² and Juan Romo¹

¹ Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain.

² Universidad Pablo de Olavide, 41013 Sevilla, Spain

Abstract: A new classification method for time series is proposed. A series is assigned to a class after comparing distances between its integrated periodogram and the mean of the integrated periodograms in each group. The approach can be used with nonstationary time series by computing the periodogram locally. Depth based techniques are used to make the classification robust. The method provides small error rates both with simulated and real data; it also shows good computational behavior.

Keywords: time series; classification; integrated periodogram; depth.

1 Introduction

Classifying time series is an important task in several real problems. Previous work has considered both the time and frequency domains to discriminate and classify time series. We propose a frequency domain technique based on the integrated periodogram. We assign a new time series by considering the distance between its integrated periodogram and the mean of integrated periodograms in each group. Since these means are highly sensitive to outliers, we replace them by the corresponding α -trimmed means, where the $100\alpha\%$ excluded data are chosen using the idea of statistical depth extended to functional data by López-Pintado and Romo (2006).

2 The classification method

One of the main points of our classification proposal is that we turn the time series problem into a functional data problem by considering the integrated periodogram of each time series. The *periodogram* $I(\omega)$ is the sample version of the *spectral density* and expresses the contribution of Fourier analysis frequencies to the series total variance. Its cumulative version is the *integrated periodogram* $F_Z(\omega_m) = \sum_{i=1}^m I(\omega_i)$. Though the periodogram is properly defined only for stationary series, we shall assume that the series are approximately locally stationary in order to classify nonstationary time series. We shall split them into k blocks and compute the integrated periodogram of each block.

When functions—instead of time series—need to be classified, a natural criterion is to assign them to the class minimizing some distance from the new function to the

group. As a reference function of each group we take the mean of its elements; since the mean is not robust to the presence of outliers, robustness can be added to the process by considering the α -trimmed mean instead. Both means can be expressed as follows. Let $\Psi_{g(i)}(\omega)$, $i = 1, \dots, N$, be functions of the class g ordered by decreasing depth; the α -trimmed mean is:

$$\bar{\Psi}_g^\alpha(\omega) = \frac{1}{N - [N\alpha]} \sum_{i=1}^{N-[N\alpha]} \Psi_{g(i)}(\omega), \quad (1)$$

where $[\cdot]$ is the integer part function. When $\alpha = 0$, the whole sample is taken, while if $\alpha > 0$ the $100\alpha\%$ of the less depth data are leaved out. We use the definition of functional *generalized band depth* proposed by López-Pintado and Romo (2006). We have taken the L_1 distance between two functions.

Let $\{X_1, \dots, X_{N_x}\}$ be a sample containing time series from the population P_X , and let $\{Y_1, \dots, Y_{N_y}\}$ be a sample from P_Y . The classification method follows the next steps:

1. For each time series in the samples, the integrated periodogram of the k blocks is obtained, i.e., we have $\{\Psi_{X_1}, \dots, \Psi_{X_{N_x}}\}$ and $\{\Psi_{Y_1}, \dots, \Psi_{Y_{N_y}}\}$, where $\Psi_{X_i} = (F_{X_i}^{(1)} \dots F_{X_i}^{(k)})$, $\Psi_{Y_i} = (F_{Y_i}^{(1)} \dots F_{Y_i}^{(k)})$ and $F_{X_i}^{(j)}$ is the integrated periodogram of the j -th block of the i -th series of the population X ; $F_{Y_i}^{(j)}$ is the analogous function for the population Y .
2. For both P_X and P_Y samples the α -trimmed class mean is computed: $\bar{\Psi}_X^\alpha$ and $\bar{\Psi}_Y^\alpha$.
3. Let Ψ_Z be the curve associated to a new series Z , that is $\Psi_Z = (F_Z^{(1)} \dots F_Z^{(k)})$. Then Z is classified in the group P_X if $d(\Psi_Z, \bar{\Psi}_X^\alpha) < d(\Psi_Z, \bar{\Psi}_Y^\alpha)$, and in the group P_Y otherwise.

To apply the algorithm to stationary series take k as 1. The extension to more than two groups is straightforward.

3 Simulations results

We evaluate our algorithms for $\alpha = 0$ (DbC) and $\alpha = 0.2$ (DbC- α). Also, we take the method proposed in Huang et al. (2004) as a reference (SLEXbC). For the last method we have used an implementation provided by the authors. We have analyzed the three experimental settings considered by these authors. Next, we present the results for the first of them:

- Gaussian white noise (P_X) is compared with an autoregressive process of order 1 (P_Y). Here $N_x = N_y = 8$ and time series lengths are $T_x = T_y = 1024$. We also apply both weak and strong contaminations to the training sets. See the details in Alonso et al (2007).

Series are stationary in setting 1, composed of stationary parts in setting 2 and non-stationary in setting 3. For each comparison we have run 1000 times the training-testing processes. The three algorithms are called with exactly the same data sets.

TABLE 1. Estimated means of misclassification rates in setting 1 with and without contaminations.

	$\phi = -0.3$	$\phi = -0.1$	$\phi = +0.1$	$\phi = +0.3$
DbC	0.000	0.063	0.060	0.000
DbC-α	0.000	0.065	0.062	0.000
SLEXbC	0.000	0.131	0.127	0.000
Weak contamination				
DbC	0.000	0.077	0.074	0.000
DbC-α	0.000	0.064	0.062	0.000
SLEXbC	0.000	0.175	0.172	0.000
Strong contamination				
DbC	0.001	0.512	0.300	0.000
DbC-α	0.000	0.064	0.062	0.000
SLEXbC	0.002	0.490	0.377	0.001

In table 1 we present some results for simulation setting 1. When contamination is not present, DbC provides slightly better results than DbC- α , and about half of the errors of SLEXbC. The DbC and SLEXbC errors increase slightly with the weak contamination and substantially with the strong one, while errors do not change for DbC- α , because its trimming keep contamination out; this shows its robustness. Similar results are obtained from the two remaining settings (see Alonso et al, 2007).

4 Real data example

We have evaluated our proposal in a benchmark data set containing 8 explosions, 8 earthquakes and an extra series—known as NZ event—not classified (but being an earthquake or an explosion). Each series contents 2048 points in two different parts: the first half is the P wave, and the second is the S wave. For each series we have considered the curve formed by merging the non-normalized integrated periodograms of parts P and S. Considering the 8 earthquakes as group 1 and the 8 explosions as group 2, and applying leave-one-out cross validation, both of our algorithms misclassify only the first series of the group 2. With respect to the NZ event, both algorithms agree on assigning it to the explosions group, as other authors do, for example, Kakizawa et al. (1998) and Huang et al. (2004).

Now we have carried out an additional exercise. Since many methods classify the NZ event as an explosion, we consider an artificial data set constructed by the 8 earthquakes plus the NZ event as group 1, and the 8 explosions as group 2. In this situation, the result for DbC is that it misclassifies the first and the third—not only the first—elements of the group 2. But DbC- α again misclassifies only the first series of group 2, even though the NZ event was included in group 1 (earthquakes). This illustrates the robustness of our proposed second algorithm.

Acknowledgments: This research was partially supported by project SEJ 2005–06454. The first author acknowledges support by a “Juan de la Cierva” grant.

References

- Alonso, A.M., Casado, D., López–Pintado, S., and Romo, J. (2007). A functional data based method for time series classification. Submitted.
- Huang ,H., Ombao, H., and Stoffer, D. (2004). The SLEX approach to discrimination and classification of nonstationary time series. *Journal of the American Statistical Association* **99**.
- Kakizawa, Y., Shumway, R.H. and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association* **93**.
- López–Pintado, S. and Romo, J. (2006). On the concept of depth for functional data. *Working Paper*. Departamento de Estadística. Universidad Carlos III de Madrid.

Forecasting the Spanish mortality rates

Andrés M. Alonso¹, Daniel Peña¹ and Julio Rodríguez²

¹ Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain

² Universidad Autónoma de Madrid, 28049 Madrid, Spain

Abstract: This paper looks at projections for the Spanish mortality rates by sex and age for the period of 2005 to 2050. These forecast are calculated using two main sources of information. First, a multivariate time series model was applied for the series from 1950 to 2004 period. Second a model was estimated for life expectancy at birth. Both sources of information were combined to obtain the forecasts for the rates. The results are compared with the life expectancy projections proposed by the National Statistical Institute (2004).

Keywords: mortality time series; factorial model; bootstrap.

1 Introduction

The cohort component method, which is used in population forecasts, requires the establishment of future paths for the three basic components of the population changes: mortality, fertility and migration. In this work we consider the mortality component. We use a dynamic factor model similar to that proposed by Lee and Carter (1992), but introducing restrictions in the first common factor, to model the mortality rates by age groups. The prediction of these rates is based on a modification of the sieve bootstrap procedure proposed by Alonso et al (2002).

2 Modelling and forecasting mortality rates

2.1 Dynamic factor model

In this section we follow the presentation of the dynamic factor model in Peña and Poncela (2004). Let $\{\mathbf{y}_t\}_{t \in Z}$ be a vector series of dimension m , for example, male mortality rates of m age groups. The dynamic factor model assumes that \mathbf{y}_t can be written as a linear combination of r common factors plus an error term:

$$\mathbf{y}_t = \mathbf{P} \mathbf{f}_t + \mathbf{e}_t, \quad (1)$$

where \mathbf{f}_t is the r -dimensional vector of common factors, \mathbf{P} is the weight matrix of factors, and \mathbf{e}_t is the specific factors vector or error term. Additionally, it is assumed that \mathbf{f}_t follows a VARIMA(p, d, q) model defined by:

$$\Phi(B) \mathbf{f}_t = \Theta(B) \mathbf{v}_t, \quad (2)$$

where B is the backward shift operator, $\Phi(B)$ and $\Theta(B)$ are the autoregressive and moving average polynomial matrices, respectively. Moreover, in this paper, specific factors are allowed to follow stationary univariate models.

2.2 Bootstrap procedure for forecasting

In this section we present a bootstrap procedure for obtaining forecasts paths based on a modification of the procedure proposed by Alonso et al (2002).

1. The estimations for the r common factors and the weight matrix in model (1) are obtained using the singular values decomposition as in Lee and Carter (1992).
2. The residuals of model (1) are calculated using $\widehat{\mathbf{e}}_t = \mathbf{y}_t - \widehat{\mathbf{P}} \widehat{\mathbf{f}}_t$.
3. An AR(p_a) model is chosen for $\widehat{\mathbf{e}}_{a,\cdot}$ with $a \in \{1, 2, \dots, m\}$ and an ARI (p_s, d_s) model is chosen for common factors with $s \in \{1, 2, \dots, r\}$ using the BIC criteria.
4. The empirical distribution function are obtained for the centered residuals of the AR and ARI models: $\widehat{F}_{\epsilon_e}^*$ and $\widehat{F}_{v_s}^*$, respectively.
5. A resample ϵ_t^* of i.i.d. observations from \widehat{F}_{ϵ}^* and a resample v_t^* of i.i.d. observations from \widehat{F}_v^* were selected.
6. The future bootstrap observations are calculated for common and specific factors using the relations: and

$$f_{s,T+h}^* = \sum_{j=1}^{p_s+d_s} \widehat{\phi}_{s,j} f_{s,T+h-j}^* + v_{s,T+h}^*, \quad (3)$$

and

$$e_{a,T+h}^* = \bar{e}_a + \sum_{j=1}^{p_a} \widehat{\phi}_{a,j} (e_{a,T+h-j}^* - \bar{e}_a) + \epsilon_{a,T+h}^*, \quad (4)$$

where $h > 0$, $f_{s,t}^* = \widehat{f}_{s,t}$ for $t \leq T$ and $e_{a,t}^* = \widehat{e}_{a,t}$ for $t \leq T$, with T being the last available year.

7. The future bootstrap observations are calculated for vector \mathbf{y} using the relation:

$$\mathbf{y}_{T+h}^* = \widehat{\mathbf{P}} \mathbf{f}_{T+h}^* + \mathbf{e}_{T+h}^*. \quad (5)$$

In case of mortality rates we have seen that there is a high correlation between the first factor of model (1) and the life expectancy at birth. This allows us to establish a simple model between the first factor, $f_{1,t}$, and the synthetic index, i_t :

$$f_{1,t} = \alpha_0 + \alpha_1 i_t + \iota_t, \quad (6)$$

where ι_t assumes that it follows an AR(p_ι) model. The previous model together with a specific modelling of the life expectancy at birth allow us to make forecasts for future values of this factor.

3 Spanish mortality rates forecast

3.1 Life expectancy modelling

In this section we propose a method for establishing an upper bound for life expectancy. It is important to point out the close relationship between life expectancy and the first factor of mortality; the correlation between both is: -0.9951 in men and -0.9974 in women. This allows us to establish restrictions on the mortality factor through restrictions on life expectancy.

The following transformation is considered for life expectancy at birth, EVN_t :

$$Y_{t,A} = \ln \frac{EVN_t}{A - EVN_t}, \quad (7)$$

where A is the upper bound for life expectancy. As opposed to the procedures described in IEA (1995) and Blanes et al (2004), no single value was set for A , instead, A is considered a parameter of the model. In Figure 1 we present the estimated distribution of the maximum likelihood estimators for the upper bounds of life expectancy in men and women. In Figure 2 we present the fan chart (see, Wallis 1999) of the bootstrap distribution of life expectancy forecasts. Specifically, in Figure 2 we represent the 20%, 40%, 60%, 80% and 90% forecast intervals, in addition to the median of the forecasts. The projections proposed by the INE (2004) fall in the 40% or 60% forecast intervals for men and 60% - 70% for women in the period from 2005 to 2035. Their projections are close to the median of the our forecasts at the end of the prediction forecast horizons.

3.2 Forecasts of Mortality Rates by Age and Sex

Once we have obtained the predictions for life expectancy at birth we can find the corresponding predictions for the first mortality factor using model (6) and, using relation (5), we obtain the forecast paths for mortality rates by age. To illustrate this further, in Figure 3 we show a fan chart of mortality rates during the first year of life. A clear reduction is observed in mortality in this age group. The remaining ages can be obtained using the routines developed in this paper and which are available from the authors upon request.

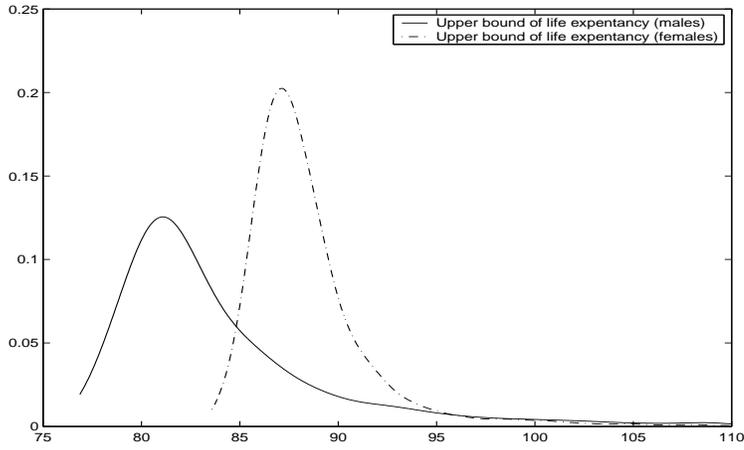


FIGURE 1. Estimated distribution of the MLE for the upper bounds of life expectancy.

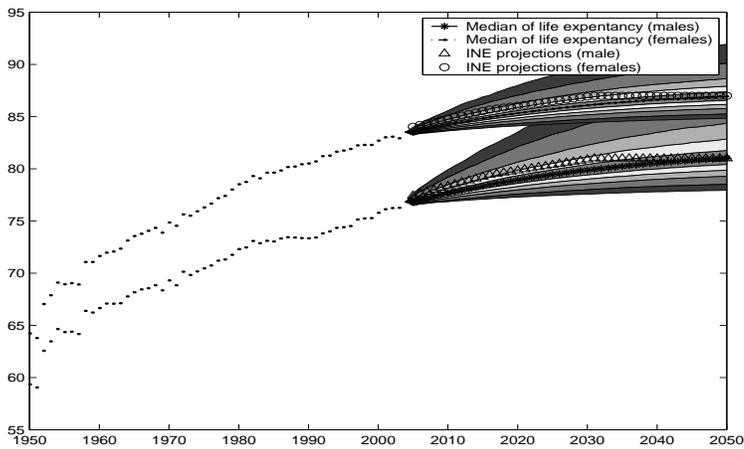


FIGURE 2. Fan chart and the observed values of life expectancy at birth, Spain 1950–2050.

Acknowledgments: This research was partially supported by the Fundación BBVA and project SEJ2005-06454. The first author acknowledges support by a “Juan de La Cierva” grant. The authors wish to thank Juan Bógalo for the interface modifications in MATLAB/TRAMO.

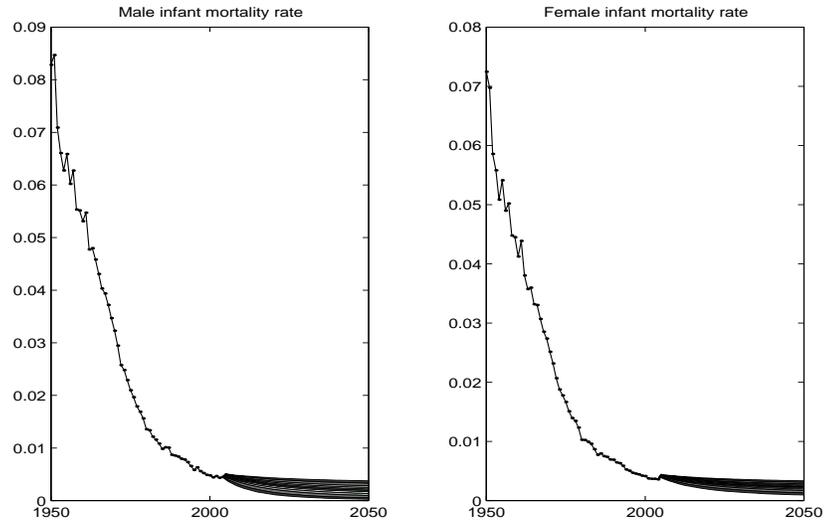


FIGURE 3. Fan chart and observed values of infant mortality rates. Spain 1950–2050.

References

- Alonso, A.M., Peña, D. and Romo, J. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference* **100**, 1–11.
- Blanes, A., Recaño, J, and Menacho, T. (2004). Proyección de población de la Comunidad de Madrid, 2002–2017. Instituto de Estadística de la Comunidad de Madrid, Madrid.
- IEA: Instituto de Estadística de Andalucía (1995). Proyección de la población de Andalucía 1991-2006, Sevilla.
- INE: Instituto Nacional de Estadística (2004). Proyecciones de la población de España calculadas a partir del censo de población de 2001, Madrid.
- Lee, R.D. and Carter, L. (1992). Modelling and forecasting the time series of U.S. mortality. *The Journal of the American Statistical Association* **87**, 659–671.
- Peña, D. and Poncela, P. (2004). Forecasting with nonstationary dynamics factor models. *Journal of Econometrics* **119**, 291–321.
- Wallis, K.F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart, *National Institute Economic Review* **167**, 106–112.

Modelling of local elections in Portugal

Anibal Areira¹, Manuela M. Oliveira² and João T. Mexia³

¹ Department of Economy and Management, College of Business Sciences, Setúbal Polytechnic Institute, Setúbal, Portugal

² Department of Mathematics, University of Évora, Colégio Luis António Verney, Rua Romão Ramalho 59. 7002 Évora, Portugal

³ FCT Nova University of Lisbon. Department of Mathematics. Quinta da Torre, 2825 Monte da Caparica, Portugal.

Abstract: Common structure matched series are modelled in the framework of the STATIS method. For that purpose, the information of each series is summarized in a structure vector and in a sum of squares of residues vector. These vectors are used as input to an ANOVA like analysis. This approach encompasses both a transversal and a longitudinal analysis. An application to the results of local elections in 8 districts of mainland Portugal is presented. Results show that the proposed approach may be successfully used to carry out inference about both political geography and political evolution at local level.

Keywords: ANOVA; F tests; Hilbert-Schmidt product; STATIS method.

1 Introduction

In this study we use ANOVA like analysis in the framework of the STATIS method to carry out inference in series of studies. Originally the STATIS method was used to elaborate algebraic results to describe data organized in series of studies.

In the next section we will review briefly this method and it will be shown how inference may be carried out. Nextly we present the data used in this study and discuss results of the application of this method.

2 STATIS method and inference

The STATIS method was introduced by L'Hermier des Plantes (1976) and developed by several authors, in particular by Lavit (1988) to analyze data organized in series of studies. A study consists of a matrix triplet $(\mathbf{X}_i, \mathbf{D}_{n_i}, \mathbf{D}_{m_i})$, $i = 1, \dots, k$, with \mathbf{X}_i being a data array and \mathbf{D}_{n_i} and \mathbf{D}_{m_i} being two weight matrices for objects and variables. To obtain a geometrical representation of the studies, Escoufier (1976), developed the operators $\mathbf{A}_i = \mathbf{X}_i \mathbf{D}_{m_i} \mathbf{X}_i^t \mathbf{D}_{n_i}$, $i = 1, \dots, k$. Let (θ_j, γ_j) , $j = 1, \dots, k$, be the pairs of eigenvalues and eigenvectors of matrix $\mathbf{S} = [S_{ij}]$, $i = 1, \dots, k; j = 1, \dots, k$, with $S_{ij} = \text{tr}(\mathbf{A}_i \mathbf{A}_j^t)$, $i = 1, \dots, k; j = 1, \dots, k$. Then the coordinates of the point representing the i -th study are the i -th components of vectors, $\theta_j^{\frac{1}{2}} \gamma_j$, $j = 1, \dots, k$. When these points are along the first axis the series has a common structure (e.g. Lavit, 1988).

To carry out inference for series with a common structure Oliveira and Mexia (2006) proposed the model $\mathbf{S} = \lambda \boldsymbol{\alpha} \boldsymbol{\alpha}^t + \mathbf{E}$ where $\mathbf{E} = \frac{1}{2}(\mathbf{E}^0 + \mathbf{E}^{0t})$ and $\text{vec}(\mathbf{E}^0)$ is normal, with null mean

vector and covariance matrix $\sigma^2 \mathbf{I}_{k,2}$. When preponderance $\tilde{\tau} = \frac{\theta_1^2}{\sum_{j=2}^k \theta_j^2} \geq 200$ and $k \leq 20$, simulation studies showed that we may use θ_1 and γ_1 to estimate λ and α (e.g. Oliveira and Mexia, 2004). The information summarized in a series of studies may then be contained in an adjusted structure vector $\tilde{\beta} = \theta_1 \gamma_1$ and a sum of squares $V_i = \sum_{j=2}^k \theta_j^2, i = 1, 2, \dots, k$. When we have matched series of studies two types of ANOVA like analysis may be carried out (e.g. Oliveira and Mexia, 2007):

- Transversal Analysis in which corresponding studies are compared. This analysis rests on homologue components of the adjusted structure vectors;
- Longitudinal Analysis in which the evolution of the series is compared. For this purpose contrasts on the components of the structure vectors are obtained and compared.

3 Case study

Portugal territory is classified into districts. These are divided into townships. Each township is further divided into parishes. For the purpose of this study four districts stretching from North to South along the coast and other four the border with Spain were selected. Local elections for the city board were considered in each township. These are by far the most important local elections since the head of the most voted list becomes the mayor. The ANOVA like approach was used to model the results of the elections carried out in 1985, 1989, 1993, 1997 and 2001.

Namely we wanted to study the influence of the factors:

- Administrative relevance - distinguishing the district capitals from the other townships ($C_p - 1^{st}$ factor);
- Longitude ($L_g - 2^{nd}$ factor) - distinguishing between the districts along the coast and those along the border with Spain;
- Latitude ($L_t - 3^{rd}$ factor) - with four levels: North, Center North, Center and South;

on corresponding studies and on the evolution of the local elections.

For each election in each township we had a data (objects x variables) array. The objects were the civil parishes and the variables were the main parties participating in the electoral process: Social Democratic Party (PSD); Portuguese Socialist Party (PS); Central Social Democratic/Popular Party (CDS/PP); United Democratic Coalition where the Communist Party has a predominant role (CDU); Other parties (OUT); Blank and invalid ballots (B/N); Abstentions (ABS).

The data array was D_n - centered. We considered identity weight matrices $\mathbf{D}_{n_i} = \mathbf{I}_{n_i}$ for parishes (objects) and $\mathbf{D}_{m_i} = \mathbf{I}_{m_i}$ for parties (variables).

4 Results and discussion

For the 16 series of studies we have the preponderance, $\tilde{\tau}$ and sum of square, V_i (Table 1).

Viana do Castelo (V.Cas), Ponta da Barca (P.Bar), Braganca (Bra), Alfandega da Fé (A.Fe), Aveiro (Ave), Santa Maria da Feira (S.M.F.), Guarda (Gua), Pinhel (Pin), Leiria (Lei), Alcobaça (Alc), Portalegre (Por), Elvas (Elv), Faro (Far), Silves (Sil), Beja (Bej) and Mertola (Mer).

TABLE 1. Values of the preponderance and sum of square.

	V.Cas	P. Bar	Bra	A. Fé	Ave	S.M.F	Gua	Pin
$\tilde{\tau}$	11401.4	29599.8	54427.3	7039.4	36052.9	24746.5	19042.7	14960.0
V_i	0.575	0.538	0.387	1.583	0.401	0.424	1.090	0.834
	Lei	Alc	Por	Elv	Far	Sil	Bej	Mer
$\tilde{\tau}$	131484.1	33739.7	98166.1	62641.7	121701.2	73105.6	392064.8	280434.8
V_i	0.101	0.458	0.111	0.336	0.165	0.129	0.043	0.073

Since our factors had 2, 2, and 4 levels we used the Yates algorithm to carry out the ANOVA-like Analysis. First we considered four 2 levels factors and them merged the last two ones. Firstly we considered transversal analysis in which the action of the factors on corresponding studies was studied. Nextly we carried out a longitudinal analysis in which the action of the studies in the evolution of the series was considered.

To express this evolution we used the contrasts L_1 , L_2 , and Q , to measure:

- evolution between 2001 and 1985 ($L_1 = \frac{1}{\sqrt{2}}(-1, 0, 0, 0, 1)$);
- evolution between 1997 and 1989 ($L_2 = \frac{1}{\sqrt{2}}(0, -1, 0, 1, 0)$);
- non-linearity of the evolution ($Q = \frac{1}{\sqrt{6}}(1, 0, -2, 0, 1)$).

In both analysis we tested the significance of the factors: C_p , L_g and L_t and of these interactions $C_p \times L_g$, $C_p \times L_t$, $L_g \times L_t$ and $C_p \times L_g \times L_t$. As we had 160 degrees of freedom for error the tests for C_p , L_g and $C_p \times L_g$ had 1 degrees of freedom for the numerator while the remaining had 3 degrees of freedom for the numerator. The F tests for Transversal Analysis and Longitudinal Analysis are presented in Table 2 and Table 3, respectively.

TABLE 2. F test for Transversal Analysis.

E. year	C_p	L_g	$C_p \times L_g$	L_t	$C_p \times L_t$	$L_g \times L_t$	$C_p \times L_g \times L_t$
1985	436.5	19394.0	2552.7	13632.1	47841.4	54093.8	45798.1
1989	65846.5	13377.9	6443.1	25319.6	26150.1	4201.9	84866.6
1993	10053.5	37382.5	41777.8	9131.6	34649.0	16213.9	51565.4
1997	3967.0	147754.8	9368.8	6938.1	41540.5	6309.8	102012.8
2001	0.67	105226.1	22006.7	11401.9	55847.2	9284.8	92735.3

TABLE 3. F test for Longitudinal Analysis.

Contracts	C_p	L_g	$C_p \times L_g$	L_t	$C_p \times L_t$	$L_g \times L_t$	$C_p \times L_g \times L_t$
L_1	235.7	17141.9	19778.1	1657.7	6468.5	52551.8	39400.8
L_2	18755.6	36114.8	15678.9	6607.9	19500.1	6733.4	9418.0
Q	8113.1	986.9	42786.8	15910.4	26878.5	3929.6	2216.4

Given the very highly preponderance almost all results are extremely highly significant. Thus the most interesting insights are given by decreasing of significance of the first factor (administrative relevance) as we consider successive elections. This would point out towards a political homogenization of the districts. Another relevant aspect is the fact that interaction predominate over effects. This points to a complex pattern of cross influence of the factors: administrative relevance, longitude and latitude. From the statistical view point we must stress the great condensation of information achieved. Thus a series of studies is condensed into a structure vector and a sums of squares of residues. Moreover the adjustment of the individual models was excellent as shown by the smallness of the sum of squares of residues.

References

- Escoufier, Y. (1973). Le Traitement des Variables Vectorielles. *Biometrics* **29**(4), 751 – 760.
- Lavit (1988) Lavit (1988) Lavit C., 1988. *Analyse Conjointe de Tableaux Quantitatifs*. Collection Méthods+Programmes, Masson, Paris. 91-262.
- L’Hermier des Plantes H. (1976) L’Hermier des Plantes H. (1976) L’Hermier des Plantes H., 1976. *Structuration des Tableaux à Trois Indices de la Statistique: théorie et application d’une méthode d’analyse conjointe*. Thèse de 3^{eme} cycle. Université de Montepplier II.
- Oliveira, M.M., Mexia, J. T. (2004). AIDS in Portugal: endemic versus epidemic forecasting scenarios for mortality. *International Journal of Forecasting* **20**, 131 – 137.
- Oliveira, M.M., Mexia, J. T. (2007). ANOVA like Analysis of Matched Series of Studies with a Common Structure. *Journal of Statistical Planning and Inference* **137**, 1862 – 1870.
- Oliveira and Mexia (2006) Oliveira Mexia 2006 Oliveira M. M., Mexia T. J., *Modelling series of studies with a common structure*. *Computational Statistoical & Data Analysis* (2006), doi:10.1016/j.csda.2006.11.003.

Alternative modelling approaches for the SF–36 health questionnaire

Inmaculada Arostegui¹ and Vicente Núñez-Antón²

¹ Departamento de Matemática Aplicada y Estadística. Universidad del País Vasco, Aptdo. 644, 48080 Bilbao, Spain, inmaculada.arostegui@ehu.es

² Departamento de Econometría y Estadística. Universidad del País Vasco, Lehendakari Agirre, 83, 48015 Bilbao, Spain, vicente.nunezanton@ehu.es

Abstract: Health Related Quality of Life (HRQoL) is an important indicator of health status and the Short Form – 36 (SF-36) is a generic instrument to measure it. Multiple Linear Regression (MLR) is often used to study the relationship of HRQoL with patients’ characteristics, though HRQoL outcomes tend to be not normally distributed, skewed and bounded. Bootstrap, ordinal methods and Beta-Binomial Regression (BBR) are tested as alternative methods to analyze the SF-36 and their performance is illustrated with an example and simulations. The BBR approach is shown to have a better behavior in the HRQoL domains with few ordered categories and a very similar one in the more continuous domains. A common technique of statistical analysis is preferable for all the domains present in the HRQoL instrument. Therefore, the BBR approach is recommended to analyze and interpret the effect of several explanatory variables on the SF–36.

Keywords: Beta-Binomial Model; Goodness of Fit; Health Related Quality of Life; SF–36.

1 Introduction and Motivation

HRQoL measures are becoming frequently used in clinical epidemiology and health services research. Statistical modelling is playing an important role on planning and analyzing HRQoL measures. One of the most widely used HRQoL questionnaires is the Short Form – 36 (SF–36), consisting of 36 questions and eight domains: *Physical Functioning* (PF), *Role Physical* (RP), *Bodily Pain* (BP), *General Health* (GH), *Vitality* (VT), *Social Functioning* (SF), *Role Emotional* (RE) and *Mental Health* (MH).

Since the early nineties, a lot of research has focused on the statistical analysis of HRQoL assessment in clinical trials. However, there are many settings where it is not possible to design a clinical trial and, thus, decisions on treatments are sometimes based on observational studies, where HRQoL has been measured as outcome and confounders are not controlled by the researcher during the study design. We concentrate primarily in this context. MLR is frequently used for analyzing the effect of several explanatory variables on HRQoL. HRQoL instruments comprise items on domains by addition, and thus, the measure for each domain is an ordered categorical scale. These scores tend to be skewed, *J*-shaped, or even *U*-shaped. Since HRQoL outcome measures may not meet the distributional requirements of MLR, other authors suggest alternative methods of analysis, like ordinal methods (Lall et al., 2002) or bootstrap

(Walters and Campbell, 2004). From a clinical point of view, it is desirable to have standard measurements of the different domains of a HRQoL instrument. Therefore, a common technique of statistical analysis is preferable for all the domains present in the HRQoL instrument.

We have analyzed HRQoL data from the SF-36 questionnaire in real patients with four approaches: MLR, bootstrap, Ordinal Logistic Regression (OLR) and BBR. The BBR is recommended as a method of analysis for the SF-36 outcome (Arostegui et al., 2007). Benefits of the BBR approach with respect to the other approaches have been discussed based on clinical consequences of conclusions addressed from the results of the data analysis and the easiness of the implementation in clinical epidemiology.

2 Methods of Analysis

We describe four methods of analysis of the SF-36 as the main outcome. Method 1 assumes the scores of the SF-36 are continuous and normally distributed. Method 2 assumes a continuous score, without making any other distributional assumption. Method 3 assumes the SF-36 is an ordered categorical outcome. The fourth method assumes that the SF-36 scores follow a beta-binomial distribution.

Method 1: Multiple Linear Regression

Data were analyzed using the MLR approach. Therefore, the model is given by:

$$y_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j + \epsilon_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, 8, \quad (1)$$

where y_{ij} represents the response, for subject i on domain j ; \mathbf{x}_i is a k -vector of explanatory variables observed on subject i ; $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jk})^T$ are k -vectors of unknown regression coefficients on domain j and $\epsilon_{ij} \sim$ independent $N(0, \sigma_j^2)$.

Method 2: Bootstrap

Data were analyzed using resampling bootstrap technique to estimate $\boldsymbol{\beta}_j$, their standard errors and confidence intervals in model (1).

Method 3: Ordinal Logistic Regression

The SF-36 scores have been recoded as ordinal scales that take integer values from 0 to n_j , $j = 1, \dots, 8$. Data were analyzed using the proportional odds ordinal regression model, a cumulative logit model that assumes that the odds ratio for each predictor is constant across all possible collapsings of the response variable. The model is given by:

$$\text{logit}(\pi_{ijk}) = \beta_{0jk} + \mathbf{x}_i^T \boldsymbol{\beta}_j, \quad k = 1, \dots, n_j - 1, \quad (2)$$

where $\pi_{ijk} = \theta_{ij1} + \dots + \theta_{ijk}$ represents cumulative probabilities of obtaining k points or less on the j th HRQoL domain and θ_{ijk} are the probabilities of obtaining k points on the j th HRQoL domain, both for subject i . The specific intercepts for each cumulative logit are β_{0jk} , and \mathbf{x}_i and $\boldsymbol{\beta}_j$ are defined as in model (1).

Method 4: Beta-Binomial Regression

Fitting to a beta-binomial distribution was performed based on the ordinal scale used for method 3. Data were analyzed using the BBR approach. Therefore, the model is given by:

$$\text{logit}(\theta_{ij}) = \mathbf{x}_i^T \boldsymbol{\beta}_j + \sigma_j u_i, \quad i = 1, \dots, N, \quad j = 1, \dots, 8, \quad (3)$$

where θ_{ij} is the probability of obtaining one point on the j th HRQoL domain for subject i , and \mathbf{x}_i and $\boldsymbol{\beta}_j$ are defined as in model (1). The u_i 's are i.i.d. random variables with mean zero and variance one, and $\sigma_j > 0$.

Finally, a simulation study was conducted to be able to assess the robustness of the two methods with distributional constraints, methods 1 and 4. Simulated values of the SF-36 have been obtained using the normal and the beta-binomial distribution. Simulated data were also analyzed using these approaches. In each case, a separate analysis was performed for each of the eight HRQoL domains.

3 Main Results and Conclusions

Goodness-of-fit to a normal distribution was rejected for seven out of the eight domains of HRQoL. Comparison between results obtained from real and simulated data showed that there were differences in the significance of a covariate in the model and in the magnitude of the effect of such covariate in HRQoL. Therefore, we conclude that data analysis of HRQoL with the MLR approach, under non normality of the response variable, does not support clinical conclusions obtained from it. The bootstrap method, when compared to the MLR approach, produced the same significant covariates, as well as very similar standard errors for the regression coefficient estimates.

After recoding to ordinal scales, the number of categories was more than 8 in six out of eight domains. Thus, it is not worthy to use OLR with outcome variables which have 9, 11 or 21 ordered categories. Therefore, OLR was only applied to two SF-36 domains. The proportional odds assumption was not rejected for RP, whereas it was for RE for some covariates. Therefore, the proportional odds model was applied to RP and partial proportional odds model was applied to RE.

Goodness-of-fit to a beta-binomial distribution was rejected for one of the eight HRQoL domains. There were very few differences between the results obtained from real and simulated data when the BBR model was used.

Results from BBR and OLR were very similar for those domains where both approaches were used. The BBR and the OLR approaches were more powerful than the MLR and the bootstrap approaches on detecting statistically significant covariates for these two domains. Moreover, the magnitude of the relationship between the covariates and HRQoL, and its interpretation, is quite different depending upon the selected method of analysis and different conclusions could be reached based on which one is considered clinically significant by researchers in the area. The main advantage of BBR over OLR is that the number of parameters to be estimated is smaller for the former. For the remaining HRQoL domains, the BBR approach detects the same significant covariates as MLR and bootstrap and, although the magnitude of the relationship between the covariates and these HRQoL domains is quite differently interpreted, it does not affect clinical significance.

In real applications of HRQoL studies, a researcher is interested not only on detecting statistically significant relationships between HRQoL and other covariates, but also in the interpretation of the results and the clinical significance of such relationships. Considering that the eight domains of the SF-36 are analyzed all together, the same method of analysis is preferable for all of them. In conclusion, the BBR approach is recommended to analyze HRQoL data measured by the SF-36 questionnaire.

Our work has concentrated primarily on observational studies. It does not mean that BBR is not a valid methodology to analyze data from clinical trial, but the performance of standard methods in this setting has been broadly studied, and the benefits of using a more complicated method of analysis should be tested against simplicity in this particular context. Strictly speaking, our conclusions only apply to the SF-36, further work is required to test the performance of BBR for other HRQoL outcomes.

References

- Arostegui, I., Núñez-Antón, V. and Quintana, J.M. (2007). Analysis of the Short Form - 36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine* **26**, 1318-1342.
- Lall, R., Campbell, M. J., Walters, S. J., Morgan, K. and MRC CFAS (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical Methods in Medical Research* **11**, 49-67.
- Walters, S. J. and Campbell, M. J. (2004). The use of bootstrap methods for analyzing health related quality of life outcomes (particularly the SF-36). *Health and Quality of Life Outcomes* **2:70**.

A Bayesian Approach to the Estimation of Technical Efficiency in Health Care Foodservice Operations

A. Assaf¹ and K.M. Matawie¹

¹ University of Western Sydney, PO Box 10, Kingswood NSW 2747, Australia,
a.assaf@uws.edu.au, k.matawie@uws.edu.au

Abstract: This study uses a Bayesian stochastic frontier model to analyze the level of technical efficiency in health care foodservice operations. Various distributional assumptions are assumed for the random component part of the model including the half-normal, gamma, truncated normal and exponential distributions. The results from a sample of 101 health care foodservice operations show the existence of the technical inefficiency in the sample and identified the statistical differences between the average efficiency generated by the various distributional assumptions. Finally, it was concluded that the rank of technical efficiency is statistically invariant to the various distributions.

Keywords: Bayesian stochastic frontier; technical efficiency

1 Statement of the Problem

The important role of efficiency in the health care foodservice sector has been widely addressed in the literature. Different methods for assessing economic performance have been proposed. In general, most measures are calculated as simple ratios and key performance indicators such as food and labor cost per meal. A weakness of these approaches is that they are calculated using only a subset of the data available on the firm. This is problematic because a foodservice operation may perform well using one measure (e.g. food cost per employee) but badly using another (e.g. labor cost per meal). What is needed is a single measure of relative performance (or efficiency) that is calculated using all the input and output variables available on the firm. Stochastic Frontier is a statistical technique for obtaining such a measure - it is applied in this study to combine all the input and output variables of health foodservice operations into a single measure of productive efficiency that will take a value between zero (implying the firm is performing extremely poorly) and one (implying the firm is fully efficient).

2 Methodology

The stochastic frontier model denoted in logs can be expressed as:

$$\ln(y_i) = \ln x_i \beta + v_i + u_i \quad (1)$$

where y_i represents the output of the i -th firm, x_i denotes an input vector, β is a vector of unknown parameters, v_i depicts random error commonly represented as

$N(0, \sigma_u^2)$ and u_i is a non-negative random error introduced to account for technical inefficiency of firms and it is commonly selected from the following set of alternatives:

Exponential: $u \sim \text{Exp}(\sigma_u)$ for $\sigma_u > 0$

Half-normal: $u \sim N_+(0, \sigma_u^2)$ for $\sigma_u^2 > 0$

Truncated normal: $u \sim N_+(\mu, \sigma_u^2)$ for $\mu \in R$ and $\sigma_u^2 > 0$

Gamma: $u \sim \text{Gamma}(m, \sigma_u^2)$ for $m > -1$ and $\sigma_u > 0$

Aigner and Chu (1968) presented the non-stochastic frontier model (i.e., $v=0$) and they also used the exponential model to capture the negative deviations from the frontier. Aigner et al (1977) adopted the exponential form for their seminal work on the stochastic frontier model and they also introduced the half-normal model with Meeusen and van den Broeck (1977). The exponential and half-normal models have modes at zero which may be unrealistic for many production functions. Stevenson (1980) proposed the truncated normal model in order to allow for a positive mode in the distribution of u . He also introduced the gamma model which was later extended by Greene (1980). Further, the truncated normal model reduces to the half-normal model if $\mu=0$ and the gamma model reduces to the exponential model if $m=0$.

Estimators based on the above models have mainly been based on the maximum likelihood method, and studies that examined the distributional assumption of on sample mean efficiencies have not converged to the same answer (Kumbhakar and Lovell, 2000). We estimate in this study the frontier following the four distributional assumptions, but improve on the Maximum likelihood assumption by using Bayesian statistics. There are several advantages to incorporating the Bayesian approach rather than using the classical approach. First, it is possible to include "prior" information about parameters in our inferences. Second, sampling theory inferences in stochastic frontier models are based on asymptotic standard errors, whereas Bayesian can be used when working with finite sample.

In this section we limit the Bayesian discussion of the stochastic production frontier to the exponential model in which we assume that the inefficiency term follows an exponential distribution:

$$p(\mathbf{u}|\lambda) = \prod_{i=1}^l \lambda^{-1} \exp(-u_i/\lambda) \quad (2)$$

where λ is a shape parameter which defines both the mean and the variance of the exponential distribution. For a Bayesian treatment of the other distributions used in this study (gamma, truncated and half-normal distributions) see Koop et al. (1997). More specifically in our model, we choose a flat (constant) prior for β (i.e. that is, we assume no prior knowledge about these parameters) and gamma prior for σ^2 and λ . As van den Broeck et al. (1994) noted, choosing an informative prior λ^{-1} and σ^2 ensures that the posterior is proper, and defines the complete prior as:

$$p(\beta, \sigma^2, \lambda^{-1}) = \sigma^{-2} f_G(\lambda^{-1} | 1, \ln(r^*)) \quad (3)$$

where $f_G(\cdot | a, b)$ is the gamma distribution with a degree of freedom and mean b , and $f_N(\cdot | a, \mathbf{b})$ indicates a multivariate normal distribution with mean a and a covariance matrix \mathbf{b} . The posterior corresponding to this prior is completely intractable and must be analyzed using simulation methods. In particular a Gibbs sampler with data augmentation can be set-up for this model involving the following conditional distributions:

$$p(\beta | \mathbf{y}, \sigma^{-2}, \mathbf{u}, \lambda) = f_N(\beta | \hat{\beta}, h^{-1}(\mathbf{X}'\mathbf{X})^{-1}) \times I(\beta) \quad (4)$$

$$p(\sigma^{-2} | \mathbf{y}, \beta, \mathbf{u}, \lambda) = f_G(\sigma^{-2} | I/(\mathbf{y} + \mathbf{u} - \mathbf{X}'\beta)(\mathbf{y} + \mathbf{u} - \mathbf{X}'\beta), I) \quad (5)$$

$$p(\lambda^{-1}|\mathbf{y}, \beta, \sigma^{-2}, \mathbf{u}) = f_G(\lambda^{-1}|(I+1)/\mathbf{u}'\mathbf{j}_I - \ln(\tau^*), 2(I+1)) \quad (6)$$

$$p(u_i|\mathbf{y}, \beta, \sigma^{-2}, \lambda) = f_N(u_i|\mathbf{x}_i'\beta - y_i - (h\lambda)^{-1}, h^{-1}) \times I(u_i) \quad (7)$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_I)'$ is an $I \times K$ matrix, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} + \mathbf{u})$ is a least square estimator, and $I(u_i)$ is an indicator function that takes the value 1 if $u_i \geq 0$ and 0 otherwise. Note that the technical efficiency (TE) of the i -th firm is measured as $TE = \exp(-u_i)$. Using these conditional densities the Gibbs sampler follows. As the iterations approaches to infinity the Gibbs sampling methods converges to the actual joint posterior density function. In this paper, we generate 25,000 parameter vectors and we drop the first 5,000 to avoid sensitivity of starting values.

3 Data Collection

The sample for this study contained data from 101 health care foodservice operations representing both the private and the public sector. All data were collected by means of a web questionnaire. In line with the literature we defined three inputs and one output. Inputs variables are number of full time equivalent employees, amount of energy, and total square are of the foodservice department. As output we defined the volume of production, measured as the annual number of meals produced. Additionally, we also included in our model three environmental variables (age of equipment, degree of readiness of raw materials and skill level of employees) which are in nature neither inputs nor inputs but deemed to indirectly the efficiency of the health care foodservice operations.

More specifically, the logarithmic stochastic frontier model specified is the cross-sectional case is defined as:

$$\ln q_i = \beta_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \beta_3 \ln x_{3i} + \beta_4 \ln z_{1i} + \beta_5 \ln z_{2i} + \beta_6 \ln z_{3i} + v_i + u_i \quad (8)$$

where q_i is the number of meal produced per year; x_{1i} = the number of full time equivalent employees; x_{2i} = the amount of energy; x_{3i} = the total square area of the department; z_{1i} = the age of equipment; z_{2i} = the skill level of employees; z_{3i} = the degree of readiness of raw materials. β_0 through β_3 are input coefficients, and β_4 through β_6 are environmental variables coefficients. The disturbance v_i represents the symmetric statistical noise component and u_i is the one-sided inefficiency component.

4 Results

The results showed that the posterior means across the different distributions of the inefficiency term are almost of equal magnitude, which implies some consistency in the estimation of the different frontier models. An ANOVA comparison between the efficiency estimates ($F = 17.63$) indicate that the efficiency estimates yielded by the different distributions are significantly different at the 1 % confidence level. The effect from the distributional assumptions was examined further using the Spearman's rank order correlation between the efficiency rankings derived from the four models. All of the estimated coefficients were found to be significantly different from zero at the 1 % level indicating that the rank of each foodservice operation derive from applying the different distributions is similar. A combination of ANOVA and Spearman's rank order correlation coefficient yields to the conclusion that the efficiency estimates yielded by

the different distributions follow the same pattern across the foodservice operations, making it therefore feasible to draw inference from the efficiency results of this study.

References

- Aigner, D. J. and Chu, S. F. (1968). On Estimating the industry production function. *American Economic Review* **58**(4), 826-836.
- Greene, W. H. (1980). Maximum likelihood estimation of econometric frontier functions. *Journal of Econometrics* **13**(1), 27-56.
- Koop, G., Osiewalski, J. and Steel, M. F. J. (1997). Bayesian efficiency analysis through individual effects: Hospital cost frontiers. *Journal of Econometrics* **76**(1-2), 77-105.
- Kumbhakar, S. and Lovell, C.A.K. (2000). *Stochastic frontier analysis*. New York, Cambridge University Press.
- Meeusen, W., and Van Den Broeck, J. (1977). . Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. *International Economic Review* **18**(2), 435-444.
- Stevenson, R. E. (1980). Likelihood functions for generalized stochastic frontier estimation. *Journal of Econometrics* **13**(1), 57-66.
- Van Den Broeck, J., Koop, G., Osiewalski, J. and Steel, M.F.J. (1994). Stochastic frontier models : A Bayesian perspective. *Journal of Econometrics* **61**(2), 273-303.

A Bayesian Hierarchical spatial model for the bioclimatic classification of Cyprus island

A. Barber¹, X. Barber², A. M. Mayoral² and J. Morales²

¹ IDENTIA Institut

² Applied Statistical Unit-CIO, Miguel Hernández University

Abstract: We propose a Bayesian geostatistical model to establish a Bioclimatic classification of the Island of Cyprus, based the current Worldwide Bioclimatic Classification System used to define and describe the bioclimates and bioclimatics belts (thermo-ombrotypes).

Keywords: Bayesian Hierarchical Model; Bioclimatic classification; Geostatistics.

1 Introduction

Bioclimatology or Phytoclimatology are important sciences for the comprehension of the close relationship between climate and vegetation, and therefore, the plant landscape. Thus, the better knowledge of the interrelationship climate-vegetation we get, the better management of plant resources, landscape, and environment can we unmistakably develop.

The main aim of this work is to establish a Bioclimatic classification of the Island of Cyprus, based on the current Worldwide Bioclimatic Classification System used to define and describe the bioclimates and bioclimatics belts (thermo-ombrotypes). Relevance of this approach concerns the richness and flora diversity of Cyprus.

2 Data and Bioclimatic index

Available data consists on different measures from 58 meteorological stations distributed all over the island (mean monthly and mean annual precipitation, mean daily maximum, mean daily minimum, mean monthly maximum, mean monthly minimum, etc...). Geographical UTM coordinates and altitude are also available.

From the information above, several bioclimatic indexes derive. One of them is the Ombrothermic Index (OI):

$$OI = (Pp/Tp) \times 10,$$

where Pp is the Annual Positive Precipitation (sum –in tenths of Celsius degrees– of the monthly mean temperature of those months whose average temperature is higher than 0C), and Tp is the Annual Positive Temperature (total average precipitation of those months whose mean temperature is higher than 0C).

Based on ranges for the values of the OI , a classification of the ombrotypes or horizons emerges: Ultrahyperarid (<0.2), Hyperarid (0.2-0.4), Arid (0.4-1.0), Inferior Semiarid (1.0-1.5), Superior Semiarid (1.5-2.0), Inferior Dry (2.0-2.8), Superior Dry (2.8-3.6), Inferior Subhumid (3.6-4.8), Superior Subhumid (4.8-6.0), Inferior Humid (6.0-9.0).

3 The model

We propose a Hierarchical Spatial Bayesian model to predict the OI on the whole island, just by knowing its value at the 58 meteorological stations. We assume a Normal model to relate OI with the coordinates and altitude for each meteorological station. The geostatistical model is given by:

$$\begin{aligned} \log(OI)|\theta, W &\sim N(X\beta + W, \tau^2 I) \\ W|\sigma^2, \phi &\sim N(0, \sigma^2 H(\phi)) \\ &p(\beta)p(\tau^2)p(\sigma^2)p(\phi), \end{aligned}$$

where $X = (1, \textit{Elevation})$, W is the vector of spatial random effects, H is a correlation matrix between spatial locations with isotropic correlation function ρ , τ^2 is the nugget (non-spatial variance), σ^2 is the partial sill (spatial variance) and ϕ is the decay parameter. Also, we define the range (R) as $1/\phi$ and the effective range (ER) as the distance at which the correlation has dropped to only 0.05. See Banerjee, Carlin and Gelfand (2004) for a detailed explanation of the parameters involved in this model.

We assume independent priors for all the parameters: $p(\beta) \propto 1$, $p(\sigma^2) = IG(2, 0001)$, $p(\tau^2) = IG(2, 0001)$, $p(\phi) = LUnif(5e - 5, 0.003)$, where IG denotes the inverse gamma distribution and $Lunif$ denotes de log-uniform distribution. The prior for ϕ induced the approximate prior for ER as $LUnif(1000, 60000)$.

We consider the Matern family of correlations as a general correlation function (see Banerjee, Carlin and Gelfand, 2004). This family is indexed by a parameter (ν) controlling the smoothness of the realized random field, and special cases of the above are the exponential ($\nu = 0.5$) and the Gaussian ($\nu \rightarrow \infty$). Different values $\nu = 0.5; 1; 1.5; \infty$ are used in order to compare models.

3.1 Selection criteria

Gelfand and Ghosh (1998) criteria, denoted by D , is used to select the degree of smoothness of the realized random field controlled by the ν parameter. The model with the smallest D is preferred.

4 Results and Conclusions

In Table 1 we present the values of D -criteria for the different correlation functions considered. The Gaussian correlation function is preferred, and thus used for inferences. In Table 2 posterior distributions of the parameters involved are summarized. Figure 1 represents the predictive posterior median of OI at any point of the Cyprus island for each of the four considered models. The Bayesian framework provides interesting summaries as probabilities for any location of belonging to some specific ombrotype, which are displayed in Figure 2, for the model with Gaussian correlation function.

TABLE 1. D values for the different correlation functions considered.

	Exponential	Matern 1	Matern 1.5	Gaussian
D	1.132	1.077	1.099	1.049

TABLE 2. Posterior median and credible region (CR) with probability 0.95 for parameters in model with Gaussian correlation function.

	Median	CR 0.95		Median	CR 0.95
β_0	5.18e-01	(4.16e-01,6.26e-01)	σ^2	2.68e-02	(1.59e-02,4.70e-02)
β_1	8.76e-04	(7.01e-04,1.05e-03)	ϕ	5.36e-05	(5.02e-05,6.71e-05)
τ^2	1.01e-02	(6.43e-03,1.64e-02)	ER	32262	(24439,34504)

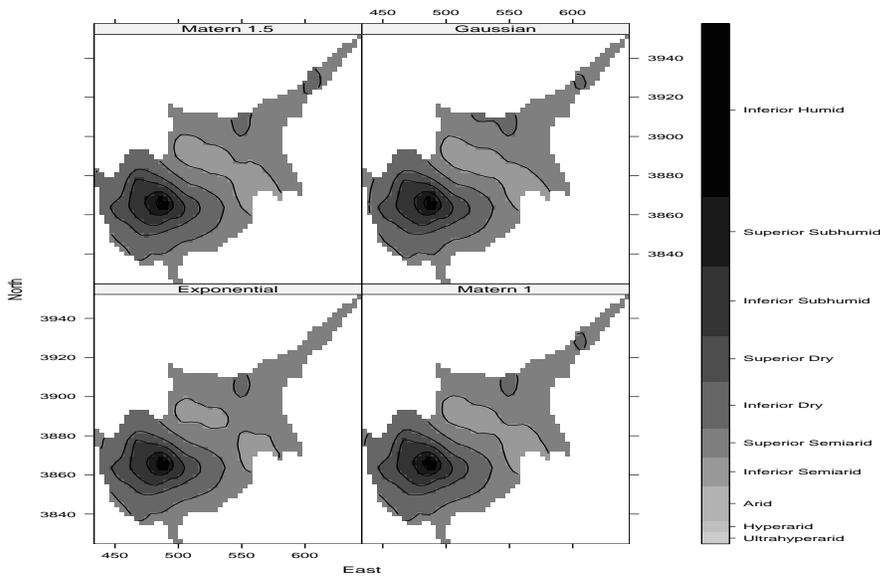


FIGURE 1. Image-contour displaying the predictive posterior median of OI for each correlation function.

Acknowledgments: This research was supported by the Spanish Ministry of Education and Science, under Grant MTM2004-03290, and by the Autonomous Valencia Government, under Grant ACOMP06/2005.

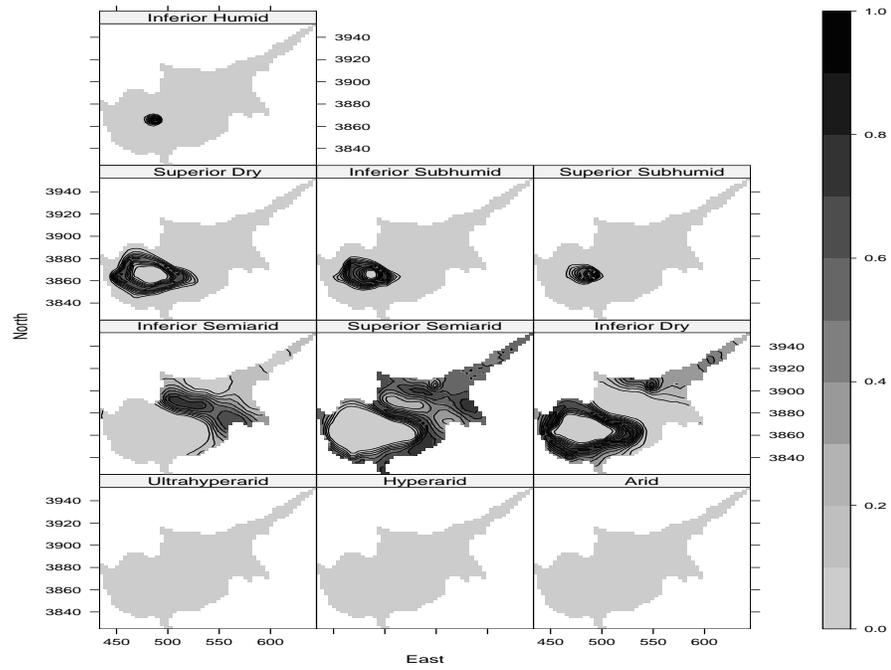


FIGURE 2. Image-contour displaying the posterior probability of each ombrotype.

References

- Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical modelling and analysis for spatial data*. Boca Raton: Chapman and Hall.
- Gelfand, A.E., and Ghosh (1998). Model Choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1-11.

Compositional Time Series: A First Approach

C. Barceló-Vidal¹, L. Aguilar² and J.A. Martín-Fernández¹

¹ Dept. Informàtica i Matemàtica Aplicada, Campus de Montilivi, Univ. de Girona, E-17071 Girona, Spain

² Dept. de Matemàtiques, Escuela Politécnica, Univ. de Extremadura, E-10071 Cáceres, Spain

Abstract: Compositional time series, i.e., multivariate time series of vectors of D proportions, arise in many areas of application where the focus of attention is on the relative, rather than the absolute, values of their components. Such series are characterized by components which are positive and sum to one at each instance in time. Although data of this type constitute multivariate time series, standard modelling techniques are not applicable due to the positivity of the components and the constant sum constraint. In other words, problems arise because its sample space is not the D -dimensional real space, nor the positive real space, but the $(D - 1)$ -dimensional simplex space. We consider basic concepts regarding the Euclidean structure of the simplex, and the alr, clr and ilr transformations on it are introduced to present compositional ARIMA models.

Keywords: ARIMA models, Compositional time series, Simplex

1 The simplex \mathcal{S}^D as a compositional space

1.1 The simplex as a real vector space

A D -part composition $\mathbf{x} = (x_1, \dots, x_D)'$ is any element of the simplex

$$\mathcal{S}^D = \{(x_1, \dots, x_D)' : x_1 > 0, \dots, x_D > 0; x_1 + \dots + x_D = 1\}.$$

Basic operations on \mathcal{S}^D have been introduced by Aitchison (1986) and Barceló-Vidal et al (2002). The *perturbation* operation is defined as

$$\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(x_1 x_1^*, \dots, x_D x_D^*)' \text{ for any } \mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D,$$

and the *power transformation*, defined for any $\mathbf{x} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$ as

$$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, \dots, x_D^\alpha)',$$

where \mathcal{C} denotes the *closure* operator defined for any $\mathbf{z} \in \mathbb{R}_+^D$ as $\mathcal{C}\mathbf{z} = \mathbf{z} / \sum_{i=1}^D z_i$. In this manner $(\mathcal{S}^D, \oplus, \odot)$ becomes a real vector space of dimension $D - 1$. The composition $\mathbf{0}_{\mathcal{C}} = (1/D, \dots, 1/D)'$ is the neutral (zero) element, and the inverse (opposite) of $\mathbf{x} \in \mathcal{S}^D$ is the composition $\mathbf{x}^{-1} = \mathcal{C}(1/x_1, \dots, 1/x_D)'$.

Provided that $(\mathcal{S}^D, \oplus, \odot)$ is a real vector space, it can be viewed as an affine space when the group (\mathcal{S}^D, \oplus) operates on \mathcal{S}^D as a group of transformations. Perturbations in the compositional space plays the same role as translations in the real space. The

assumption that the group of perturbations is the operating group on the compositional space is the keystone of the methodology introduced by Aitchison (1986). In fact, it means accepting that the "difference" between two compositions \mathbf{x} and \mathbf{x}^* is the composition $\mathbf{x} \ominus \mathbf{x}^* = \mathcal{C}(x_1/x_1^*, \dots, x_D/x_D^*)'$, based on the ratios x_j/x_j^* between parts instead of on the subtraction $x_j^* - x_j$.

1.2 Transformations on the simplex

Let $\mathcal{A}_{D \times D}$ denote the family of all real $D \times D$ matrices such that $\mathbf{A}\mathbf{1}_D = \mathbf{A}'\mathbf{1}_D = \mathbf{0}_D$. Let $\mathbf{x} \in \mathcal{S}^D$ and $\mathbf{A} \in \mathcal{A}_{D \times D}$. We define the *product* $\mathbf{A} \odot \mathbf{x}$ as

$$\mathbf{A} \odot \mathbf{x} = \mathcal{C} \left(\prod_{j=1}^D x_j^{a_{1j}}, \dots, \prod_{j=1}^D x_j^{a_{Dj}} \right)'.$$

The function $\mathbf{x} \rightarrow \mathbf{A} \odot \mathbf{x}$ is an endomorphism of the vector space $(\mathcal{S}^D, \oplus, \odot)$. Moreover, any endomorphism of \mathcal{S}^D can be written in this form. The matrix associated to identity endomorphism is the well-known *centering matrix* $\mathbf{G}_D = \mathbf{I}_D - D^{-1}\mathbf{J}_D$ of order $D \times D$. The *additive logratio transformation* of index j ($j = 1, \dots, D$)—denoted by alr_j —is the one-to-one transformation from \mathcal{S}^D to \mathbb{R}^{D-1} defined as $\mathbf{x} \rightarrow \mathbf{y} = \text{alr}_j \mathbf{x} = \log(\mathbf{x}_{-j}/x_j)$ where \mathbf{x}_{-j} denotes the vector \mathbf{x} with the component x_j deleted. In particular, we use alr —without any subindex—to denote the transformation alr_D . The inverse transformation of alr_j is the well known *additive logistic transformation*.

The *centered* (or *symmetric*) *logratio transformation*—denoted by clr —is the function from the compositional space \mathcal{S}^D to \mathbb{R}^D , defined by $\mathbf{x} \rightarrow \mathbf{z} = \text{clr} \mathbf{x} = \log(\mathbf{x}/g(\mathbf{x}))$, where $g(\mathbf{x})$ is the geometric mean of the components of \mathbf{x} , i.e., $g(\mathbf{x}) = (x_1 x_2 \dots x_D)^{1/D}$. This transformation maps \mathcal{S}^D in the subspace $V = \{\mathbf{z} \in \mathbb{R}^D : z_1 + \dots + z_D = 0\}$ of \mathbb{R}^D , which can be seen to be a hyperplane through the origin of \mathbb{R}^D , orthogonal to $\mathbf{1}_D$ (vector of units). This subspace has dimension $D - 1$. Let $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ be any orthonormal basis of V , and let \mathbf{V} be the $D \times (D - 1)$ matrix $[\mathbf{v}_1 : \dots : \mathbf{v}_{D-1}]$. Then, the *isometric logratio transformation*—denoted by ilr_V —associated with this matrix \mathbf{V} , is the one-to-one transformation from \mathcal{S}^D to \mathbb{R}^{D-1} which assigns to each composition \mathbf{x} the components of $\text{clr} \mathbf{x}$ in the basis $\mathbf{v}_1, \dots, \mathbf{v}_{D-1}$ of V . It can be proved that $\mathbf{x} \rightarrow \mathbf{u} = \text{ilr}_V \mathbf{x} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F} \log \mathbf{x}$, for any $\mathbf{x} \in \mathcal{S}^D$, where \mathbf{F} is the $(D - 1) \times D$ matrix $[\mathbf{I}_{D-1} : -\mathbf{1}_{D-1}]$.

It is very important to emphasize that all these transformations— alr_j , clr , ilr_V , and its inverses—are one-to-one linear transformations between the compositional vector space $(\mathcal{S}^D, \oplus, \odot)$ and the real vector space $(\mathbb{R}^{D-1}, +, \cdot)$ (or $V \subset \mathbb{R}^D$) with the natural structure. Vectors $\mathbf{u} = \text{ilr}_V \mathbf{x}$, $\mathbf{y} = \text{alr}_D \mathbf{x}$ and $\mathbf{z} = \text{clr} \mathbf{x}$ associated with the same composition $\mathbf{x} \in \mathcal{S}^D$ are related by the following linear relationships expressed in matrix form:

1. $\mathbf{u} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{y}$, and $\mathbf{u} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F}\mathbf{z}$.
2. $\mathbf{y} = \mathbf{F}\mathbf{V}\mathbf{u}$, and $\mathbf{y} = \mathbf{F}\mathbf{z}$.
3. $\mathbf{z} = ((\mathbf{F}\mathbf{V})^{-1}\mathbf{F})'\mathbf{u}$, and $\mathbf{z} = \mathbf{F}'\mathbf{H}^{-1}\mathbf{y}$, where \mathbf{H} is the $(D - 1) \times (D - 1)$ matrix $\mathbf{F}\mathbf{F}' = \mathbf{I}_{D-1} + \mathbf{J}_{D-1}$, with $\mathbf{J}_{D-1} = \mathbf{1}_{D-1}\mathbf{1}'_{D-1}$.

1.3 The simplex as a metric space

The one-to-one linear transformation clr allows one to transfer the real Euclidean structure defined on \mathbb{R}^{D-1} to \mathcal{S}^D . Thus the compositional norm (\mathcal{C} -norm) of $\mathbf{x} \in \mathcal{S}^D$ is equal to the Euclidean norm in \mathbb{R}^D of the clr -transformed vector, i.e., $\|\mathbf{x}\|_{\mathcal{C}} = \|\text{clr } \mathbf{x}\|$, and the \mathcal{C} -distance between two compositions $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ is given by the \mathcal{C} -norm of the difference $\mathbf{x} \ominus \mathbf{x}^*$. Thus the \mathcal{C} -distance just defined converts \mathcal{S}^D into a Euclidean space and the transformation clr is the natural isometry between \mathcal{S}^D and the subspace V of \mathbb{R}^D . Moreover, as the $D - 1$ columns of matrix \mathbf{V} used in the transformation ilr_V constitute, by definition, an orthonormal basis of V , this transformation is also an isometry between \mathcal{S}^D and \mathbb{R}^{D-1} . The same cannot be said for the additive logratio transformations alr_j .

1.4 The covariance structure of the simplex

Let \mathbf{x} be a random D -part composition defined on \mathcal{S}^D . According to the metric structure defined on \mathcal{S}^D , the \mathcal{C} -mean of \mathbf{x} , symbolized by $\boldsymbol{\xi}$ or $E_{\mathcal{C}}\{\mathbf{x}\}$, is defined as $\boldsymbol{\xi} = \text{clr}^{-1}E\{\text{clr } \mathbf{x}\}$ and the \mathcal{C} -covariance matrix $\boldsymbol{\Sigma}^{\mathcal{C}}$ of \mathbf{x} as

$$\boldsymbol{\Sigma}^{\mathcal{C}} = \left[\text{cov} \left\{ \log \frac{x_i}{g(\mathbf{x})}, \log \frac{x_j}{g(\mathbf{x})} \right\} \right]_{i,j=1}^D = [\sigma_{ij}^{\mathcal{C}}]_{i,j=1}^D,$$

i.e., by the covariance matrix $\boldsymbol{\Sigma}^{\mathcal{Z}}$ of the random vector $\mathbf{z} = \text{clr } \mathbf{x}$, known as *centred logratio matrix*. The consequent singularity of the distribution $\mathbf{z} = \text{clr } \mathbf{x}$ is reflected in the singularity of $\boldsymbol{\Sigma}^{\mathcal{C}}$, since $\boldsymbol{\Sigma}^{\mathcal{C}} \mathbf{1}_D = \mathbf{0}_D$.

Aitchison (1986) defines other matrices to determine the \mathcal{C} -covariance structure of a random composition \mathbf{x} . The *variation matrix* \mathbf{T} is defined as

$$\mathbf{T} = [\text{var} \{\log (x_i/x_j)\}]_{i,j=1}^D = [\tau_{ij}]_{i,j=1}^D,$$

and the *logratio covariance matrix* $\boldsymbol{\Sigma}^{\mathcal{Y}}$ as

$$\boldsymbol{\Sigma}^{\mathcal{Y}} = \left[\text{cov} \left\{ \log \frac{x_i}{x_D}, \log \frac{x_j}{x_D} \right\} \right]_{i,j=1}^{D-1} = [\sigma_{ij}^{\mathcal{Y}}]_{i,j=1}^{D-1},$$

i.e., by the covariance matrix of the random vector $\mathbf{y} = \text{alr } \mathbf{x}$ on \mathbb{R}^{D-1} . It is clear that $\boldsymbol{\Sigma}^{\mathcal{Y}}$ will depend on the denominator used in the alr -transformation.

Finally, the covariance matrix of the random vector $\mathbf{u} = \text{ilr } \mathbf{x}$ on \mathbb{R}^{D-1} will be denoted by $\boldsymbol{\Sigma}^{\mathcal{U}} = [\sigma_{ij}^{\mathcal{U}}]_{i,j=1}^{D-1}$. This covariance matrix will depend on the matrix \mathbf{V} used in the ilr -transformation. Although the \mathcal{C} -covariance structure of \mathbf{x} is given by $\boldsymbol{\Sigma}^{\mathcal{C}}$, the relationships between all these matrices allow one to deduce $\boldsymbol{\Sigma}^{\mathcal{C}}$ from any of the other matrices.

1.5 Joint distribution on the simplex

Let $(\mathbf{x}_1, \mathbf{x}_2)$ be a bivariate random compositional vector defined on $\mathcal{S}^D \times \mathcal{S}^D$. If $\boldsymbol{\xi}_i = E_{\mathcal{C}}\{\mathbf{x}_i\}$ ($i = 1, 2$), the \mathcal{C} -covariance matrix $\boldsymbol{\Gamma}^{\mathcal{C}}(\mathbf{x}_1, \mathbf{x}_2) = [\gamma_{ij}^{\mathcal{C}}]_{i,j=1}^D$ of $(\mathbf{x}_1, \mathbf{x}_2)$ is defined

as

$$\mathbf{\Gamma}^C(\mathbf{x}_1, \mathbf{x}_2) = \left[\mathbb{E} \left\{ \left(\log \frac{x_{1i}}{g(\mathbf{x}_1)} - \log \frac{\xi_{1i}}{g(\boldsymbol{\xi}_1)} \right) \left(\log \frac{x_{2j}}{g(\mathbf{x}_2)} - \log \frac{\xi_{2j}}{g(\boldsymbol{\xi}_2)} \right) \right\} \right]_{i,j=1}^D.$$

Therefore, $\mathbf{\Gamma}^C(\mathbf{x}_1, \mathbf{x}_2)$ coincides with the covariance matrix $\mathbf{\Gamma}^Z(\mathbf{z}_1, \mathbf{z}_2)$ of $(\mathbf{z}_1, \mathbf{z}_2) = (\text{clr } \mathbf{x}_1, \text{clr } \mathbf{x}_2)$ defined on $V \times V \subset \mathbb{R}^D \times \mathbb{R}^D$. The matrix $\mathbf{\Gamma}^C(\mathbf{x}_1, \mathbf{x}_2)$ is not symmetric but is singular because $\mathbf{\Gamma}^C \mathbf{1}_D = (\mathbf{\Gamma}^C)' \mathbf{1}_D = \mathbf{0}_D$.

We denote by $\mathbf{\Gamma}^Y(\mathbf{y}_1, \mathbf{y}_2) = [\gamma_{ij}^Y]_{i,j=1}^{D-1}$ the covariance matrix of $(\mathbf{y}_1, \mathbf{y}_2) = (\text{alr } \mathbf{x}_1, \text{alr } \mathbf{x}_2)$, and by $\mathbf{\Gamma}^U(\mathbf{u}_1, \mathbf{u}_2) = [\gamma_{ij}^U]_{i,j=1}^{D-1}$ the covariance matrix of $(\mathbf{u}_1, \mathbf{u}_2) = (\text{ilr } \mathbf{x}_1, \text{ilr } \mathbf{x}_2)$. As before, there exists matrix relationships between the covariance matrices $\mathbf{\Gamma}^C$, $\mathbf{\Gamma}^Y$ and $\mathbf{\Gamma}^U$.

2 Compositional time series models

Let $\mathbf{x}_t = (x_{t1}, \dots, x_{tD})'$, $t = 0, \pm 1, \pm 2, \dots$ be a compositional process (\mathcal{C} -time series process) defined on \mathcal{S}^D for any t . The compositional second-order properties of \mathbf{x}_t are then specified by the \mathcal{C} -mean vectors, $\boldsymbol{\xi}_t = \mathbb{E}_C\{\mathbf{x}_t\} = (\xi_{t1}, \dots, \xi_{tD})'$, and the \mathcal{C} -autocovariances matrices,

$$\mathbf{\Gamma}^C(t+h, t) = \mathbb{E} \left\{ (\text{clr } \mathbf{x}_{t+h} - \text{clr } \boldsymbol{\xi}_{t+h}) (\text{clr } \mathbf{x}_t - \text{clr } \boldsymbol{\xi}_t)' \right\} = [\gamma_{ij}^C(t+h, t)]_{i,j=1}^D,$$

which belong to the family of $\mathcal{A}_{D \times D}$ matrices.

Notice that in the compositional context, given a \mathcal{C} -time series $\{\mathbf{x}_t\}$ it makes no sense to analyze any of the individual parts $\{x_{ti}\}$ as univariate time series. However, in some cases one might be interested in analyzing the relative behavior of two parts i and j ($i \neq j$) or, in general, of a sub-compositional time series $\{\mathbf{x}_{St}\}$, where S symbolizes any subset of two or more parts $1, \dots, D$ of \mathbf{x}_t .

The clr, alr and ilr transformations applied to a compositional process $\{\mathbf{x}_t\}$ induce three processes $\{\mathbf{z}_t\}$, $\{\mathbf{y}_t\}$ and $\{\mathbf{u}_t\}$, respectively. The former, $\{\mathbf{z}_t\}$, defined on \mathbb{R}^D , is restricted to the hyperplane V because $\mathbf{z}_t' \mathbf{1}_D = 0$. The other two time series processes are defined on \mathbb{R}^{D-1} but $\{\mathbf{y}_t\}$ depends on the denominator used in the alr-transformation and $\{\mathbf{u}_t\}$ on the matrix \mathbf{V} used in the ilr-transformation. We denote by $\boldsymbol{\mu}_t^Z$, $\boldsymbol{\mu}_t^Y$ and $\boldsymbol{\mu}_t^U$ the mean vectors of $\{\mathbf{z}_t\}$, $\{\mathbf{y}_t\}$ and $\{\mathbf{u}_t\}$, respectively, and by $\mathbf{\Gamma}^Z(t+h, t)$, $\mathbf{\Gamma}^Y(t+h, t)$ and $\mathbf{\Gamma}^U(t+h, t)$ the autocovariance matrices of these time series processes. Observe that, by definition, $\boldsymbol{\mu}_t^Z = \text{clr } \boldsymbol{\xi}_t$ and $\mathbf{\Gamma}^Z(t+h, t) = \mathbf{\Gamma}^C(t+h, t)$. The mean vectors $\boldsymbol{\mu}_t^Y$ and $\boldsymbol{\mu}_t^U$, and the covariance matrices $\mathbf{\Gamma}^Y(t+h, t)$ and $\mathbf{\Gamma}^U(t+h, t)$ are related to $\text{clr } \boldsymbol{\xi}_t$ and $\mathbf{\Gamma}^C(t+h, t)$ by the equations given in 2.2.

2.1 Stationary \mathcal{C} -time series processes

The \mathcal{C} -time series process $\{\mathbf{x}_t\}$ is said to be (weakly) \mathcal{C} -stationary if $\boldsymbol{\xi}_t$ and $\mathbf{\Gamma}^C(t+h, t)$, $h = 0, \pm 1, \dots$ are independent of t . For a \mathcal{C} -stationary process we use the notation

$$\boldsymbol{\xi} = \mathbb{E}_C\{\mathbf{x}_t\}; \quad \mathbf{\Gamma}^C(h) = \mathbb{E} \left\{ (\text{clr } \mathbf{x}_{t+h} - \text{clr } \boldsymbol{\xi}) (\text{clr } \mathbf{x}_t - \text{clr } \boldsymbol{\xi})' \right\} = [\gamma_{ij}^C(h)]_{i,j=1}^D.$$

We shall refer to $\boldsymbol{\xi}$ as the \mathcal{C} -mean of $\{\mathbf{x}_t\}$ and to $\boldsymbol{\Gamma}^{\mathcal{C}}(h)$ as the \mathcal{C} -autocovariance at lag h , and $\boldsymbol{\Gamma}^{\mathcal{C}}(h)_{h=0,1,\dots}$ as the \mathcal{C} -autocovariance function. The \mathcal{C} -autocorrelation function $\mathbf{R}(h)_{h=0,1,\dots}$ is defined by

$$\mathbf{R}^{\mathcal{C}}(h) = \left[\gamma_{ij}^{\mathcal{C}}(h) / \sqrt{\gamma_{ii}^{\mathcal{C}}(0)\gamma_{jj}^{\mathcal{C}}(0)} \right]_{i,j=1}^D = [\rho_{ij}^{\mathcal{C}}(h)]_{i,j=1}^D.$$

The \mathcal{C} -time series process $\{\mathbf{w}_t\}$ is said to be \mathcal{C} -white noise with \mathcal{C} -mean $\mathbf{0}_{\mathcal{C}} = (1/D, \dots, 1/D)'$ and \mathcal{C} -covariance matrix $\boldsymbol{\Sigma}^{\mathcal{C}}$ —written as $\{\mathbf{w}_t\} \sim \text{WN}^{\mathcal{C}}(\mathbf{0}_{\mathcal{C}}, \boldsymbol{\Sigma}^{\mathcal{C}})$ —if and only if $\{\mathbf{w}_t\}$ is \mathcal{C} -stationary with \mathcal{C} -mean vector $\mathbf{0}_{\mathcal{C}}$ and \mathcal{C} -autocovariance function

$$\boldsymbol{\Gamma}^{\mathcal{C}}(0) = \boldsymbol{\Sigma}^{\mathcal{C}}; \quad \boldsymbol{\Gamma}^{\mathcal{C}}(h) = \mathbf{0}_{D \times D}, \text{ if } h \neq 0.$$

The \mathcal{C} -stationary property of $\{\mathbf{x}_t\}$ is equivalent to the stationary property of any of the transformed processes $\{\mathbf{z}_t\}$, $\{\mathbf{y}_t\}$ and $\{\mathbf{u}_t\}$. Moreover, $\{\mathbf{x}_t\}$ is \mathcal{C} -white noise if and only if $\{\mathbf{z}_t\}$ —or $\{\mathbf{y}_t\}$, or $\{\mathbf{u}_t\}$ —is white noise.

2.2 \mathcal{C} -ARIMA processes

A \mathcal{S}^D -variate \mathcal{C} -time series process $\{\mathbf{x}_t\}$ is a \mathcal{C} -ARMA(p, q) process if

$$\begin{aligned} (\mathbf{x}_t \ominus \boldsymbol{\xi}) \ominus (\boldsymbol{\Phi}_1 \odot (\mathbf{x}_{t-1} \ominus \boldsymbol{\xi})) \ominus \dots \ominus (\boldsymbol{\Phi}_p \odot (\mathbf{x}_{t-p} \ominus \boldsymbol{\xi})) = \\ \mathbf{w}_t \ominus (\boldsymbol{\Theta}_1 \odot \mathbf{w}_{t-1}) \ominus \dots \ominus (\boldsymbol{\Theta}_q \odot \mathbf{w}_{t-q}), \end{aligned}$$

where $\boldsymbol{\Phi}_1, \dots, \boldsymbol{\Phi}_p, \boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_q$ are $\mathcal{A}_{D \times D}$ -matrices and $\mathbf{w}_t \sim \text{WN}^{\mathcal{C}}(\mathbf{0}_{\mathcal{C}}, \boldsymbol{\Sigma}^{\mathcal{C}})$. These equations can be written in the more compact form

$$\boldsymbol{\Phi}^{\mathcal{C}}(L_{\mathcal{C}})(\mathbf{x}_t \ominus \boldsymbol{\xi}) = \boldsymbol{\Theta}^{\mathcal{C}}(L_{\mathcal{C}})\mathbf{w}_t, \quad \{\mathbf{w}_t\} \sim \text{WN}^{\mathcal{C}}(\mathbf{0}_{\mathcal{C}}, \boldsymbol{\Sigma}^{\mathcal{C}}),$$

where $\boldsymbol{\Phi}^{\mathcal{C}}(z) = \mathbf{G}_D \ominus (\boldsymbol{\Phi}_1 \odot z) \ominus \dots \ominus (\boldsymbol{\Phi}_p \odot z^p)$ and $\boldsymbol{\Theta}^{\mathcal{C}}(z) = \mathbf{G}_D \ominus (\boldsymbol{\Theta}_1 \odot z) \ominus \dots \ominus (\boldsymbol{\Theta}_q \odot z^q)$ are $\mathcal{A}_{D \times D}$ -matrix-valued polynomials, \mathbf{G}_D is the centering matrix and $L_{\mathcal{C}}$ the backshift operator. In the compositional context, the operator $1 - L_{\mathcal{C}}$ represents the \mathcal{C} -difference operator, i.e., $(1 - L_{\mathcal{C}})\mathbf{x}_t = \mathbf{x}_t \ominus \mathbf{x}_{t-1}$. Applying $1 - L_{\mathcal{C}}$ to $\{\mathbf{x}_t\}$ is equivalent to apply $1 - L$ to the transformed processes $\{\mathbf{z}_t\}$, $\{\mathbf{y}_t\}$ and $\{\mathbf{u}_t\}$.

If $\{\mathbf{x}_t\}$ is \mathcal{C} -ARMA(p, q) process then $\{\mathbf{z}_t\}$ is an ARMA(p, q) process because

$$\boldsymbol{\Phi}^{\mathcal{Z}}(L)(\mathbf{z}_t - \boldsymbol{\mu}^{\mathcal{Z}}) = \boldsymbol{\Theta}^{\mathcal{Z}}(L)\mathbf{w}_t^{\mathcal{Z}}, \quad \{\mathbf{w}_t^{\mathcal{Z}}\} \sim \text{WN}(\mathbf{0}_D, \boldsymbol{\Sigma}^{\mathcal{Z}}),$$

where $\boldsymbol{\Phi}^{\mathcal{Z}}(z) = \mathbf{I}_D - \sum_{i=1}^p \boldsymbol{\Phi}_i z^i$; $\boldsymbol{\Theta}^{\mathcal{Z}}(z) = \mathbf{I}_D - (\sum_{i=1}^q \boldsymbol{\Theta}_i z^i)$; and $\boldsymbol{\Sigma}^{\mathcal{Z}} = \boldsymbol{\Sigma}^{\mathcal{C}}$. Equally, $\{\mathbf{y}_t\}$ will be an ARMA(p, q) process because

$$\boldsymbol{\Phi}^{\mathcal{Y}}(L)(\mathbf{y}_t - \boldsymbol{\mu}^{\mathcal{Y}}) = \boldsymbol{\Theta}^{\mathcal{Y}}(L)\mathbf{w}_t^{\mathcal{Y}}, \quad \{\mathbf{w}_t^{\mathcal{Y}}\} \sim \text{WN}(\mathbf{0}_{D-1}, \boldsymbol{\Sigma}^{\mathcal{Y}}),$$

where

$$\boldsymbol{\Phi}^{\mathcal{Y}}(z) = \mathbf{I}_{D-1} - \left(\sum_{i=1}^p \mathbf{F} \boldsymbol{\Phi}_i \mathbf{F}' \mathbf{H}^{-1} z^i \right), \quad \boldsymbol{\Theta}^{\mathcal{Y}}(z) = \mathbf{I}_{D-1} - \left(\sum_{i=1}^q \mathbf{F} \boldsymbol{\Theta}_i \mathbf{F}' \mathbf{H}^{-1} z^i \right)$$

and $\Sigma^{\mathcal{Y}} = \mathbf{F}\Sigma^{\mathcal{C}}\mathbf{F}'$. And $\{\mathbf{u}_t\}$ will be an ARMA(p, q) process because

$$\Phi^{\mathcal{U}}(L)(\mathbf{u}_t - \boldsymbol{\mu}^{\mathcal{U}}) = \Theta^{\mathcal{U}}(L)\mathbf{w}_t^{\mathcal{U}}, \quad \{\mathbf{w}_t^{\mathcal{U}}\} \sim \text{WN}(\mathbf{0}_{D-1}, \Sigma^{\mathcal{U}}),$$

where $\Phi^{\mathcal{U}}(z) = \mathbf{I}_{D-1} - (\sum_{i=1}^p \mathbf{U}'\Phi_i\mathbf{U}z^i)$; $\Theta^{\mathcal{U}}(z) = \mathbf{I}_{D-1} - (\sum_{i=1}^q \mathbf{U}'\Theta_i\mathbf{U}z^i)$ —with $\mathbf{U} = \mathbf{F}'\mathbf{H}^{-1}\mathbf{F}\mathbf{V}$ —; and $\Sigma^{\mathcal{U}} = (\mathbf{F}\mathbf{V})^{-1}\mathbf{F}\Sigma^{\mathcal{C}}((\mathbf{F}\mathbf{V})^{-1}\mathbf{F})'$.

If d is a non-negative integer, it is natural to define $\{\mathbf{x}_t\}$ as a \mathcal{C} -ARIMA(p, d, q) processes if $(1 - L_{\mathcal{C}})^d\mathbf{x}_t$ is a \mathcal{C} -ARMA(p, q) processes. This definition means that $\{\mathbf{x}_t\}$ satisfies a \mathcal{C} -difference equation of the form

$$\Phi^{\mathcal{C}}(L_{\mathcal{C}})(1 - L_{\mathcal{C}})^d\mathbf{x}_t = \Theta^{\mathcal{C}}(L_{\mathcal{C}})\mathbf{w}_t, \quad \{\mathbf{w}_t\} \sim \text{WN}^{\mathcal{C}}(\mathbf{0}_{\mathcal{C}}, \Sigma^{\mathcal{C}}),$$

where $\Phi^{\mathcal{C}}(z) = \mathbf{G}_D \ominus (\Phi_1 \odot z) \ominus \dots \ominus (\Phi_p \odot z^p)$ and $\Theta^{\mathcal{C}}(z) = \mathbf{G}_D \ominus (\Theta_1 \odot z) \ominus \dots \ominus (\Theta_q \odot z^q)$ are $\mathcal{A}_{D \times D}$ -matrix-valued polynomials.

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, New York: Chapman & Hall. 416 pp. Reprinted in 2003 by Blackburn Press.
- Barceló-Vidal, C., Martín-Fernández, J.A. and Pawlowsky-Glahn, V. (2001). Mathematical foundations of compositional data analysis. In: *Proceedings of IAMG'01 – The sixth annual conference of the International Association for Mathematical Geology*, p. 20. CD-ROM.

Modelling longitudinal spatial curve data

Sarah Barry¹ and Adrian Bowman¹

¹ Department of Statistics, University of Glasgow, 15 University Gardens, Glasgow, G12 8QW, sarah@stats.gla.ac.uk

1 Introduction

Shape data arise in several different fields, such as computer science, medicine and statistics. Though much work has been done in the area of modelling two-dimensional shape data in statistics, there has been relatively little analysis of three-dimensional shape data, particularly when they are of a longitudinal nature.

We present a pairwise mixed effects modelling approach for longitudinal data of high dimension, introduced in Fieuws and Verbeke (2006), and apply it to data from a study of the facial shapes of infants suffering from cleft-lip and palate. Both landmarks and curves have been used to describe the facial shapes, and to demonstrate the applicability and benefits of using the pairwise approach for such kinds of data, but here we focus solely on curves. The approach of Fieuws and Verbeke (2006) is extended to include a quadratic test of model fixed effects, which may be applied to the analysis of either landmarks or curves, and parametric bootstrapping is employed to verify the accuracy of such a test.

Analysis of the facial curves proceeds by fitting a B-spline to the data and using the spline coefficients as the model responses. Informal 95% confidence intervals for the curve model estimates are presented, followed by a discussion of the appropriateness of this approach and a comparison with the results obtained from the quadratic test of the fixed effects.

2 Modelling longitudinal cleft-lip and palate data

2.1 Cleft-lip and palate data

The data arise from a study comparing the facial shapes of 49 children with unilateral cleft-lip and palate to 100 age-matched controls (Hood et. al, 2004). Each child had a facial image captured at 3 months of age (before surgical repair on the cleft children) and subsequently at 6, 12 and 24 months.

2.2 Pairwise mixed effects modelling

The pairwise modelling approach of Fieuws and Verbeke (2006) involves fitting linear mixed effects models to each of the $m(m-1)/2$ pairwise combinations of responses

$$(Y_1, Y_2), (Y_1, Y_3), \dots, (Y_1, Y_m), (Y_2, Y_3), \dots, (Y_2, Y_m), \dots, (Y_{m-1}, Y_m),$$

where Y_r is the vector of all responses (across individuals and time) for response r . The sum of the log-likelihoods across individuals is maximized for each combination (r, s) of responses and since fitting all of the pairwise models is equivalent to fitting a pseudo-likelihood, asymptotically

$$\sqrt{N}(\hat{\theta} - \theta) \sim MVN(0, J^{-1}KJ^{-1}),$$

where N is the number of individuals, J and K are matrices of second and first derivatives of the log-likelihood, respectively, and θ is a vector of all parameter estimates across all models. Since θ may contain some repetitions, the requisite vector containing one estimate per parameter, θ^* , may be calculated as $\theta^* = A\theta$ with, approximately, $\text{Var}(\hat{\theta}^*) = AJ^{-1}KJ^{-1}A'/N$, where A is a matrix of appropriate coefficients.

2.3 Curve analysis

The shapes are described by curves placed on each of the images. Procrustes analysis is applied to the full set of curves (examples of which are displayed in Figure 1), describing the entire face, in order to remove the effects of location, rotation and scale in the images, so that only the shape information remains. We use the resulting Procrustes coordinates to describe the curves on each image. Full details on Procrustes analysis may be found in Dryden and Mardia (1998).

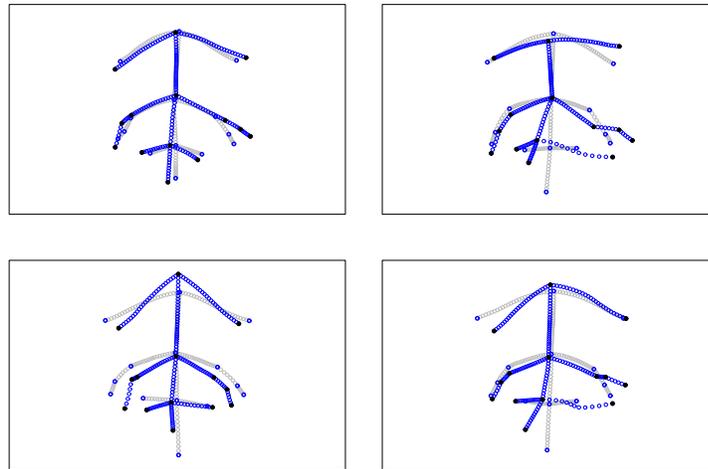


FIGURE 1. Frontal view of the full set of mean curves for a selection of cleft and control individuals at 3 months, with landmarks superimposed. Control group - grey points with black circles for landmarks; cleft group - dark points with solid circles.

Each curve is represented by many points placed very close together. We have analysed midline curves, which trace the line from the midpoint between the eyes down the nose to the middle of the upper lip and these are displayed in Figure 2, with measured landmarks superimposed.

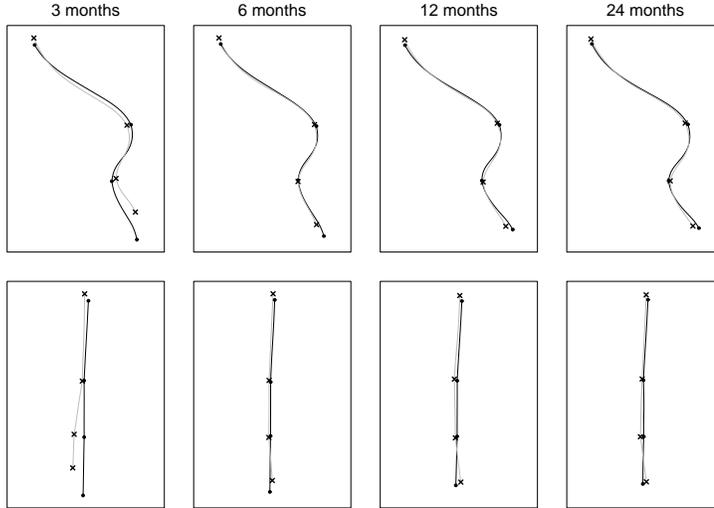


FIGURE 2. Mean curves for each group at 3, 6, 12 and 24 months, with landmarks superimposed. Upper - profile view; lower - frontal view. Control group - black line and solid circles; cleft group - grey lines and crosses.

To reduce the number of points used, the curves were parameterized as $(x(d), y(d), z(d))$, where the variable d represents the proportion of the distance travelled along the curve. This information was used to fit a B-spline and a linear model was fitted to the x , y and z points on the curves with the resulting bases as predictors. For any individual i , therefore, the midline curve at a time t and in a particular dimension is described by the following vector of points:

$$y_i(t) = s_{0i}(t) + \sum_{r=1}^k s_{ir}(t)b_{ir}(t),$$

where the s_{ir} are spline coefficients, k is the number of knots chosen and $b_{ir}(t)$ is the basis vector describing the part of the curve corresponding to knot r . This holds for each dimension, so there are $3k$ spline coefficients describing each person's midline curve at any time t . The spline coefficients, s_{ir} , were then extracted from the model and used as the responses in the linear mixed effects model below. Nine knots were used for the B-spline as that provided an acceptably close fit to the mean curves.

The spline coefficient traces over time can be found in Figure 3 below. It is clear that there is a large amount of variation between individuals and across spline coefficients. The assumption that each subject follows a similar trajectory seems reasonable but it is clearly necessary to include a random intercept that is allowed to differ across coefficients.

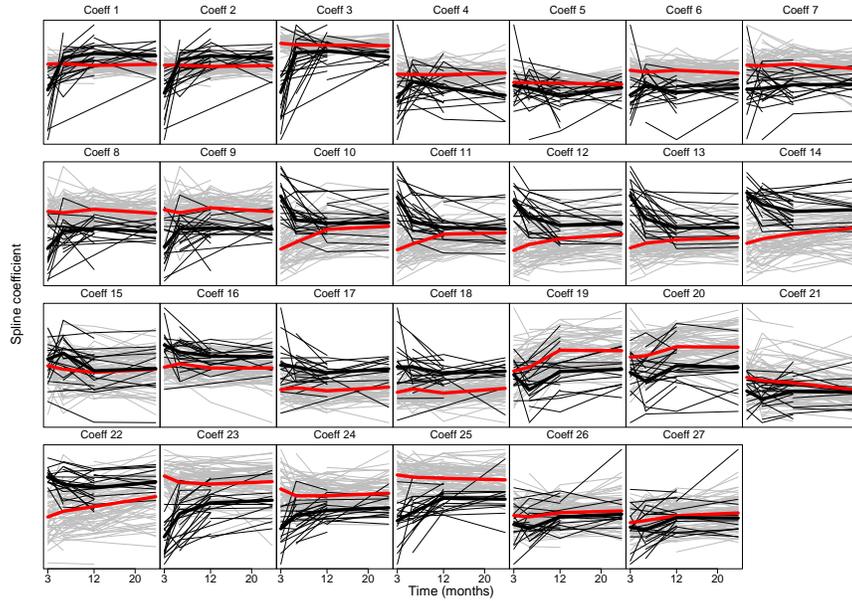


FIGURE 3. Traces over time of the spline coefficients for each individual. The black lines represent the cleft individuals and the grey the controls, with the thicker lines representing the group means. The scales are arbitrary and different for each plot.

3 Applying the pairwise modelling approach

3.1 Curves

The models were fitted to the data using a program written in R (R Development Core Team, 2004), making use of the `deriv` and `lme` packages. For $s_{ir}(t)$ the vector of spline coefficients $r = u, v$ ($u = 1, \dots, m, v = 2, \dots, m - 1$) from individual i at time t , the model is given by

$$s_{ir}(t) = \beta_{0r} + b_{ir} + \beta_{1r}gp_i + \beta_{2r}p(t) + \beta_{3r}t + \beta_{4r}p(t) * gp_i + \beta_{5r}gp_i * t + \epsilon_i(t),$$

where $p(t)$ takes the value zero at 3 months and one otherwise, and gp_i takes value one if subject i is in the cleft group and zero otherwise. The random intercepts $b_i = (b_{iu}, b_{iv})$ have variance matrix V , and $\text{var}(\epsilon_i(t)) = \sigma^2 I$. All fixed and random effects are therefore coordinate-specific.

The model estimates of the mean curve positions are displayed in Figure 4, for both the cleft and control groups. Bivariate 95% confidence intervals (in both the frontal and profile views) were plotted at equally spaced points along the curve, to give the impression of an overall “confidence region” for the curve in two dimensions. Despite potential issues with multiple comparisons, this gives a subjective view of where the differences between the groups lie. Use of a test of the model effects will allow global inference about the differences between the groups.

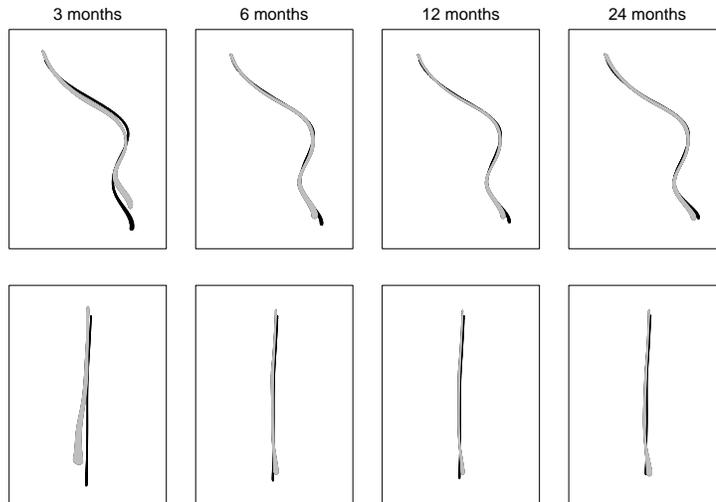


FIGURE 4. Model estimates with 95% bivariate confidence regions for cleft (grey) and control (black) groups for the mean midline curves at each time point in the profile (upper) and frontal (lower) views

3.2 Model comparison

We suggest a method for testing for fixed effects in the pairwise mixed modelling approach. For any particular fixed effect, we must test the null hypothesis that all coordinate-specific regression coefficients pertaining to that fixed effect are equal to zero.

We assume that $\hat{\theta}_S^*$, the vector containing the estimates for all parameters corresponding to the relevant fixed effect (say, the group:time interaction), is distributed as multivariate normal with mean θ_S^* and variance V_S , where V_S is a sub-matrix of V . Therefore, under the null hypothesis, $H_0 : \theta_S^* = 0$:

$$\hat{\theta}_S^{*T} V_S^{-1} \hat{\theta}_S^* \sim \chi_p^2,$$

where p denotes the number of parameters in θ_S^* .

If this test is applied to the group:time interaction in the model fitted to the curve data, a χ^2 -statistic of 101.8 is obtained on 27 degrees of freedom (nine spline coefficients for each of three dimensions). This provides us with strong evidence to reject the null hypothesis that there is no difference between the cleft and control groups from 6 to 24 months.

References

- Dryden, I.L., & Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester: Wiley.
- Fieuws, S., & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* **62**(2), 424–431.

Hood, C.A., Hosey, M.T., Bock, M., White, J., Ray, A., & Ayoub, A.F. (2004). Facial characterization of infants with cleft lip and palate using a three-dimensional capture technique. *Cleft Palate-Craniofacial Journal* **41**, 27-35.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

The multilevel latent Markov model

Francesco Bartolucci¹ and Monia Lupparelli¹

¹ Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli 20, 06123 Perugia, Italy

Abstract: We introduce a multilevel version of the latent Markov model with covariates which is suitable for the analysis of binary longitudinal data when subjects are grouped in a large number of clusters. For the maximum likelihood estimation of this model we introduce an EM algorithm which can be implemented by means of certain recursions well known in the hidden Markov literature. The approach is illustrated through the application to a dataset deriving from the administration of a set of items to a sample of patients suffering from cancer who were admitted to different hospitals.

Keywords: Binary longitudinal data; Clustered data; Latent variable models.

1 Introduction

The Latent Markov model (LM; Wiggins, 1973; Langeheine and van de Pol, 2002) has become a standard tool for the analysis of binary longitudinal data, especially when the main aim of the analysis is describing individual change with respect to a certain latent status over the time. This model assumes that the response variables corresponding to the different time occasions are conditionally independent given a discrete latent process which follows a first-order Markov chain. Maximum likelihood estimation of the parameters of this model may be obtained by using the EM algorithm of Dempster *et al.* (1977); see Bartolucci (2006) and the references therein.

The LM model has been extended in several ways. One of the most interesting extensions is for allowing the distribution of the latent process to depend on individual covariates; see, in particular, Vermunt *et al.* (1999) and Bartolucci and Pennoni (2007). However, the extension to the case of clustered data, i.e. when subjects are grouped according to some criteria such as being admitted to the same hospital, does not seem to be already considered in the literature. In fact, an assumption common to all versions of the LM model is that subjects are independent of each other and, therefore, the correlation that may arise between those subjects belonging to the same cluster is ignored. This correlation could be taken into account by including, among the covariates, one dummy variable for each cluster. Obviously, this approach is not viable when the number of clusters is large.

In this paper, we illustrate an extended version of the LM model with covariates to deal with clustered data. As in the LM model of Vermunt *et al.* (1999), we assume that the covariates affect the initial and the transition probabilities of the latent process. However, we also assume that these probabilities depend on random parameters which capture the effect of the cluster to which the subject belongs. A set of restrictions on the

parameters is also considered. For the maximum likelihood estimation of the resulting model, we introduce an EM algorithm in which the E-step is based on certain recursions well known in the hidden Markov literature (MacDonald and Zucchini, 1997).

The paper is organized as follows. Next section illustrates the proposed model. Likelihood inference for this model is dealt with in Section 3. Finally, in Section 4 we describe an application based on a dataset concerning a sample of cancer patients admitted to different hospitals.

2 The model

Let y_{hit} denote the binary response variable for the i -th subject in the h -th cluster at the t -th occasion, with $h = 1, \dots, H$, $i = 1, \dots, n_h$ and $t = 1, \dots, T_{hi}$. Also let \mathbf{x}_{hit} be a corresponding vector of covariates and let $\mathbf{y}_{hi} = (y_{hi1}, \dots, y_{hiT_{hi}})'$ be a vector which collects the response variables for the same subject. The covariates are assumed as fixed and so, in the following, we will omit to explicitly conditioning on them.

The LM model assumes that the response vectors \mathbf{y}_{hi} are independent and that, for every subject i in cluster h , the variables y_{hit} are conditionally independent given a sequence of latent variables $\mathbf{s}_{hi} = (s_{hi1}, \dots, s_{hiT_{hi}})'$ which follows a first-order Markov chain (Vermunt *et al.*, 1999). The distribution of y_{hit} depends only on the corresponding state of the latent process and then parameters of the model are also the *success probabilities* $\lambda_{st} = p(y_{hit} = 1 | s_{hit} = s)$, with $s = 1, \dots, S$ and $t = 1, \dots, T$, where $T = \max_{hi} T_{hi}$. The model also assumes that

$$\log \frac{\pi_{hi1}(s)}{\pi_{hi1}(1)} = \mathbf{x}'_{hi1} \boldsymbol{\beta}_s, \quad s = 2, \dots, S,$$

where $\pi_{hi1}(s) = p(s_{hi1} = s)$ is the initial probability of state s , and that

$$\log \frac{\pi_{hit}(s_1 | s_0)}{\pi_{hit}(s_0 | s_0)} = \mathbf{x}'_{hit} \boldsymbol{\gamma}_{s_0 s_1}, \quad s_0, s_1 = 1, \dots, S, \quad s_1 \neq s_0,$$

for $t = 2, \dots, T_{hi}$, where $\pi_{hit}(s_1 | s_0) = p(s_{hit} = s_1 | s_{hi,t-1} = s_0)$ is the probability of transition from state s_0 to state s_1 .

The above model does not explicitly consider the cluster effect. In the proposed multilevel LM model, MLM for short, we take this effect into account by assuming that, for every pair of subjects (i_1, i_2) in the same cluster h , \mathbf{y}_{hi_1} and \mathbf{y}_{hi_2} are conditionally independent given a latent vector $\boldsymbol{\alpha}_h$ having a discrete distribution. The elements of $\boldsymbol{\alpha}_h$ are random parameters that capture the effect of cluster h on the initial and the transition probabilities and are denoted, respectively, by α_{hs} , for $s = 2, \dots, S$, and $\alpha_{hs_0 s_1}$, for $s_0, s_1 = 1, \dots, S$ with $s_0 \neq s_1$. To include these effects in the model, we assume that

$$\log \frac{\pi_{hi1}(s | \boldsymbol{\alpha}_h)}{\pi_{hi1}(1 | \boldsymbol{\alpha}_h)} = \alpha_{hs} + \mathbf{x}'_{hi1} \boldsymbol{\beta}_s, \quad s = 2, \dots, S,$$

where $\pi_{hi1}(s | \boldsymbol{\alpha}_h) = p(s_{hi1} = s | \boldsymbol{\alpha}_h)$, and that

$$\log \frac{\pi_{hit}(s_1 | \boldsymbol{\alpha}_h, s_0)}{\pi_{hit}(s_0 | \boldsymbol{\alpha}_h, s_0)} = \alpha_{hs_0 s_1} + \mathbf{x}'_{hit} \boldsymbol{\gamma}_{s_0 s_1}, \quad s_0, s_1 = 1, \dots, S, \quad s_1 \neq s_0,$$

for $t = 2, \dots, T_{hi}$, where $\pi_{hit}(s_1|\alpha_h, s_0) = p(s_{hit} = s_1|\alpha_h, s_{hi,t-1} = s_0)$. The support points of the distribution of each latent vector α_h are denoted by ξ_c and the corresponding probabilities by ρ_c , with $c = 1, \dots, C$.

Under the above assumptions, the marginal (or *manifest*) distribution of the response vectors \mathbf{y}_{hi} for the subjects in the same cluster h may be expressed as

$$p(\mathbf{y}_{h1}, \dots, \mathbf{y}_{hn_h}) = \sum_{\alpha_h} p(\alpha_h) \prod_i \sum_{s_{hi}} p(\mathbf{y}_{hi}|\alpha_h, s_{hi}) p(s_{hi}|\alpha_h), \quad (1)$$

where \sum_{α_h} stands for the sum over all the possible configurations of α_h and $\sum_{s_{hi}}$ for that over all the possible configurations of s_{hi} . Moreover,

$$\begin{aligned} p(\mathbf{y}_{hi}|\alpha_h, s_{hi}) &= \prod_t p(y_{hit}|\alpha_h, s_{hit}), \\ p(s_{hi}|\alpha_h) &= \pi_{hi1}(s_{hi1}|\alpha_h) \prod_{t>1} \pi_{hit}(s_{hit}|\alpha_h, s_{hi,t-1}), \end{aligned}$$

where $p(y_{hit}|\alpha_h, s_{hit})$ depends on the parameters λ_{st} . In practice, we compute the probability in (1) by exploiting a recursion derived from the hidden Markov literature (MacDonald and Zucchini, 1997).

3 Likelihood inference

For an observed set of data $(\mathbf{x}_{hit}, y_{hit})$, $h = 1, \dots, H$, $i = 1, \dots, n_h$, $t = 1, \dots, T_{hi}$, the log-likelihood of the model illustrated in Section 2 is given by

$$\ell(\boldsymbol{\theta}) = \sum_h \log[p(\mathbf{y}_{h1}, \dots, \mathbf{y}_{hn_h})],$$

where $p(\mathbf{y}_{h1}, \dots, \mathbf{y}_{hn_h})$ is computed according to (1) as a function of the parameters of the model which are collected in the vector $\boldsymbol{\theta}$. We recall that these parameters are β_s , $\gamma_{s_0 s_1}$ (for the distribution of the subject-specific latent process), ξ_c and ρ_c (for the distribution of the cluster-specific latent vector) and λ_{st} (for the conditional distribution of the response variables given the latent process).

To estimate $\boldsymbol{\theta}$, we maximize $\ell(\boldsymbol{\theta})$ by using an EM algorithm (Dempster *et al.*, 1977). This algorithm alternates two steps, indicated by E and M, until convergence in $\ell(\boldsymbol{\theta})$. At the E-step, we compute the conditional expected value of the *complete data log-likelihood* $\ell^*(\boldsymbol{\theta})$ given the observed data. At the M-step, the expected value of $\ell^*(\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ and the estimate of this parameter vector is then updated.

The function $\ell^*(\boldsymbol{\theta})$ is the log-likelihood that we could compute if we knew the latent class of each cluster and the latent state of each subject at every occasion. More precisely, let $w_h(c)$ be a dummy variable equal to 1 if cluster h belongs to latent class c , let $z_{hit}(s)$ be a dummy variable equal to 1 if subject i in cluster h is in latent state s at occasion t and let $z_{hit}(s_0, s_1) = z_{hi,t-1}(s_0)z_{hit}(s_1)$. This function may be expressed as

$$\ell^*(\boldsymbol{\theta}) = \sum_h \sum_c w_h(c) [\log(\rho_c) + m_{hc}^*(\boldsymbol{\theta})]$$

where

$$\begin{aligned}
m_{hc}^*(\boldsymbol{\theta}) &= \sum_i \sum_s z_{hi1}(s) \log[\pi_{hi1}(s|\boldsymbol{\alpha}_h = \boldsymbol{\xi}_c)] + \\
&+ \sum_i \sum_{s_0} \sum_{s_1} \sum_{t>1} z_{hit}(s_0, s_1) \log[\pi_{hit}(s_1|\boldsymbol{\alpha}_h = \boldsymbol{\xi}_c, s_0)] + \\
&+ \sum_i \sum_s \sum_t z_{hit}(s) [y_{hit} \log(\lambda_{st}) + (1 - y_{hit}) \log(1 - \lambda_{st})].
\end{aligned}$$

Computing the conditional expected value of $\ell^*(\boldsymbol{\theta})$, given the observed data, is equivalent to computing the expected value of the above dummy variables. This is performed at the E-step on the basis of certain recursions taken from the hidden Markov literature (MacDonald and Zucchini, 1997).

In our framework, we also address the problem of testing hypotheses on $\boldsymbol{\theta}$ and, in particular, on the structure of the transition matrix. For this aim, we use the likelihood ratio statistic whose null asymptotic distribution may be derived by extending the results of Bartolucci (2006).

4 An application

As an illustrative example, we consider a dataset derived from a survey carried out in Italy in the 90's about 516 patients suffering from cancer who were admitted to 58 different hospitals. In order to assess the physical and mental status of the patients, a set of 36 items were administered to them at different occasions. The number of occasions is not the same for all subjects and it goes from 1 to 15. The response to any item was coded in four ordinal categories. For each subject and each time occasion, we summarized the responses to these items by one binary response variable which is equal to 1 for a patient with a bad physical and/or mental status. Finally, as individual covariates we used *gender* (dummy variable equal to 1 for a female), *age* (dummy variable equal to 1 for an over-70 subject) and *time* (interval of time between occasions at which the questionnaire was administered). We analyzed the resulting dataset by applying the LM model and the MLM model with $C = 2$ latent classes. For both models, we used $S = 2$ latent states and we assumed that $\lambda_{st} = \lambda_s$ for all t , so that the distribution of each response variable depends on the occasion only through the corresponding state of the latent process.

For the ML model we obtained a maximum log-likelihood of -1282.4 and an estimate of the probability of success equal to $\hat{\lambda}_1 = 0.013$ for the first latent state and to $\hat{\lambda}_2 = 0.880$ for the second latent state. The estimates of the regression parameters affecting the initial and the transition probabilities of the latent process are reported in Table 1.

We can observe that the two latent states are well separated, with the first corresponding to patients in better conditions with respect to those in the second latent state. On the basis of the estimates of the regression parameters, we can also observe that the initial health status tends to get worse for males and for over-70 patients, whereas the probability of changing status over the time is larger for males and for subjects under-70.

TABLE 1. Estimates of the regression parameters under the LM and the MLM models (the covariate time does not affect the initial probabilities).

covariate	LM			MLM		
	$\hat{\beta}_2$	$\hat{\gamma}_{12}$	$\hat{\gamma}_{21}$	$\hat{\beta}_2$	$\hat{\gamma}_{12}$	$\hat{\gamma}_{21}$
intercept	-0.249	-2.634	-1.970	-0.756	-2.878	-0.667
gender	-0.401	-0.130	-0.384	-0.433	-0.179	-0.381
age	0.079	-0.111	-0.148	-0.124	-0.132	0.083
time	-	0.027	0.055	-	0.025	0.051

The MLM model has a maximum log-likelihood of -1263.4 that, with only four more parameters, is considerably higher than that of the LM model. The estimates of the probabilities λ_s are very similar to those obtained with the ML model: $\hat{\lambda}_1 = 0.008$ and $\hat{\lambda}_2 = 0.899$. For the distribution of the latent classes we have the following estimates:

$$\hat{\xi}_2 = \begin{pmatrix} 0.888 \\ 0.803 \\ -1.673 \end{pmatrix}, \quad \hat{\rho} = \begin{pmatrix} 0.383 \\ 0.617 \end{pmatrix},$$

with ξ_1 constrained to $\mathbf{0}$ to ensure identifiability. Finally, the estimates of the regression parameters are reported in Table 1.

Introducing the random effect allows us to distinguish between two kinds of hospitals. With respect to hospitals in the first group (38%), hospitals in the second group (62%) tend to admit patients in worse conditions. These patients also show a faster worsening of their health status. Note that the estimates of the regression parameters do not dramatically change with the inclusion of the cluster effect. An exception is represented by the estimates of the coefficients for the covariate age, the signs of which change when we introduce the cluster effect. However, the signs of these estimates under the LM model are not completely reasonable. This is probably due to the fact that this model ignores the selection bias arising from the inclusion of only certain types of subjects to certain hospitals. By considering the cluster effect, the proposed model can correct for this bias.

Acknowledgments. We wish to thank Guido Miccinesi of CSPO (Firenze, IT) for providing us with the dataset analyzed in this paper. We acknowledge the financial support from MIUR (PRIN 2005 - “Modelli marginali per variabili categoriche con applicazioni all’analisi causale”)

References

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, Series B* **68**, 155-178.

- Bartolucci, F., and Pennoni, F. (2007). A class of latent Markov models for capture-recapture data allowing for time, heterogeneity and behavior effects. *Biometrics*. To appear.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1-38.
- Langeheine, R., and van de Pol, F. (2002). Latent Markov chains. In: J.A. Hagenaars and A.L. McCutcheon (Eds.), *Applied Latent Class Analysis*. 304-341, Cambridge University Press, Cambridge.
- MacDonald, I.L., and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. London: Chapman & Hall.
- Vermunt, J.K., Langeheine, R., and Böckenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics* **24**, 179-207.
- Wiggins, L.M. (1973). *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. Amsterdam: Elsevier.

Nonlinear discrete-time hazard models for the rate of first marriage

Andy Batchelor¹, Heather L Turner¹ and David Firth¹

¹ Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom, heather.turner@warwick.ac.uk.

Abstract: We seek to model the hazard of entry into marriage for a sample of women in Ireland born between 1950 and 1973. Motivated by the work of Blossfeld and Huinink (1991), we propose a nonlinear discrete-time hazard model, which estimates the risk period and allows the effect of covariates on both the scale of risk and the age of maximum risk to be investigated.

Keywords: discrete-time survival analysis; non-proportional hazards; aliasing.

1 Introduction

In this paper we investigate the timing of first marriage for women in Ireland based on the Living in Ireland Surveys conducted by the Economic and Social Research Institute between 1994 and 2001. We limit our analysis to women born between 1950 and 1973, giving five, five-year cohorts who have passed the mean age at marriage for women in the full data set.

2 Linear Discrete-time Hazard Models

We first use the approach of Blossfeld and Huinink (1991), who proposed an exponential model for the hazard of first marriage, with baseline variables to control for the non-monotonic dependence of marriage rate on age:

$$r(t) = r_0 \exp\{\beta_L \log(\text{age} - 15) + \beta_R \log(45 - \text{age}) + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2(t) \boldsymbol{\beta}_2\}. \quad (1)$$

Here r_0 is the constant, baseline hazard; \mathbf{x}_1 and \mathbf{x}_2 are time-constant and time-varying covariates respectively, whilst $\log(\text{age} - 15)$ and $\log(45 - \text{age})$ are the baseline variables that combine to produce a bell-shaped curve.

We only have the year of marriage, so we use episode splitting to generate yearly life course data, making appropriate adjustment for the month of birth. We then use the following discrete-time equivalent of Model 1:

$$C(r(t)) = \beta_0 + \beta_L \log(\text{age} - 15) + \beta_R \log(45 - \text{age}) + \mathbf{x}'_1 \boldsymbol{\beta}_1 + \mathbf{x}'_2(t) \boldsymbol{\beta}_2, \quad (2)$$

where $C(r)$ is the complementary log-log transformation. Here age ranges from 15.04 to 44.96 years, so we keep the endpoints fixed at 15 and 45.

TABLE 1. Linear discrete-time hazard models.

Variables	Model					
	1	2	3	4	5	6
Intercept	-2.81	-17.92	-17.90	-19.31	-17.27	-17.21
Log(age - 15)		2.13	2.14	2.26	1.91	1.89
Log(45 - age)		3.63	3.67	4.14	3.70	3.67
Class s/skilled manual			-0.13	-0.10	-0.08	
Class skilled manual			-0.13	-0.06	-0.03	
Class non manual			-0.26	-0.22	-0.16	
Class low professional			-0.21	-0.18	-0.10	
Class high professional			-0.48	-0.43	-0.29	
Class missing			-0.07	-0.08	-0.02	
Cohort (54,59]				0.03	0.03	0.03
Cohort (59,64]				-0.08	-0.07	-0.07
Cohort (64,69]				-0.58	-0.55	-0.55
Cohort (69,74]				-1.30	-1.23	-1.23
In education					-1.52	-1.56
Deviance	13483	12414	12388	12086	11971	11981
Residual df	29866	29864	29858	29854	29853	29859

As far as possible, we follow Blossfeld and Huinink (1991) in building a model for our data, adding the baseline variables first, then social class, cohort and education variables. Our results are presented in Table 1. We find that women in later cohorts are less likely to marry and that the risk of marriage is significantly less whilst women are in education. Social class becomes insignificant when the education status is taken into account. Adding the final level of education does not significantly improve the model.

2.1 Nonlinear Discrete-time Hazard Models

We first consider extending Model 2 by defining the endpoints of the bell curve as parameters to be estimated:

$$\beta_0 + \beta_L \log(\text{age} - \alpha_L) + \beta_R \log(\alpha_R - \text{age}) \quad (3)$$

However we find that there is aliasing amongst the parameters in Equation 3, such that perturbations of one parameter can be compensated for by changes in the other parameters.

We therefore consider the following re-parameterization in which the aliasing is reduced:

$$\gamma - \exp(\delta) \left\{ \frac{(\nu - \alpha_L) \log\left(\frac{\nu - \alpha_L}{\text{age} - \alpha_L}\right) + (\alpha_R - \nu) \log\left(\frac{\alpha_R - \nu}{\alpha_R - \text{age}}\right)}{(\nu - \alpha_L) \log\left(\frac{\nu - \alpha_L}{\nu - D - \alpha_L}\right) + (\alpha_R - \nu) \log\left(\frac{\alpha_R - \nu}{\alpha_R - \nu + D}\right)} \right\} \quad (4)$$

Now the rate of marriage has a maximum of $C^{-1}(\gamma)$ at age ν and tends to zero as the age approaches α_L or α_R . The sharpness of the peak is captured by δ , since the rate

TABLE 2. Nonlinear discrete-time hazard models.

Variables	Model				
	6	7	8	9	10
Intercept(γ)	-2.12	-1.96	-1.68	-1.81	-2.31
Peak age (ν)					
Intercept	25.11	25.09	24.76	24.61	16.00
Education level (years)					0.78
Peakedness (δ)	-0.47	-0.45	-0.31	-0.41	-0.20
Left endpoint (α)	13.77	13.74	13.40	12.04	12.35
Class s/skilled manual		-0.13	-0.10		
Class skilled manual		-0.14	-0.06		
Class non manual		-0.26	-0.22		
Class low professional		-0.21	-0.19		
Class high professional		-0.49	-0.43		
Class missing		-0.07	-0.09		
Cohort (54,59]			0.03	0.03	0.06
Cohort (59,64]			-0.08	-0.07	-0.04
Cohort (64,69]			-0.58	-0.56	-0.53
Cohort (69,74]			-1.31	-1.25	-1.19
In education				-1.55	-0.65
Education level (years)					0.05
Deviance	12394	12368	12060	11960	11813
Residual df	29863	29857	29853	29858	29856

of marriage will be $C^{-1}(\gamma - \exp(\delta))$ at age $\nu - D$, where D is a fixed distance from ν , which we take to be 5 years.

Using the re-parameterization, we find that the fitted models are not significantly different from models in which $\alpha_R \rightarrow \infty$, where the baseline model is then:

$$\gamma - \exp(\delta) \left\{ \frac{(\nu - \alpha) \log \left(\frac{\nu - \alpha}{age - \alpha} \right) + age - \nu}{(\nu - \alpha) \log \left(\frac{\nu - \alpha}{\nu - D - \alpha} \right) - D} \right\} \quad (5)$$

Repeating the analysis of the previous section with this baseline model leads to a significant improvement over the equivalent fixed endpoint models (Models 6 to 9 in Table 2).

We can improve the model further by including the additive effect of education level on both the maximum rate of marriage (γ) and the age at which this maximum is reached (ν), leading to a non-proportional hazard model (Model 10, Table 2). We represent the level of education by the equivalent years spent in education, based on averages from the data. We can see from the corresponding hazard and survival curves in Figure 1 that an increase in education level delays the age at which the marriage rate peaks and increases the maximum marriage rate, so that women with a higher education level eventually overtake those with a lower education level in terms of the proportion that marry.

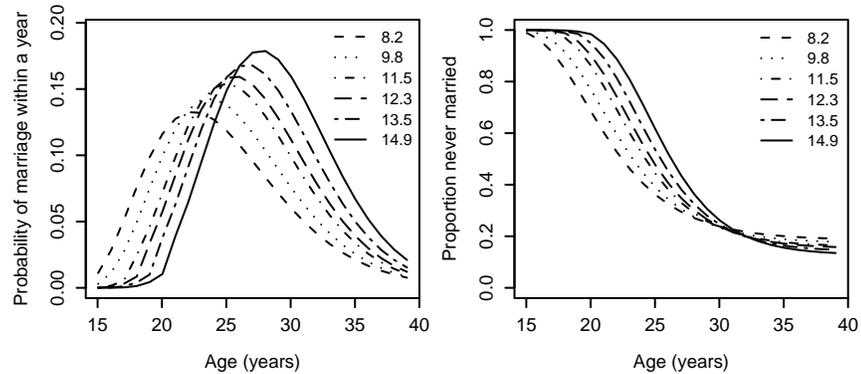


FIGURE 1. Hazard and survival curves under Model 10, for the (59, 64] year cohort and skilled manual class, by increasing education level (equivalent years).

3 Summary

The nonlinear discrete-time hazard models we propose allow the risk period to be estimated and the effect of covariates on both the scale of risk and the age of maximum risk to be investigated. We find the latter to be important in describing the effect of education level on the risk of entry into marriage.

Software: The generalized nonlinear models described in this paper were fitted using the *R* package *gnm* (Turner and Firth, 2007).

Acknowledgments: The work of Heather Turner and David Firth was supported by ESRC Professorial Fellowship RES-051-27-0055 and by the ESRC National Centre for Research Methods, Lancaster-Warwick Node (ref RES-576-25-5020). The data are from The Economic and Social Research Institute Living in Ireland Survey Microdata File (©Economic and Social Research Institute). We acknowledge Carmel Hannan for introducing us to this application and providing background on the data.

References

- Blossfeld, H.-P., and Huinink, J. (1991). Human Capital Investments or Norms of Role Transition? How Women's Schooling and Career Affect the Process of Family Formation. *American Journal of Sociology* **97**, 143-168.
- Turner, H.L. and Firth, D. (2007). Generalized nonlinear models in R: An overview of the *gnm* package. Documentation in the *gnm* package, <http://cran.r-project.org>.

Mixed Text and Data Mining through a principal axes method. Application to legal documents

Mónica Bécue-Bertaut¹ and Marta Poblet²

¹ EIO, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain,
monica.becue@upc.edu

² Institute of Law and Technology, UAB, 08193 Bellaterra (Cerdanyola del Vallès),
Spain, marta.poblet@uab.cat

Keywords: Text Mining; Free-texts; Correspondence Analysis; Multiple Factor Analysis

1 Introduction

Correspondence Analysis (CA) is widely used in Text Mining to give account of the structure of a documents-by-words matrix (Lebart et al., 1998). We propose to incorporate contextual and/or metadata and to globally analyze the texts and these complementary data.

2 Data and objectives

The data used as example consist of a corpus of 430 legal judgements issued by the Spanish Supreme Court (Tribunal Supremo) during the 1979 to 1996, and related to prostitution offences. Since the democratic constitution (December 1978), both the Spanish political and legal systems undergone profound transformations. Referring to the regulation of justice, the basic norm was passed in 1985. In the following, by the term *judgment* we refer to the document published by the court at the end of a trial, which contains the verdict as well as the other parts arguing the verdict. When analyzing these data, our aim is double: first, to study the relationship between the actual chronology, as indicated by the year of the publication, and the vocabulary used in the judgements; second, to detect if there are judgements which do not follow the common rule, in the sense that the vocabulary is behind the times or, on the contrary, more advanced than the date of publication suggests.

3 Methodology

In our case, the rows refer to the judgements, the frequency variables represent the words and we consider only one quantitative variable, the year of the publication. Nevertheless, we present the methodology in the general case of one quantitative table. We want to keep a CA-like approach by using an extension of multiple factor analysis (MFA, Bécue-Bertaut & Pages, *to be published*). We summarize the main features of this methodology, which is a particular weighted principal component analysis (PCA).

Notation

A judgments-by-words frequency table (with J columns) and one judgments-by-quantitative variables table (with K columns) are juxtaposed row-wise. In the frequency table, the cell (i,j) contains f_{ij} , the relative frequency with which judgment i ($i = 1, \dots, I$) uses word j ($j = 1, \dots, J$); ($\sum_{j=1}^J f_{ij} = 1$). In the case of the quantitative table, x_{ik} indicates the value of quantitative variable k as measured on judgment i ($k = 1, \dots, K$).

CA as a specific non-standardized weighted PCA

The results of CA can be obtained by performing a non-standardized PCA on the table having the general term:

$$(f_{ij} - f_{i.}f_{.j})/(f_{i.}f_{.j}) \quad (1)$$

using $\{f_{i.}; i = 1, \dots, I\}$ as row weights (and metric in the column space) and $\{f_{.j}; j = 1, \dots, J\}$ as column weights (and metric in the row space) (Escofier & Pagès, 1998, p. 96).

MFA of the multiple mixed table

By using the latter property, the extended MFA applied to a table juxtaposing a frequency table (with J columns) and a quantitative variables table (with K columns) is equivalent to perform a weighted PCA: first, on the multiple table juxtaposing both the table issued from the frequency table suitably transformed, i.e. with general term indicated by (1) and the quantitative table (every variable is centred and, generally, standardized); second, giving to the rows the weights imposed by CA (i.e. $\{f_{i.}; i = 1, \dots, I\}$); third, giving to the frequency columns the *a priori* weights induced by CA $\{f_{.j}; j = 1, \dots, J\}$ and to the quantitative columns the *a priori* weight 1. Those weights are divided, in both cases, by the first eigenvalue issued from the separate principal axes method applied to the corresponding table, denoted λ_1^J in the case of CA (frequency table) and λ_1^K in the case of Principal Component Analysis (PCA, quantitative table). This overweighting standardizes to 1 the highest axial inertia of each table (Escofier & Pagès, 1998, p. 132) and thus balances their influence on the first global principal axis; and finally, in our data, there is only one quantitative variable; this feature limits the separate PCA of the quantitative table, reduced to one column in this case, to the only standardization of the score and to adopt $\lambda_1^K=1$.

4 Results

We select the words used at least 50 times in all the judgements (in total, 961 different words). The resulting multiple mixed table has 430 judgement-rows and 961 words-column +1 quantitative column. The high value of the first eigenvalues (1.76, not far from the maximum value which is 2; Escofier & Pagès, 1998, p.161) indicates that the first principal component issued of MFA is an important dispersion axis that is common to both the chronology and the vocabulary.

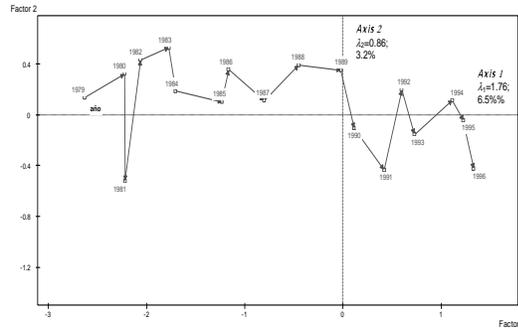


FIGURE 1. Projection of the different years as supplementary categories on the first principal plane issued from MFA.

Figure 1 shows the chronological trajectory; every year lies at the centroid of the judgements published during it. This trajectory underlines that the vocabulary evolves at different rates depending on the temporal period. Important vocabulary shifts are detected between 1985 and 1991 on the first axis. Two main reasons account for this shift: the renewal of the Supreme Court magistrates starting in 1985 as a result of forced retirement of members older than 65 years-old, on the one hand, as well as significant reforms in criminal legislation and its institutions during this period. The word mapping, not represented here, shows that the vocabulary evolves from moral arguments (at the beginning of the period), to neutral and technical arguments, looking for evidence and worried about the defence wrights (during the last years). MFA also allows for pointing out the judgements that show a deviation from the common structure. Figure 2 presents the two judgements that present more discordance between chronology and vocabulary. Judgement 207, published in 1989 is already technically argued and gives a great importance to evidence and testimonies, while judgement 328, written in 1993, turns back to arguments corresponding to the former years, linked to moral considerations. In a similar way, the judge-redactors who show a great deviation between period and vocabulary can be pointed out, which allows us to detect those who lead the reform process.

5 Conclusions

The possibility of simultaneously taking into account quantitative information and text offers interesting prospects in Text Mining, in particular in the case of legal studies. Such a tool allows for studying the interconnections between texts and contextual data and can be used with different perspectives.

Acknowledgments: Grant SEJ2005-00741/ECON and Grant SGR 00004/2005.

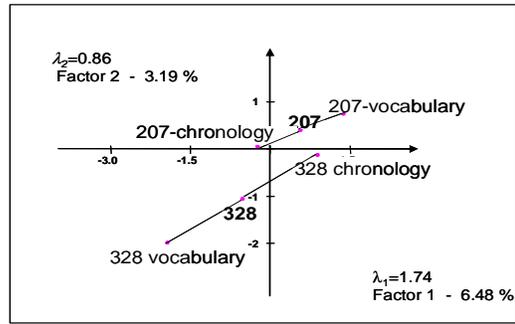


FIGURE 2. Superposed representation of the two judgments showing more discordance between vocabulary and chronology.

References

- Bécue-Bertaut M. and Pagès J. (2007) Analysis of a mixture of quantitative, categorical and frequency data through an extension of multiple factor analysis. Application to survey data. *Compt. Stat. and Data Analysis*. To appear.
- Escofier B. and Pagès J. (1998). *Analyses factorielles simples et multiples* Paris: Dunod.
- Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*. Kluwer.

CTDL-Positive Stable Frailty Model

M. Blagojevic¹

¹ Department of Mathematics, Keele University, Staffordshire ST5 5BG, UK

Abstract: The non-PH Canonical Time Dependent Logistic (CTDL) survival regression model is extended by incorporating a positive stable frailty component into the hazard function within the Bayesian framework. The resulting model is compared numerically with the Weibull-positive stable frailty model, using data from a placebo controlled randomized trial of gamma interferon in chronic granulomatous disease (CGD). Moreover, supremum bounds of the ratio-of-uniforms (ROU) algorithm, used for sampling from complete conditional distributions, are obtained analytically thus yielding a more efficient form of the algorithm.

Keywords: CTDL, Frailty Models, Positive Stable, Non-PH, Ratio-of-Uniforms.

1 Introduction

A flexible non-PH model is the Canonical Time-Dependent Logistic (CTDL) model described by MacKenzie (1996). In our earlier work, the CTDL model was extended to univariate and multivariate gamma frailty models within the frequentist framework using a marginal Likelihood approach and its properties compared with the Weibull-Gamma models analytically and numerically. It was revealed via an extensive simulation study that the Weibull and Weibull-gamma models gave more precise results, in terms of standard errors, than the CTDL and CTDL-gamma models. However, analysis of real data revealed that the CTDL based models provided superior fits to the data.

Allowance is made for a higher degree of heterogeneity among subjects by using infinite variance frailty distributions than would be possible using finite variance frailty (Qiou *et al*, 1999). A positive stable frailty has previously been mixed with PH models, but never with non-PH basic models. Therefore, it was timely to develop a such a new model for the multivariate shared frailty setting using the CTDL hazard as the basic model and comparing its performance with that of the Weibull-positive stable frailty model. Inference was carried out in a Bayesian framework, and bounds of components necessary for the ratio-of-uniforms (ROU) algorithm for sampling from complete conditional distributions were derived successfully, thus improving the efficiency of the algorithm.

2 Positive stable frailty models

Let the random variable U denote unobservable individual frailties. Buckle (1995) gives the joint density of n iid 4-parameter stable distributed rv's by using the joint

pdf of U_i and Y , $f(u_i, y|\omega)$, from which the marginal density of U_i turns out to be the stable pdf.

$$f(u_i|\omega) = \frac{\omega|u_i|^{1/(\omega-1)}}{|\omega-1|} \int_{-1/2}^{1/2} \exp\left[-\left|\frac{u_i}{\tau_\omega(y)}\right|^{|\omega/(\omega-1)|}\right] \left|\frac{1}{\tau_\omega(y)}\right|^{|\omega/(\omega-1)|} dy \quad (1)$$

where

$$\tau_\omega(y) = \frac{\sin(\pi\omega y + \psi_\omega)}{\cos\pi y} \left[\frac{\cos\pi y}{\cos(\pi(\omega-1)y + \psi_\omega)} \right]^{(\omega-1)/\omega}$$

$\omega \in (0, 1)$, $y \in (-1/2, 1/2)$ and $\psi_\omega = \min(\omega, 2-\omega)\pi/2$.

The observed data for the j th time observation for the i th individual is $(t_{ij}, \delta_{ij}, x_{ij})$. Let D_{obs} denote all such triplets. The unobserved data are the frailties $u = (u_1, \dots, u_n)$, so the complete data is $D = (D_{obs}, u)$. Note that u is based on a vector of auxiliary variables $y = (y_1, \dots, y_n)$ in equation (1). So given the data D_{obs} and the parameters of interest, a likelihood and prior for parameters are needed so that a posterior density may be obtained.

2.1 CTDL-positive stable frailty model

A non-PH CTDL regression model is defined by the hazard function

$$\lambda(t|x) = \lambda \exp(t\alpha + x'\beta) / \{1 + \exp(t\alpha + x'\beta)\} \quad (2)$$

where $\lambda > 0$ is a scalar, α is a scalar measuring the effect of time and β is a $p \times 1$ vector of regression parameters associated with fixed covariates $x' = (x_1, \dots, x_p)$. The corresponding survival function is $S(t|x) = \{(1 + \exp(t\alpha + x'\beta)) / (1 + \exp(x'\beta))\}^{-\lambda}$. The observed data likelihood, which is simply the marginal model once the frailty components have been integrated out, is:

$$\begin{aligned} L(\lambda, \alpha, \beta, \omega | D_{obs}) &= \prod_{i=1}^n \int \prod_{j=1}^{m_i} \left\{ \frac{u_i \lambda \exp(t_{ij}\alpha + x'_{ij}\beta)}{1 + \exp(x'_{ij}\beta)} \right\}^{\delta_{ij}} S(t_{ij}|x_{ij})^{u_i} \\ &\times \frac{\omega|u_i|^{1/(\omega-1)}}{|\omega-1|} \int_{-1/2}^{1/2} \exp\left[-\left|\frac{u_i}{\tau_\omega(y_i)}\right|^{|\omega/(\omega-1)|}\right] \\ &\times \left|\frac{1}{\tau_\omega(y_i)}\right|^{|\omega/(\omega-1)|} dy_i du_i \end{aligned} \quad (3)$$

The posterior density, expressed in terms of the observed data likelihood and the joint prior for the parameters is:

$$\pi(\lambda, \alpha, \beta, \omega | D_{obs}) \propto L(\lambda, \alpha, \beta, \omega | D_{obs}) p(\lambda) p(\alpha) p(\beta) p(\omega) \quad (4)$$

where $\pi(\cdot)$ denotes the posterior and $p(\cdot)$ the prior distribution. Note that independence among all parameters is assumed.

The integrals in equation (3) do not have a closed form, so instead the unknown parameter vector $(\lambda, \alpha, \beta, \omega)$ is augmented with vectors u and y and MCMC methods are used to obtain samples for $(\lambda, \alpha, \beta, \omega, u, y)$.

Complete conditional distributions are needed for $\lambda, \alpha, \beta, \omega, u$ and y from which the corresponding samples are to be drawn; They are derived as being proportional to (4). The choice of priors is given in table 1. All resulting complete conditional distributions are of non-standard forms and we use ROU algorithm for generation of all quantities apart from y and ω . We have developed a more efficient way of implementing the ROU algorithm, namely deriving its components analytically instead of performing numerical bisection. The detailed outline of the approach is too extensive to be re-produced here and will be presented elsewhere.

The familiar Weibull regression distribution has the hazard and survival functions given by $\lambda(t|x) = \lambda\rho(t\lambda)^{\rho-1} \exp(x'\beta)$ and $S(t|x) = \exp(-(t\lambda)^\rho e^{x'\beta})$, respectively. Model development follows in the same steps as for the CTDL model with MCMC methods being used to obtain samples for $(\lambda, \rho, \beta, \omega, u, y)$.

3 Example Data Analysis

The models outlined above are now applied to a data set from a placebo controlled randomized trial of gamma interferon in chronic granulomatous disease (CGD). Treatment was given to each of the 128 patients at the first scheduled visit for that patient. The data for each patient give the time to first and any recurrent serious infections. For bivariate data, only information pertaining to first 3 records is needed from which the gap times and censoring indicators are calculated. Only one factor, gender, is included in the analysis.

Prior and hyperparameter specifications for the two models are given in table 1. Gibbs sampler is used to generate samples from the derived complete conditional distributions; S-Plus (V4.5) was used for programming. 5000 iterations were taken as "burn-in" and a further 5000 iterations taken for inference purposes. Gelman and Rubin convergence statistic for the parameters of the two models indicated successful convergence for all parameters (table not shown). Posterior distributions are summarized in terms of means and standard errors of each parameter in the two models in table 2. Gender is significant in both models, its negative estimated effect meaning that females are at a lower risk of being infected. Values $\omega = 0.751$ in the CTDL case and $\omega = 0.622$ in the Weibull case correspond to a reasonable degree of dependence between times of each patient (note that the value of 1(0) implies maximum dependence(independence)), more so in the CTDL case.

4 Remarks and Future work

We have shown that the CTDL-positive stable frailty model confirmed a higher degree of dependence between individual observations than the corresponding Weibull model. Sensitivity to prior specifications for models considered here will be the subject of future work. We have already developed the CTDL and Weibull models with gamma

TABLE 1. Prior and Hyperparameter specifications

Parameter	Prior	Hyperparameter specifications
$\lambda_{weibull}$ λ_{ctdl}	Gamma(γ, γ)	$\gamma = 0.001$
ρ	Gamma(μ, μ)	$\mu = 0.001$
α	Normal(ξ, ν)	$\xi = 0, \nu = 1000$
β	Normal(ϵ, m)	$\epsilon = 0, m = 1000$
ω	$p(\omega) = 1$	$0 < \omega < 1$

TABLE 2. Posterior Summary of Model Parameters

CTDL + Positive Stable Frailty		
Parameter	Posterior Mean	Posterior S.E.
λ_{ctdl}	0.236	0.017
α	-0.097	0.032
β	-0.098	0.039
ω	0.751	0.082
Weibull + Positive Stable Frailty		
Parameter	Posterior Mean	Posterior S.E.
$\lambda_{weibull}$	0.058	0.015
ρ	1.342	0.321
β	-0.073	0.021
ω	0.622	0.066

frailty in Bayesian framework and a comparison of these with their corresponding “frequentist” counterparts will be the subject of future work.

References

- Buckle, D.J. (1995) Bayesian Inference for Stable Distributions. *Journal of the American Statistical Association* **90**, 605-613.
- MacKenzie, G. (1996) Regression models for survival data: the generalised time dependent logistic family. *JRSS Series D* **45**, 21-34.
- Wakefield, J.C. *et al.* (1991) Efficient generation of random variates via the ratio-of-uniforms method. *Stat. Comput.* **1**, 129-133.

Beyond Kappa: Use of multifaceted RASCH analysis and multilevel modelling to investigate observer effects

A. Blance^{1,2}, J. Carvalho³ and M.S. Gilthorpe¹

¹ Biostatistics Unit, Centre for Epidemiology & Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, University of Leeds, 30-32 Hyde Terrace, Leeds, LS2 9LN, UK

² Leeds Dental Institute, University of Leeds, Clarendon Way, Leeds, LS2 9LU, UK

³ School of Dentistry, Catholic University of Louvain, Brussels, Belgium

Abstract: Kappa is often adopted to demonstrate a 'sufficient' level of calibration amongst observers. Observer effects are then erroneously ignored in subsequent analyses. Multifaceted RASCH analysis and multilevel modelling are employed to illustrate how observer effects can be investigated and incorporated into the analysis. This situation arises commonly in calibration exercises and is illustrated with data investigating observer effects in dental caries.

Keywords: Kappa; multifaceted RASCH analysis; multilevel modelling; calibration.

1 Introduction

It is common in (health) research to calculate some measure of agreement. For the case of categorical observations, Cohen's Kappa statistic is frequently utilized. Kappa can be expanded from the simplest case of a binary outcome with two observers, to the situation of many categories and observers. Although documented (Feinstein and Cicchetti, 1990), difficulties in the interpretation of Kappa are still not widely appreciated. The 'standard' hypothesis - that the agreement is better than expected by chance alone - is erroneous. Rather, Kappa should be assessed for non-inferiority to one. This requires typically very large sample sizes: with the minimum sample size required being 250 pairs of observations, though typically the minimum number is at least one order of magnitude greater than this (Blance and Gilthorpe, Submitted). Further, the interpretation of Kappa depends on the study scenario. Kappa should not be blindly assessed against a 'one size fits all' criterion. Sample sizes required to estimate Kappa robustly are rarely achievable and estimates of Kappa are usually underpowered. Kappa values close to one are rarely achieved and research analyses that subsequently ignore the observer effect may lead to erroneous conclusions being drawn. This paper looks beyond the use of Kappa and considers two solutions to this issue (1) Multifaceted RASCH modelling and (2) Multilevel modelling.

1.1 Example data

The following analyses draw on a dataset recording the carious status of 26 subjects, each observed by 26 observers (clinical dentists). Data collection began once the 26

observers were judged to be calibrated. The carious status (measured on a 10 point ordinal scale) of each surface (5), around each tooth (28) was recorded for all 26 subjects by all 26 observers. Thus, each observer made 3640 observations (140 observations per subject). All observations were recorded on a standard file.

2 Methods

2.1 Multifaceted RASCH analysis

The package 'RUMM2020' (Rummlab, 1998) was used to fit the multifaceted RASCH model (Rasch, 1960). The number of facets was 26 (number of observers). Each surface was considered as an individual item, thus each of the 26 subjects had 140 items. A suitable fit to the RASCH model was first sought. This necessitated collapsing some of the categories where disordered thresholds existed. Once an acceptable fit to the RASCH model was achieved, the observer effect (facet) was investigated.

2.2 Multilevel modelling

Multilevel modelling was first conducted using a two-level model, observer at the lowest level (level 1) and surface at level 2 (Hox, 2002). The model allowed random variation at both levels. Thus, the variation at level 1 is a measure of observer variation. Subsequent extensions involved specifying three-level (subject, surface, observer) and four-level models (subject, tooth, surface, observer). Neither extension provided further insight into the observer variation, so the two-level model was adopted for simplicity. All analyses were performed in MLwiN (Centre for Multilevel Modelling, 2007).

3 Results

It is illustrated that Kappa has issues pertaining to its calculation that lead to difficulties in its interpretation. Pairwise agreement amongst the observers (measured using Kappa) ranged from 0.41 to 0.80. It was clear that agreement was at best variable; some pairs having very poor agreement. Further, investigation suggested that certain pairs had (reasonable) agreement, suggesting possible 'traits' amongst the observers. Agreement was not sufficient to ignore observer effects. Multifaceted RASCH analysis revealed several interesting features. In ranking the items in level of difficulty, it could be seen that not all disagreements lay at the difficult end of the spectrum. A small group of observers disagreed with the majority at the 'easier' lower end. Observers diverged across the centre of the scale, where the majority of the observations happened to lie. Multilevel modelling was successful in explicitly incorporating the observer effects into the model. Observer effects were substantial. The flexibility of this approach is being exploited with further investigation currently been undertaken.

4 Conclusions

The dataset analyzed is typical of calibration exercises and research involving more than one observer. This work demonstrates the benefits of adopting alternative approaches to Kappa: observer effects should not be ignored as this can potentially lead to erroneous conclusions being drawn; costly calibration exercises are not (always) necessary.

References

- Blance, A., and Gilthorpe, M.S. Do we categorically agree?: A consideration of Kappa. *BMC Medical Research Methodology*. Submitted.
- Centre for Multilevel Modelling (2007). <http://www.cmm.bristol.ac.uk/MLwiN/index.shtml>. Bristol: England. [15 Feb 2007]
- Feinstein, A.R., and Cicchetti D.V. (1990). High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* **43**, 543-549.
- Hox, J. (2002). *Multilevel Analysis Techniques and Applications*. Manwah, New Jersey: Lawrence Erlbaum Associates, Inc.
- Rummlab (1998). <http://www.rummlab.com/>. Australia. [15 Feb 2007]
- Rasch G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Asymmetric distributions generated by the normal distribution function

Heleno Bolfarine¹

¹ University of São Paulo, Department of Statistics, Rua do Matão, 1010, CEP 05508 - 090, São Paulo, S.P., Brasil

Abstract: In this paper we propose an alternative asymmetric model to the usual skew normal model proposed in Azzalini (1985). The main advantages of the alternative model are the fact that it encompasses more general asymmetry data, that is, the asymmetry range is wider than the one resulting for the skew normal and the Fisher information matrix is not singular in situations where the asymmetry parameter is null. Estimation is developed by using the maximum likelihood approach.

Keywords: skew normal; skew t-normal; asymmetry range, maximum likelihood

1 Asymmetric models

The usual standard skew normal distribution defined by Azzalini (1985) presents a density function given by

$$f(x) = 2\phi(x)\Phi(\alpha x).$$

This model can be easily extended to location-scale, regression and multivariate settings. It is also well known that it presents some inference difficulties. The maximization with respect to α can be infinity (all observations are negative or positive) and it may be very large (likelihood unbounded) very often. The asymmetry range is $(-0.956, 0.956)$, which can not capture all the asymmetry when data presents higher degree of it. Further, information matrix is singular when $\alpha = 0$. EM algorithm can be tried but it also presents difficulties (see Arellano-Valle et al., 2005). A more general family of models can be define by replacing ϕ and Φ above by any density g and any distribution function G , ending up with the general family

$$f(x) = 2g(x)G(\alpha x),$$

which can also be extended to location-scale, regression and multivariate settings. By making use of this general representation, Nadarajah and Kotz (2003) introduced several asymmetric models by taking $g(\cdot)$ as the density of the normal distribution and $G(\cdot)$ as the distribution function of the normal, logistic, uniform and exponential power distributions. We consider instead, $g(\cdot)$ as the density of the Student-t distribution and $G(\cdot)$ as the distribution function of the normal distribution. More details related to model properties are given in Gomez et al. (2007).

2 The skew t-normal model

The main object of this paper is to consider the asymmetric model where g above is the Student-t density with ν degrees of freedom and G is the distribution function of the normal distribution. The location-scale representation of this model presents a density function given by

$$f(x) = 2K(\sigma, \nu) \left(1 + \frac{(x - \mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \Phi\left(\alpha \frac{x - \mu}{\sigma}\right),$$

where $k(\sigma, \nu)$ is a constant depending on ν and σ .

By computing moments for this model, it can be shown that the asymmetry range for the above model can be much wider than the one for the usual skew normal distribution. We have shown that for 5 degrees of freedom, the range for the asymmetry parameter is $(-2.55, 2.55)$, much wider than the corresponding one for the skew normal models. Estimation is developed by using maximum likelihood, which can be done by direct maximization using routines such as BFGS in Ox or by working with the EM algorithm, which can be implemented by using the fact that the Student-t distribution can be obtained as a mixture of the normal and chi-square distributions so that in each step, the complete likelihood reduces to the likelihood corresponding to a skew symmetric distribution (Arellano-Valle et al., 2005). It can also be shown that the Fisher information matrix for the location-scale situation, with $\alpha = 0$ is given by

$$\mathbf{I}_F(\theta) = \begin{pmatrix} \frac{\nu+1}{\sigma^2(\nu+3)} & 0 & 0 & \frac{2}{\sigma\sqrt{2\pi}} \\ 0 & \frac{2\nu}{\sigma^2(\nu+3)} & \frac{-2}{\sigma(\nu+1)(\nu+3)} & 0 \\ 0 & \frac{-2}{\sigma(\nu+1)(\nu+3)} & -h(\nu) - \frac{\nu-3}{2\nu^2(\nu+1)(\nu+3)} & 0 \\ \frac{2}{\sigma\sqrt{2\pi}} & 0 & 0 & \frac{2\nu}{\pi(\nu-2)} \end{pmatrix},$$

where

$$h(\nu) = \frac{1}{4} \left[\frac{2}{\nu^2} + \frac{\partial \Psi(\frac{\nu+1}{2})}{\partial \nu} - \frac{\partial \Psi(\frac{\nu}{2})}{\partial \nu} \right],$$

with $\Psi(\cdot)$ as the digamma function, implying that the Fisher information matrix is not singular for ν finite. Hence, large sample confidence intervals can be constructed by using the inverse of the Fisher information matrix evaluated at the maximum likelihood estimators. Moreover, likelihood ratio statistics, Wald and score statistics will have the usual chi-square asymptotic distribution, which can be used for model comparison.

References

- Arellano-Valle, R., Bolfarine, H., Ozan, S. and Lachos, V.H. (2005). Skew normal measurement error models. *Journal of Multivariate Analysis* **96**, 265-281.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171-178.

Gómez, H.W., Venegas, O. and Bolfarine, H. (2006). Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics*. To appear.

Nadarajah, S. and Kotz, S. (2003). Skewed distributions generated by the normal kernel. *Statistics and Probability Letters* **65**, 269-277.

Spatio-Temporal Bayesian Model for studying waterbird biodiversity in artificial ponds

F. Botella¹, X. Barber³, A. López-Quílez², A.M. Mayoral³, J. Morales³, J.A. Sánchez-Zapata¹ and E. Sebastián¹

¹ Dept. Applied Biology, Miguel Hernández University

² Dept. Statistics and Operations Research, University of Valencia

³ Applied Statistical Unit-Operations Research Center, Miguel Hernández University

Abstract: Agriculture practices linked to intensification processes have been identified as the main forces driving biodiversity loss in European countries. Wetlands and associated biota are among the ecosystems more affected by drainage and water use for intensive irrigation. The Vega Baja valley region, in southeastern Spain, is a typical example of this irrigation transformation. As a consequence, thousands of ponds have been constructed for drip irrigation and could play an important role as alternative or surplus habitat offer for wintering waterbirds. We propose a Bayesian hierarchical model considering geographical situation, to study the abundance and richness of waterbirds in a sample of the artificial ponds of the Vega Baja valley.

Keywords: Artificial ponds; Hierarchical Bayesian model; Model based Geostatistics..

1 Introduction

The transfer of water from basins of surplus supply to those in deficit has become an increasingly common answer to the redistribution of water. This is specially so in arid and semi-arid areas where human activities largely rely on such water supplies. The ecological impacts of such inter-basin water transfers include the introduction of exotic species, the loss of biogeographical integrity, the alteration of hydrological regimes, including marine and estuarine processes and water quality, among others. Some of these impacts have been documented for the water transfer between Tajo and Segura Rivers from central to southeastern Spain, including the transformation of about 200000 ha of extensive agriculture into intensive irrigation crops.

The Vega Baja valley, in southeastern Spain, is an important agricultural production area at EU and counts with important natural and seminatural wetlands and salines with different international status because of their international importance for waterbird conservation. Some of these places are *El Hondo*, *Salinas de Santa Pola*, *Salinas de La Mata and Torrevieja*, and *Salinas de San Pedro*. As a consequence of intensive irrigation, thousands of artificial ponds have been constructed for drip irrigation. With an adequate management and structural design, could play a role for some waterbird species conservation, because they can be used as alternative or surplus habitat for wintering waterbirds.

Artificial ponds vary in size, construction materials and distance to natural or semi-natural wetlands. The objective of this work is to determine wintering waterbirds use of these agricultural facilities and to explore the effect of structural attributes and geographical situation on pond performance as alternative habitat for waterbirds. Our aim is to provide essential elements to the regional and European authorities in order to conceal agricultural production with biodiversity conservation and management.

2 Description of Data

A sample of 204 ponds from the Vega Baja valley was obtained from a total of 3000 in the region. The censuses of birds was carried out during 2002, 2003, 2004 and 2005 winters by at least one observer using binoculars and scopes (Koskimies and Väisänen, 1991). Each pond was observed for 8 minutes and the number of different species was identified and marked. The poor vegetation cover of the ponds and their small size reduces the census error (Dawson, 1985). From this information it was obtained the abundance and richness for each pond. These variables are usual in biodiversity studies. The ponds were classified into two categories depending on their construction materials. One kind was built using low-density polyethylene (LDP) and was then covered with sand and gravel to prevent from external aggressions. The other kind was constructed using high-density polyethylene (HDP) like PVC or other materials without the cover. We used digitalized aerial photographs, and a geographic information system to calculate the position and the size of the ponds and the distance to the closest wetland.

3 Bayesian Spatio-Temporal Model

Statisticians are increasingly faced with the task of analyzing data that are temporal and geographically referenced (Besag, York and Moillé, 1991, and Banerjee, Carlin, Gelfand, 2004). Let NAV_{ij} and E_{ij} be the observed and expected abundance of waterbirds for pond i at year j , standardized by size of the pond and proximity to the coast. Let TC_i represent the construction material for pond i , with value 1 for the HDP material (0 for LDP), and TDH_i be the distance to the closest wetland for pond i (in excess of the mean of all distances). Our spatio-temporal proposal links the spatial and temporal components through an autoregressive model, in the following way:

$$NAV_{ij} \sim \text{Poisson}(E_{ij}RA_{ij}), \quad i = 1, \dots, 204; j = 1, \dots, 4$$

where RA_{ij} stands for the relative abundance at each pond and time interval under study. We define the log-relative abundance first time observed as:

$$\log(RA_{i1}) = \mathbf{X}_i\boldsymbol{\beta} + temp_1 + s_{i1} + b_{i1}$$

where \mathbf{X}_i is the i th-row of the design matrix with columns $(1, tc_i, tdh_i, tc_i * tdh_i)$, $\boldsymbol{\beta}$ is the vector of parameters, $temp_1$ is the temporal effect first time observed, b_{i1} is the vector of heterogenous random effects given by

$$b_{i,1} \sim N(0, \sigma_{b(1)}^2),$$

and s_{i1} is the vector of random spatial effects defined by:

$$S_1 \sim N(0, \sigma_{e(1)}^2 H(\phi_1)); \quad (H(\phi_{e(1)}))_{i,i'} = \exp(-(\phi_{e(1)} d_{i,i'})^{\kappa_{e(1)}}),$$

where $d_{i,i'}$ is the distance between pounds i and i' .

At successive time intervals we define the relative abundance as:

$$\log(RA_{ij}) = \mathbf{X}_i \boldsymbol{\beta} + temp_j + s_{ij} + b_{ij} + \rho(s_{i(j-1)} + b_{i(j-1)}), \quad j = 2, \dots, 4$$

where

$$b_{i,j} \sim N(0, \sigma_{b(2)}^2), \quad j = 2, \dots, 4$$

$$S_j \sim N(0, \sigma_{e(2)}^2 H(\phi_j)); \quad (H(\phi_{e(2)}))_{i,i'} = \exp(-(\phi_{e(2)} d_{i,i'})^{\kappa_{e(2)}}), \quad j = 2, \dots, 4$$

$$temp_{1:4} \sim CAR(\tau_t).$$

The quite noninformative hyperpriors are given by:

$$\rho \sim Un(-1, 1); \quad \phi_j \sim Un(0.00005, 0.1282), \quad j = e(1), e(2)$$

$$\boldsymbol{\beta} \sim N(0, 10000); \quad 1/\sigma_k^2 \sim Ga(0.5, 0.005), \quad k = e(1), e(2), b(1), b(2).$$

$$\kappa_j \sim Un(0.05, 1.95), \quad j = e(1), e(2).$$

We use Gelfand and Goshn criteria to select the best model. Table 1 shows the posterior median and posterior credible interval with probability 0.95 for the parameters of this model.

TABLE 1. Posterior summaries of selected model.

Parameter	Median	CR 0.95	Parameter	Median	CR 0.95
β_0	-0.750	(-1.134 , -0.468)	ρ	0.540	(0.419 , 0.655)
β_{tc}	-0.525	(-0.949 , -0.107)	$\sigma_{e(1)}^2$	3.095	(2.232 , 4.413)
β_{tdh}	-0.299	(-1.217 , 0.471)	$\sigma_{e(2)}^2$	2.113	(1.650 , 2.711)
$\beta_{tc:tdh}$	-0.404	(-1.681 , 1.081)	$temp_1$	0.026	(-0.144 , 0.209)
$\kappa_{e(1)}$	1.043	(0.092 , 1.909)	$temp_2$	0.025	(-0.076 , 0.175)
$\kappa_{e(2)}$	0.973	(0.212 , 1.899)	$temp_3$	0.001	(-0.114 , 0.130)
$\phi_{e(1)}$	0.066	(0.014 , 0.125)	$temp_1$	-0.051	(-0.284 , 0.099)
$\phi_{e(2)}$	0.083	(0.021 , 0.126)			

4 Conclusions

- Spatial random effects are very important in terms of goodness of fit, and heterogeneous random effects are non relevant in this model. Temporal random effects decrease with the passing of the time.
- Abundance of waterbirds is higher for HPD construction and closeness to some wetland.

Acknowledgments: This research was supported by the Spanish Ministry of Education and Science, under Grant MTM2004-03290, and by the Autonomous Valencia Government, under Grant ACOMP06/2005.

References

- Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical modelling and analysis for spatial data*. Boca Raton: Chapman and Hall.
- Besag, J., York, J., and Moillé, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-21.
- Dawson D.G. (1985). A review of methods for estimating bird numbers. In: *Taylor K., Fuller R.J. and Lack P.C. (eds), Bird Census and Atlas Studies. BTO, Tring.* 27-33.
- Koskimies, P. and Väisänen, R.A. (1991). *Monitoring bird populations*. Finish Museum of Natural History. Helsinki. Finland.

Temporal Smoothing of Compositional Data on Water Quality

Mark J. Brewer¹, Dörthe Tetzlaff², Susan Waldron³ and Chris Soulsby²

¹ Biomathematics and Statistics Scotland, Macaulay Institute, Craigiebuckler, Aberdeen, Scotland, AB15 8QH, UK

² Geography & Environment, School of Geosciences, University of Aberdeen, Aberdeen, AB24 3UF, UK

³ Department of Geographical and Earth Sciences, University of Glasgow, Glasgow, G12 8QQ, UK

Abstract: We consider temporal smoothing at two different timescales within a compositional analysis of water quality monitoring data. Our aims are (i) to account for autocorrelation to enable better estimation of the relationship between water quality and flow rates, an important issue with changing weather patterns, and (ii) to disentangle between- and within-day autocorrelation.

Keywords: Compositional analysis; Temporal smoothing; Water quality.

1 Introduction

Motivated by the Water Framework Directive (WFD), there is an increasing level of water quality monitoring in UK rivers. Deployment of continuous water quality loggers in river catchments has provided data sets at very high temporal resolutions. We consider a data set at 15-minute resolution, from a stream in a sub-catchment of the River Feugh in north-east Scotland, collected over one year from October 2003 (see Tetzlaff *et al.*, 2007).

Given (derived) readings on (Gran) alkalinity from stream water and similar readings from samples of the different sources of the water, we can estimate the proportions of each source present in each stream via the method of Brewer *et al.* (2005). By including the recorded flow rate as a covariate in the model, we can assess the impact of flow on the source fractions. This is an important consideration since changing patterns in flow rate may be a consequence of climate change.

Since the data are now recorded at such a high frequency, it is especially important to account for any temporal correlation. We would like to differentiate between very short-term autocorrelation within the readings themselves and the slightly longer-term “trend” of day-to-day variation. With this in mind, we consider introducing *two* levels of random effects for temporal smoothing using random-walk priors. These both represent RW(1) priors, but the day-to-day trends correspond to a set of random effects using an index denoting the days; the short-term smoothing is the more standard implementation described in the GeoBUGS manual (Thomas *et al.*, 2004) connecting the individual 15-minute time points.

2 Compositional analysis

As our model has been introduced elsewhere (see Brewer *et al.*, 2002, and Brewer *et al.*, 2005, for example), we provide only a brief description here. We assume there are N sources, for which there are A markers. We have direct measurements of markers for each source, denoted x_j^i for $j = 1, \dots, N$ with $i = 1, \dots, n_j$ samples from each source, and where x_j is a vector of length A . We assume multivariate normality of x_j with mean vectors μ_j and covariance matrices Σ_j , $j = 1, \dots, N$, hence

$$x_j^i \sim \text{MVN}_A(\mu_j, \Sigma_j), \quad j = 1, \dots, N, \quad i = 1, \dots, n_j. \quad (1)$$

Also, we have measurements on mixture (stream) samples, denoted similarly by z^k for $k = 1, \dots, n_z$ samples (z^k being another A -vector).

Further, we assume that each mixture sample is comprised of a weighted combination of *latent* draws from the source distributions at (1); we denote these latent quantities by y_j^k , and hence

$$y_j^k \sim \text{MVN}_A(\mu_j, \Sigma_j), \quad j = 1, \dots, N, \quad k = 1, \dots, n_z;$$

in addition, the weights are the source proportions, and hence we can model the mixture samples via

$$z^k = \sum_{j=1}^N p_j^k y_j^k + \epsilon^k, \quad k = 1, \dots, n_z;$$

where ϵ^k is an A -vector of measurement errors having zero mean (vector) and covariance Σ_ϵ . To model the composition proportions p_j , we use *log-ratios* (Aitchison, 1986) $q_j^k \equiv \log(p_j^k/p_N^k)$ for $j = 1, \dots, N-1$, and the overall log-ratio composition $(N-1)$ -vector q^k is also assumed multivariate normal, with

$$q^k \sim \text{MVN}_{N-1}(\mu_q, \Sigma_q), \quad k = 1, \dots, n_z,$$

for the mean vector μ_q on the log-ratio scale, and covariance Σ_q .

We complete the model definition by defining appropriate diffuse priors: multivariate normal for the mean vectors μ ; and Wishart priors on the reciprocals of the covariance matrices Σ . We make inferences on the model via Markov chain Monte Carlo methods using the WinBUGS (Spiegelhalter *et al.*, 2004) package.

3 Flow-Rate Covariate and Temporal Smoothing

The model as proposed permits extensions to, for example, the inclusion of covariates and random effect terms. We have flow rate information in the form of a covariate v . We also wish to include random effects; in our current context we account for temporal autocorrelation with a standard Gaussian intrinsic autoregression Markov random field (MRF) prior (Besag *et al.*, 1991) which, for a vector of univariate random variables $u = (u_1, u_2, \dots, u_n)^T$, we define as

$$f(u | \tau) \propto \exp \left\{ -\frac{\tau}{2} \sum (u_{i+1} - u_i)^2 \right\}$$

where τ is a parameter controlling the level of smoothing. In fact, we define two such terms—we allow random effects u to represent very short-term smoothing between the 15-minute readings, but we also create a dummy variable d (having values 1 to D) to index the calendar days and define random effect $w = (w_1, w_2, \dots, w_D)^T$ to model smoothing over a longer time frame. Thus we allow the log-ratio means to vary by observation:

$$\mu_q^k = \mu_q^* + v_k \beta + u_k + w_{d(k)}$$

for $k = 1, \dots, n_z$, where β is an $(N - 1)$ -vector of parameters, $d(k)$ is the day for observation k and μ_q^* is now the overall mean on the log-ratio scale. Finally, we redefine the distribution for the log-ratios like so:

$$q^k \sim \text{MVN}_{N-1}(\mu_q^k, \Sigma_q).$$

4 Results

Figure 1 shows the fitted proportion of groundwater in the stream over time, which unsurprisingly is strongly inversely related to the flow rate (first two plots). We also see plots of the smoothed random effects; the longer-term daily trend shows, for example, a gradual increase in groundwater proportion (on the log-ratio scale) from November to July over and above the effect of flow. The RW(1) effects are effectively detrended compared to results from fitting a similar model without the daily trend term; the smoothing is not so evident here since there are over 35000 observations. The regression coefficient for flow is estimated as 0.491 (95% interval [0.473, 0.509]) in the smoothing model, interpreted as a factor increase of 1.63 in the ratio of groundwater to other water per $\log 1 \text{ s}^{-1} \text{ km}^{-2}$; this compares with a figure of 1.296 (95% interval [1.283, 1.308], and a factor increase per $\log 1 \text{ s}^{-1} \text{ km}^{-2}$ of 3.65) for a model not accounting for autocorrelation. This difference highlights the need to use an appropriate model.

5 Further work

The model suggested here is rather simplistic, in that there will be “steps” in the fitted proportions over time caused by the coarse-scale random effects for each day. Of course, when data sets are large, as in this case, and when the relative movements of the curve or surface are small, this will not be too much of a concern.

One way to resolve this problem might be to allow the RW(1) process to have a (much) larger variance at the transitions between days; this might then allow the RW(1) values at the transition to be further apart, helping to counteract the effect of the step.

Perhaps a neater way to solve the problem would be to fit a model such that the daily random effects are modelled, but what feeds into the linear predictor is in fact an interpolated version, avoiding the step directly. Our limited experiments in this regard have proved promising, using the linear interpolation facility in WinBUGS. Sampling, as one might expect, is very slow in such a model, but improvements appear possible if we centre the RW(1) process on the interpolated version of the daily process. This avoids steps, but of course a linear interpolation is still jagged. A smooth interpolation would appear preferable, and this is especially true if the ideas here are to be translated

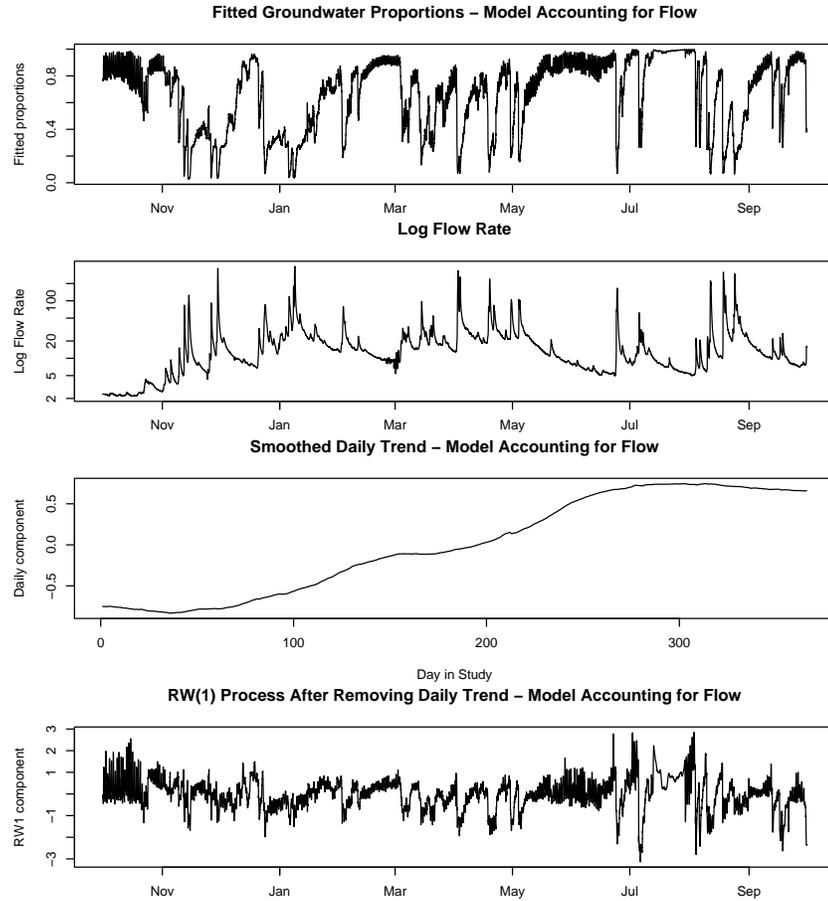


FIGURE 1. From top to bottom, time series plots of: fitted groundwater proportions; log-flow rate in $\log \text{l s}^{-1} \text{ km}^{-2}$; smoothed daily trend from the RW(1) process on days; and the RW(1) process on the 15-minute scale.

to the spatial rather than temporal domain. Kriging would seem to be the obvious choice, and this will be the subject of future work.

Also of concern are issues such as coherence and identifiability. The latter is easily corrected for if required by the use of appropriate constraints, whereas the issue of coherence was addressed by De Iorio and Lavine (2005), who proposed an algorithm for obtaining nearly-coherent MRFs at different scales. Future work will use a simulation study to compare simple use of random effects at multiple scales with the algorithmic approach of De Iorio and Lavine (2005).

Acknowledgments: M.J. Brewer is funded by the Scottish Executive Environment and Rural Affairs Department. S. Waldron is funded by the Natural Environment Research Council.

References

- Aitchison J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-59.
- Brewer, M.J., Dunn, S.M., and Soulsby, C. (2002). A Bayesian model for compositional data analysis. In: *Proceedings of COMPSTAT 2002*, Physica-Verlag, Heidelberg, 105-110.
- Brewer, M.J., Filipe, J.A.N., Elston, D.A., Dawson, L.A., Mayes, R.W., Soulsby, C., and Dunn, S.M. (2005). A hierarchical model for compositional data analysis. *Journal of Agricultural, Biological and Environmental Statistics* **10**, 19-34.
- De Iorio, M., and Lavine, M. (2005). Intrinsic autoregressions at multiple resolutions. *Journal of Statistical Planning and inference* **134**, 102-115.
- Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2004). *WinBUGS user manual*. MRC Biostatistics Unit, Cambridge, UK.
- Tetzlaff, D., Waldron, S., Brewer, M.J., and Soulsby, C. (2007). Assessing nested hydrological and hydrochemical behaviour of a mesoscale catchment using continuous tracer data. *Journal of Hydrology* **336**, 430-443.
- Thomas, A., Best, N., Lunn, D., Arnold, R., and Spiegelhalter, D. (2004). *GeoBUGS user manual, version 1.2*. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK.

Explaining electoral participation by an economic capacity index in Barcelona

Sonia Broner¹ and Pedro Delicado¹

¹ Universitat Politècnica de Catalunya

Abstract: This work shows an application of spatial data analysis techniques to the results of the 2004 General Elections in the city of Barcelona. Specifically we study the relationship between electoral participation rate and an economic capacity index computed by the Barcelona city council.

Keywords: Electoral results; Lattice data; Spatial Data Analysis; Spatial regression model.

1 Introduction

In Political Science voting in elections is considered as the most relevant way of participation in the political system (Anduiza and Bosch 2004). Therefore the participation rate in a given election has deserved special attention by political scientist. Specialists agree that the participation rate in different electoral districts is affected by social and economic conditions of their inhabitants. Usually people in higher social classes tend to vote more than people in lower classes, and economic capacity is also positively correlated with participation. In order to establish such kind of relations based on data, the geographic distribution of electoral districts has to be taken into account because electoral behaviour, as well as other potential explanatory variables, can be affected by spatial dependence.

In this work we analyze the electoral results corresponding to general elections held in Spain in March 14th, 2004. We concentrate on the city of Barcelona, at level of Small Research Zones (ZRP, from their initials in Catalan), an official division of the city in 248 areas with almost equal number of inhabitants. We study the relationship between participation rate and an household economic capacity index (ICEF, from their initials in Catalan) computed by the Barcelona city council. Data for both variables of interest are available for the 248 areas. Spatial dependence is taken into account by using regression models specifically designed for data with this kind of dependence.

In Section 2 we describe the data we are working with. Then a Spatial Lag Model is fitted in Section 3. Conclusions are summarized in Section 4

2 Description of the data

We analyze the structure of spatial dependency of the two variables of interest: participation rate and economic capacity (ICEF). The Exploratory Analysis is made with *GeoDATM*, a free software for the analysis of data of areas (Anselin, 2003). Figure 1

shows some of the results. Box-plot maps for both variables are included. They show that high values of both variables are concentrated in the upper part of the city and in some center neighbourhoods. This was an expected result, based on the known relationship between participation and social classes distribution, as well as on the knowing of Barcelona idiosyncrasy.

The Moran's index I , a global index of association, is calculated for both variables and the Moran's scatter-plots are shown in Figure 1 (see there the values of Moran's I). It is observed that there is a systematic variation in both Moran's scatter-plots, implying that both variables exhibit space autocorrelation.

3 Data modelling

The observed dependence between participation rate and economic capacity, including, spatial dependence, can be modelled with a Spatial Lag Model (Anselin, 1988), also known as spatial regression. The general form of the Spatial Lag Model is

$$y = \rho W y + X \beta + \xi,$$

where y is the vector of observations on the dependent variable, W is a spatial weights matrix that contains information on the neighborhood structure for each location, $W y$ is the spatially lagged dependent variable for weights matrix W , X is the matrix of observations on the explanatory variables, ξ is a vector of i.i.d. error terms, and ρ and β are parameters.

In our case, y is the participation rate, X is the economic capacity index (ICEF) and we define the spatial weights matrix W by $W_{i,j} = 1$ if and only if areas i and j are neighbors. The model is fitted by maximum likelihood (assuming normality for the error term ξ) using *GeoDATM*. The resulting estimated model is

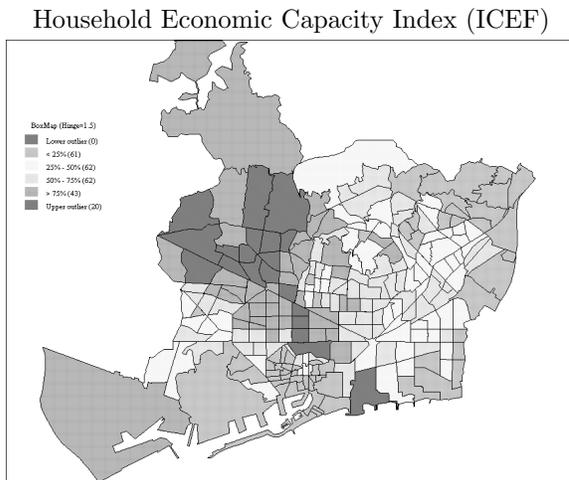
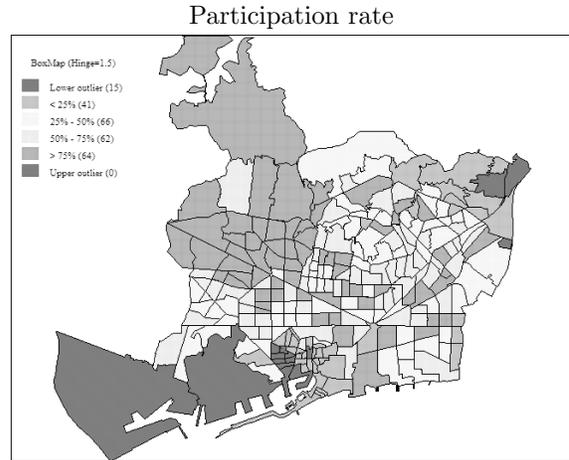
$$\text{PARTIC} = 0.2412073 + 0.5922076 \text{ W_PARTIC} + 0.0007057174 \text{ ICEF}$$

$$(0.0401384) \quad (0.0597108) \quad (0.0001163594)$$

where standard errors of the estimated coefficients are indicated in brackets. The p-values associated with the individual test of zero coefficients are lower than $1e-7$. The estimated residual standard deviation is 0.045619. The R -squared is 0.572901 for this model.

4 Conclusions

We conclude that, in the city of Barcelona, participation rate is positively affected by the economic capacity index. Spatial dependence of both variables of interest has been detected and taken into account in the model fitted to the data.



Variable	Min.	Q1	Median	Mean	Q3	Max.
Participation	0.33	0.74	0.79	0.76	0.81	0.87
ICEF	48.70	82.50	94.50	99.09	106.90	309.90

Acknowledgments: Research partially supported by the Spanish Ministry of Education and Science and FEDER, MTM2006-09920, and by the EU PASCAL Network

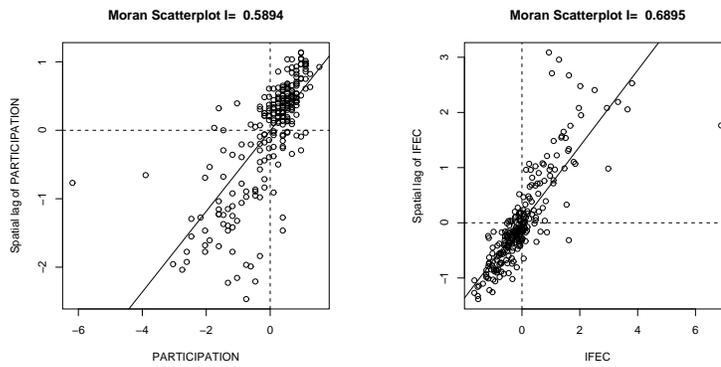


FIGURE 1. Graphical and numerical description of variables *Participation rate* and *Household Economic Capacity Index (ICEF)*.

of Excellence, IST-2002-506778.

References

- Anduiza, E. and A. Bosch (2004). *Comportamiento político y electoral*. Barcelona: Ariel.
- Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis* **2**, 93–115.
- Anselin, L. (2003). *GeoDATM 0.9 User's Guide*. Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, IL. www.geoda.uiuc.edu.

Modeling long memory time series: the Shihua Cave speleothems

Jennifer A. Brown¹, William S. Rea¹ and Marco Reale¹

¹ Mathematics and Statistics Department, University of Canterbury, Private Bag 4800 Christchurch, New Zealand, jennifer.brown@canterbury.ac.nz

Abstract: In the literature many papers state that long-memory time series models such as Fractional Gaussian Noises (FGN) or Fractionally Integrated series (FI(d)) are empirically indistinguishable from structural break, or regime switching, models. We address this issue analyzing the Shihua Cave speleothems in China.

Keywords: Self-similarity; Regression Trees; Structural Breaks.

1 Introduction

Many time series exhibit the property of statistical long memory, also known as long-range dependence, strong dependence, global dependence, or the Hurst phenomenon. While these types of series had occasionally been observed, they were brought to prominence by Hurst (1951) who studied a large number of hydrological records and reported long-memory to be pervasive in them. Mandelbrot and Wallis (1969) reported long-memory in many geophysical records such as river flows, precipitation records, mud varve sequences, tree-ring indices, earthquake frequencies and the solar sunspot cycle. Subsequently long-memory has been reported in many diverse series. Long memory series are characterized in the time domain by serial correlations which decay at a hyperbolic rate, and in the frequency domain, by a power law spectrum,

$$P(f) \propto f^\alpha, \tag{1}$$

where the exponent α is the index representing the irregularity of the series. Any model of long-memory time series must account for both the slow decay of the serial correlation and the power law spectrum.

It is known that long memory and structural change are easily confused (Granger and Hyung, 2004; Diebold and Inoue, 2001; Smith, 2005).

Distinguishing between long memory and structural breaks is difficult because their finite sample properties are similar and so standard methodologies fail (Sibbertsen, 2004). Structural break detection and location techniques report breaks when only long memory is present. Similarly, long memory measurement techniques report long memory when only structural breaks are present even if the series is Markovian.

It is of interest to both theoreticians and practitioners to know the statistical properties of procedures for detecting and quantifying long memory when only structural change is present. Similarly, it is of interest to know the statistical properties of procedures

for detecting and locating structural change when, in fact, there is only long memory. The former problem has received considerable attention while the literature on the latter problem is somewhat sparse, see Sibbertsen (2004) for a survey.

This paper addresses the latter problem by examining the statistical properties of the “regimes” reported by of Atheoretical Regression Trees (ART) (Cappelli and Reale, 2005), a structural break technique, when ART was applied to a pure long memory time series.

2 Method

In Section (3) we give details of our real data set for which we had obtained an estimate of H and d . We simulated by computer up to 10,000 FGN and FI(d) series for the length and value of H and d as estimated for our example data set. The standard deviation was standardized so the series standard deviation was one in all cases. We broke these simulated long memory series into “regimes” using Atheoretical Regression Trees (ART) (see the appendix for a sketched description of ART). For each “regime” we estimated the length, mean, standard deviation, skewness, kurtosis, normality by the Jarque-Bera test, linear trend, H using the Whittle estimator, and the AR order. In addition, for the whole series we estimated H , the number of breaks detected by ART and the CUSUM range.

We obtained empirical (usually bivariate) distributions of the above quantities (e.g. regime length against standard deviation). We then compared the (usually bivariate) distribution obtained from the simulated series with the real data set to see if the real data set also resembled incorrectly analysed FGN or FI(d) processes.

For the real data set the whole series, the regimes and sometimes aggregations of the regimes discovered by ART were subjected to the Beran (1992) goodness-of-fit test for time series with long-range dependence. We used the Beran test using functions implemented in the R package `longmemo`.

The structural break method we used, ART, breaks the series into regimes based on local changes in the levels. There is no *a priori* reason to suspect that any other statistical property of the regimes should change with the level in an H -self-similar series.

3 Modeling the Shihua Cave Speleothems

Shihua Cave Speleothems are a set of thicknesses (in microns) of the annual layers of a stalagmite from Shihua Cave, Beijing, China. It covers a 2650-year period from 665BC to 1985AD (Tan *et al.*, 2001). Figure (1) presents the raw data and a log transform together with the ART breaks. The dates are in years before present.

We obtained $H = 0.830$ for the FGN and $d = 0.319$ for an FI(d). Beran’s goodness-of-fit test showed the FGN model provided an adequate fit to the data ($p = 0.15$).

The results for the Shihua Cave are presented in Figure (3). The “S”-symbols are the H estimates for each of the Shihua Cave regimes. The percent interval (PI) lines were obtained empirically from 10,000 simulated series by taking vertical cuts through the data at intervals of 100. The dots represent data points obtained from one-tenth of the

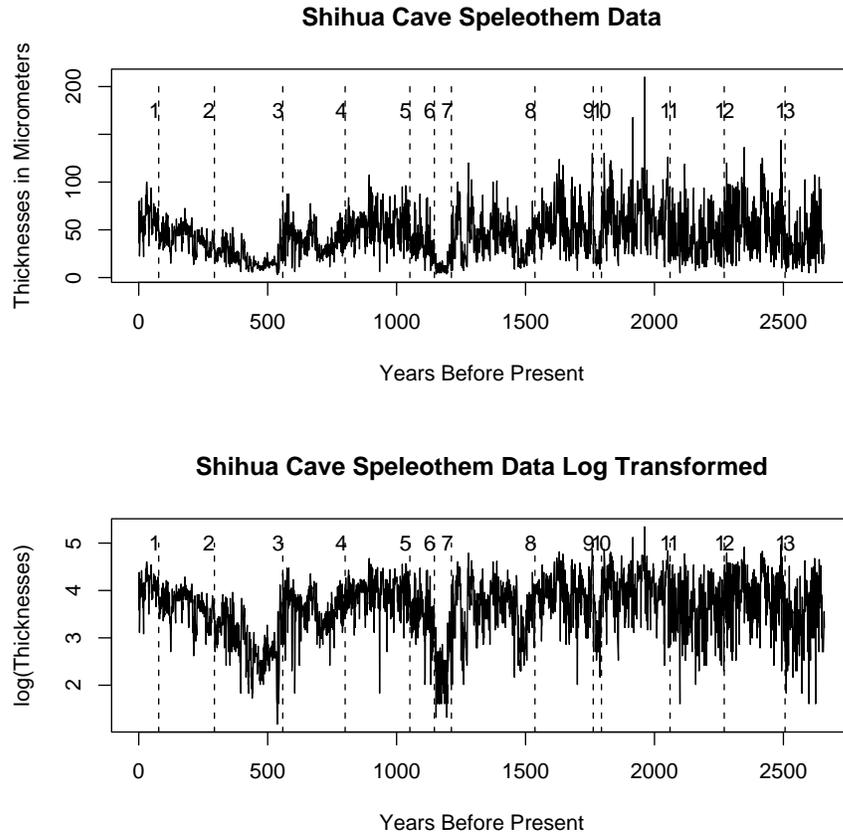


FIGURE 1. Plot of Shihua Cave speleothem data with breaks discovered by ART.

simulated FGN series. This gives a visual presentation of the distribution of regime lengths (the horizontal axis) and the H values (the vertical axis) of the regimes. ART found 13 breaks giving 14 regimes. Of the 14 regimes three estimates of H for that regime lies outside the 95% PI.

For a null hypothesis of H constant for the series we can conservatively estimate the p -value by assuming the H -estimates are binomially distributed. The probability of obtaining at least three points outside the 95% PI is thus

$$1 - \sum_{i=0}^2 \binom{14}{i} 0.95^{14-i} 0.05^i.$$

We obtain $p < 0.004$.

We subjected each regime to Beran's goodness-of-fit test. The second to last regime (2271-2506BP) is statistically significantly different from an FGN with the value of H

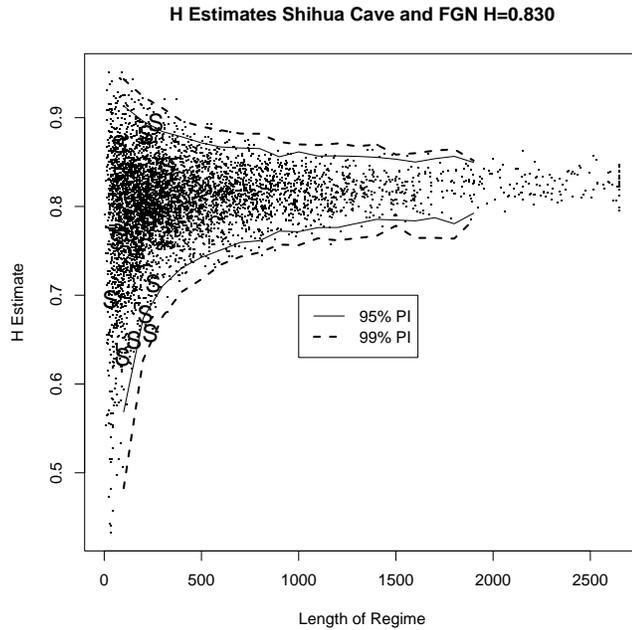


FIGURE 2. Empirical distribution and percentage intervals of H estimates against regime length for FGNs with $H=0.83$. “S” indicates a Shihua Cave data point.

estimated from either the whole series or the regime using the Whittle estimator. We obtain $p = 0.007$ for H estimated from the regime and $p = 0.0001$ for H estimated from the series.

It seems unreasonable to claim that single value of the H parameter characterizes the long memory properties of this series. This suggests the series is not H -self-similar.

4 Appendix: Atheoretical Regression Trees

Regression trees are now a well established non parametric modelling tool.

Let (Y, X) be a random vector, with $Y \in R$ and $X \in R^p$, regression trees seek a function $f(X)$, for predicting the response variable Y given values of the predictor variables X . The choice of the mean squared error $E(Y - f(X))^2$ as error function of the predictor $f(X)$ leads to least square regression trees (LSRT) in which $f(X)$ is the conditional expectation $E(Y|X = x)$. Thus, LSRT fit to each tree node the group mean, i.e. the mean of the Y 's values falling into the node, because this represents the optimal (or Bayes) prediction minimizing the mean squared error (for a complete discussion on this issue see Breiman *et al.*, 1984 Chap.9). Based on a training set $(y_i, x_{i1}, \dots, x_{ip})_{i=1}^n$, the algorithm proceeds by recursively splitting the data into two subsets. Any split is a binary question of the form: 'Is $x_j \in A$?', so that, in the case of a

numeric predictor variable, the set of possible splits includes all questions: Is $x_j \leq c$?, for c ranging over the domain of x_j . The split induces a partition of the observations y_i : the left descendant nodes h_l satisfying $\{x_{ij} \leq c\}$ and the right descendant node h_r satisfying $\{x_{ij} > c\}$. Thus, at any node h the algorithm selects the split s which maximally distinguishes the response variable in the left and the right descendant nodes providing the highest reduction in deviance

$$SS(h) - [SS(h_l) + SS(h_r)] \quad (2)$$

where $SS(h) = \sum_{y_i \in h} (y_i - \bar{y}(h))^2$, ($i = 1, \dots, n$), is the sum of squares for node h , and $SS(h_l)$ and $SS(h_r)$ are the sums of squares for the left and right descendants, respectively. As h_l and h_r are an exhaustive partition of h and $SS(h)$ can be thought of as the sum of squares at node h . The splitting criterion consists of minimizing, over all binary partitions of h , the within-group sum of squares:

$$WSS_{y|s}(h) = [SS(h_l) + SS(h_r)]. \quad (3)$$

Apart from having the advantage of being non parametric, LSRT is competitive when compared to linear regression. With a often a better performance in non linear problems while has a tendency to underperform in the presence of a good linear structure. The tree structure also may reveal patterns in the analysis that would not be so obvious with linear regression.

References

- Beran, J. (1992). A goodness of fit test for time series with long range dependence. *Journal of the Royal Statistical Society, Series B* **3**, 749-760.
- Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Monterey (CA): Wadsworth & Brooks.
- Cappelli, C., and Reale, M. (2004). Detecting multiple structural breaks in the mean via Atheoretical Regression Trees. In *Proceedings of the 20th International Workshop on Statistical Modelling*. 131-134, Sydney, Australia.
- Diebold, F.X., and Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics* **105**, 131-159.
- Granger, C.W.J., and Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance* **11**, 213-228.
- Hurst, H.E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* **116**, 770-808.
- Mandelbrot, B.B., and Wallis, J.R. (1969). Global dependence in geophysical records. *Water Resources Research* **5**, 321-340.
- Sibbertsen, P. (2004). Long memory versus structural breaks: an overview. *Statistical Papers* **45**, 465-515.

- Smith, A.D. (2005). Level Shifts and the Illusion of Long Memory in Economic Time Series. *Journal of Business & Economic Statistics* **23**, 321-335.
- Tan, M., Liu, T.S., Hou, J., Qin, X., Zhang, H., and li, T. (2003). Cyclic rapid warming on centennial-scale revealed by a 2650-year stalagmite record of warm season temperature. *Geophysical Research Letters* **30**, 19/1-19/4.

Mixed-Models for Genetic Linkage Analysis of Quantitative Traits: Analysis of APTT in the GAIT Project

A. Buil^{1,2}, J.C. Souto¹, J. Fontcuberta¹ and J.M. Soria¹

¹ Unitat de Hemostasia i Trombosis. Hospital de la Santa Creu i Sant Pau, Barcelona, Spain.

² Communicating author: Alfonso Buil. e-mail: abuil@santpau.es

Abstract: We present an application of mixed-effects models to find the genes that underly variability of quantitative traits. As an example, we present the results of the analysis of the trait Activated Partial Thromboplastin Time (APTT) in the Genetic Analysis Idiopathic Thrombosis (GAIT) Project.

Keywords: Mixed Models; Variance Components; Genetic Linkage Analysis.

1 Introduction

The search of quantitative trait loci (QTL)—genes that control the variability of quantitative traits, has become an active area of research in genetics. Genetic studies often use families and that means that the individual observations are not independent. A way of circumvent this difficulty is to use mixed-effect models—usually called variance component models in genetics. These models allow to specify the correlations between individuals due to their genetic similarities.

The GAIT Project started in 1995 with the goal of finding new genes related to Thrombosis. In this paper we describe the basic variance component genetic linkage model and we present its application to analyze the APTT trait in the GAIT Project.

2 Variance Component-Based Linkage Model

Modelling the trait

Let the measure of a quantitative trait for the i th individual, y_i , be modelled as a combination of fixed and random effects as shown in Equation 1.

$$y_i = \mu + \sum_{k=1}^p \beta_k x_{ik} + q_j + g + e \quad (1)$$

The fixed effects are represented by the βx terms. The random effects are: the effect of QTL j (q_j), the average effect of the remaining genetic contributions (g), and the environmental residual effects (e). We assume that q_j , g and e are uncorrelated random variables with expectation zero, so that the variance of y_i is:

$$\sigma_{yi}^2 = \sigma_{qj}^2 + \sigma_g^2 + \sigma_e^2 \quad (2)$$

Following the classical quantitative genetics formulation developed by Fisher (1918) the expected phenotypic covariance among the members of a family can be represented in matrix form as follows:

$$\Omega = \hat{\Pi}_j \sigma_{qj}^2 + 2\Phi \sigma_g^2 + I \sigma_e^2 \quad (3)$$

where $\hat{\Pi}_j$ is a matrix whose elements $\hat{\pi}_{jkl}$ are the ibd coefficients between individuals k and l —a measure of its genetic similarity at locus j ; 2Φ is the kinship coefficient—a measure of the average genetic similarity between individuals; and I is the identity matrix. Matrix 2Φ is determined by the familiar relationships between pairs of individuals, while matrix $\hat{\Pi}_j$ is estimated by using DNA markers at the specific genomic location j , as described in Almasy and Blangero (1998).

Maximum likelihood estimation

If multivariate normality of the trait vector for a family (\mathbf{y}) is assumed, the likelihood of any family can be expressed easily, and standard numerical methods can be used to estimate the model parameters. For the covariance model in Equation 3, the log-likelihood of a family of t individuals is:

$$\ln L(\mu, \beta_1, \dots, \beta_p, \sigma_q^2, \sigma_g^2, \sigma_e^2 | y) = -\frac{t}{2} \ln(2\pi) - \frac{1}{2} \ln |\Omega| - \frac{1}{2} \Delta' \Omega^{-1} \Delta \quad (4)$$

where $\Delta = \mathbf{y} - \mu - \beta \mathbf{x}$

The likelihood ratio statistic and the lod score

In the variance component model, the null hypothesis that the genetic variance due to the j -th QTL is zero (i.e., that there is no linkage at genomic location j) can be tested by using the likelihood ratio statistic:

$$\Lambda = 2 \left[\ln L(\hat{\theta}) - \ln L(\tilde{\theta}) \right] \quad (5)$$

where $L(\hat{\theta})$ is the likelihood under the alternative hypothesis of linkage and $L(\tilde{\theta})$ is the likelihood under the null hypothesis of linkage.

The difference between the two \log_{10} likelihood yields a lod score that is equivalent to the classical lod score in linkage analysis, widely used by geneticist. That is,

$$\text{LOD} = \log_{10} \frac{L(\hat{\theta})}{L(\tilde{\theta})} = \log_{10} L(\hat{\theta}) - \log_{10} L(\tilde{\theta}) = \frac{\Lambda}{2 \ln 10} \quad (6)$$

In our test, the value of the parameter under the null is on the boundary of the parameter space defined by the alternative hypothesis. This is a nonstandard condition, and thus the test is distributed as a $1/2 : 1/2$ mixture of a chi-square distribution with one degree of freedom (χ_1^2) and a unit point mass at the origin (χ_0^2), rather than a simple χ^2 with one degree of freedom (Self and Liang 1987). However, for the fixed effects, the likelihood ratio statistic fits the standard conditions.

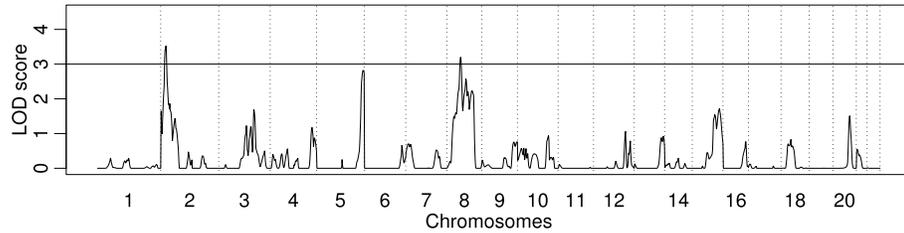


FIGURE 1. Genome-wide linkage result for the APTT trait in the GAIT sample. The lod score is represented along the whole genome. The vertical lines divide the chromosomes. The horizontal line indicates the significance threshold level.

The genome-wide scan strategy

The model described above tests the relation between the trait and the specific genomic location j . As we do not know where the real QTL is located, we have to perform a search along the whole genome. The genome-wide strategy consists of repeating the linkage model described above for 500 different genomic locations. Since 500 tests are performed, a genome-wide p-value of 0.05 will be equivalent to a p-value of 0.0001 in every individual test. This p-value corresponds to a lod score of 3.

Software

The software SOLAR (Almasy and Blangero 1998) has been used to perform all of the calculations.

3 Example: The GAIT Project

The GAIT Project has been extensively described in previous publications (Souto et al., 2000). The sample consists of 398 individuals in 21 extended Spanish families composed of 3 to 5 generations. Subjects were genotyped for a genome-wide scan including 500 DNA markers that were used to estimate the ibd coefficients of every pair of individuals at every location along the genome. Moreover, more than 50 quantitative traits were measured in the entire GAIT sample. In this paper we present the analysis of the APTT trait, a measure of coagulation time that has been related with the risk of thrombosis (Tripodi et al., 2004).

Analysis of the APTT trait

The APTT trait follows a normal distribution in our sample. We tried sex, age, smoking behavior and contraceptive use as fixed effects, but only age was significant ($p = 2E - 7$). Figure 1 shows the genome-wide linkage result. There are two significant lod scores, suggesting linkage on chromosomes 2 and 8, and an almost significant lod score on chromosome 5.

4 Conclusions

The variance-component framework is a flexible way of modelling quantitative traits in families. It allows to test the significance of fixed effects as well as to perform genetic linkage analysis in the search of QTL involved in complex human diseases.

In our example, the analysis of the variability of APTT, we found two significant QTL on chromosomes 2 and 8. That means that there are genes in these genomic areas that affect this trait. The next step in our research will be to identify the specific genes and to determine if they affect the risk of thrombosis.

References

- Almasy, L., and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics* **62**, 1198-1211.
- Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399-433.
- Self, S.G., and Liang, K. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* **82**, 605-610.
- Souto, J.C. et al. (2000). Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation* **101**, 1546-1551.
- Tripodi, A. et al. (2004). A shortened activated partial thromboplastin time is associated with the risk of venous thromboembolism. *Blood* **104**, 3631-3634.

Estimation of signals transmitted by different randomly delayed sensors using covariance information

R. Caballero-Águila¹, A. Hermoso-Carazo², J.D. Jiménez-López¹, J. Linares-Pérez² and S. Nakamori³

¹ Dpto. Estadística e I.O., Universidad de Jaén, 23071 Jaén, Spain
(raguila@ujaen.es, jdomingo@ujaen.es)

² Dpto. Estadística e I.O., Universidad de Granada, 18071 Granada, Spain
(ahermoso@ugr.es, jlinares@ugr.es)

³ Department of Technology, Faculty of Education, Kagoshima University, 1-20-6, Kohri-moto, Kagoshima, 890-0065 Japan (nakamori@edu.kagoshima-u.ac.jp)

Abstract: The least-squares linear estimation problem of a signal from noisy measurements coming from multiple randomly delayed sensors, with different delay characteristics, is approached. Assuming that the state-space model of the signal is unknown and using only covariance information, recursive filtering and smoothing algorithms are derived by an innovation approach.

Keywords: Least-squares estimation; Randomly delayed observations.

1 Introduction

In some practical situations, such as in communication networks, the data coming from measurements devices or sensors may be randomly delayed due to heavy network traffic, so the measurements available of a signal may not be up-to-date. In the last years, the state estimation problem for system models with randomly time-varying delays has been widely investigated. Among others, Yaz and Ray (1998) approached the state estimation problem in a model involving randomly varying bounded sensor delays reformulating it as an estimation problem in systems with stochastic parameters. Nakamori et al. (2005) studied the least-squares (LS) linear estimation problem from observations with stochastic delays assuming that the state-space model is not completely known but using only covariance information. These papers solve situations in which there is random sensor delays, but all the sensors have the same delay characteristics. Recently, Hounkpevi and Yaz (2007) have generalized this situation by considering multiple delayed sensors with different delay characteristics. In this paper, this general situation is considered and the LS linear estimation problem is addressed assuming that the state-space model of the signal to be estimated is not known and only covariance information is available.

2 Observation model and problem statement

Consider m scalar sensors whose real measurements, \tilde{y}_k^i , of the signal, z_k , are perturbed by additive noise vectors v_k^i ; that is,

$$\tilde{y}_k^i = H_k^i z_k + v_k^i, \quad k \geq 1, \quad i = 1, \dots, m. \quad (1)$$

We assume that at time $k = 1$ the real measurements, \tilde{y}_1^i , are always available for the estimation, but at each time $k > 1$, the available measurement coming from each sensor may be randomly delayed by one sampling time, independently of the others, according to different delay characteristic, and that a delay at time k is independent of a delay at time s . Therefore, if $\{\gamma_k^i; k > 1\}$, $i = 1, \dots, m$, denote mutually independent sequences of independent Bernoulli random variables, the available measurement of the i th sensor, y_k^i , is described by

$$y_k^i = (1 - \gamma_k^i) \tilde{y}_k^i + \gamma_k^i \tilde{y}_{k-1}^i, \quad k > 1; \quad y_1^i = \tilde{y}_1^i, \quad i = 1, \dots, m. \quad (2)$$

The following hypotheses about the processes involved in (2) are assumed:

- (I) $\{z_k; k \geq 1\}$ has zero mean and factorizable covariance function $K_{k,s}^z = E[z_k z_s^T] = A_k B_s^T$, $s \leq k$, with A_k and B_s known $n \times M$ matrices.
- (II) For $i = 1, \dots, m$, the sensor noises, $\{v_k^i; k \geq 1\}$, are zero-mean white sequences with known variances $Var[v_k^i] = R_k^i$, $\forall k \geq 1$.
- (III) For $i = 1, \dots, m$, $\{\gamma_k^i; k > 1\}$ are sequences of independent Bernoulli random variables with known probabilities, $P[\gamma_k^i = 1] = p_k^i$, $\forall k > 1$.
- (IV) The signal process, $\{z_k; k \geq 1\}$, and the noise processes, $\{v_k^i; k \geq 1\}$ and $\{\gamma_k^i; k > 1\}$, for $i = 1, \dots, m$, are mutually independent.

To simplify the notation, we rewrite (1) and (2) as follows:

$$\begin{aligned} \tilde{y}_k &= H_k z_k + V_k, \quad k \geq 1, \\ y_k &= (I_m - \Gamma_k) \tilde{y}_k + \Gamma_k \tilde{y}_{k-1}, \quad k > 1; \quad y_1 = \tilde{y}_1, \end{aligned} \quad (3)$$

where $\tilde{y}_k = (\tilde{y}_k^1, \dots, \tilde{y}_k^m)^T$, $H_k = ((H_k^1)^T, \dots, (H_k^m)^T)^T$, $V_k = (v_k^1, \dots, v_k^m)^T$, $y_k = (y_k^1, \dots, y_k^m)^T$, $\Gamma_k = \text{Diag}(\gamma_k^1, \dots, \gamma_k^m)$ and I_m is the $m \times m$ identity matrix.

Clearly from hypotheses (II)-(IV), the following properties hold:

- (i) $\{V_k; k \geq 1\}$ is a zero-mean white sequence with variance $E[V_k V_k^T] = R_k$, where $R_k = \text{Diag}(R_k^1, \dots, R_k^m)$.
- (ii) The matrices $\{\Gamma_k; k \geq 1\}$ are independent and $E[\Gamma_k] = P_k$, with $P_k = \text{Diag}(p_k^1, \dots, p_k^m)$.
- (iii) $\{z_k; k \geq 1\}$, $\{V_k; k \geq 1\}$ and $\{\Gamma_k; k > 1\}$ are mutually independent.

3 Linear LS estimation problem

Our aim is to address the LS linear estimation problem of the signal, z_k , based on the randomly delayed observations $\{y_1, \dots, y_L\}$ given in (3); more specifically, recursive algorithms for the filtering ($L = k$), fixed-point (k fixed and $L > k$) and fixed-interval (L fixed and $k < L$) smoothing problems will be derived by an innovation approach. Starting from the following general expression for the linear LS estimator of the signal, obtained from the Orthogonal Projection Lemma,

$$\widehat{z}_{k/L} = \sum_{j=1}^L S_{k,j} \Pi_j^{-1} \nu_j$$

with ν_j the innovation at time j , $\Pi_j = E[\nu_j \nu_j^T]$ and $S_{k,j} = E[z_k \nu_j^T]$, the following algorithms for the filter and the smoothers are derived.

Filter: $\widehat{z}_{k/k} = \mathbf{A}_k \mathbf{O}_k$, $k \geq 1$, with O_k given by

$$O_k = O_{k-1} + J_k \Pi_k^{-1} \nu_k, \quad k \geq 1; \quad O_0 = 0,$$

and the matrix J_k , the innovation ν_k and its covariance matrix Π_k , by

$$\begin{aligned} J_k &= G_{B_k}^T - r_{k-1} G_{A_k}^T - J_{k-1} \Pi_{k-1}^{-1} F_{k-1}, \quad k \geq 2; \quad J_1 = B_1^T H_1^T, \\ \nu_k &= y_k - G_{A_k} O_{k-1} - F_{k-1} \Pi_{k-1}^{-1} \nu_{k-1}, \quad k \geq 2; \quad \nu_1 = y_1, \\ \Pi_k &= (I_m - P_k) [H_k A_k B_k^T H_k^T + R_k] + P_k [H_{k-1} A_{k-1} B_{k-1}^T H_{k-1}^T + R_{k-1}] \\ &\quad - G_{A_k} [G_{B_k}^T - J_k] - F_{k-1} \Pi_{k-1}^{-1} [G_{A_k} J_{k-1} + F_{k-1}]^T, \quad k \geq 2; \\ \Pi_1 &= H_1 A_1 B_1^T H_1^T + R_1 \end{aligned}$$

with $r_k = E[O_k O_k^T]$ recursively obtained from

$$r_k = r_{k-1} + J_k \Pi_k^{-1} J_k^T, \quad k \geq 1; \quad r_0 = 0.$$

The matrices G_{A_k} , G_{B_k} and F_{k-1} are given by

$$\begin{aligned} G_{A_k} &= (I_m - P_k) H_k A_k + P_k H_{k-1} A_{k-1}, \\ G_{B_k} &= (I_m - P_k) H_k B_k + P_k H_{k-1} B_{k-1} \\ F_{k-1} &= P_k (I_m - P_{k-1}) R_{k-1}, \quad k > 2; \quad F_1 = P_2 R_1. \end{aligned}$$

Fixed-point smoother: $\widehat{z}_{k/L} = \widehat{z}_{k/L-1} + \mathbf{S}_{k,L} \Pi_L^{-1} \nu_L$, $L > k$, with initial condition given by the filter, $\widehat{z}_{k/k}$, and

$$S_{k,L} = (B_k - E_{k,L-1}) G_{A_L}^T - S_{k,L-1} \Pi_{L-1}^{-1} F_{L-1}, \quad L > k; \quad S_{k,k} = A_k J_k,$$

where the matrices $E_{k,L}$ satisfy the following recursive formula

$$E_{k,L} = E_{k,L-1} + S_{k,L} \Pi_L^{-1} J_L^T, \quad L > k; \quad E_{k,k} = A_k r_k$$

with Π_L , ν_L , G_{A_L} , F_L , J_L and r_L given in the filtering algorithm.

Fixed-interval smoother: $\widehat{z}_{k/L} = \widehat{z}_{k/k} + \Upsilon_k \mathbf{q}_{k/L}$, $k < L$,

where the vectors $q_{k/L}$ are backward recursively obtained from

$$q_{k/L} = \Xi_{k+1} \begin{pmatrix} q_{k+1/L} \\ \Pi_{k+1}^{-1} \nu_{k+1} \end{pmatrix}, \quad k < L; \quad q_{L/L} = 0.$$

and the matrices Υ_k and Ξ_k are calculated as

$$\Upsilon_k = \left(B_k - A_k r_k \mid -A_k J_k \right)$$

$$\Xi_k = \begin{pmatrix} I_M - G_{A_k}^T \Pi_k^{-1} J_k^T & -G_{A_k}^T & G_{A_k}^T \\ -\Pi_{k-1}^{-1} F_{k-1} \Pi_k^{-1} J_k^T & -\Pi_{k-1}^{-1} F_{k-1} & \Pi_{k-1}^{-1} F_{k-1} \end{pmatrix}$$

with Π_L , ν_L , r_L , J_L , G_{A_L} and F_L given in the filtering algorithm.

The precision of the LS estimators is measured by the covariance matrices of the corresponding estimation errors, $\Sigma_{k/L} = E \left[\{z_k - \hat{z}_{k/L}\} \{z_k - \hat{z}_{k/L}\}^T \right]$, which, for the problems at hand, are calculated as follows:

- *Filtering*: $\Sigma_{k/k} = A_k B_k^T - A_k r_k A_k^T$, $k \geq 1$.
- *Fixed-point smoothing*: $\Sigma_{k/L} = \Sigma_{k/L-1} - S_{k,L} \Pi_L^{-1} S_{k,L}^T$, $L > k$,
with initial condition $\Sigma_{k/k}$.
- *Fixed-interval smoothing*: $\Sigma_{k/L} = \Sigma_{k/k} - \Upsilon_k Q_{k/L} \Upsilon_k^T$, $k < L$,
where $Q_{k/L} = E[q_{k/L} q_{k/L}^T]$ satisfies the following recursive relation

$$Q_{k/L} = \Xi_{k+1} \begin{pmatrix} Q_{k+1/L} & 0 \\ 0 & \Pi_{k+1}^{-1} \end{pmatrix} \Xi_{k+1}^T, \quad k < L; \quad Q_{L/L} = 0.$$

Acknowledgments: This work is partially supported by the ‘Ministerio de Ciencia y Tecnología’ through the project MTM2005-03601.

References

- Yaz, E.E., and Ray, A. (1998). Linear unbiased state estimation under randomly varying bounded sensor delay. *Applied Mathematical Letters* **11**, 27-32.
- Nakamori, S., Caballero-Águila, R., Hermoso-Carazo, A., and Linares- Pérez, J. (2005). Recursive estimators of signals from measurements with stochastic delays using covariance information. *Applied Mathematics and Computation* **162**, 65-79.
- Houkpevi, F.O., and Yaz, E.E. (2007). Minimum variance generalized state estimators for multiple sensors with different delay rates. *Signal Processing* **87**, 602-613.

Estimation from observations with randomly missing signals using an innovation approach

R. Caballero-Águila¹, A. Hermoso-Carazo², J.D. Jiménez-López¹, J. Linares-Pérez² and S. Nakamori³

¹ Dpto. Estadística e I.O., Universidad de Jaén, 23071 Jaén, Spain, raguila@ujaen.es, jdomingo@ujaen.es

² Dpto. Estadística e I.O., Universidad de Granada, 18071 Granada, Spain, ahermoso@ugr.es, jlinares@ugr.es

³ Department of Technology, Faculty of Education, Kagoshima University, 1-20-6, Kohrimoto, Kagoshima, 890-0065 Japan, nakamori@edu.kagoshima-u.ac.jp

Abstract: The least-squares linear estimation problem of a discrete-time signal from noisy observations in which the signal can be randomly missing is considered. The uncertainty about the signal being present or missing at each observation is characterized by a set of Bernoulli variables which are correlated when the difference between sampling times is equal to a certain value m . Recursive filtering and fixed-point smoothing algorithms are obtained using an innovation approach.

Keywords: Least-squares estimation; Uncertain observations.

1 Introduction

There are many practical situations in which the signal appears in the observation in a random manner, for instance, systems where there are intermittent failures in the observation mechanism; in this paper the problem of estimating a signal from noisy observations in which the signal can be randomly missing is considered. To describe this uncertainty, the observation equation, with the usual additive noise, is formulated by multiplying the signal, at any sampling time, by a Bernoulli random variable; the value one of this variable indicates that the signal is present in the observation, whereas the value zero reflects the fact that it is missing.

In some cases, the variables modelling the uncertainty in the observations can be assumed to be independent and, then, their distribution is fully determined by the probability that each observation contains the signal. A different situation, in which the Bernoulli variables are correlated at consecutive instants, is considered by Jackson and Murthy (1976) who, using a state-space approach, derived a least-squares linear filtering algorithm.

On the other hand, in some situations, the state-space model of the signal is not available and another type of information must be used to address the estimation problem. In the last years, the estimation problem from uncertain observations has been investigated using covariance information and algorithms with a simpler structure than those obtained when the state-space model is known have been derived for the

filtering and fixed-point smoothing problems (see Nakamori et al. (2003) when the uncertainty is modelled by independent random variables and Nakamori et al. (2005), for the model considered by Jackson and Murthy (1976)).

In this paper, we consider the situation of unknown state-space model and covariance information is used; the aim is to obtain least-squares linear estimators using uncertain observations when the uncertainty in a time instant k depends only on the uncertainty in the previous time $k - m$; this correlation structure, which allows us to consider certain models where the signal cannot be missing in $m + 1$ consecutive observations, is more general than that considered in Nakamori et al. (2005) and, consequently, the algorithms proposed in this paper generalize those of the latter.

2 Observation model and estimation problem

Consider that, at each time $k \geq 1$, the observation, y_k , of the n -dimensional random signal, z_k , is described by

$$y_k = \theta_k z_k + v_k, \quad k \geq 1 \quad (1)$$

where the involved processes satisfy the following hypotheses:

(H1) $\{z_k; k \geq 1\}$ has zero mean and separable autocovariance function, $K_{k,s}^z = E[z_k z_s^T] = A_k B_s^T$, $s \leq k$, where A_k and B_s are $n \times M$ matrices.

(H2) $\{v_k; k \geq 1\}$ is a zero-mean white sequence with $E[v_k v_k^T] = R_k$.

(H3) $\{\theta_k; k \geq 1\}$ is a sequence of Bernoulli variables with $P[\theta_k = 1] = \bar{\theta}_k$, and its autocovariance function, $K_{k,s}^\theta = E[(\theta_k - \bar{\theta}_k)(\theta_s - \bar{\theta}_s)]$, vanishes for $|k - s| \neq 0, m$ but it can be nonzero for $|k - s| = m$.

(H4) $\{z_k; k \geq 1\}$, $\{v_k; k \geq 1\}$ and $\{\theta_k; k \geq 1\}$ are mutually independent.

The purpose is to address the least-squares (LS) linear estimation problem of the signal z_k based on the available observations up to a certain time L ; specifically, our aim is to obtain the filter, $\hat{z}_{k/k}$, and the smoother $\hat{z}_{k/k+l}$, at the fixed point k , for any $l \geq 1$, from recursive algorithms. For this purpose we use an innovation approach.

The innovation at time k is defined as the difference between the observation at time k and its one-stage predictor, $\nu_k = y_k - \hat{y}_{k/k-1}$. Since the LS linear estimator of z_k based on the observations $\{y_1, \dots, y_L\}$ is equal to the LS linear estimator based on the innovations $\{\nu_1, \dots, \nu_L\}$, the estimation problem will be approached by replacing the observation process by the innovation one. Then, by denoting $\Pi_i = E[\nu_i \nu_i^T]$ and $S_{k,i} = E[z_k \nu_i^T]$, from the Orthogonal Projection Lemma (OPL), it is easy to see that

$$\hat{z}_{k/L} = \sum_{i=1}^L S_{k,i} \Pi_i^{-1} \nu_i. \quad (2)$$

3 Filter and fixed-point smoother

In view of expression (2), the first step to obtain the linear filter and fixed-point smoother is to establish an explicit formula for the innovations; then recursive filtering and fixed-point smoothing algorithms are easily derived.

3.1 Innovations: $\nu_k = y_k - \hat{y}_{k/k-1}$

As in (2), noting $T_{k,i} = E[y_k \nu_i^T]$, the OPL leads to the following expression for the observation predictor:

$$\hat{y}_{k/k-1} = \sum_{i=1}^{k-1} T_{k,i} \Pi_i^{-1} \nu_i, \quad k \geq 2; \quad \hat{y}_{1/0} = 0. \quad (3)$$

From hypotheses (H1)-(H4), and since $K_{k,i}^\theta = \delta_{i,k-m} K_{k,k-m}^\theta$, by denoting $K_{k,k-m}^\theta = 0$ for $k \leq m$, we obtain that

$$T_{k,i} = \bar{\theta}_k A_k J_i + K_{k,k-m}^\theta A_k B_{k-m}^T \Psi_{k,i}$$

where

$$J_i = \bar{\theta}_i B_i^T - \sum_{j=1}^{i-1} J_j \Pi_j^{-1} T_{i,j}^T \quad \text{and} \quad \Psi_{k,i} = \begin{cases} \delta_{i,k-m} I_n, & i \leq k-m, \\ - \sum_{j=k-m}^{i-1} \Psi_{k,j} \Pi_j^{-1} T_{i,j}^T, & i > k-m. \end{cases}$$

Now, by substituting $T_{k,i}$ in (3) and denoting $O_k = \sum_{i=1}^k J_i \Pi_i^{-1} \nu_i$, we have

$$\nu_k = y_k - \bar{\theta}_k A_k O_{k-1} - K_{k,k-m}^\theta A_k B_{k-m}^T \sum_{i=k-m}^k \Psi_{k,i} \Pi_i^{-1} \nu_i, \quad k \geq 1. \quad (4)$$

3.2 Linear filtering algorithm

The linear filter, $\hat{z}_{k/k}$, of the signal z_k is given by

$$\hat{z}_{k/k} = A_k O_k, \quad k \geq 1,$$

where the vectors O_k are recursively calculated from

$$O_k = O_{k-1} + J_k \Pi_k^{-1} \nu_k, \quad k \geq 1; \quad O_0 = 0.$$

The innovation ν_k is given in (4) and the matrix J_k satisfies

$$J_k = \bar{\theta}_k [B_k^T - r_{k-1} A_k^T] - K_{k,k-m}^\theta \sum_{i=k-m}^{k-1} J_i \Pi_i^{-1} \Psi_{k,i}^T B_{k-m} A_k^T$$

with $r_k = E[O_k O_k^T]$ being recursively obtained from

$$r_k = r_{k-1} + J_k \Pi_k^{-1} J_k^T, \quad k \geq 1; \quad r_0 = 0,$$

and

$$\Psi_{k,i} = \begin{cases} \delta_{i,k-m} I_n, & i \leq k-m, \\ - \sum_{j=k-m}^{i-1} \Psi_{k,j} \Pi_j^{-1} [\bar{\theta}_i A_i J_j + K_{i,i-m}^\theta A_i B_{i-m}^T \Psi_{i,j}]^T, & i > k-m. \end{cases}$$

The innovation covariance matrix, Π_k , verifies

$$\Pi_k = \bar{\theta}_k (1 - \bar{\theta}_k) A_k B_k^T + R_k + \bar{\theta}_k A_k J_k + K_{k,k-m}^\theta A_k B_{k-m}^T \Psi_{k,k}.$$

3.3 Linear fixed-point smoothing algorithm

By starting from the filter, $\hat{z}_{k/k}$, the fixed-point smoothers, $\hat{z}_{k/k+l}$, $l \geq 1$, are recursively obtained by

$$\hat{z}_{k/k+l} = \hat{z}_{k/k+l-1} + S_{k,k+l} \Pi_{k+l}^{-1} \nu_{k+l}, \quad k, l \geq 1,$$

where

$$S_{k,k+l} = [B_k - E_{k,k+l-1}] \bar{\theta}_{k+l} A_{k+l}^T - K_{k+l,k+l-m}^\theta \sum_{i=k+l-m}^{k+l-1} S_{k,i} \Pi_i^{-1} \Psi_{k+l,i}^T B_{k+l-m} A_{k+l}^T$$

with $S_{k,i} = A_k J_i$, for $i \leq k$, and $E_{k,k+l} = E[z_k O_{k+l}^T]$ satisfying

$$E_{k,k+l} = E_{k,k+l-1} + S_{k,k+l} \Pi_{k+l}^{-1} J_{k+l}^T, \quad l > 1; \quad E_{k,k} = A_k r_k.$$

3.4 Error covariance matrices

The LS method uses the estimation error covariance matrices, $P_{k/L} = E[(z_k - \hat{z}_{k/L})(z_k - \hat{z}_{k/L})^T]$, to measure the goodness of the estimators. From the OPL and taking into account hypothesis (H1), it is easy to verify that $P_{k/L} = A_k B_k^T - E[\hat{z}_{k/L} \hat{z}_{k/L}^T]$.

Now, using the recursive relation for $\hat{z}_{k/k+l}$ and the uncorrelation property between ν_{k+l} and $\hat{z}_{k/k+l-1}$, we obtain

$$P_{k,k+l} = P_{k,k+l-1} - S_{k,k+l} \Pi_{k+l}^{-1} S_{k,k+l}^T, \quad k, l \geq 1.$$

The initial condition of this relation is $P_{k/k}$, the filtering error covariance matrix which, since $\hat{z}_{k/k} = A_k O_k$ and $r_k = E[O_k O_k^T]$, satisfies

$$P_{k/k} = A_k [B_k^T - r_k A_k^T], \quad k \geq 1.$$

Acknowledgments: This work is partially supported by the ‘Ministerio de Ciencia y Tecnología’ through the project MTM2005-03601.

References

- Jackson, R.N., and Murthy, D.N.P. (1976). Optimal linear estimation with uncertain observations. *IEEE Transactions on Information Theory*, 376-378.
- Nakamori, S., Caballero-Águila, R., Hermoso-Carazo, A., and Linares- Pérez, J. (2003). Linear estimation from uncertain observations with white plus colored noises using covariance information. *Digital Signal Processing* **13**, 552-568.
- Nakamori, S., Caballero-Águila, R., Hermoso-Carazo, A., and Linares- Pérez, J. (2005). New recursive estimators from correlated interrupted observations using covariance information. *International Journal of Systems Science* **36**, 617-629.

Modelling General Patterns of Digit Preference

Carlo Giovanni Camarda¹, Paul H. C. Eilers² and Jutta Gampe¹

¹ Max Planck Institute for Demographic Research, Konrad-Zuse-Str. 1 D-18057 Rostock, Germany. email: camarda@demogr.mpg.de, gampe@demogr.mpg.de

² Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands, P.H.C.Eilers@uu.nl

Abstract: Digit preference is a typical example of indirect observations of a latent distribution. A variation on the composite link model is a useful framework to uncover such latent distributions and it allows to estimate the proportions of counts that were transferred to neighbouring digits. We impose smoothness assumptions on the latent distribution since the estimating equations generally are singular or severely ill-conditioned. To estimate the misreported proportions we use a weighted least-squares regression with an added L_1 -penalty. The optimal smoothing parameters are found by minimizing the AIC. A simulation study and an actual application are presented.

Keywords: Composite link model; digit preference; penalized likelihood; smoothing.

1 Introduction

When people read analog scales or report numeric results from memory, a commonly found effect is that certain preferred end-digits are reported substantially more often than the general pattern of the distribution suggests. These digits are typically multiples of 5 and 10, possibly combined with tendencies to avoid certain unpleasant numbers like, e.g., 13. This type of misreporting leads to unusual heapings at the preferred digits and the observed data actually present a biased, though well understood image of the true distribution. This tendency is called digit preference or age heaping, if the reported numbers refer to ages.

2 The Composite Link Model

The digit preference problem can be viewed as an inverse problem where the actually observed values are linear compositions of a latent sequence representing the true distribution. This sequence is to be estimated and the composition pattern reveals the amount of misreporting. The composite link model (CLM), proposed by Thompson and Baker (1981), provides an elegant framework for modelling indirect observations of counts.

We assume a smooth discrete sequence $\gamma = \exp(\mathbf{X}\beta)$, with β smooth. The elements $\gamma_j, j = 1, \dots, J$ are the counts that would be expected, if there were no digit preferences. However, the mechanism that actually generates observations operates by

linearly composing the values in $\boldsymbol{\gamma}$ to a vector $\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$. The observed counts \mathbf{y} are realizations from Poisson variables with $E(\mathbf{y}) = \boldsymbol{\mu}$.

The composition matrix \mathbf{C} embodies the digit preference mechanism by partly redistributing certain elements of $\boldsymbol{\gamma}$ to neighbouring, preferred values in $\boldsymbol{\mu}$. Generalizing Eilers and Borgdorff (2004), we allow that counts at any digit can be redistributed to the immediate neighbouring categories and we will estimate the misreporting pattern along with the latent smooth distribution $\boldsymbol{\gamma}$. Therefore we obtain the $J \times J$ composition matrix \mathbf{C} :

$$\mathbf{C} = \begin{pmatrix} 1 - p_{21} & p_{12} & 0 & 0 & 0 \\ p_{21} & 1 - p_{12} - p_{32} & p_{23} & \cdots & \vdots \\ 0 & p_{32} & 1 - p_{23} - p_{43} & p_{34} & \vdots \\ 0 & 0 & p_{43} & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & p_{J-1,J} \\ 0 & \cdots & \cdots & 0 & 1 - p_{J-1,J} \end{pmatrix} \quad (1)$$

where p_{jk} denote the proportion of γ_k that is moved from category k to category j . Allowing only one-step transitions implies that $p_{jk} = 0$ for $|j - k| > 1$.

3 Estimating the CLM and the preference pattern

In case of no digit preference we would be able to directly observe counts z_j , $j = 1, \dots, J$, following a Poisson distribution with mean γ_j .

Since in our applications the only covariate is the sequence of age- or measurement-values, the matrix \mathbf{X} will simply be the identity matrix, i.e. $\gamma_j = \exp\{\beta_j\}$. Estimates of $\boldsymbol{\beta}$ in this generalized linear model (GLM) would be obtained by the iteratively reweighted least squares (IRWLS) algorithm. If we, however, do not observe \mathbf{z} , but realizations of the composed counts $\mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu})$, with $\boldsymbol{\mu} = E(\mathbf{y}) = \mathbf{C}\boldsymbol{\gamma}$, or $\mu_i = \sum_j c_{ij}\gamma_j$, $i = 1, \dots, J$, we can easily adapt the IRWLS-scheme.

By defining $\check{x}_{ik} = \sum_j c_{ij}x_{jk}\gamma_j/\mu_i$, the system of equations becomes, in matrix notation,

$$\check{\mathbf{X}}'\check{\mathbf{W}}\check{\mathbf{X}}\boldsymbol{\beta} = \check{\mathbf{X}}'\check{\mathbf{W}}\{\check{\mathbf{W}}^{-1}(\mathbf{y} - \check{\boldsymbol{\mu}}) + \check{\mathbf{X}}\check{\boldsymbol{\beta}}\}, \quad (2)$$

where $\check{\mathbf{W}} = \text{diag}(\check{\boldsymbol{\mu}})$. A detailed derivation of (2) can be found in Eilers (2007).

Both in GLMs and CLMs we can force the solution vector $\boldsymbol{\beta}$ to be smooth by subtracting a roughness penalty from the log-likelihood (Eilers and Marx, 1996). If we introduce this penalty into the likelihood for the CLM, we obtain the following system of equations

$$(\check{\mathbf{X}}'\check{\mathbf{W}}\check{\mathbf{X}} + \lambda \mathbf{P})\boldsymbol{\beta} = \check{\mathbf{X}}'\check{\mathbf{W}}\{\check{\mathbf{W}}^{-1}(\mathbf{y} - \check{\boldsymbol{\mu}}) + \check{\mathbf{X}}\check{\boldsymbol{\beta}}\}, \quad (3)$$

where $\mathbf{P} = \mathbf{D}'_d\mathbf{D}_d$ and $\mathbf{D}_d \in \mathbb{R}^{(J-d) \times J}$ is the matrix that computes d -th order differences. The smoothing parameter λ balances model fidelity and smoothness of the parameter estimates, as expressed by the penalty term.

In order to estimate the proportions p_{jk} of misreported counts in the matrix \mathbf{C} , we solve a constrained weighted least-squares regression within the IRWLS procedure.

From the structure of the composition matrix \mathbf{C} and since $\mathbf{y} \sim \text{Poisson}(\boldsymbol{\mu})$ we can approximate the distribution of $(\mathbf{y} - \boldsymbol{\gamma})$ as

$$(\mathbf{y} - \boldsymbol{\gamma}) \approx N(\mathbf{\Gamma}\mathbf{p}, \text{diag}(\boldsymbol{\mu})). \quad (4)$$

where $\mathbf{p} = (p_{12}, p_{23}, \dots, p_{J-1,J}; p_{21}, \dots, p_{J,J-1})^T$ is the upper and lower subdiagonal of \mathbf{C} , concatenated into a vector of length $2 \cdot (J - 1)$. The $J \times 2 \cdot (J - 1)$ -matrix $\mathbf{\Gamma}$ is the associated model matrix.

As the number of unknowns in \mathbf{p} , namely $2 \cdot (J - 1)$, is considerably larger than the number J of available data points, additional restrictions have to be imposed on \mathbf{p} .

We introduce a L_1 -penalty into the weighted least-squares problem (4). As pointed out by Tibshirani (1996), this penalty tends to select a small number of elements p_{jk} that exhibit the strongest effects, while shrinking (many) other elements to zero. In this way we obtain the following system of equations for the preference pattern \mathbf{p} :

$$(\mathbf{\Gamma}'\mathbf{V}\mathbf{\Gamma} + \kappa\mathbf{Q})\mathbf{p} = \mathbf{\Gamma}'\mathbf{V}(\mathbf{y} - \boldsymbol{\gamma}), \quad (5)$$

where $\mathbf{V} = \text{diag}(1/\boldsymbol{\mu})$ and the matrix \mathbf{Q} denotes the matrix $\text{diag}(1/|\mathbf{p}| + \epsilon)$. ϵ is a small number, introduced to prevent numerical instabilities when elements of \mathbf{p} become very small; in our experience $\epsilon = 10^{-6}$ worked well.

With the additional parameter κ in (5), which tunes the L_1 -penalty on the misreporting proportions p_{jk} , the estimating equations depend on the combination of the two smoothing parameters λ and κ . To choose the optimal (λ, κ) -combination we minimize Akaike's Information Criterion (AIC), where the effective dimension is taken as the sum of the effective dimensions of the two model components, i.e. the penalized CLM and the penalized WLS-regression.

4 Simulation and Application

To demonstrate the performance of the approach we applied it to several simulated scenarios. The upper-left panel in Figure 1 shows one of the true distributions, i.e. the vector $\boldsymbol{\gamma}$, together with the simulated \mathbf{y} such that $E(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{C}\boldsymbol{\gamma}$ and the estimated values $\hat{\boldsymbol{\gamma}}$. The assumed digit preference in this example attracted additional observations to 5 and 15, from both neighbouring categories.

These estimates were obtained from the optimal combination of (λ, κ) as picked from the AIC-profile shown in Figure 1, upper-right image. The image on the bottom-right panel demonstrates the effect of the L_1 -penalty. On the horizontal axis the value of $\log \kappa$, i.e. the weight of the L_1 -penalty, is given. The optimally chosen $\hat{\kappa}$ practically selects the true proportions, which are depicted by the horizontal dashed lines. As pointed out in Section 3, the model actually estimates $2 \cdot (J - 1)$ misreporting proportions, which have not been restricted to be positive.

Nevertheless we would like to see as final results the net-proportions as positive numbers. This can be easily achieved by simple transformation, which converts $2 \cdot (J - 1)$ parameters to $J - 1$ positive proportions.

The bottom-right image in Figure 1 shows these transformed and hence positive estimates. Additionally, the bottom-left panel in Figure 1 summarizes true and estimated misreporting probabilities for the simulation example.

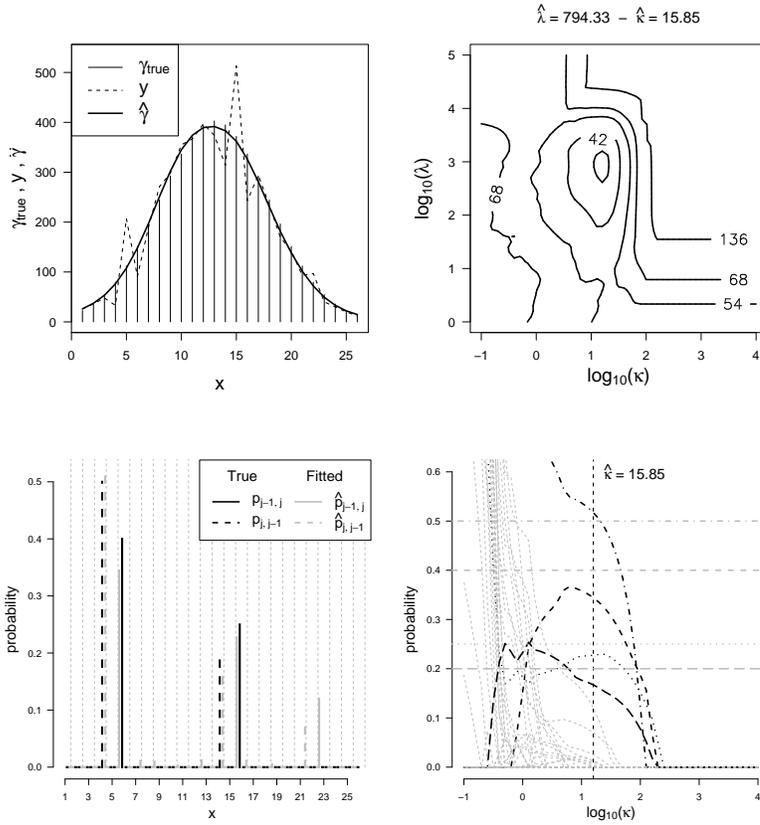


FIGURE 1. Simulated data. Raw data, true values and estimates (top-left). AIC-contour plot (top-right). True misreporting probabilities and estimates (bottom-left). Misreporting probabilities over the grid of κ (bottom-right)

Furthermore we applied our approach to an actual data set taken from the National Health and Nutrition Examination Survey (NHANES, 1980). The survey contains both the self-reported weight (in pounds) of $n = 11,614$ women and men as well as measured weight during the examination.

Figure 2 shows the raw data of both weight variables and the fitted distribution, based on the self-reported data only. The digit preference pattern is clearly seen from the peculiar spiky shape of the self-reported distribution. Even though measured weights may also show minor digit preferences, we may treat them as a proxy to the true ones. Figure 2 demonstrates the close agreement of the estimated latent distribution with the measured one, despite the severe preference pattern present in the original data.

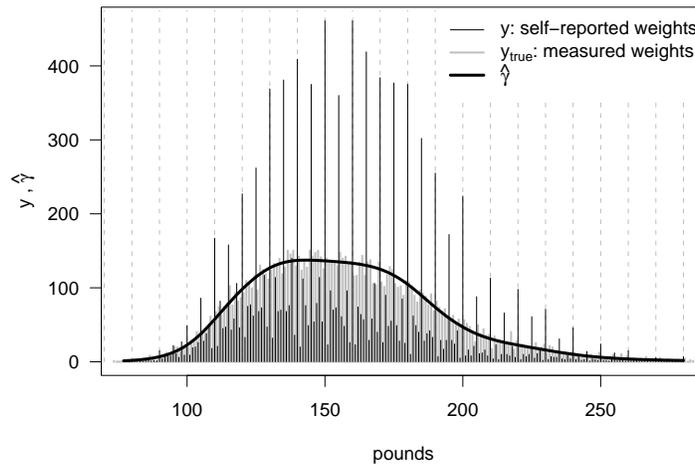


FIGURE 2. Self-reported and measured weight (in pounds) and fitted values for NHANES II data

5 Concluding remarks

The method we have presented in this paper demonstrates how one can deal with digit preferences by combining the concept of penalized likelihood with the composite link model. The only assumption that is made about the underlying true distribution is smoothness.

The misreporting pattern was allowed to partly redistribute observations from any digit to its adjacent neighbours. Again a penalty, in this case a L_1 -penalty, restrains the problem and makes estimation feasible. By allowing this rather flexible preference pattern the tendency to misreport need not be the same for identical end-digits, but may vary over the measurement range, which is often seen in real data.

Extracting the latent distribution will be most important in many applications, however, the pattern of misclassification may also be of interest in itself. The proposed model, which goes beyond the mere quantification of digit preference that is provided by many indexes, allows the analysis of both aspects.

As a future extension of the model we plan to include even more general patterns of misreporting, i.e. allow for exchanges between digits that are more than one category apart. Moreover a challenging area is the use of additional roughness penalties to connect similar transfer probabilities, like from 39 to 40, 49 to 50, etc. If several measurement occasions are available one can envisage digit preferences that improve over time, leading to transfer probabilities that are expected to change smoothly. With additional smoothing parameters to be optimized more speedy algorithms to search for the optimal combination will be advisable.

References

- Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science* **11**, 89-121.
- Eilers, P. H. C. (2007). Ill-posed Problems with Counts, the Composite Link Model, and Penalized Likelihood. *Statistical Modelling* (to appear).
- Eilers, P. H. C. and Borgdorff, M. W. (2004). Modeling and correction of digit preference in tuberculin surveys. *International Journal of Tuberculosis and Lung Diseases* **8**, 232-239.
- NHANES (1976-1980). *National Health and Nutrition Examination Survey (NHANES)*. US National Center for Health Statistics. Available at www.cdc.gov/nchs/nhanes.htm
- Thompson, R. and Baker, R. J. (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics* **30**, 125-131.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* **58**, 267-288.

Following brake reaction time in total knee arthroplasty: analysis of variance for repeated measures

Ana Isabel Carita¹, Carlos J. Marques², João Barreiros³ and Jan Cabri⁴

¹ Departamento de Métodos Matemáticos/CIPER, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa, Estrada da Costa, 1495-688 Cruz Quebrada-Dafundo, Portugal, acarita@fmh.utl.pt

² ENDO-Clinic, Hamburg

³ Departamento de Ciências da Motricidade, Faculdade de Motricidade Humana

⁴ Unidade de Fisioterapia, Faculdade de Motricidade Humana

Abstract: Total knee arthroplasty (TKA) is a common solution for patients with arthritis of the knee. After surgery, patients often ask when they can safely start driving again. The aim of this study was to investigate the effects of primary right and left TKA on brake response time (BRT) and its components - reaction time (RT) and movement time (MT) - of right leg. Furthermore, the effect of gender on BRT, RT and MT were investigated. The BRT, RT and MT of the patients were assessed in a car simulator before de surgery and then ten and third days after surgery. Analysis of variance (ANOVA) for repeated measures with two between-subjects factor (gender and side of surgery) and one within-subject factor (measurement time) was used to analyze differences in the mean BRT, RT and MT values between groups, across the three measurements times.

Keywords: Repeated measures; total knee arthroplasty; brake response time.

1 Brake Response Time after Primary Knee Arthroplasty

Total knee arthroplasty (TKA) reduces pain, enhances function and improves the health-related quality of life in individuals with knee impairment and disability secondary to osteoarthritis or rheumatoid arthritis. After surgery, patients often ask their physicians when they can safely start driving again. This is a relevant question because an early return to car driving can improve mobility in daily activities during first stages of the rehabilitation process, reducing dependency and social isolation.

Driving is a complex activity involving cognitive and motor skills, and brake response time (BRT) has been widely used to measure drivers' performance (Hau *et al.*, 2000). The aim of this study was to investigate the effects of primary right and left TKA on brake response time (BRT) and its components, reaction time (RT) and movement time (MT), in order to find out when patients can safely return to car driving. Furthermore, the question whether gender affects BRT and BRT recovery after TKA was also investigated.

1.1 Study Design

To study the effects of TKA on BRT a prospective study with repeated measures design was used. BRT performances from each patient were measured at three points in time: one day before surgery and ten and thirty days post-operatively (Marques, 2006).

The study population consisted of active automobile drivers who underwent left or right primary TKA with osteoarthritis of the knee joint as the indication for surgery. To be eligible all patients had to meet three criteria: they were admitted for primary right or left knee arthroplasty, they assured that they drove frequently (at least once a week), and they would be available for the third measurement. Sixty-four patients signed the informed consent and were examined pre-operatively. Ten days after surgery 51 patients were followed-up and 30 days after the operation 32 patients returned to the clinic for the third measurement.

1.2 Task Description

The components of BRT (RT and MT) were assessed as main dependent variables in a car simulator. The data acquisition system consisted of a BIOPAC MP100A-CE, connected to a laptop with processing software, one red LED lamp located at patients' eye level, one trigger and switches on the accelerator and brake pedals. Sample rate was set at 1000Hz.

On each measurement day the patients performed five practice and ten test trials in the BRT task in car simulator. The patients were instructed to press the accelerator pedal as if they were driving. When the red LED switched on the patients had to brake as quickly as possible. After pressing the brake pedal patients were instructed to move their foot back to the accelerator and to wait for the next stimulus. Stimulus triggering varied randomly from one to five seconds.

RT was defined as the time interval (ms = milliseconds) between the lighting up of the LED and initiation of the movement of the foot. MT was defined as the time (ms) between initiation of the movement of the foot and the first contact with the brake pedal. BRT (ms) was defined as the sum of RT and MT.

2 Statistical Analysis

Analysis of variance (ANOVA) for repeated measures (Crowder *et al*, 1990) with two between-subjects factor (gender and side of surgery) and one within-subject factor (measurement time) was used to analyze differences in the mean BRT, RT and MT values between groups, across the three measurements times.

All statistical tests were carried out using the software SPSS 14.5.

3 Results

The data of the thirty-two patients, who remained in the study throughout all three measurements, were pooled for analysis.

TABLE 1. Analysis of Variance: gender \times side of surgery \times time.

	Sum of squares	df	Mean square	F	p
Time	14084.1	2	7042.0	4.297	0.018
Time \times gender	1570.0	2	785.0	0.479	0.622
Time \times sidesur	10279.5	2	5138.3	3.135	0.051
Time \times gender \times sidesur	1858.7	2	929.4	0.567	0.570
Gender	39888.1	1	39888.1	4.209	0.050
Sidesur	8991.7	1	8991.7	0.949	0.338
Gender \times sidesur	13777.7	1	13777.7	1.454	0.238

3.1 Analysis of variance for repeated measures

The results of ANOVA for repeated measures for BRT differences between gender and side of surgery across time are shown in table 1. The Mauchly test of sphericity showed that the required assumptions for ANOVA were met ($p = 0.136$).

A 3.135 F ratio was significant for the within-subjects interaction factor time \times sidesur ($p = 0.051$). With this result we must do careful interpretation to significant F ratio for the main factor time ($p = 0.018$). There were no significant interactions between factors for both, within or between-subjects effects, and there were a significant main effect for the between-subjects factor gender ($p = 0.05$).

Male and female patients had different BRT profiles of change across the time and side of surgery as shown in figure 1.

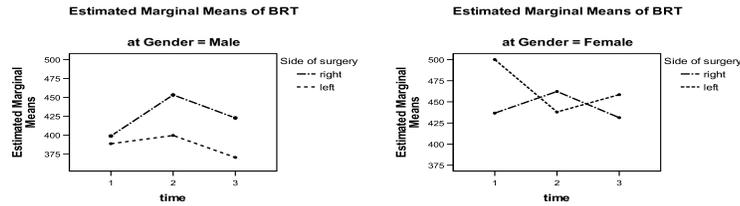


FIGURE 1. BRT profiles for male and female patients.

RT and MT differences between gender and side of surgery across time were analyzed. ANOVA for repeated measures showed significant interactions time \times gender for RT and time \times sidesur for MT and a significant main effect for the between-subjects factor gender for MT.

3.2 Conclusions

Significant interactions between time and side of surgery and significant main effect for gender, for the variables BRT and MT, indicates that male and female patients did not have the same BRT profile neither the same MT profile across the three measurements, for right and left side of surgery. Significant interaction between time and gender for RT indicates different RT profiles for male and female across the three measurements.

Gender and time affect BRT and its components. Side of surgery affects BRT and MT. Pré-operatively there were BRT, RT and Mt statistically significant differences for males and females ($p < 0.05$) in a independent sample t -test. The effect of gender on BRT an its components was clear before surgery. For that reason and because only two females patients in the left TKA group remained in the study throughout all three measurements we decided to do a new ANOVA for repeated measures with one between-subject factor, side of surgery, and one within-subject factor, measurement time. There was a significant interaction between time and side of surgery only for MT ($F = 5.975; p = 0.004$) and a significant main effects for the within-subject factor time for BRT ($F = 4.26; p = 0.019$). These results indicated that the right and left TKA groups did not have the same MT pattern of change across the three measurements. Once more we concluded that time and side of surgery affect BRT and its components. A further analysis with a linear mixed model can be helpful to establish relations between variables.

References

- Crowder, M. J. and Hand, D.J. (1990). *Analysis of Repeated Measures*. London: Chapman & Hall.
- Hau, R., Csongvay, S., & Bartlett, J. (2000). Driving reaction time after right knee arthroscopy. *Knee Surg Sports Traumatol Arthrosc* **8(2)**, 89-92.
- Marques, C. J. (2006). *Brake Response Time after Primary Total Knee Arthroplasty*. Master thesis, Faculty of Human Movement Sciences at the Technical University of Lisbon.

Testing Markov Chain Lumpability

Manuela Cazzaro¹, Roberto Colombi² and Sabrina Giordano³

¹ Università di Milano-Bicocca, (Milano) - Italy, manuela.cazzaro@unimib.it

² Università di Bergamo, Dalmine (Bergamo) - Italy, colombi@unibg.it

³ Università della Calabria, (Cosenza) - Italy, esabrina.giordano@unical.it

Abstract: When the states of a Markov chain are aggregated the new process is not necessarily a Markov process with transition probabilities that do not depend on the past of the original Markov chain. When this happens the original Markov process is called lumpable. In this work we show how to test the hypotheses of lumpability. The relevance of the previous problem is highlighted for bivariate Markov chains. We point out that the lumpability hypotheses are equivalent to linear constraints on some interaction parameters defined on the transition probabilities. The methodology is finally applied to some real data.

Keywords: Markov Chains; Lumpability; Granger Noncausality; Marginal Models; Generalized Marginal Interactions.

1 Introduction

In the study of Markov chains the lumpability property allows the states to be grouped into subsets which are identified with the states of another Markov chain, called *lumped chain*. This is extremely advantageous because the lumped process reduces the analytical and computational complexity and, at the same time, it inherits the Markov property of the original process. Moreover, the identification of proper aggregations may provide new insights about the process and sometimes studying the simpler lumped process is enough when a coarse analysis is needed.

In section 2 we firstly give a definition of the lumping property equivalent to that proposed by Kemeny and Snell (1960). The new definition is based on the Granger noncausality condition (Chamberlain, 1982) that allows us to point out some aspects of the involved processes that are neglected in the Kemeny and Snell definition. Then we focus on the relevance of lumpability in multivariate Markov processes when it is useful to verify if a marginal process of the original multivariate Markov chain is markovian as well. The hypotheses of lumpability are expressed in section 3 through some linear constraints on suitable interactions of a parameterization of the transition probabilities. We finally test some lumpability hypotheses on bivariate categorical time series of soft-drinks sales demand (section 4).

2 Lumped Markov chains

We assume that $\{A_t\}$ is a first order Markov chain which means that the condition $A_t \perp\!\!\!\perp \mathbf{A}_{-\infty}^{t-2} | A_{t-1}$ holds for all t . Let $h(\cdot)$ be a function that maps the state space $\mathcal{A} =$

$\{a_1, a_2, \dots, a_r\}$ of $\{A_t\}$ onto a set $\mathcal{G} = \{g_1, g_2, \dots, g_s\}$ of smaller cardinality, $1 < s < r$. Obviously the sets $\{a : g_k = h(a), a \in \mathcal{A}\}$, $k = 1, 2, \dots, s$, define a partition of \mathcal{A} . The stochastic process $\{G_t\}$ with state space \mathcal{G} and such that $G_t = h(A_t)$, is a *lumped version* of $\{A_t\}$ if and only if the following two conditions hold:

- i) *Granger noncausality condition*: the process $\{A_t\}$ does not Granger cause $\{G_t\}$ which means that $G_t \perp\!\!\!\perp \mathbf{A}_{-\infty}^{t-1} | \mathbf{G}_{-\infty}^{t-1}$ or equivalently it is true that $Pr(G_t | \mathbf{A}_{-\infty}^{t-1}) = Pr(G_t | \mathbf{A}_{-\infty}^{t-1}, \mathbf{G}_{-\infty}^{t-1}) = Pr(G_t | \mathbf{G}_{-\infty}^{t-1})$, for all t ;
- ii) *Markovianity condition*: $\{G_t\}$ is a Markov chain that is $G_t \perp\!\!\!\perp \mathbf{G}_{-\infty}^{t-2} | G_{t-1}$ or equivalently $Pr(G_t | \mathbf{G}_{-\infty}^{t-1}) = Pr(G_t | G_{t-1})$ hold for all t .

In general, the previous conditions imply that the lumpability property is stronger than the markovianity of the process $\{G_t\}$, obtained by aggregating the states of $\{A_t\}$. We now state a theorem which shows that i) and ii) are equivalent to the simpler condition, for all t ,

$$Pr(G_t | A_{t-1}) = Pr(G_t | A_{t-1}, G_{t-1}) = Pr(G_t | G_{t-1}). \quad (1)$$

Moreover, the theorem implies that our definition of lumpability (conditions i) and ii)) is equivalent to the one given by Kemeny and Snell (1960).

Theorem 1 *If $\{A_t\}$ is a Markov chain then $\{G_t\}$ is a lumped version of $\{A_t\}$ if and only if $G_t \perp\!\!\!\perp A_{t-1} | G_{t-1}$ for all t .*

The theorem straightly follows from the well-known properties of the conditional independence relation.

Now we show the importance of lumpability when the Markov chain is multivariate and, without loss of generality, we will discuss the lumpability property for the bivariate case only.

If $\{A_t\} = \{E_t, F_t\}$ is a first order bivariate Markov chain with state space $\mathcal{E} \times \mathcal{F} = \{(e_i, f_j), i = 1, 2, \dots, r, j = 1, 2, \dots, c\}$, and if the function $h(\cdot)$ is such that $e_i = h(e_i, f_j)$, the marginal process $\{E_t\}$, with state space \mathcal{E} , can be seen as a lumped version of $\{A_t\}$. In this case, which we call *marginal lumpability*, the conditions i) and ii) state that the process $\{F_t\}$ does not Granger cause $\{E_t\}$ and that $\{E_t\}$ is a univariate Markov chain. Here, the necessary and sufficient condition of the theorem 1 for $\{A_t\} = \{E_t, F_t\}$ to be lumpable in $\{G_t\} = \{E_t\}$ becomes $E_t \perp\!\!\!\perp F_{t-1} | E_{t-1}$ and the (1) is rewritten as follows:

$$Pr(E_t | E_{t-1}, F_{t-1}) = Pr(E_t | E_{t-1}). \quad (2)$$

Thus, marginal lumpability is useful when we are interested in testing if the marginal process $\{E_t\}$ of the Markov chain $\{E_t, F_t\}$ retains the markovian property with transition probabilities independent of F_{t-1} .

We consider another example of lumpability in which all pairs of states in $\mathcal{E} \times \mathcal{F}$ of the original bivariate chain are lumped by aggregating the categories of a component process and marginalizing with respect to the other one. A simple way to achieve this is to adopt a function $g = h(e, f) = \tilde{h}(e)$, $g \in \mathcal{G}$, $(e, f) \in \mathcal{E} \times \mathcal{F}$ that defines a partition of the set of categories $\{e_1, e_2, \dots, e_r\}$ of the marginal process $\{E_t\}$. The image space \mathcal{G} is the state space of the lumped chain $\{G_t\}$ where $G_t = h(E_t, F_t) = \tilde{h}(E_t)$.

In this case the condition of the theorem 1 is equivalent to the independence $G_t \perp\!\!\!\perp E_{t-1}, F_{t-1} | G_{t-1}$ that is clearly the same as

$$Pr(G_t | E_{t-1}, F_{t-1}) = Pr(G_t | G_{t-1}). \quad (3)$$

Note that the chain $\{G_t\}$ just presented and the marginal process $\{E_t\}$ of case (2) are both lumped versions of the chain $\{E_t, F_t\}$, but in (3) the marginal $\{E_t\}$ is not necessarily a Markov chain, whereas $\{G_t\}$ it is.

Another lumpability hypothesis may be specified by aggregating the states of $\{E_t\}$ of the bivariate chain $\{E_t, F_t\}$ in pairwise disjoint subsets whatever category of $\{F_t\}$ is observed at each time. The chain $\{G_t, F_t\}$, which is a lumped version of $\{E_t, F_t\}$, is still bivariate, unlike the previous cases. Technically, the chain with joint state space $\mathcal{E} \times \mathcal{F}$ is lumped into the chain with state space given by the image space $\{(g_k, f_j), k = 1, 2, \dots, s, j = 1, 2, \dots, c\}$ of the function $(g, f) = h(e, f)$, $(e, f) \in \mathcal{E} \times \mathcal{F}$. For the resulting lumped process $\{G_t, F_t\}$, where $(G_t, F_t) = h(E_t, F_t)$, the condition of the theorem 1 is $G_t, F_t \perp\!\!\!\perp E_{t-1} | G_{t-1}, F_{t-1}$ that is equivalent to

$$Pr(G_t, F_t | E_{t-1}, F_{t-1}) = Pr(G_t, F_t | G_{t-1}, F_{t-1}). \quad (4)$$

Many other possibilities to apply the property of lumpability exist and they could be an interesting subject for further researches.

The previous conditions of lumpability hold for both homogeneous and non-homogeneous Markov chains, however, in the next sections we will focus only on time-homogeneous Markov chains with strictly positive transition probabilities.

3 Constraints on generalized marginal interactions

Let us assume that the joint transition probabilities of a first order bivariate Markov chain $\{E_t, F_t\}$ may be thought of as the conditional joint probabilities of E_t, F_t given E_{t-1}, F_{t-1} in a contingency table. Note that joint probabilities of $E_t, F_t, E_{t-1}, F_{t-1}$, (variables 1, 2, 3, 4), can be obtained by multiplying the transition probabilities of E_t, F_t , given E_{t-1}, F_{t-1} , for example, with the probabilities of the invariant distribution. However the interaction parameters we are going to introduce do not depend on the choice of the marginal distribution of E_{t-1}, F_{t-1} .

The similarity with contingency tables justifies a parameterization of the joint distribution for the four variables in terms of *generalized marginal interactions* (Bartolucci et al., 2007). In particular, we will show that the lumpability hypotheses of interest (conditions 2-4) are equivalent to linear constraints on some generalized marginal interactions.

Note that generalized marginal interactions are standard log-linear interactions which are computed in tables obtained by marginalizing with respect to some variables and by aggregating the categories of some other variables and they can be seen as contrasts of well-known types of generalized logits and log odds ratios.

Any generalized marginal interaction, $\eta_{\mathcal{I}, \mathcal{M}}(\mathbf{m}_{\mathcal{I}})$, is specified by the interaction set \mathcal{I} of the variables involved, by the set \mathcal{M} of the variables of the marginal distributions

within which the interactions are defined, by the logit type assigned to each variable of \mathcal{M} and by the multi-index $\mathbf{m}_{\mathcal{I}}$ which the interaction depends on.

The generalized marginal interactions include also the *recursive* or *nested* logits and log-odds ratios, introduced by Cazzaro and Colombi (2006). In order to define the recursive interactions, that we will adopt, the following definition of a *Coherent Complete Hierarchy of Sets* (CCHS) is needed.

Let A be a variable with r categories in the set \mathcal{A} ; by *coherent complete hierarchy of sets* $\mathcal{H}(\mathcal{A})$ a family of subsets of \mathcal{A} is intended. It is characterized by the following properties: (i) $\mathcal{A} \in \mathcal{H}(\mathcal{A})$; (ii) $\{a_i\} \in \mathcal{H}(\mathcal{A})$, $i = 1, 2, \dots, r$; (iii) if $P, Q \in \mathcal{H}(\mathcal{A})$ then $P \cap Q \in \{P, Q, \emptyset\}$; (iv) let M_m be a non-minimal set (node) of $\mathcal{H}(\mathcal{A})$ then for $m = 1, 2, \dots, r - 1$:

$$M_m = \mathcal{B}(a_{i(m)}, 0) \cup \mathcal{B}(a_{i(m)}, 1), \quad \mathcal{B}(a_{i(m)}, 0), \mathcal{B}(a_{i(m)}, 1) \in \mathcal{H}(\mathcal{A})$$

$$i(m) = \sup\{i : a_i \in \mathcal{B}(a_{i(m)}, 0)\}, \quad i(m) + 1 = \inf\{i : a_i \in \mathcal{B}(a_{i(m)}, 1)\}.$$

The previous definition implies that each set M_m contains contiguous categories. This restriction is natural for ordered categorical variables but however it is not a limitation for non ordered variables because the categories can always be relabelled in such a way to meet the restriction.

The generalized marginal interactions are defined, from now on, by choosing local logits for the variables F_t, F_{t-1} . For the variables E_t, E_{t-1} recursive logits are chosen, based on the same $\mathcal{H}(\mathcal{E})$ such that $\mathcal{H}(\mathcal{E}) \supset \mathcal{S}(\mathcal{E})$, where $\mathcal{S}(\mathcal{E})$ is the family of the sets $\{e : g = \tilde{h}(e)\}$, $\forall g \in \mathcal{G}$, defined before formula (3). Let $\mathcal{N}(\mathcal{E}) = \{N_m, m = 1, 2, \dots, r - 1\}$ be the family of nodes of $\mathcal{H}(\mathcal{E})$ and let $\mathcal{S}^*(\mathcal{E}) = \{N_m : N_m \in \mathcal{N}(\mathcal{E}), N_m \not\subseteq N_t, \forall N_t \in \mathcal{S}(\mathcal{E})\}$ be the family of nodes not contained in any set of $\mathcal{S}(\mathcal{E})$.

In order to test the lumpability conditions (2) and (3), generalized marginal interactions defined within the marginal distribution of (E_t, E_{t-1}, F_{t-1}) have to be constrained. In particular, the marginal lumpability condition (2), $E_t \perp F_{t-1} | E_{t-1}$, holds when the following interactions are equal to zero:

$$\eta_{14;134}(m, q), \quad \eta_{134;134}(m, p, q)$$

$$m = 1, 2, \dots, r - 1, \quad p = 1, 2, \dots, r - 1, \quad q = 1, 2, \dots, c - 1. \quad (5)$$

The marginal interactions

$$\eta_{14;134}(m, q) = \ln \left(\frac{P(E_t \in \mathcal{B}(e_{i(m)}, 1), F_{t-1} \in \{f_{q+1}\}, E_{t-1} \in \mathcal{B}(e_{i(1)}, 0))}{P(E_t \in \mathcal{B}(e_{i(m)}, 1), F_{t-1} \in \{f_q\}, E_{t-1} \in \mathcal{B}(e_{i(1)}, 0))} \cdot \frac{P(E_t \in \mathcal{B}(e_{i(m)}, 0), F_{t-1} \in \{f_q\}, E_{t-1} \in \mathcal{B}(e_{i(1)}, 0))}{P(E_t \in \mathcal{B}(e_{i(m)}, 0), F_{t-1} \in \{f_{q+1}\}, E_{t-1} \in \mathcal{B}(e_{i(1)}, 0))} \right),$$

$$m = 1, 2, \dots, r - 1, \quad q = 1, 2, \dots, c - 1,$$

are recursive-local log-odds ratios and the variable E_{t-1} is set to the reference category. The three-order interactions $\eta_{134;134}(m, p, q)$ are log-contrasts between recursive-local odds ratios obtained varying the level of the variable E_{t-1} . Note that all the generalized marginal interactions that follow are similarly interpreted.

The following interactions are constrained to be zero by condition (3), $G_t \perp\!\!\!\perp E_{t-1}, F_{t-1} | G_{t-1}$:

$$\begin{aligned} & \eta_{13;134}(m, p), \quad \eta_{14;134}(m, q), \quad \eta_{134;134}(m, p, q) \\ \forall m : N_m \in \mathcal{S}^*(\mathcal{E}) \quad \forall p : N_p \in \mathcal{N}(\mathcal{E}) \setminus \mathcal{S}^*(\mathcal{E}) \quad q = 1, 2, \dots, c-1. \end{aligned} \quad (6)$$

In fact, when the interactions (6) are null, there is stochastic independence between E_t and E_{t-1}, F_{t-1} in all the three way sub-tables where the categories of E_t are lumped or aggregated into the sets of $\mathcal{S}(\mathcal{E})$ and the variable E_{t-1} is conditioned to a set of $\mathcal{S}(\mathcal{E})$.

In order to consider the lumpability hypothesis $G_t, F_t \perp\!\!\!\perp E_{t-1} | G_{t-1}, F_{t-1}$, that is condition (4), we use a set of generalized interactions defined on the joint probabilities of $(E_t, F_t, E_{t-1}, F_{t-1})$.

The following interactions are constrained to be zero by condition (4):

$$\begin{aligned} & \eta_{13;1234}(m, p), \quad \eta_{23;1234}(n, p), \quad \eta_{123;1234}(m, n, p), \\ & \eta_{134;1234}(m, p, q), \quad \eta_{234;1234}(n, p, q), \quad \eta_{1234;1234}(m, n, p, q) \\ \forall m : N_m \in \mathcal{S}^*(\mathcal{E}) \quad \forall p : N_p \in \mathcal{N}(\mathcal{E}) \setminus \mathcal{S}^*(\mathcal{E}), \quad n, q = 1, 2, \dots, c-1. \end{aligned} \quad (7)$$

In fact, the just mentioned constraints on the interactions (7) imply that E_t, F_t are stochastically independent of E_{t-1} given F_{t-1} in all the four way sub-tables where the categories of E_t are lumped or aggregated into the sets of $\mathcal{S}(\mathcal{E})$ and the variable E_{t-1} is conditioned to a set of $\mathcal{S}(\mathcal{E})$.

4 Example

The methodology proposed in this work is used to analyze the data of a soft-drink company, Ching et al. (2002). The data are a one-year time series of daily sales demand of two soft-drinks: orange juice, $\{E_t\}$, and apple juice, $\{F_t\}$, both with categories: n -no demand, l -low, m -medium, h -high level. Under the assumption that a first order bivariate Markov chain well describes the behavior of the joint time series, we will show how the hypotheses presented in section 2 can be verified on these data. These hypotheses have been tested by using the package *hmmm*, *Hierarchical Multinomial Marginal Models*, developed by Cazzaro and Colombi in R. The package and the data are available from the authors. At first we consider the condition (2) of marginal lumpability which means that the process $\{F_t\}$ does not Granger cause $\{E_t\}$ and that the process $\{E_t\}$ is marginally a Markov chain. This signifies that the demand of the orange juice, given its previous level, is independent of the demand of the apple juice occurred the day before. Moreover, the demand of orange juice does not depend on its overall past sales, once known its last sales demand. Testing this model against the saturated one gives this result: $G^2 = 39.23$, $df = 36$, $p\text{-value} = 0.327$. The condition (3), instead, means that the bivariate process $\{E_t, F_t\}$ does not Granger cause $\{G_t\}$ and that the lumped process $\{G_t\}$ is markovian. The states of $\{G_t\}$, we considered, are the lumped categories nl (no demand-low) and mh (medium-high) of orange juice demand. In this case (3) indicates that the present level of demand of orange juice (nl or mh) depends on the last level (nl or mh) of orange juice sales, regardless of

the exact level (n, l, m, h) of last sales of orange and apple juices. Testing this model against the saturated one gives: $G^2 = 15.62$, $df = 11$, $p\text{-value} = 0.156$. If we need both the lumped process $\{G_t\}$ and the marginal one $\{E_t\}$ to be markovian, we will test simultaneously the conditions (2) and (3), $G^2 = 48.61$, $df = 38$, $p\text{-value} = 0.116$. Finally, $G^2 = 52.29$, $df = 56$, $p\text{-value} = 0.616$ are the results of testing the model for the condition (4) which signifies that $\{E_t, F_t\}$ does not Granger cause $\{G_t, F_t\}$ and $\{G_t, F_t\}$ is a bivariate Markov chain. Note that all tested models show a good fit.

References

- Bartolucci, F., Colombi, R., and Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica* **17**, 2.
- Cazzaro, M., Colombi, R. (2006). Modelling two way contingency tables with recursive logits and odds ratios. Submitted.
- Chamberlain, G. (1982). The General equivalence of Granger and Sims causality. *Econometrica* **50**, 569-581.
- Ching, W.K, Fung, E.S., Ng, M.K. (2002). A multivariate Markov chain for categorical data sequences and its applications in demand predictions. *IMA Journal of Management Mathematics* **13**, 187-199.
- Kemeny, J.C., Snell, J.L. (1960). *Finite Markov Chains*. D. Van Nostrand.

Note on A-optimal chemical balance weighing design

Bronisław Ceranka¹ and Małgorzata Graczyk²

¹ Department of Mathematical and Statistical Methods Agricultural University Wojska Polskiego 28, 60-637 Poznań, bronicer@au.poznan.pl

² Department of Mathematical and Statistical Methods Agricultural University Wojska Polskiego 28, 60-637 Poznań, magra@au.poznan.pl

Abstract: The A-optimal criterion of the existing of the chemical balanced weighing design is considered.

Keywords: A-optimal chemical balance weighing design; balanced bipartite weighing design; ternary balanced block design.

1 Introduction

Suppose specifically that there are p objects of true unknown weights w_1, w_2, \dots, w_p , respectively, and we wish to estimate them employing n measuring operations using a chemical balance. Let y_1, y_2, \dots, y_p denote the recorded observations in these n operations, respectively. It is assumed that the observations follow the standard linear model

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}_n, \quad E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}_n, \quad (1)$$

where \mathbf{X} is of order $n \times p$ and is called the weighing design matrix. The elements of \mathbf{X} are x_{ij} , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$ and a typical element x_{ij} is -1 if the j th object is placed on the left pan during the i th weighing operation, $+1$ if the j th object is placed on the right pan during the i th weighing operation and 0 if the j th object is not utilized in either pan during the i th weighing operation. Hence $\mathbf{w} = (w_1, w_2, \dots, w_p)'$ is the vector of true unknown weights (of parameters). The vector \mathbf{e} is the so-called vector of error components satisfying the usual homoscedasticity condition. The inference problem centers around the estimation of true individual weights of all the objects. The optimality problem is concerned with efficient estimation in some sense by a proper choice of the design matrix \mathbf{X} among designs at our disposal. The model (1) is the standard Gauss - Markoff model and the following results are well known. The parameter vector \mathbf{w} is estimable if and only if $r(\mathbf{X}) = p$ in which case

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \mathbf{V}(\hat{\mathbf{w}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \quad (2)$$

where $\hat{\mathbf{w}}$ is the blue and \mathbf{V} is the dispersion matrix.

Among all possible designs an A-optimal design minimizes the sum (or equivalently the average) of the variances of $\hat{w}_1, \hat{w}_2, \dots, \hat{w}_p$. Wong and Masaro (1984) gave the lower bound for $\text{tr}(\mathbf{X}'\mathbf{X})^{-1}$ and some construction methods of the A-optimal chemical balance weighing designs.

In the next section we give the new lower bound for $\text{tr}(\mathbf{X}'\mathbf{X})^{-1}$ and the necessary and sufficient conditions to this lower bound to be attained under the given restriction on the number of objects included in the particular measurement operation. We present a new method of construction of the A-optimal design based on the balanced bipartite weighing designs and the ternary balanced block designs.

2 Some results on variance limit of estimated weights

Let \mathbf{X} be an $n \times p$ design matrix of a chemical balance weighing design. The following result from the paper Wong and Masaro (1984) gives a lower bound for $\text{tr}(\mathbf{X}'\mathbf{X})^{-1}$

Lemma 2.1. For an $n \times p$ design matrix \mathbf{X} of rank p we have inequality

$$\text{tr}(\mathbf{X}'\mathbf{X})^{-1} \geq \frac{p^2}{\text{tr}(\mathbf{X}'\mathbf{X})},$$

the equality being attained if and only if $\mathbf{X}'\mathbf{X}$ is equal to the $p \times p$ identity matrix \mathbf{I}_p multiplied by scalar, i.e. $\mathbf{X}'\mathbf{X} = z\mathbf{I}_p$.

In many problems concerning weighing experiments the A-optimal designs are considered. There are designs in which $\text{tr}(\mathbf{X}'\mathbf{X})^{-1}$ attain the lower bound. The lower bound of $\text{tr}(\mathbf{X}'\mathbf{X})^{-1}$ is attained if and only if the elements of the design matrix \mathbf{X} are equal to -1 or 1 , only. It implies that in each measurement operation all objects must be included in different combinations. That is the reason why in the present paper we consider the situation the elements of the design matrix \mathbf{X} are equal to 0 , either. It is equivalent in each weighing not all objects are included. Thus we give new lower bound of $\text{tr}(\mathbf{X}'\mathbf{X})^{-1}$. We have

Theorem 2.1. For any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} = (x_{ij})$ we have

$$\text{tr}(\mathbf{X}'\mathbf{X})^{-1} \geq \frac{p^2}{q \cdot n}, \quad (3)$$

where $q = \max(q_1, q_2, \dots, q_n)$, $q_i = \sum_{j=1}^p x_{ij}^2$, $i = 1, 2, \dots, n$. In the case $q = p$ we get the inequality given in Wong and Masaro (1984).

Definition 2.1. Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} = (x_{ij})$ is said to be A-optimal if

$$\text{tr}(\mathbf{X}'\mathbf{X})^{-1} = \frac{p^2}{q \cdot n}.$$

Theorem 2.2. Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} = (x_{ij})$ is A-optimal if and only if

$$\mathbf{X}'\mathbf{X} = \frac{q \cdot n}{p} \mathbf{I}_p.$$

In the present paper we will construct an A-optimal chemical balance weighing design under the restriction $p_1 + p_2 = q \leq p$, where p_1 and p_2 represent the number of objects placed on the left and on the right pan, respectively, in each of the measurement operations. The construction is based on the incidence matrices of the balanced bipartite weighing designs and the ternary balanced block designs.

3 Construction of the design matrix

Let \mathbf{N}^* be the incidence matrix of the balanced bipartite weighing design with the parameters $v, b, r, k_1, k_2, \lambda_1, \lambda_2$ (See Huang (1976)). From the matrix \mathbf{N}^* we form the matrix \mathbf{N} by replacing k_1 elements equal to $+1$ of each column which correspond to the elements belonging to the first subblock by -1 . Thus each column of the matrix \mathbf{N} will contain k_1 elements equal to -1 and k_2 elements equal to $+1$. From the matrix \mathbf{N} we construct the design matrix \mathbf{X} of the chemical balance weighing design in the form $\mathbf{X} = \mathbf{N}'$. In this design $p = v$ and $n = b$.

Theorem 3.1. Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} = \mathbf{N}'$ is A-optimal if and only if

$$\lambda_2 = \lambda_1 \quad (4)$$

and

$$q = k. \quad (5)$$

Now, we consider the chemical balance weighing design with the design matrix $\mathbf{X} = \mathbf{N}' - \mathbf{1}_b \mathbf{1}'_v$, where \mathbf{N} is the incidence matrix of the ternary balanced block design with the parameters $v = b, r = k, \lambda, \rho_1, \rho_2$ (See Billington and Robinson (1983)). In this design $p = v$ and $n = b$.

Theorem 3.2. Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} = \mathbf{N}' - \mathbf{1}_b \mathbf{1}'_v$ is A-optimal if and only if

$$b + \lambda - 2r = 0 \quad (6)$$

and

$$q = b - \rho_1. \quad (7)$$

4 Balanced bipartite weighing designs leading to the A-optimal design

We have seen in the Theorem 3.1 that if the parameters of the balanced bipartite weighing design satisfy the condition (5) then a chemical balance weighing design with the design matrix $\mathbf{X} = \mathbf{N}'$ is A-optimal. Under this condition we have the following theorem

Theorem 4.1. The existence of the balanced bipartite weighing design with the parameters $v, b = \frac{2sv(v-1)}{c^2(c^2-1)}, r = \frac{2s(v-1)}{c^2-1}, k_1 = \frac{c(c-1)}{2}, k_2 = \frac{c(c+1)}{2}, \lambda_1 = s, \lambda_2 = s, c = 2, 3, \dots, s = 1, 2, \dots$ implies the existence of the A-optimal chemical balance weighing design, $v \geq c^2, q = c^2$.

5 Ternary balanced block designs leading to the A-optimal design

We have seen in the Theorem 4.2 that if the parameters of the ternary balanced block design satisfy the condition (7) then a chemical balance weighing design with

the design matrix $\mathbf{X} = \mathbf{N}' - \mathbf{1}_b \mathbf{1}'_v$ is A-optimal. Under this condition we have the following theorem

Theorem 5.1. The existence of the ternary balanced block design with the parameters

- (i) $v = b = s, r = k = s - 2, \lambda = \rho_1 = s - 4, \rho_2 = 1, s = 6, 7, \dots,$
- (ii) $v = b = s, r = k = s - 3, \lambda = s - 6, \rho_1 = s - 9, \rho_2 = 3, s = 10, 11, \dots,$
- (iii) $v = b = s, r = k = s - 4, \lambda = s - 8, \rho_1 = s - 16, \rho_2 = 6, s = 17, 18, \dots,$

implies the existence of the A-optimal chemical balance weighing design, $q = 4, 6, 8,$ respectively.

References

- Wong, C.S. and Masaro, J.C. (1984). A-optimal design matrices $\mathbf{X} = (x_{ij})_{N \times n}$ with $x_{ij} = -1, 0, 1$. *Linear and Multilinear Algebra* **15**, 23-46.
- Huang, Ch. (1976). Balanced bipartite block designs. *Journal of Combinatorial Theory (A)* **21**, 20-34.
- Billington, E.J. and Robinson, P.J. (1983). A list of balanced ternary designs with $R \leq 15$ and some necessary existence conditions. *Ars Combinatoria* **16**, 235-258

Optimum chemical balance weighing design for $p+1$ objects

Bronisław Ceranka¹ and Małgorzata Graczyk²

¹ Department of Mathematical and Statistical Methods Agricultural University Wojska Polskiego 28, 60-637 Poznań, bronicer@au.poznan.pl

² Department of Mathematical and Statistical Methods Agricultural University Wojska Polskiego 28, 60-637 Poznań, magra@au.poznan.pl

Abstract: The problem of estimation of unknown weights of objects in the model of the chemical balanced weighing design under the condition errors are uncorrelated and have different variances is considered.

Keywords: chemical balance weighing design; balanced bipartite weighing design.

1 Introduction

Let us suppose we want to estimate the weights of p objects by weighing them n times using a chemical balance, $p \leq n$. The manner of allocation of objects on the pans is described through columns of the $n \times p$ matrix \mathbf{X} . Its elements are equal to -1 , 1 or 0 if the object is kept on the left pan, right pan or is not included in the particular measurement operation, respectively. It is assumed that $n \times 1$ random column vector of errors \mathbf{e} is such that $E(\mathbf{e}) = \mathbf{0}_n$ and $E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{G}$, where $\mathbf{0}_n$ is an $n \times 1$ column vector of zeros, \mathbf{G} is an $n \times n$ positive definite diagonal matrix of known elements, $E(\cdot)$ stands for the expectation of (\cdot) and $(\cdot)'$ is used for the transpose of (\cdot) . For the estimation of unknown weights of objects we use the weighed least squares method and we get

$$\hat{\mathbf{w}} = \left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y}$$

and the dispersion matrix of $\hat{\mathbf{w}}$ is

$$V(\hat{\mathbf{w}}) = \sigma^2 \left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1}$$

provided \mathbf{X} is full column rank, $(r(\mathbf{X}) = p)$, where \mathbf{w} and \mathbf{y} are column vectors of unknown weights of p objects and of the recorded results in n weighings, respectively. The problem connected with the optimality of chemical balance weighing design is the choosing of a design matrix \mathbf{X} which minimizes $\phi\left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)$ for some real-valued function ϕ . ϕ is called an optimality criterion. In this paper we consider the optimality criterion as minimum variance for each of the estimated weights.

2 Some results on variance limit of estimated weights

We assume that matrix \mathbf{G} is given in the form

$$\mathbf{G} = \begin{bmatrix} \frac{1}{a}\mathbf{I}_{n_1} & \mathbf{0}_{n_1}\mathbf{0}'_{n_2} \\ \mathbf{0}_{n_2}\mathbf{0}'_{n_1} & \mathbf{I}_{n_2} \end{bmatrix}, \quad (1)$$

where $n = n_1 + n_2$, $a > 0$, and \mathbf{I}_{n_h} is the $n_h \times n_h$ identity matrix, $h = 1, 2$. This structure of the dispersion matrix of errors may be useful in the following situation. Suppose that are two kinds of chemical balances of different precision. Let n_1 and n_2 be the numbers of times in which the respectively balances are used. Suppose further that the matrix \mathbf{X} is partitioned correspondingly to the matrix \mathbf{G} , i.e.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}. \quad (2)$$

Ceranka and Graczyk (2004) showed that the minimum attainable variance for each of the estimated weights for a chemical balance weighing design with the design matrix \mathbf{X} given by (2) and the dispersion matrix of errors $\sigma^2\mathbf{G}$, where \mathbf{G} is given in (1), is

$$V(\hat{w}_j) \geq \frac{\sigma^2}{am_1 + m_2}, \quad j = 1, 2, \dots, p \quad (3)$$

where m_1 and m_2 is the number of elements equal to -1 and 1 in the j th column of the matrix \mathbf{X}_1 and \mathbf{X}_2 , respectively.

Definition 2.1 Any nonsingular chemical balance weighing design with the design matrix \mathbf{X} given in (2) and with the dispersion matrix $\sigma^2\mathbf{G}$, where \mathbf{G} is given by (1), is called optimal for the estimation of individual weights if in (3) the equality is fulfilled for each j , $j = 1, 2, \dots, p$.

Theorem 2.1 Any nonsingular chemical balance weighing design with the design matrix \mathbf{X} given in (2) and with the dispersion matrix $\sigma^2\mathbf{G}$, where \mathbf{G} is given by (1), is optimal for the estimated individual weights if and only if

$$\mathbf{X}'\mathbf{G}^{-1}\mathbf{X} = (am_1 + m_2)\mathbf{I}_p.$$

3 Optimum chemical balance weighing design for $p+1$ objects

Let \mathbf{X} given in (2) be the $n \times p$ matrix of the chemical balance weighing design. Based on this matrix we want to construct matrix \mathbf{T} of the chemical balance weighing design for $p+1$ objects in the form

$$\mathbf{T} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{n_1} \\ \mathbf{X}_2 & \mathbf{0}_{n_2} \end{bmatrix}, \quad (4)$$

where $\mathbf{1}_{n_1}$ is the $n_1 \times 1$ vector of units.

Theorem 3.1 If \mathbf{X} given in (2) is the $n \times p$ matrix of the chemical balance weighing design with the dispersion matrix $\sigma^2\mathbf{G}$, where \mathbf{G} is given by (1), then the \mathbf{T} given (4)

is the $n \times (p + 1)$ matrix of the optimum chemical balance weighing design with the same dispersion matrix $\sigma^2\mathbf{G}$ if and only if

$$\mathbf{X}'_1\mathbf{X}_1 + \mathbf{X}'_2\mathbf{X}_2 = an_1\mathbf{I}_p \quad (5)$$

and

$$\mathbf{X}'_1\mathbf{1}_{n_1} = \mathbf{0}_p. \quad (6)$$

4 Construction of the design matrix

Let \mathbf{N}_h^* be the incidence matrix of a balanced bipartite weighing design with the parameters $v, b_h, r_h, k_{1h}, k_{2h}, \lambda_{1h}, \lambda_{2h}, h = 1, 2$, (see Huang, 1976 and Swamy, 1982). From the matrix \mathbf{N}_h^* we construct the matrix \mathbf{N}_h by replacing k_{1h} elements equal to 1, which corresponds to the elements belonging to the first subblock by elements equal to -1 . Thus each column of the matrix \mathbf{N}_h will contain k_{1h} elements equal to -1 , k_{2h} elements equal to 1 and $v - k_{1h} - k_{2h}$ elements equal to 0. , Now we define the matrices \mathbf{X}_1 and \mathbf{X}_2 of the chemical balance weighing designs in the form

$$\mathbf{X}_1 = \begin{bmatrix} \mathbf{N}'_1 \\ -\mathbf{N}'_1 \end{bmatrix}, \quad (7)$$

$$\mathbf{X}_2 = \mathbf{N}'_2. \quad (8)$$

Then

$$\mathbf{T} = \begin{bmatrix} \mathbf{N}'_1 & \mathbf{1}_{b_1} \\ -\mathbf{N}'_1 & \mathbf{1}_{b_1} \\ \mathbf{N}'_2 & \mathbf{0}_{b_2} \end{bmatrix}. \quad (9)$$

In this design $n_1 = 2b_1, n_2 = b_2$, each of the v first columns contains $2r_{11} + r_{12}$ elements equal to -1 , $2r_{21} + r_{22}$ elements equal to 1 and $2b_1 + b_2 - 2r_1 - r_2$ elements equal to 0, $(v + 1)th$ column contains $2b_1$ elements equal to 1 and b_2 elements equal to 0. Let the dispersion matrix of errors $\sigma^2\mathbf{G}$ be in the form

$$\mathbf{G} = \begin{bmatrix} \frac{1}{a}\mathbf{I}_{b_1} & \mathbf{0}_{b_1}\mathbf{0}'_{b_1} & \mathbf{0}_{b_1}\mathbf{0}'_{b_2} \\ \mathbf{0}_{b_1}\mathbf{0}'_{b_1} & \frac{1}{a}\mathbf{I}_{b_1} & \mathbf{0}_{b_1}\mathbf{0}'_{b_2} \\ \mathbf{0}_{b_2}\mathbf{0}'_{b_1} & \mathbf{0}_{b_2}\mathbf{0}'_{b_1} & \mathbf{I}_{b_2} \end{bmatrix}. \quad (10)$$

Theorem 4.1 Any nonsingular chemical balance weighing design with the design matrix \mathbf{T} in the form (9) and with the dispersion matrix of errors $\sigma^2\mathbf{G}$, where \mathbf{G} is of the form (10), is optimal if and only if

$$2a(\lambda_{21} - \lambda_{11}) + (\lambda_{22} - \lambda_{12}) = 0 \quad (11)$$

and

$$2a(b_1 - r_1) - r_2 = 0. \quad (12)$$

Theorem 4.2 For a given $a = \frac{uc^2}{2s(v-c^2)}$ the existence of the balanced bipartite weighing design with the parameters $v, b_1 = \frac{2sv(v-1)}{c^2(c^2-1)}, r_1 = \frac{2s(v-1)}{c^2-1}, k_{11} = \frac{c(c-1)}{2}, k_{21} =$

$\frac{c(c+1)}{2}$, $\lambda_{11} = s$, $\lambda_{21} = s$ and v , $b_2 = \frac{2uv(v-1)}{c^2(c^2-1)}$, $r_2 = \frac{2u(v-1)}{c^2-1}$, $k_{12} = \frac{c(c-1)}{2}$, $k_{22} = \frac{c(c+1)}{2}$, $\lambda_{11} = u$, $\lambda_{21} = u$, $c = 2, 3, \dots$, $s, u = 1, 2, \dots$, $v > c^2$ implies the existence of the optimum chemical balance weighing design with the design matrix \mathbf{T} in the form (9) and with the dispersion matrix of errors $\sigma^2\mathbf{G}$, where \mathbf{G} is of the form (10).

Theorem 4.3 For a given $a = \frac{1}{2}$ the balanced bipartite weighing designs with the parameters

- (i) $v = 13$, $b_1 = 78$, $r_1 = 36$, $k_{11} = 2$, $k_{21} = 4$, $\lambda_{11} = 8$, $\lambda_{21} = 7$ and $v = 13$, $b_2 = 78$, $r_2 = 42$, $k_{12} = 2$, $k_{22} = 5$, $\lambda_{11} = 10$, $\lambda_{21} = 11$
- (ii) $v = 13$, $b_1 = 78$, $r_1 = 42$, $k_{11} = 2$, $k_{21} = 5$, $\lambda_{11} = 10$, $\lambda_{21} = 11$ and $v = 13$, $b_2 = 78$, $r_2 = 36$, $k_{12} = 2$, $k_{22} = 4$, $\lambda_{11} = 8$, $\lambda_{21} = 7$
- (iii) $v = 17$, $b_1 = 68$, $r_1 = 20$, $k_{11} = 1$, $k_{21} = 4$, $\lambda_{11} = 2$, $\lambda_{21} = 3$ and $v = 17$, $b_2 = 136$, $r_2 = 48$, $k_{12} = 2$, $k_{22} = 4$, $\lambda_{11} = 8$, $\lambda_{21} = 7$
- (iv) $v = 21$, $b_1 = 42$, $r_1 = 12$, $k_{11} = 1$, $k_{21} = 5$, $\lambda_{11} = 1$, $\lambda_{21} = 2$ and $v = 21$, $b_2 = 210$, $r_2 = 30$, $k_{12} = 1$, $k_{22} = 2$, $\lambda_{11} = 2$, $\lambda_{21} = 1$

give the optimum chemical balance weighing design with the design matrix \mathbf{T} in the form (9) and with the dispersion matrix of errors $\sigma^2\mathbf{G}$, where \mathbf{G} is of the form (10).

References

- Ceranka, B. and Graczyk, M. (2004). Optimum chemical balance weighing designs with diagonal variance-covariance matrix of errors. *Discussiones Mathematicae - Probability and Statistics* **24**.
- Huang, Ch. (1976). Balanced bipartite block designs. *Journal of Combinatorial Theory (A)* **21**, 20-34.
- Swamy, M.N. (1982). Use of balanced bipartite weighing designs as chemical designs. *Comm. Stat. Theory Methods* **11**, 769-785.

Small Sample Properties of Maximum Likelihood Estimators for Type II Censored Data

S. J. Chua¹, H. K. Finselbach¹ and A. J. Watkins¹

¹ School of Business and Economics, University of Wales Swansea, Singleton Park, Swansea SA2 8PP, United Kingdom, a.watkins@swansea.ac.uk

Abstract: This paper discusses the use of relative likelihood function and related contour plots to obtain confidence regions of parameters in relatively small samples drawn from a Weibull distribution under Type II censoring. In particular, the effect of the amount of censoring on the contours is illustrated. Examples are given using published data and simulation experiments.

1 Introduction

The two-parameter Weibull probability density function is

$$f(x; \beta, \theta) = \frac{\beta}{\theta^\beta} x^{\beta-1} \exp \left\{ - \left(\frac{x}{\theta} \right)^\beta \right\}, \quad (1)$$

for $x \geq 0$, where β and θ are, respectively, shape and scale parameters. When $n (> 0)$ items are independently tested, and the experiment is terminated after some (pre-specified) number r ($1 \leq r \leq n$) of failures, the data available for analysis is said to be Type II censored, and comprises the r order statistics $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{r:n}$, and $n - r$ lifetimes censored at $X_{r:n}$. The distinction between Type II censoring and complete sampling decreases as $r \rightarrow n$, and vanishes when $r = n$. We can now write down the corresponding likelihood $L(\beta, \theta)$, and maximum likelihood estimation for the parameters in (1) for both complete and censored samples is widely discussed in the literature; see, for instance, Lawless (1982) and Cohen (1991). For Type II censoring, we denote the maximum likelihood estimator by $(\hat{\beta}_r, \hat{\theta}_r)$, and use the popular ball-bearings data set, found, *inter alia*, in Kalbfleisch (1980), to illustrate this experimental set-up. Table 1 summarises the estimates calculated for various r , and we note that $(\hat{\beta}_r, \hat{\theta}_r)$ converge to their complete counterparts as $r \rightarrow n$.

Furthermore, it is known that $(\hat{\beta}_r, \hat{\theta}_r)$ is asymptotically Normally distributed with mean (β, θ) and covariance matrix equal to the inverse of the expected Fisher information matrix. This large sample result is often used in inference from small to moderate samples, despite the drawback that it is not always accurate with such sample sizes. As an alternative, we can, however, employ the relative likelihood function of (β, θ) , defined as

$$R(\beta, \theta) = \frac{L(\beta, \theta)}{L(\hat{\beta}_r, \hat{\theta}_r)}, \quad (2)$$

so that $0 < R(\beta, \theta) \leq 1$ for all (β, θ) . The relative likelihood thus ranks possible parameter pairs according to their consistency with the observed data, and, as Kalbfleisch

TABLE 1. Weibull MLEs calculated at varying r using the ball-bearings data.

r	11	14	17	20	23
$\hat{\beta}_r$	3.715	2.995	2.293	2.354	2.102
$\hat{\theta}_r$	62.943	70.483	79.438	78.967	81.878

(1980) has discussed, contour plots of $R(\beta, \theta)$ may be used to obtain confidence regions for a sample. In this paper, we consider two issues emerging from the above. In Section 2 we study the extent to which asymptotic Normality of the maximum likelihood estimator applies in finite samples based on Type II censored data from (1); then, in Section 3, we consider the use of (2) as a method for obtaining confidence regions of the sampling distribution of maximum likelihood estimators in small samples of varying size. We are also interested in the effects of varying r on the convergence to asymptotic Normality, and on the shape and size of contours of $R(\beta, \theta)$.

2 Asymptotic Normality of Estimators under Type II Censoring

2.1 Tests of Normality

With two parameters, it is usual to start with univariate tests for marginal Normality, since detection of one non-Normal marginal implies that the joint distribution is non-Normal. As the literature features tests of Normality based on measures of skewness (b_1) and kurtosis (b_2), we first consider some measures of the Normality (or otherwise) of the distributions of $\hat{\beta}_r$ and $\hat{\theta}_r$, based on samples from those distributions obtained via simulation experiments. In particular, following D'Agostino & Pearson (1973), we define

$$K^2 = \left\{ Z \left(\sqrt{b_1} \right) \right\}^2 + \left\{ Z \left(b_2 \right) \right\}^2,$$

where $Z(\sqrt{b_1})$ and $Z(b_2)$ are Normalised measures of skewness and kurtosis. Then, under the hypothesis that the marginal distribution of a maximum likelihood estimator is Normal, we have $K^2 \sim \chi_2^2$, so that we can assess the marginal Normality of $\hat{\beta}_r$ and $\hat{\theta}_r$ using the critical value $-2 \ln(p)$ for an upper tail probability of p . For tests of multivariate Normality, Mardia & Foster (1983) have proposed similar statistics, including

$$S_W^2 = \left\{ W \left(b_{1,2} \right) \right\}^2 + \left\{ W \left(b_{2,2} \right) \right\}^2,$$

in which $W(b_{1,2}), W(b_{2,2})$ are standardised multivariate measures of skewness and kurtosis, defined at equations (1.1) and (1.2) therein. Again, under the hypothesis that the joint distribution of the estimators is multivariate Normal, we have $S_W^2 \sim \chi_2^2$, with a corresponding assessment of joint Normality of $(\hat{\beta}_r, \hat{\theta}_r)$.

TABLE 2. K^2 statistics for the marginal Normality for $\hat{\beta}_r$ (upper entries) and $\hat{\theta}_r$ (lower entries), for various r, n .

r	n					
	25	50	100	1000	2500	5000
0.2n	8509.099	3604.381	1263.878	137.534	40.057	21.146
	4910.147	3199.557	1460.428	94.061	48.985	34.868
0.4n	3613.723	1498.453	758.525	65.430	45.613	28.233
	545.873	260.101	113.544	14.879	10.705	0.074
0.6n	2655.990	975.793	345.441	26.975	10.778	13.627
	64.238	28.990	25.464	15.426	1.630	0.270
0.8n	1696.814	652.694	283.699	17.841	5.565	4.624
	18.321	26.121	15.004	9.016	0.015	1.305
1.0n	931.695	470.907	301.582	6.453	1.450	5.537
	19.684	30.530	19.649	9.840	0.630	2.084

TABLE 3. S_W^2 statistics for the joint Normality for $(\hat{\beta}_r, \hat{\theta}_r)$, for various r, n .

r	n				
	100	500	1000	2500	5000
0.2n	2030.456	277.930	140.305	61.032	32.885
0.4n	346.992	33.627	20.156	15.362	4.851
0.6n	103.648	12.108	7.507	0.047	0.466
0.8n	69.645	7.147	5.026	0.612	0.694
1.0n	101.442	5.345	1.195	0.109	1.528

2.2 Simulation Results

Here, we take $\beta = 2$ and $\theta = 100$, and, for each combination of r and n , replicate 10^4 sets of data; this yields 10^4 values from the sampling distribution of $(\hat{\beta}_r, \hat{\theta}_r)$, and Table 2 summarises the K^2 statistic for these 10^4 values. In general, and entirely as expected, we obtain smaller K^2 values with increasing n and r . Table 3 tabulates S_W^2 for testing multivariate Normality, and again, confirms the lack of Normality in small samples.

3 Relative Likelihood and Contour Plots

We use (2) to obtain, via an intuitive interpretation of $R(\beta, \theta)$, confidence regions for maximum likelihood estimates in small samples; if $R(\beta, \theta) \geq \gamma$, then the pair (β, θ) is said to have at least $100\gamma\%$ of the maximum consistency possible under the model. With two parameters, a contour map of $R(\beta, \theta)$ portrays this consistency over the

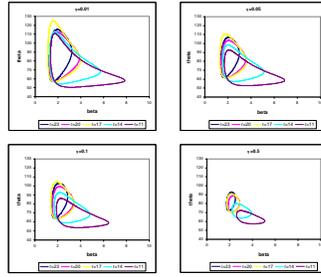


FIGURE 1. Four sets of relative likelihood contour plots for the ball-bearings data.

parameter space; for instance, points inside the 0.5-contour constitute fairly plausible parameter pairs, whereas values outside the 0.01-contour are very implausible. Watkins & Leech (1989) outline an automatic algorithm for drawing contours; we show here contour maps for the ball bearings data for censoring as in Table 1 above, with $\gamma = 0.01, 0.05, 0.1$ and 0.5 . Thus, the first case yields approximate 99% confidence regions for (β, θ) .

Figure 1 shows the effect of r on the contours. In general, for given γ , we see that the contours get smaller as r increase; we also note that contours extend over larger values in the β -axis, but over smaller values in the θ -axis. It is also clear that, as γ increases, contour areas drop dramatically and the contour shapes become more elliptical. Further, the shift in location, in line with the values in Table 1, is now more apparent.

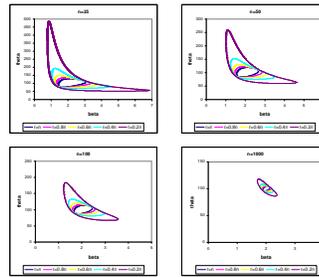
Although the above discussion is based on a single set of data, we can adapt the approach to provide confidence regions for the sampling distribution of $(\hat{\beta}_r, \hat{\theta}_r)$; this involves specifying - for any (β, θ) , sample size and censoring regime - an *idealised* sample, which can be produced by using the corresponding expected order statistics as data values, and then calculating and plotting the contours for that idealised sample. For illustration, we assume $\gamma = 0.05$, and show in Figure 2, the contour maps for some ideal samples for various r and n ; this yields the approximate 95% confidence regions for (β, θ) . To validate these contours, we plot the 10^4 simulated observations of $(\hat{\beta}_r, \hat{\theta}_r)$, and expect to find $95\% \times 10^4$ of $(\hat{\beta}_r, \hat{\theta}_r)$ within the corresponding contour area; results in Table 4 compare favourably with expected values.

4 Conclusion

We conclude that where asymptotic Normality assumption is implausible in finite samples, the relative likelihood and its contour plots provide an alternative to obtain approximate confidence regions of parameters in relatively small samples, subject to Type II censoring.

TABLE 4. The observed number of pairs of $(\hat{\beta}_r, \hat{\theta}_r)$ within the 0.05-contour for Weibull data generated with $\beta = 2, \theta = 100$.

r	n			
	25	50	100	1000
$0.2n$	7949	8826	9168	9450
$0.4n$	8798	9138	9344	9466
$0.6n$	9063	9267	9411	9492
$0.8n$	9193	9352	9444	9491
$1.0n$	9257	9415	9434	9508

FIGURE 2. Four sets of relative likelihood contour plots for Weibull data generated with $\beta = 2, \theta = 100$.

References

- Cohen, A.C. (1991). *Truncated and Censored Samples: Theory and Applications*. New York: Marcel Dekker.
- D'Agostino, R.B., and Pearson, E.S. (1973). Testing for departures from normality. I. Fuller empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika*, **60**, 613-622.
- Kalbfleisch, J.G. (1980). *Probability and Statistical Inference II*. New York: Springer-Verlag.
- Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley&Sons.
- Mardia, K.V., and Foster, K. (1983). Omnibus tests of multinormality based on skewness and kurtosis. *Communications in Statistics - Theory and Methods*, **12**, 207-221.
- Watkins, A.J., and Leech, D.J. (1989). Towards automatic assessment of reliability for data from a Weibull distribution. *Reliability Engineering and System Safety*, **24**, 343-350.

Modelling High Dimensional Sets of Binary Co-morbidities

Susana Conde¹ and Gilbert MacKenzie¹

¹ Centre of Biostatistics, Department of Mathematics & Statistics, The University of Limerick, Limerick, Ireland, susana.conde@ul.ie, gilbert.mackenzie@ul.ie

Abstract: The construction of classical co-morbidity indices is described. When the co-morbidities are binary we advocate the use of log-linear models which better capture the dependence structure in the data. We use R to implement new search strategies which enable us to analyze, sparse, high dimensional contingency tables rapidly and hence identify the best fitting models. We apply our new algorithms to a set of real medical data.

Keywords: Co-morbidity index; binary data; hierarchical log-linear model

1 Introduction

A co-morbidity is a coexisting (or additional) medical condition co-occurring with a primary disease of interest. In phase four studies, for example, when patients are on medication, the scientific interest is often in outcome - recurrence or death. Then, the burden of co-morbidity may be an important contributory determinant of outcome - one which is often overlooked in headline reporting attributing adverse events erroneously to the original treatment.

A number of solutions have been proposed in the medical literature. For example Charlson (1987) developed a Co-morbidity Index (a CCI) based on all patients admitted to the New York Hospital-Cornell Medical Center during a 1-month period in 1984. It comprises a linear combination of the co-morbidities with (age-adjusted) weights derived from a multivariate proportional hazards model of mortality. More recently Davis (1996) working with patients on dialysis derived another score based on clinical insight into the role of co-morbidity.

The construction of such indices (or so-called *risk-scores*) by divers methods is common in the medical literature and a fundamental concern is the optimality of such techniques. Below, we criticise classical methods of CCI construction and propose alternative methods of analyzing multivariate binary co-morbidities, especially when p is large.

2 Classical Indices

We define a co-morbidity index as $I = w'X$ where $w' := (w_1, w_2, \dots, w_p)$ is the weight vector and X is the corresponding co-morbidity vector. The expected value of I is $E(I) = w'E(X)$ where, for binary co-morbidities, $E(X) = [Pr(X_1 = 1), \dots, Pr(X_p =$

1)]'. The variance is $V(I) = w'\Sigma w$ with $\Sigma = V(X)$, where (r_{th}, s_{th}) element is $\sigma_{rs} = Pr(X_r = 1 \cap X_s = 1) - Pr(X_r = 1)Pr(X_s = 1)$. An interesting case is $w = 1$ ie, the unit vector, whence $I = w'X$ is a simple count of the co-morbidities, it being assumed clinically (and erroneously in many cases) that the risk of outcome is an increasing function of I . The assumption that $w_u > 0, \forall u, u = 1, \dots, p$ can also be rather dubious in practice. A key point is, that because the variables are binary and not MV Normal, their dependence structure is not summarized appropriately in the $p \times p$ variance covariance matrix, Σ .

3 Model Formulation

Given p binary co-morbidities we consider a p -dimensional contingency table with exactly $n = 2^p$ cells. Let n_j be the observed frequency (the count) in the j th. cell, $j = 1, \dots, n$, where the cells are ordered lexicographically in Fortran major order and we have the bijective mapping $j \mapsto (i_1, \dots, i_p)$ with each i_1, \dots, i_p taking the value 0 (absent) or 1 (present), MacKenzie & O'Flaherty (1982). Then our basic model is the usual log-linear model for contingency tables in which:

$$E(N_j) = \mu_j = \exp(a'_j\theta) \quad (1)$$

where N_j is the random variable denoting the number in the j th. cell, a'_j is the j th. row of the $(n \times n)$ saturated design matrix, A , and θ is the $(n \times 1)$ vector of unknown parameters measuring the influence of the constant, main effects and interactions on the response. From the last equation we have:

$$\log \mu_j = a'_j\theta = \alpha_0 + \alpha_{1i_1} + \alpha_{2i_2} + \dots + \alpha_{pi_p} + (\alpha_{1i_1} \alpha_{2i_2}) + (\alpha_{1i_1} \alpha_{3i_3}) + \dots + (\alpha_{1i_1} \alpha_{2i_2} \alpha_{3i_3}) + \dots + (\alpha_{1i_1} \alpha_{2i_2} \dots \alpha_{pi_p})$$

For inference we use the conditional Poisson model (Birch, 1963), so that $Pr(N_j = n_j) = \exp\{-\mu_j\} \mu_j^{n_j} / n_j!$, leading to:

$$\ell(\theta) \propto \sum_{j=1}^k [-\exp(a'_j\theta) + n_j a'_j\theta] \quad (2)$$

$$i_{r,s}(\theta) = \frac{\partial^2}{\partial \theta_r \partial \theta_s} \ell(\theta) = \sum_{j=1}^k a_{jr} \cdot a_{js} \exp(a'_j\theta_j) \quad (3)$$

where $1 \leq r, s \leq k$ and $k = n$ in the saturated case. We consider the class of hierarchical log-linear models (HLLMs) as a first step (Goodman, 1971).

4 Paradigms & Problems

Our adoption of this framework is predicated on the need to address some open problems in different, but related, modelling areas. For example, much original log-linear modelling was formulated in a model development environment dating back to the

TABLE 1. Tests of m - way effects are zero with 15 co-morbidities. The best fitting models must contain some 3-way interaction terms.

m	df	LR	P
1	15	689703.2000	0.0000
2	105	5998.8920	0.0000
3	445	480.3214	0.1198

1970's where $p = 10$ was considered very large. Then, today's data-mining paradigm was not envisaged and the original ideas have become ossified in legacy code in the major software packages. Accordingly, one objective of the current research is to relax these constraints by developing a new package in R. Yet another challenge is the ability to address the analysis of sparse, high-dimensional, contingency tables which might arise, for example, in thresholded micro-array data. The ability to search within these high dimensional spaces efficiently and so identify the model best supported by the data is a key objective of this research. Such searches may be facilitated by sacrificing high order interaction terms, replacing them by random effects terms instead, thereby extending the model class from a GLM to a GLMM.

5 Results

A dataset, comprising 48,158 subjects, half of whom had Chronic Obstructive Pulmonary Disease (COPD) and an equal number who were COPD-free, was analyzed. A total of $p = 15$ co-morbidities were recorded. These included the presence or absence of: Myocardial Infarction, Congestive Heart Failure, Peripheral Vascular Disease, Cerebrovascular Disease, Dementia, Rheumatologic Disease, Peptic Ulcer, Mild Liver Disease, Diabetes, Hemiplegia or Paraplegia, Lung Cancer, Other Cancers, Other Respiratory Disease, Nervous System Disorder and Psychiatric Disorder.

We implemented a backwards elimination search algorithm in R, using the Iterative Proportional Fitting algorithm (Haberman, 1972) to identify the best fitting class of models. This algorithm, which tests whether the m -way interactions are exactly zero, identified the class of HLLM models including as a maximum the 3-way interactions (Table 1). This set was also identified by another algorithm which tested whether the m -way or higher order effects were zero. One of the models involved in the comparison in the 3rd row of Table 1 contains exactly all possible 3-way interaction terms, namely 455. The best fitting model(s) in this class have yet to be identified. In total, there are $2^{455} - 1$ possible models containing at least one 3-way interaction and no higher order terms. The set of all possible 3-way interactions may now be viewed as defining another hierarchical (sub-)class of models, which can be searched (backwards or forwards) using a variant of our existing algorithms. In this way the best-fitting model(s) can be identified rapidly and compared with the results of conventional (e.g., best-subset) search strategies. At the time of writing, we are developing our new search strategies in R and will present these and other methodological innovations in the poster.

6 Discussion

We have outlined herein the construction of conventional co-morbidity indices and highlighted some limitations of interpretation, especially in relation to dependence structures. For binary co-morbidities we propose a log-linear modelling approach which more appropriately captures the dependence between the measured co-morbidities. The method facilitates implementation in R which is free of the many restrictions imposed by existing algorithms in mainstream software packages (eg, in SPSS $p=10$ maximally, or $p = 8$ when generating flat contingency tables). In the R environment we have been able to develop new search strategies which allow us to identify best-fitting models efficiently.

References

- Birch, M.W. (1963). Maximum Likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B* **25**, 220-233.
- Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R. (1987). , A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronical Disease* **40**(5), 373-383.
- Goodman, L.A.(1971). The Analysis of Multidimensional Contingency Tables: Step-wise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. *Technometrics* **13**(1), 33-61.
- Haberman, S.J. (1972). Algorithms AS 51: Log-linear Fit for Contingency Tables. *Applied Stats.* **21**, 2, 218-225.
- O'Flaherty, M. and MacKenzie, G. (1982). Direct Simulation of Nested Fortran DO-LOOPS. *Statistical Algorithms. Applied Statistics* **31**(1).

Bayesian Markov switching models for epidemiologic surveillance

David Conesa¹, Antonio López-Quílez¹ and Miguel A. Martínez-Beneito^{1,2}

¹ Department of Statistics and Operations Research, Universitat de València

² Conselleria de Sanitat, Generalitat Valenciana

Abstract: Early detection of disease outbreaks is one of the most challenging objectives of epidemiologic surveillance systems. In this work, a Markov switching model is introduced to determine the epidemic and non-epidemic periods in different kinds of surveillance data. In particular, the process of differenced incidence rates is modelled either with a first-order autoregressive process or with a Gaussian white noise process depending if the system is in epidemic or non epidemic phase. The transition between phases of the disease is modelled as a Markovian process. Bayesian inference is carried out on the former model to detect outbreaks at the very moment of its beginning. Moreover, the proposal provides at every moment the probability of being in epidemic state. Methodology is evaluated on influenza illness data obtained from the epidemiologic Sentinel Network of the Comunitat Valenciana, one of the 17 autonomous regions in Spain.

Keywords: Autoregressive modelling; Bayesian inference; Hidden Markov models; Outbreak detection.

1 Introduction

An important matter of concern when dealing with surveillance of infectious diseases is that of detecting the onset of an outbreak as soon as possible. This would imply the early beginning of timely interventions which could suppose a great impact, for example, on the number of lives saved. Unfortunately, surveillance systems have also recently gained an increasing importance due to the threat of emerging infections (like the outbreaks caused by the H5N1 bird-flu strain) and the increased potential for bioterrorist attacks. Our main goal in this paper is to introduce an approach to surveillance that does not fall in some of the disadvantages of previous works.

2 The model

As it can be seen at the top of Figure 1, typical surveillance data are a collection of time series formed by weekly incidence rates. Nevertheless, these series are usually not stationary which could lead to some difficulties in the data analysis. This suggests to work with the first order differenced series (formed by the differences of rates between weeks) displayed at the bottom of Figure 1.

Note that, in the differenced series, non-epidemic dynamics are characterized by small random changes around zero, while in epidemic dynamics changes are greater and related between them (positive and negative values are usually followed by positive

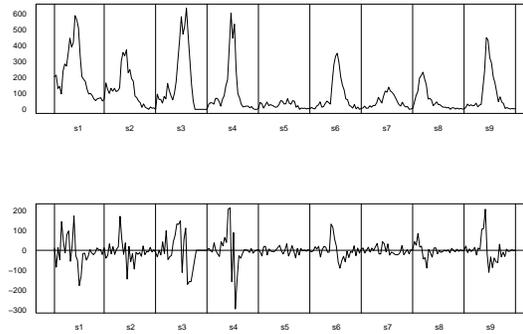


FIGURE 1. Typical surveillance data: weekly incidence rates (at the top) and increments of incidence rates between weeks during the seasons analyzed (at the bottom).

and negative values respectively). Then, our modelling is based on a segmentation of the series of differences into an epidemic and a non-epidemic phases using a two-stage Markov switching model.

Let $Y = \{Y_{i,j}, i = 1, \dots, 29; j = 1, \dots, 9\}$ denote the difference between the rates of weeks i and $i - 1$ in year j . Each $Y_{i,j}$ is associated with an unobserved random variable $Z_{i,j}$ that indicates in which phase the system is (1, epidemic; 0, non-epidemic). Moreover, the unobserved sequence of $Z_{i,j}$ follows a 2-state Markov chain of order 1 with transition probabilities:

$$P_{k,l} = P(Z_{i+1,j} = l | Z_{i,j} = k), k, l \in \{0, 1\}, i \in \{1, \dots, 29\}, j \in \{1, \dots, 9\}.$$

The conditional distribution of $Y_{i,j}$ is modelled either as a Gaussian white noise process or as an autoregressive process of order 1 depending if the system is in a non-epidemic or in an epidemic phase:

$$\begin{aligned} Y_{1,j} | (Z_{1,j} = 0) &\sim N(0, \sigma_{0,j}^2) \\ Y_{1,j} | (Z_{1,j} = 1) &\sim N(0, \sigma_{1,j}^2) \\ Y_{i,j} | (Z_{i,j} = 0) &\sim N(0, \sigma_{0,j}^2) \quad i = 2, \dots, 30, j = 1, \dots, 9, \\ Y_{i,j} | (Z_{i,j} = 1) &\sim N(\rho Y_{i-1,j}, \sigma_{1,j}^2) \quad i = 2, \dots, 30, j = 1, \dots, 9, \end{aligned} \quad (1)$$

where the first subindex of the variance $\sigma_{k,j}^2$ represents if the system is in the epidemic phase ($k = 1$) or not ($k = 0$).

Once the model is determined, the following step is to estimate its parameters. To do so, we take profit of the Bayesian paradigm, which require specification of the priors distributions of each parameter involved in the model. In this case, with the aim of expressing our initial vague knowledge about them, we consider the usual noninformative prior distributions for ρ , $P_{0,0}$ and $P_{1,1}$:

$$\rho \sim Unif(-1, 1)$$

$$\begin{aligned} P_{1,1} &\sim \text{Beta}(0.5, 0.5) \\ P_{0,0} &\sim \text{Beta}(0.5, 0.5) \end{aligned} \quad (2)$$

Moreover, taking into account that $\sigma_{0,j}^2$ should be lower than $\sigma_{1,j}^2$ as it responds to random variations instead of the effect of the epidemic, then we express our prior knowledge about $\{\sigma_{0,j}^2, \sigma_{1,j}^2; j = 1, \dots, 9\}$ via the following hierarchical structure:

$$\begin{aligned} \sigma_{0,j} &\sim \text{Unif}(\theta_{low}, \theta_{mid}) \\ \sigma_{1,j} &\sim \text{Unif}(\theta_{mid2}, \theta_{sup}) \\ \theta_{low} &\sim \text{Unif}(a, \theta_{mid1}) \\ \theta_{mid1} &\sim \text{Unif}(a, b) \\ \theta_{mid2} &\sim \text{Unif}(\theta_{mid1}, b) \\ \theta_{sup} &\sim \text{Unif}(\theta_{mid2}, b) \end{aligned} \quad (3)$$

where a and b are hyperparameters to be fixed. In order to perform inference, we have to resort to Markov chain Monte Carlo (MCMC) methods. In particular, we have worked with WinBUGS (Spiegelhalter et al., 1999).

3 Results

Figure 2 shows the results of our analysis in a particular data set based on the weekly influenza incidence rates observed in the Comunitat Valenciana in nine seasons. In particular, at the top of Figure 2 we present, for each week in the analysis, the posterior probability of being in an epidemic phase. These values correspond to the posterior mean of the state variable $Z_{i,j}$. Values exceeding 0.5 indicate that, in that week, we are observing a higher probability of being in epidemic phase than of being in non-epidemic, and so an alarm could be triggered if it is considered necessary. This information can be appreciated in more detail in the graph of the influenza incidence rates that we present at the bottom of Figure 2, in which we have plotted with black spots those weeks with a posterior probability of being in epidemic phase higher than 0.5.

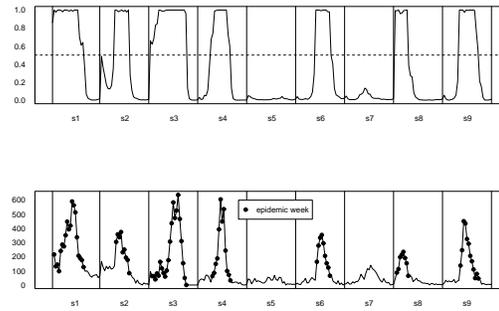


FIGURE 2. Posterior probability of being in the epidemic phase (at the top) and a representation of the influenza incidence rates per 100000 inhabitants in which the spots indicate those weeks where the posterior probability of being in an epidemic phase exceeds 0.5 (at the bottom).

Acknowledgments: Financial support from the Conselleria de Sanitat of the Generalitat Valenciana is gratefully acknowledged. Authors would like also to thank financial support from the Ministerio de Educación y Ciencia via the research grant MTM2004-03290 (jointly financed with European Regional Development Fund) and from the Generalitat Valenciana via research grants ACOMP06/205 and CS2005-049.

References

- LeStrat, Y., and Carrat, F. (1999). Monitoring epidemiological surveillance data using hidden Markov models. *Statistics in Medicine* **18**, 3463-3478.
- Stroup, D. F., Williamson, G. D., Herndon, J. L., and Karon, J. M. (1989). Detection of aberrations in the occurrence of notable diseases surveillance data. *Statistics in Medicine* **8**, 323-329.
- Spiegelhalter, D.J., Thomas, A., and Best, N.G. (1999). *Winbugs version 1.2 user manual*. MRC Biostatistics Unit.

Model Selection With Missing Covariates Under Ignorable Missingness

Fabrizio Consentino and Gerda Claeskens

¹ ORSTAT and University Center for Statistics, K.U. Leuven, Naamsestraat 69, B-3000, Leuven, Belgium

Abstract: Missing data represent a common problem in the analysis of data. The missingness can occur both for the response and the explanatory variables. This work is focused on the presence of missing covariates on some observations, when the missing data mechanism is ignorable, whereas the response is fully observed. Model selection using, for example, the Akaike Information Criterion (AIC) is not applicable anymore when missing data are present. We propose a variation of the AIC, through the utilization of the EM algorithm using the method of weights proposed by Ibrahim et al. (Biometrics, 1999). The new model selection method is investigated via a simulation study and real data analysis.

Keywords: AIC; Missing covariates; Ignorable missingness mechanism.

1 Introduction

We develop a model selection criterion similar to the AIC, which is usable for missing data situations in parametric regression models where the response is completely observed, though some of the covariates might be incomplete. To deal with the missing covariates, we follow the approach of Ibrahim et al. (1999). Their procedure obtains parameter estimates via weighting, using Gibbs sampling to draw from the distribution of the missing covariates given the observed variables in a Monte Carlo EM algorithm. In their method the missingness mechanism can be either ignorable or non-ignorable and it is valid for categorical and continuous variables, as well as for a mixture of those. Furthermore there are connections of our proposed method to the model selection criterion of Cavanaugh and Shumway (1998).

Our derivation follows that of the traditional AIC, by working with the Kullback-Leibler distance, considering the discrepancy between the true data generating mechanism and the likelihood model used in practice; secondly it uses the method of weights proposed by Ibrahim et al. (1999) and third it considers missing covariate information. The missing data mechanism is “Missing at Random” (MAR) and the ignorability assumption holds; therefore the missingness process has not to be modelled.

2 Description of the data example

We consider the Wisconsin Epidemiologic Study of Diabetic Retinopathy (Klein et al., 1984). This study provides information to study diabetic retinopathy as a function

of several measurements. The full set of data consists of patient information for 484 women and 512 men. The binary outcome variable $Y = 0$ indicates whether there is no or only mild nonproliferate retinopathy on both of the eyes. An outcome value $Y = 1$ is obtained when there is moderate to severe nonproliferate retinopathy, or proliferate retinopathy for at least one of the eyes. Other variables are: x_1 : the intraocular pressure in mmHg (maximum of the measurements for both eyes); x_2 : the age of the patient; x_3 : the duration of diabetes in years; x_4 : the percentage of glycosylated hemoglobin; x_5 : gender, using 1 for male and 0 for female, x_6 : indicator for presence of insuline protein; x_7 : area of residence (1=urban, 2=rural).

The response variable is completely observed, as are variables x_2, x_3, x_5, x_6 and x_7 . Variable x_1 contains missing entries for 51 out of the 996 cases, and variable x_4 contains 46 missing observations. For three cases both variables x_1 and x_4 are missing.

Since the response variable is binary, a logistic regression model is used for modeling the data and the full model takes the form

$$\text{logit}\{P(Y = 1)\} = \beta_0 + \beta_1 x_1 + \dots + \beta_7 x_7.$$

Our goal is to perform variable selection in the logistic regression model, including all cases. Because two covariates present missing values, we need to take this into account because we do not wish to remove the cases with missing values. Since the nature of both variables is continuous we modelled them using a bivariate normal distribution with mean $\boldsymbol{\mu}' = (\mu_{i1}, \mu_{i2})$, $\mu_{it} = \alpha_{t0} + \alpha_{t1}x_{i2} + \alpha_{t2}x_{i3} + \alpha_{t3}x_{i5} + \alpha_{t4}x_{i6} + \alpha_{t5}x_{i7}$, $t = 1, 2$ and Σ an arbitrary covariance matrix. The joint covariate distribution of x_1, x_4 is given by

$$f(x_{i1}, x_{i4} | \mathbf{v}_i, \boldsymbol{\alpha}) = f(x_{i1} | x_{i4}, \mathbf{v}_i, \alpha_1) f(x_{i4} | \mathbf{v}_i, \alpha_2)$$

with $\mathbf{v}_i = (x_2, x_3, x_5, x_6, x_7)$. While many model selection criteria could be considered, we here focus on Akaike's (1973) information criterion AIC. The criterion that we construct is directly comparable to the AIC in case no observations are missing, and easily follows by application of the EM algorithm.

3 The method of weights and the EM algorithm

Some of the explanatory variables X_{1i}, \dots, X_{pi} contain missing observations, whereas the response variable \mathbf{Y} is fully observed. For this reason the design matrix is partitioned in two parts, $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$, to distinguish between variables that are fully observed and variables with missing values.

The joint distribution of $(\mathbf{Y}_i, \mathbf{X}_i)$ is modelled by specifying the conditional distribution of $(\mathbf{Y}_i | \mathbf{X}_i)$ and the marginal distribution of (\mathbf{X}_i) . In this way the model is described as

$$f_{\theta} = f(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}) = f(\mathbf{Y} | \mathbf{X}; \boldsymbol{\beta}) f(\mathbf{X}; \boldsymbol{\alpha})$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$.

The estimation method is the EM algorithm; in our case the algorithm is used to estimate the missing part of the data. In particular we consider the method of weights as introduced by Ibrahim (1990). The EM algorithm proceeds in 2 steps, the E-step

where we estimate the expectation of $f_\theta = f(\mathbf{Y}, \mathbf{X}; \theta)$ and the M-step where the maximization is performed.

The expectation in the E-step is named the Q function, and we can write it as:

$$Q_i(\theta|\theta^{(k)}) = \int w_i \log\{f(y_i, x_i; \theta)\} dx_{\text{mis},i}$$

where $w_i = f(x_{\text{mis},i}|x_{\text{obs},i}, y_i; \theta^k)$.

In this way we can write the Q function as:

$$Q_i(\theta|\theta^{(k)}) = Q_i^{(1)}(\beta|\theta^{(k)}) + Q_i^{(2)}(\alpha|\theta^{(k)}).$$

Furthermore, due to the log concavity of the conditional distribution of \mathbf{Y} given \mathbf{X} within the exponential family, we use the Gibbs sampler along with the adaptive rejection algorithm of Gilks and Wild (1992) in order to obtain samples from $[x_{\text{mis},i}|x_{\text{obs},i}, y_i; \theta^k]$

4 The AIC for missing covariate information

In the derivation we do not assume that the true distribution of the data is known. Starting point is the Kullback-Leibler distance, with g the true pdf, defined as

$$KL(g, f_\theta) = E_g[\log\{g(\mathbf{Y}, \mathbf{X})/f(\mathbf{Y}, \mathbf{X}; \theta)\}].$$

The density f_θ can not be evaluated at (\mathbf{Y}, \mathbf{X}) , because we only observe $(\mathbf{Y}, \mathbf{X}_{\text{obs}})$. The method of weights assigns weights to the log likelihood function, and then integrates over the missing covariates. The “adjusted” likelihood function is hence defined as $\log \tilde{f}_\theta(\mathbf{y}, \mathbf{x}) = Q(\theta|\theta)$ where

$$Q(\theta_1|\theta_2) = \sum_{i=1}^n \int \log f(y_i, x_{\text{obs},i}, x_{\text{mis},i}; \theta_1) f(x_{\text{mis},i}|x_{\text{obs},i}, y_i, \theta_2) dx_{\text{mis},i}.$$

After some derivation we obtain the Takeuchi information criterion for missing covariate values as

$$\text{TIC} = 2 Q(\hat{\theta}|\hat{\theta}) - 2 \text{tr}\{\hat{J}(\hat{\theta})\hat{I}^{-1}(\hat{\theta})\}$$

where

$$\hat{I}(\hat{\theta}) = -\frac{1}{n} \ddot{Q}(\hat{\theta}|\hat{\theta}) \quad \text{and} \quad \hat{J}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^n \dot{Q}_i(\hat{\theta}|\hat{\theta}) \dot{Q}_i(\hat{\theta}|\hat{\theta})'.$$

The model with the largest value of TIC is chosen. This criterion consists of two parts. The first part is the “goodness-of-fit” term, whereas the second one is the “penalty” term, representing twice the effective number of parameters in the model. Following the Akaike’s information criterion version, a suitable criterion with missing covariates can be derived

$$\text{AIC} = 2 Q(\hat{\theta}|\hat{\theta}) - 2 \text{length}(\theta).$$

Although the ‘full’ Q function has two components, we want to stress that to allow comparison between the proposed criteria and the AIC for the complete case, the

former cannot be taken into account because the second component $Q^{(2)}$ is not comparable when the variables in the models are fully observed. So only restricting to the $Q^{(1)}$ allows a comparison of the criteria values both for the models containing missing variables and for those not containing such variables. For TIC, a similar reduced version (denoted TIC_1) is defined using $Q^{(1)}$ and contains as the penalty term the trace of the upper left submatrix of dimension $\text{length}(\boldsymbol{\beta}) \times \text{length}(\boldsymbol{\beta})$ of the matrix $\{\hat{J}(\hat{\boldsymbol{\theta}})\hat{I}^{-1}(\hat{\boldsymbol{\theta}})\}$. For the AIC_1 the penalty term is $\text{length}(\boldsymbol{\beta})$. Small corrections of the penalty term for a better approximation of the Kullback-Leibler distance for TIC_1 and AIC_1 can be derived.

5 Variable selection for the data example

To illustrate the validity of the proposed criteria, a logistic regression model is carried out, without removing the cases with missing values. An intercept is included in all of the models. Further, we performed an all subsets models search amongst the $2^7 = 128$ models. When applying this method to the data example where two continuous variables contain missing observations, we get the results presented in Table 1. Fitting the full model, including all the variables, it shows that the following variables are significant at the 5% level: age of the patient, the duration of diabetes, the percentage of glycosylated hemoglobin and gender, with the last three highly significant. The best model selected is the following

$$\text{logit}\{P(Y = 1)\} = -4.258 - 0.029x_2 + 0.126x_3 + 0.130x_4 + 0.563x_5.$$

In the table the TIC_1 , $\text{TIC}_{1,C}$, AIC_1 and $\text{AIC}_{1,C}$, and the complete cases only AIC_{cc} are displayed. While also the second best model is agreed upon by all criteria, the TIC_1 and $\text{TIC}_{1,C}$ differ in their model choice from model three onwards. The difference between the TIC_1 and AIC_1 selected models is due to the penalty term used for calculating the criteria, since the $Q^{(1)}$ function used is the same. Because of the large sample size, the corrected $\text{TIC}_{1,C}$ and the corrected $\text{AIC}_{1,C}$ do not give much different results as compared to the TIC_1 and AIC_1 , confirming that the correction in the penalty term is useful with small sample size. For TIC_1 and $\text{TIC}_{1,C}$ only the orders of models 4 and 5 are switched. The new criteria are immediately obtained from the EM algorithm and can be directly compared to the AIC and TIC in case no observations are missing. We wish to stress their ease of computation and interpretation. The results have confirmed the good performance of the criteria, in particular their efficiency to deal with the missingness. Ignoring the missing cases does not work well for model selection. Further research will extend these results to include missing response data and non-ignorable missingness schemes. For details of the derivation, a full data analysis and simulation study we refer to Claeskens and Consentino (2007).

TABLE 1. Results of variable selection for the WESDR data. The table displays the best six models selected by the different criteria, the value of the criterion for each of these models, together with the Q function and the penalty used for (a) TIC_1 , (b) $TIC_{1,C}$, (c) AIC_1 and $AIC_{1,C}$, and (d) the complete cases only AIC_{cc} . Since the models selected by AIC_1 and $AIC_{1,C}$ are the same, we only show the value of AIC_1 . All models contain an intercept.

Variables	Criterion	$Q^{(1)}$ function	penalty term	
(a) TIC_1				
x_2, x_3, x_4, x_5	-932.26	-460.79	5.34	
x_2, x_3, x_4, x_5, x_7	-932.89	-460.10	6.34	
x_2, x_3, x_4, x_5, x_6	-933.15	-460.23	6.34	
$x_2, x_3, x_4, x_5, x_6, x_7$	-933.89	-459.58	7.36	
x_1, x_2, x_3, x_4, x_5	-933.91	-460.55	6.40	
$x_1, x_2, x_3, x_4, x_5, x_7$	-934.69	-459.93	7.42	
(b) $TIC_{1,C}$				
x_2, x_3, x_4, x_5	-932.33	-460.79	5.37	
x_2, x_3, x_4, x_5, x_7	-932.99	-460.10	6.39	
x_2, x_3, x_4, x_5, x_6	-933.24	-460.23	6.39	
x_1, x_2, x_3, x_4, x_5	-934.00	-460.55	6.45	
$x_2, x_3, x_4, x_5, x_6, x_7$	-934.01	-459.58	7.42	
$x_1, x_2, x_3, x_4, x_5, x_7$	-934.82	-459.93	7.48	
(c) AIC_1 Penalty for				
			AIC_1	$AIC_{1,C}$
x_2, x_3, x_4, x_5	-931.59	-460.79	5	5.03
x_2, x_3, x_4, x_5, x_7	-932.20	-460.10	6	6.04
x_1, x_2, x_3, x_4, x_5	-933.10	-460.55	6	6.04
$x_1, x_2, x_3, x_4, x_5, x_7$	-933.86	-459.93	7	7.06
x_2, x_3, x_4, x_5, x_6	-934.47	-460.23	7	7.06
$x_2, x_3, x_4, x_5, x_6, x_7$	-935.17	-459.58	8	8.07
(d) AIC_{cc} Likelihood				
x_2, x_3, x_4, x_5	-849.43	-419.71	5	
x_2, x_3, x_4, x_5, x_7	-850.39	-419.20	6	
x_1, x_2, x_3, x_4, x_5	-850.43	-419.21	6	
$x_1, x_2, x_3, x_4, x_5, x_7$	-851.37	-418.69	7	
x_2, x_3, x_4, x_5, x_6	-852.03	-419.01	7	
$x_2, x_3, x_4, x_5, x_6, x_7$	-852.98	-418.49	8	

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, B. Petrov and F. Csáki (editors), 267–281, Akadémiai Kiadó, Budapest.
- Cavanaugh, J.E. and Shumway, R.H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference* **67**, 45-65.
- Claeskens, G. and Consentino, F. (2007). Variable selection with incomplete covariate data. Submitted.
- Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling *Applied Statistics* **41**, 337-348.
- Ibrahim, J.G. (1990). Incomplete data in generalized linear model. *Journal of the American Statistical Association* **85**, 765-769.
- Ibrahim, J.G., Chen, M.H. and Lipsitz, S.R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591-596.

Bootstrap prediction intervals: a case-study

Clara Cordeiro¹ and Maria Manuela Neves²

¹ Mathematics Department, FCT, University of Algarve, Portugal

² Mathematics Department, ISA, Technical University of Lisbon, Portugal

Abstract: In this study we propose to compare sieve bootstrap procedure for obtaining prediction intervals with an alternative approach based on choosing the best forecast model to adjust a time series. A real data example is used to illustrate the performance of the proposed procedures. Intensive computer work based on Monte Carlo experiments is carried out and R subroutines have been constructed.

Keywords: Autoregressive process; Forecast intervals; Holt-Winters method; Sieve bootstrap; Time series.

1 Introduction and motivation

A time series is a set of observations usually ordered in equally spaced intervals. One of the main goals in time series analysis is to forecast future values of the series. This interesting and ambitious task is based on recorded past observed values.

The first step in the empirical analysis of any time series is the description of the historic series. When a time series is plotted, common patterns are frequently found. These patterns may be explained by many possible cause-and-effect relationships. Common components are the trend, the seasonal effect, cyclic changes and randomness. The identification of these components is very important in the choice of a forecast model. There are several theoretical models that can be considered to fit a set of data. It is therefore extremely important to choose the model that better describes the behavior of the series in study and consequently obtain good estimates of the forecast intervals.

Classical procedures to obtain forecast intervals assume that the distribution of the error process is known. Some bootstrap approaches have been proposed to compute distribution free prediction intervals.

Alonso *et al.* (2002, 2003) used sieve bootstrap methodology to obtain forecast intervals. In section 2 a new approach for deriving those intervals is explained. In section 3 both methodologies are applied to a real air traffic data set.

2 Bootstrap forecast intervals

The bootstrap is a computer-intensive method introduced by Efron (1979) that presents solutions in situations where the traditional methods fail. Efron's bootstrap classical approach has revealed inefficient in the context of dependent data, such as in the

context of time series. Since then, a great development in the area of resampling methods for dependent data was observed (Lahiri, 2003). Several authors have proposed bootstrap methodologies for time series. Recently, Alonso *et al.* (2002, 2003) extended the sieve bootstrap approach proposed by Bühlmann (1997), in order to obtain prediction intervals, for a general class of linear models. Their approach uses $AR(\infty)$ sieve bootstrap procedure based on residual resampling from a sequence of approximating autoregressive models to the series with order $p=p(n)$ that increases as a function of the sample size n .

Here we proposed to choose first the best model that describes the data, and to proceed afterwards like in Alonso *et al.* (2002, 2003).

3 The case study

The real case study refers to monthly air traffic of Lisbon's FIR (Flight Information Regions) in the period 1985-2005 and we intend to get monthly forecast intervals for one year. In Figure 1 the data series is plotted in order to observe its behavior through

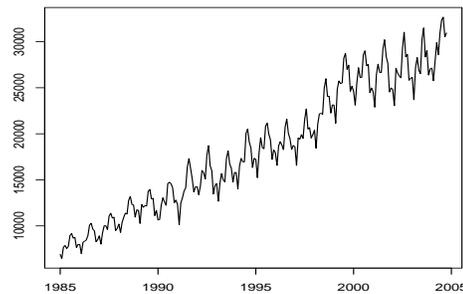


FIGURE 1. Series of monthly observed volume of air traffic.

time horizon and to identify its components. The descriptive analysis of the data suggests a model with trend and seasonality. Using the theory of forecasting models, Holt-Winters method was chosen. Those components were estimated and then they were removed from the time series. Tests of stationarity were performed.

Cordeiro and Neves (2006) compared several bootstrap methodologies in constructing forecast intervals. Here, forecast intervals using the sieve bootstrap (Alonso *et al.*, 2002, 2003) and the alternative approach based on a different initial fitting referred to above are constructed.

The results for these two methods were obtained using $B=1000$ replicates. In what concerns the new method, the Holt-Winters model was previously fitted. For the two methods, 95% monthly forecast intervals for 2006 were obtained and compared with real values. Results are presented in Table 1.

TABLE 1. Forecast intervals for 2006.

Months	Real Value	Sieve Method	New Method
January	30284	27829-30933	29351-31143
February	27908	25558-29763	27705-31679
March	31687	26484-31716	30738-33733
April	33336	25495-31508	31530-34085
May	31830	24650-31225	30547-35065
June	32212	23911-31079	30739-36046
July	35935	25408-33063	33310-37026
August	36254	25007-32732	34440-38330
September	33131	23371-31286	32358-39412
October	33721	23424-31449	32326-40494
November	31414	22059-30013	29685-40491
December	32093	22421-30791	30588-41470

In Table 1 when the Real Value is **inside** the interval bounds it is shown in **bold**. So we can see that the new method produces, in general, narrower intervals compared with sieve bootstrap, and the most important they contain the Real Value.

Interval coverage was calculated too. For each month, the B=1000 replicates were repeated 200 times. Table 2 shows the percentage of predicted intervals that contain the Real Value.

TABLE 2. Real Value Coverage (%).

Methods	Jan	Feb	Mar	Apr	May	Jun
Sieve	100	100	57,5	0	0	0
New	100	100	100	99,5	100	100
Methods	Jul	Aug	Sep	Oct	Nov	Dec
Sieve	0	0	0	0	0	0
New	66,5	93,5	100	100	99	100

In a first analysis, the choice of the model that better describes the data is very important. So, despite the nice properties that the autoregressive process presents, we propose to consider first the model that fits the data. From Table 1 and Table 2 we verify that this approach outperforms the sieve bootstrap for this case study.

4 Remarks

For the time series considered in our real data, trend and seasonality were observed. Holt-Winters method gave the best fitting. So it was used in the bootstrap approach to obtain forecast intervals and this computational alternative revealed a good performance. Accuracy measures like those given in Hyndman *et al.* (2006) will be presented in a future work. An extensive simulation study, for several models, is in progress and up to now promising results have been obtained.

Acknowledgments: Special thanks to the company **Portugal Navigation-NAV Portugal, E.P.E.**.

References

- Alonso, A.M., Peña, D. and Romo, J. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference* **100**, 1-11.
- Alonso, A.M., Peña, D. and Romo, J. (2003). On sieve bootstrap prediction intervals. *Statistical & Probability Letters* **65**, 13-20.
- Bühlmann, P. (1997). Sieve Bootstrap for Time series. *Bernoulli* **3**, 123-148.
- Bühlmann, P. (2002). Bootstrap for time series. *Statistical Science* **1**, 52-72.
- Cordeiro, C. and Neves, M. (2006). The Bootstrap methodology in time series forecasting. In: *Proceedings of CompStat2006, 17th Conference of IASC-ERS*, 1067-1073, Springer Verlag.
- Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *The Annals of Statistics* **7**, 1-26.
- Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting* **22**, 679-688.
- Lahiri, S.N. (2003). *Resampling Methods for Dependent Data*. Springer Verlag, Inc.

A model for a system of flow rivers with non-linear behavior

MariaElsa Correal¹

¹ Department of Industrial Engineering, Universidad de los Andes, Calle 19A No.1-37 Este, Bogotá COLOMBIA, South America, mcorreal@uniandes.edu.co

Abstract: This paper introduces a threshold dynamic factor model for the analysis of vector time series with non-linear behavior. The model is applied to a system of flow rivers in a region of South America. We show that those flow rivers behave differently if the Southern Oscillation Index, SOI, exceeds a specific threshold value. We propose a procedure for estimating common latent threshold factors that may explain the dynamic relationships within the group of variables. Parameter estimation is made combining the EM algorithm with a grid search procedure.

Keywords: Multivariate threshold model; dynamic factor model; EM algorithm; flow river model; SOI.

1 Introduction

The purpose of this paper is to model a system of flow rivers with common non-linear behavior of the threshold type. El Niño Phenomenon drastically alters the Colombian rainfall regime, and so it alters the hydrological system too. The Southern Oscillation Index, SOI, is a variable associated with the atmospheric component of the El Niño climatic event. We test the hypothesis that the flow rivers behaves differently if the SOI exceeds a threshold value, and we show that the system shares a common non-linear behavior. We propose a model and a procedure for estimating common latent threshold factors that may explain the dynamic relationships within the group of flow rivers.

2 The threshold dynamic factor model

The model we propose extends the dynamic factor model introduced by Peña & Box (1987) allowing the factors to follow a multivariate non-linear autoregressive threshold *TAR* model. The model represents the observable vector time series as a sum of two orthogonal latent components. The first one, common to the components of the time series vector, is described by a vector *TAR* process of small dimension. The second one is specific to each component of the vector.

Definition: Let \mathbf{Z}_t be a $k \times 1$ zero mean stationary vector time series, $\mathbf{Z}_t = (z_{1t}, z_{2t}, \dots, z_{kt})'$. We will say that \mathbf{Z}_t is represented by a *threshold dynamic*

factor model with 2 regimes of autoregressive order p and threshold variable w_t if

$$\mathbf{Z}_t = \Lambda \mathbf{f}_t + \mathbf{u}_t,$$

$$\mathbf{f}_t = \begin{cases} \sum_{i=1}^p \phi_i^{(1)} \mathbf{f}_{t-i} + \Upsilon^{(1)} \mathbf{a}_t & \text{if } w_{t-d} < \gamma \\ \sum_{i=1}^p \phi_i^{(2)} \mathbf{f}_{t-i} + \Upsilon^{(2)} \mathbf{a}_t & \text{if } w_{t-d} \geq \gamma \end{cases}$$

where w_t is an observable stationary univariate random variable, \mathbf{f}_t is a zero mean non-observable $r \times 1$ stationary random process, $r \leq k$, \mathbf{u}_t is a $k \times 1$ white noise with diagonal positive definite covariance matrix Σ_u , \mathbf{a}_t is an $r \times 1$ white noise with identity covariance matrix I_r , \mathbf{u}_t is independent of \mathbf{f}_{t-h} for $h \geq 0$, and \mathbf{a}_t is independent of \mathbf{f}_{t-h} for $h \geq 1$. It is assumed that $\{w_t\}$, $\{\mathbf{u}_t\}$ and $\{\mathbf{a}_t\}$ are pairwise independent. The parameters of the model are the so called threshold parameter, γ , the delay or threshold lag d , the factor loading matrix Λ of dimension $(k \times r)$, and the parameters of the factor process $\phi_i^{(j)}$, $\Upsilon^{(j)}$ $j = 1, 2$, of dimension $(r \times r)$. To ensure identifiability of the model, it is assumed for $j = 1, 2$ that $\Upsilon^{(j)}$ are diagonal positive definite and for Λ that $\text{rank}(\Lambda) = r$ and $\Lambda' \Lambda = I_r$, I_r being the identity matrix of order r .

3 Results

A threshold dynamic factorial model is estimated to a vector time series of five Colombian rivers. The historical data corresponds to flow monthly averages of the rivers Calima, Cauca, Grande, Ovejas and Prado, and cover a period of 36 years, from January 1955 to December 1990.

The procedure to model the vector time series consists of three steps. In the first one, we test whether there is evidence of nonlinearity in the river flows; in the second one we identify the number of common factors, if any; finally, in the third step we estimate the parameters of the model.

3.1 Linearity test

The linearity test proposed by Tsay (1989) is applied to each of the univariate time series. The test is based on arranged autoregression and orders the observations according to the size of the threshold variable w_{t-d} . The test was implemented with threshold variable $w_{t-d} = SOI_{t-d}$, for each d belonging to the set $\{1, 2, \dots, 6\}$. The conclusion is that for the Calima and Ovejas rivers there is strong evidence of nonlinearity (0.01% of significance), whereas for the other three the evidence is not so clear (12% of significance).

3.2 Identification of number of factors

The identification procedure is based on the covariance matrices of the observed processes. Let $\Gamma_Z(k) = E(Z_{t-k} Z_t')$ and $\Gamma_f(k) = E(f_{t-k} f_t')$ the covariance matrices of the observed and latent processes. If the model is correct, $\Gamma_Z(k) = \Lambda \Gamma_f(k) \Lambda'$, $k \geq 1$,

and therefore the observed covariance matrices for $k \geq 1$ have all the same eigenvectors. This property allows an explorative analysis of the number of factors in the model by computing the eigenstructure of the sample covariance matrices. This procedure leaves us with one or possibly two factors. In order to test this findings we apply two tests for the number of factors. The first one, whose details can be consulted in Peña & Poncela (2006), is based on the fact that the number of common factors, r , is equal to the number of nonzero canonical correlations between Z_{t-h} and Z_t for lags $h \neq 0$ and the second one, proposed by Hu & Chou (2004) is based on the number of canonical correlations between present and not present (past and future). Based on the tests results, we conclude that there are 2 factors.

3.3 Estimation procedure

We propose estimating the model by maximum likelihood under the normality assumption. The procedure we adopt combines the EM (Expectation - Maximization) algorithm of Dempster et. al. (1977) with a grid search method, maximizing the likelihood function L_Z^w in a sequential form, first over $\psi_1 = \{\Lambda, \Phi_{(1)}, \Phi_{(2)}, \Upsilon_1, \Upsilon_2, \Sigma_v\}$ and secondly over $\psi_2 = \{d, \gamma\}$. For d and γ fixed, the maximum over ψ_1 may be obtained by the EM algorithm as in the dynamic factor model. This is done by means of the Kalman filter and a smoothing algorithm. A simple grid search can be used to find the values \hat{d} and $\hat{\gamma}$ that maximizes the likelihood function. The grid search is taken over the pair of sets $\{1, 2, \dots, 12\}$ for the delay parameter and $\{-2.6, -2.5, \dots, 2.3\}$ for the threshold parameter. These values are chosen in such a way that there are enough observations to estimate the parameters in each regime. For each value d and γ on this grid, an EM algorithm is implemented and then the procedure must execute 60 times the EM algorithm.

3.4 Estimation results

The results for the estimated delay is $\hat{d} = 1$, the threshold parameter estimation is $\hat{\gamma} = -2.3$, and the estimated model:

$$\mathbf{Z}_t = \begin{bmatrix} 0.29 & 0.54 & 0.34 & 0.47 & 0.52 \\ 0.94 & -0.05 & -0.06 & -0.23 & -0.22 \end{bmatrix}' \begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix} + \mathbf{u}_t$$

$$\begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix} = \begin{bmatrix} 0.70 & 0.00 \\ 0.00 & 0.55 \end{bmatrix} \begin{bmatrix} f_{1,t-1} \\ f_{2,t-1} \end{bmatrix} + \Upsilon^{(1)} \mathbf{a}_t \text{ if } SOI_{t-1} < -2.3$$

$$\begin{bmatrix} f_{1t} \\ f_{2t} \end{bmatrix} = \begin{bmatrix} 0.78 & 0.00 \\ 0.00 & 0.67 \end{bmatrix} \begin{bmatrix} f_{1,t-1} \\ f_{2,t-1} \end{bmatrix} + \Upsilon^{(2)} \mathbf{a}_t \text{ if } SOI_{t-1} \geq -2.3$$

with $\hat{\Sigma}_u = \text{diag}(.016, .006, .032, .040, .201)$, $\Upsilon^{(1)} \Upsilon^{(1)} = \text{diag}(.30, .08)$, $\Upsilon^{(2)} \Upsilon^{(2)} = \text{diag}(.27, .04)$.

The model indicates that the AR(1) dependency of the factor model is stronger when $SOI_{t-1} \geq -2.3$. Since events of El Niño Phenomenon are presented along with negative values of the Southern Oscillation Index, regime 1 may be related to one of the phenomenon phases.

Acknowledgments: The results presented here are part of my Ph.D. dissertation, advised by professor Daniel Peña, Universidad Carlos III de Madrid, defended in Feb. 2006

References

- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society B* **39**, 1-38.
- Hu, Y.P. and Chou, R.J. (2004). On the Peña-Box model. *Journal of Time Series Analysis* **25**, 811-830.
- Peña, D. and Box, G. E.P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association* **82**, 836-843.
- Peña, D. and Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics* **119**, 291-321.
- Peña, D. and Poncela, P. (2006). Nonstationary dynamic factor models. *Journal of Statistical Planning and Inference* **136**, 1237-1257.
- Tsay, R.S. (1998). Testing and modelling multivariate threshold models. *Journal of the American Statistical Association* **93**, 1188-1202.

Parameterization and Penalties in Spline Models

M. J. Costa^{1 2} and J. E. H. Shaw¹

¹ Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

² Communicating author, m.j.costa@warwick.ac.uk

Abstract: We propose a parameterization for cubic splines, called value-first derivative parameterization, that facilitates both the use of unequally spaced knots and of quadratic penalty functionals. Simulation studies show that curves estimated under such parameterization tend to outperform, in terms of mean squared error, those estimated using the more standard approach based on B-spline basis functions and difference penalties.

Keywords: Bayesian inference; Penalized splines; Spline parameterization.

1 Introduction

Penalized spline models (Eilers and Marx, 1996) are a popular modelling tool in many statistical contexts where estimation of a smooth curve g is necessary. Typically, g is represented as an element in the span of a chosen cubic B-spline basis. However, B-splines and functionals of B-splines may be hard to handle analytically if unequally spaced knots are used.

We propose a local parameterization for cubic splines, termed value-first derivative parameterization (*VFDP*), that is easy to implement under any configuration of knots. We show how the *VFDP* allows us to break the effect of standard penalty functionals into interpretable quantities, an important feature when eliciting priors within a Bayesian inference framework. We compare, through simulation studies, the performance of our proposed methodology using the *VFDP* with that based on B-spline basis functions.

2 Penalized Likelihood Methods

Penalized spline methods estimate g by maximizing the penalized log-likelihood criterion:

$$l_p(g) = l(g) - P(g), \quad (1)$$

where l is the log-likelihood function of the model, and $P(g)$ is a penalty functional measuring the ‘roughness’ of g . If g is a cubic spline then usually $P(g) = \frac{\lambda}{2} \int g''^2$, with $\lambda \geq 0$ a smoothing parameter controlling the bias-variability trade-off implicit in (1); as λ tends to infinity the maximizer \hat{g} approaches a linear fit.

Alternatively, we can use a double penalty, replacing $P(g)$ in (1) by

$$P_d(g) = \frac{\lambda_1}{2} \int g'^2 + \frac{\lambda_2}{2} \int g''^2. \quad (2)$$

The term $\int g'^2$ in (2) provides a way of distinguishing, in terms of their roughness, curves that differ by a linear function. The penalty in (2) was also shown to yield estimates with good prediction performance (Aldrin, 2006). See Wood (2000) for the general case of multiple penalties.

3 The Value-First Derivative Parameterization

Let $g(x)$, $l \leq x \leq r$, be the cubic spline with knots $\{k_m\}_{m=1}^{\mathcal{K}}$. By definition g is a cubic polynomial within each knot interval. Such polynomial can be uniquely defined by four conditions over its coefficients. For each knot k_m we define:

$$a_m = g(k_m), \quad b_m = g'(k_m). \quad (3)$$

The parameters a_m , b_m , a_{m+1} and b_{m+1} define, according to (3), four equations over the coefficients of the cubic polynomial that agrees with g within $[k_m, k_{m+1})$. Take $u_m = x - k_m$ and $\Delta_m = k_{m+1} - k_m$, and consider the four polynomials:

$$\begin{aligned} \phi_{0m}(x) &= \frac{(u_m - \Delta_m)^2(2u_m + \Delta_m)}{\Delta_m^3}, & \phi_{1m}(x) &= \frac{u_m^2(3\Delta_m - 2u_m)}{\Delta_m^3}, \\ \psi_{0m}(x) &= \frac{u_m(u_m - \Delta_m)^2}{\Delta_m^2}, & \psi_{1m}(x) &= \frac{u_m^2(u_m - \Delta_m)}{\Delta_m^2}, \end{aligned}$$

It is straightforward to show that, for $x \in [k_m, k_{m+1})$,

$$g(x) = a_m\phi_{0m}(x) + b_m\psi_{0m}(x) + a_{m+1}\phi_{1m}(x) + b_{m+1}\psi_{1m}(x).$$

The complete set of parameters defining the cubic spline g in $[l, r]$ is the $2\mathcal{K}$ -dimensional vector $\boldsymbol{\alpha} = (a_1, b_1, \dots, a_{\mathcal{K}}, b_{\mathcal{K}})^T$. By defining a_m and b_m as in (3) we automatically impose that both g and g' are continuous functions in $[l, r]$. However, g'' is allowed to be discontinuous across the knots, bringing additional flexibility to the fitting process. The parameterization used in Green and Silverman (1994), based on g and g'' , constraints g to have continuous curvature throughout its domain.

4 Interpreting Penalty Functionals

The penalty functional $P(g) = \frac{\lambda}{2} \int g''^2$ can be written as:

$$P(g) = \frac{\lambda}{2} \sum_{m=1}^{\mathcal{K}-1} \int_{k_m}^{k_{m+1}} g''^2 = \sum_{m=1}^{\mathcal{K}-1} P_m(g), \quad \lambda \geq 0.$$

For each knot interval $[k_m, k_{m+1})$ we define the quantities $d_{m,m+1} = a_{m+1} - (a_m + \Delta_m b_m)$ and $d_{m+1,m} = a_m - (a_{m+1} - \Delta_m b_{m+1})$. The degree of g within $[k_m, k_{m+1})$ and the values of $d_{m,m+1}$ and $d_{m+1,m}$ are related by the following equivalence relationships:

$$\begin{aligned} d_{m,m+1} = d_{m+1,m} &\Leftrightarrow g \text{ is quadratic,} \\ d_{m,m+1} = d_{m+1,m} = 0 &\Leftrightarrow g \text{ is linear.} \end{aligned}$$

Shaw (1987) shows that we can write each of the local penalties $P_m(g)$ in terms of the quantities $d_{m,m+1}$ and $d_{m+1,m}$:

$$P_m(g) = \frac{\lambda}{2} \frac{3(d_{m,m+1} - d_{m+1,m})^2 + (d_{m,m+1} + d_{m+1,m})^2}{\Delta_m^3}. \tag{4}$$

So $P(g)$ penalizes generalizations of linear relationships, with the strength of the penalization increasing as these generalizations become more complex. Given α , evaluation of $P(g)$ is straightforward using (4), which is valid for any configuration of knots.

It turns out that we can also write the local penalty $P1_m(g) = \frac{\lambda}{2} \int_{k_m}^{k_{m+1}} g'^2$ in terms of $d_{m,m+1}$ and $d_{m+1,m}$:

$$P1_m(g) = \frac{\lambda}{2} \frac{(a_{m+1} - a_m)^2 + \frac{(d_{m,m+1} - d_{m+1,m})^2}{20} + \frac{(d_{m,m+1} + d_{m+1,m})^2}{12}}{\Delta_m}. \tag{5}$$

The term $(a_{m+1} - a_m)^2$ in (5) penalizes linear functions of x . Thus, in this case, $P1(g) = \sum_{m=1}^{K-1} P1_m(g)$ is increasingly penalizing curves that grow in complexity compared to a constant function of x .

5 Simulations

We illustrate the proposed methodology through a couple of simulation studies in different regression contexts. Since B-splines are the most popular choice in the smoothing methods literature we also compare the performance of the *VFDP* with the related approach based on B-splines and difference penalties.

5.1 Single vs Double Penalty Models

We start by comparing single and double penalty models in the following Gaussian regression setting:

$$y_i = g(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, 0.3^2), \quad i = 1, \dots, 200,$$

with $g(x) = \sin(2x) + 2 \exp(-16x^2)$, $-2 \leq x \leq 2$, a moderately smooth function. The curve g is approximated by a cubic spline with 20 knots which are taken to be equally spaced to facilitate the comparison with the B-splines methodology. Inference is performed within a Bayesian framework, making use of the natural link between quadratic penalties and Gaussian priors over α . We use Gibbs sampling to obtain posterior estimates of α and λ . Inference for double penalty models is carried out using empirical Bayes methods taking λ_1 and λ_2 in (2) as unknown constants. We draw 200 replicates. The quality of the fit and the prediction ability (for 50 ‘new’ observations) are measured by the empirical mean squared error:

$$\text{MSE}(\hat{g}) = \frac{1}{n} \sum_i (g(x_i) - \hat{g}(x_i))^2.$$

The results are shown in Figure 1. They suggest that, when a single penalty is used, curves estimated using the *VFDP* tend to capture the trend in the data more faithfully than the ones parameterized using B-spline functions. If focus is on prediction then double penalties seem to be preferable.

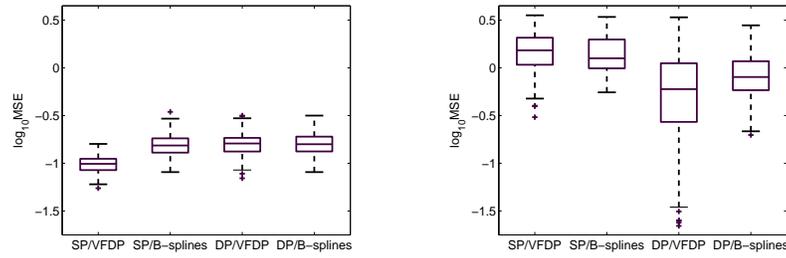


FIGURE 1. Simulation results for estimation (left) and prediction (right). ‘SP’ stands for single penalty models, ‘DP’ stands for double penalty models.

5.2 Extended Generalized Linear Models

To assess the performance of the methodology based on the *VFDP* outside Gaussian error structures we consider the logistic model:

$$y_i | x_i \sim \text{Bernoulli}(\mu_i), \quad \text{logit}\{\mu_i\} = g(x_i), \quad i = 1, \dots, 200,$$

with $g(x) = \frac{1}{0.72} \sin(6x - 3)$, $0 \leq x \leq 1$, a sinusoidal function. Again we take g to be well approximated by a cubic spline with 20 equally spaced knots. We will focus solely on single penalty models but double penalties are as easily incorporated in this setting as in the Gaussian regression case.

The Gibbs sampler cannot be used here because the full conditional for α is not of standard form. We turn to the Metropolis-Hastings sampler proposed by Brezger and Lang (2006) to obtain posterior estimates of α . A Gibbs step is then used to sample a new value of λ . The prediction ability of the model is also investigated in a manner similar to the one described for the Gaussian regression case in Section 5.1. The results for 100 replicates are presented in Figure 2. The *VFDP* and B-splines seem to perform

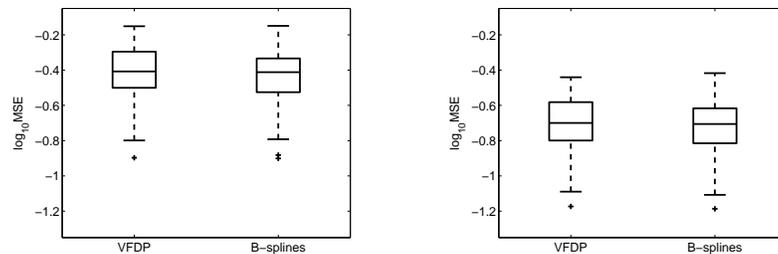


FIGURE 2. Simulation results for estimation (left) and prediction (right).

equally well in this case contrary to what happens in the Gaussian setting.

6 Application

We consider data on wages and union membership taken from a random sample of 534 individuals selected from the CPS (Current Population Survey). The data set is from 1985 and can be found in the Stalib site (URL lib.stat.cmu.edu/). The random variable `union` is the binary indicator of union membership (1 if member, 0 otherwise). `wages` (in \$/hour) is a continuous random variable. We study the regression model:

$$\text{union} \mid \text{wages} \sim \text{Bernoulli}(\mu), \quad \text{logit}\{\mu\} = g(\text{wages}).$$

The unknown smooth function g is approximated by a cubic spline with 10 equally spaced knots parameterized according to the *VFDP*. The results are presented in Figure 3.

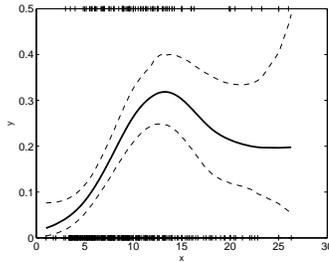


FIGURE 3. Logistic cubic spline fit to the `union` and `wages` data set (solid line) with 95% pointwise credible intervals (dashed lines). ‘x’ stands for `wages` (in \$/hour) and ‘y’ stands for $\text{Prob}(\text{union})$. The original data set is also included as pluses with values of 1 replaced by 0.5 for illustration purposes.

The estimated shape of the curve g suggests that the probability of union membership increases with wage up to an amount of \$15/hour, decreasing afterwards.

7 Conclusions

We propose a parameterization for cubic splines which is more intuitive and makes interpretation and implementation of standard quadratic penalty functionals straightforward. Our approach is competitive with the usual B-splines one both in terms of quality of fit and prediction error even in sparse data settings as is the logistic regression example of Section 5.2. We also compared double and single penalty models concluding that the former result in estimates with better predictive ability.

We are currently investigating the use of the *VFDP* in contexts such as GAMs and proportional hazards models.

Acknowledgments: M. J. Costa gratefully acknowledges financial support from Fundação para a Ciência e a Tecnologia through the grant SFRH/BD/16955/2004.

References

- Aldrin, M. (2006). Improved predictions penalizing both slope and curvature in additive models *Computational Statistics and Data Analysis* **50**, 267-284.
- Brezger, A., and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* **50**, 967-991.
- Eilers, P., and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89-121.
- Green, P., and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.
- Shaw, J. (1987). Numerical Bayesian analysis of some flexible regression models. *The Statistician* **36**, 147-153.
- Wood, M. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties *Journal of the Royal Statistical Society, series B* **62**, 413-428.

A ZIP model accounting for response bias in randomized response.

Maarten Cruyff¹, Ulf Böckenholt² and Peter van der Heijden¹

¹ Utrecht University, Netherlands

² McGill University, Canada

Abstract: Randomized response (RR) is an interview technique that ensures confidentiality when sensitive questions are asked. In RR the answer to a sensitive question is partly determined by a randomizing device. Respondent may nevertheless feel uncomfortable with the RR design and give the least sensitive response, irrespective of the outcome of the randomizing device. This kind of self-protective response behavior yields biased prevalence estimates. In this paper we present the zero-inflated Poisson model to account for self-protective response bias in the presence of an RR variable with multiple, ordered response categories. An example from a Dutch social security survey is provided.

Keywords: randomized response; self-protective response bias; zero-inflated Poisson.

1 Introduction

In many surveys sensitive questions need to be asked. Examples are questions about alcohol abuse or sexuality. Randomized response (RR) is an interview method especially designed for sensitive questions. In RR the answer depends partly on the true status of the respondent and partly on the outcome of a randomizing device. The outcome is known only to the respondent so that confidentiality is guaranteed.

It is however unrealistic to assume that all respondents comply with the RR design. Studies by Edgell, Himmelfarb and Duchan (1982), and Van der Heijden, Van Gils, Bouts and Hox (2000) show that respondents may exhibit self-protective response behavior by giving the least sensitive answer, regardless the outcome of the randomizing device. Recently models have been proposed that account for self-protective response bias in the presence of multiple binary RR variables. These models use the fact that self-protective response behavior results in an excess of observation of the response profile denying the sensitive characteristic on all questions. As an example, Böckenholt and Van der Heijden (2007) propose IRT models for RR profile data that includes a parameter accounting for the excess of observations of the response profile consisting of only *no* responses.

This paper proposes the zero-inflated Poisson (ZIP) model (Lambert, 1992) to account for self-protective response bias in the presence of a single RR variable with multiple ordered response categories. As an example we present the analysis of a variable from a Dutch social security survey. If this variable is analyzed with the general multinomial RR model we observe an unsatisfactory fit, with too many observations in the nonsensitive response category and too few observations in the more sensitive response

categories. This pattern of residuals clearly suggests the presence of self-protective response behavior. If we assume that the categories of the sensitive characteristic follow a Poisson distribution, then we can model the excess of observations in the zero count as zero-inflation due to self-protective response behavior.

The paper is structured as follows. In section 2 we present the question from the Dutch survey on social security fraud. Section 3 presents the general multinomial RR models, and derives the ZIP model for a RR variable with multiple ordered response categories under the assumptions of a Poisson distribution and the presence of self-protective response behavior. In Section 4 we compare the main results of the RR models presented in section 3. Section 5 concludes.

2 The social security survey

In 2004 the Dutch Department of Social Affairs conducts a nationwide survey to assess the level of regulatory noncompliance with the Social Security Law. In this survey 2.580 beneficiaries of social security benefits are asked the following question:

A On average, how much money a month have you earned in the past 12 months in addition to your social security benefits by working off the books ?

The question has six response categories, ranging from 0 euros to 250 euros or more. For this question according the forced-response (FR) design (Boruch, 1971) is used. In this particular application, the respondent tosses two dice and answers the question truthfully if the sum of the two dice equals 5, 6, 7, 8, 9 or 10. If the sum of the two dice is equal to 2, 3, 4, 11 or 12, the respondent tosses a single die, and answers the question by naming the number of eyes on the die. The observed response frequencies n_j^* are

TABLE 1. Response frequencies.

j	1	2	3	4	5	6
in EUR	0	1-50	51-75	76-100	101-250	250+
n_j^*	2014	245	108	74	72	67

The following eleven variables are included in this study as covariates. Beneficiaries of 3 different social security acts are distinguished by the dummy variables *AIA* (Assistance Insurance Act) and *DIA* (Disability Insurance Act), with reference category the Unemployment Insurance Act. The background variables are *gender*, *age* and (level of) *education*. The variables *costs compliance*, *benefits noncompliance*, *norm conformity* and *informal control* assess motives for noncompliance with the rules. The two variables *trust* and *understanding* assess the respondent's attitude towards the RR design.

3 The RR ZIP model

Let random variable U denote the true status on the sensitive characteristic and let U^* denote the response to the sensitive question, for $u^*, u \in \{1, \dots, 6\}$. If we define

π_{u^*} as the probability that U^* takes on the value u^* , and π_u as the probability that U takes on value u , and we assume a multinomial distribution for U , we obtain the general multinomial RR model

$$\pi_{u^*} = \sum_{u=1}^6 p_{u^*|u} \pi_u, \tag{1}$$

where $p_{u^*|u}$ denotes the conditional misclassification probability of observing response u^* given the true status u (Chaudhuri and Mukerjee, 1988). The conditional misclassification probabilities can be derived from the distribution of the dice. In the given FR design

$$p_{u^*|u} = \begin{cases} 19/24 & \text{if } u^* = u \\ 1/24 & \text{if } u^* \neq u. \end{cases}$$

We can also interpret the response categories to the sensitive question as counts of units of income. We can then define the variable $Y = U - 1$, so that for $Y = 0$ denote the event that "0 units of income" are earned, and $Y = 5$ that "5 units of income or more". If, for $y, y^* \in \{0, 1, \dots, 5\}$, we assume a censored Poisson distribution for Y we obtain the Poisson RR model

$$\pi_{y^*}^{RR} = \begin{cases} \sum_{y=0}^4 p_{y^*|y} \pi_y & \text{if } y < 5 \\ p_{y^*|5} \left(1 - \sum_{y=0}^4 p_{y^*|y} \pi_y\right) & \text{if } y = 5, \end{cases} \tag{2}$$

with $p_{y^*|y}$ equal to $p_{u^*|u}$, and $\pi_{y^*}^{RR}$ denoting the probability of observing the event $Y^* = y^*$ under the RR scheme.

We now introduce the zero-inflated Poisson model (Lambert, 1992) with the zero-inflation parameter θ to account for the effect of self-protective response behavior. Since self-protective response behavior results in an excess of zeros on the observed variable Y^* , we obtain the ZIP RR model

$$\pi_{y^*} = (1 - \theta) \pi_{y^*}^{RR} + I_{(y^*=0)} \theta, \tag{3}$$

where π_{y^*} denotes the probability of observing the event $Y^* = y^*$, and $I_{(y^*=0)}$ is an indicator variable taking on the value 1 if $Y^* = 0$, and 0 otherwise.

The covariate vectors \mathbf{x}_i explaining the RR generated count distribution and \mathbf{z}_i accounting for the self-protective response bias can be included in model (3) by respectively

$$\lambda_i = \exp(\mathbf{x}_i \beta) \text{ and } \theta_i = \frac{\exp(\mathbf{z}_i \gamma)}{\mathbf{1} + \exp(\mathbf{z}_i \gamma)}, \tag{4}$$

for $i \in \{1, 2, \dots, n\}$.

4 Main results

Table 2 compares the fit of the models presented in the previous section. Only the final model (4) includes covariates. The two variables *trust* and *understanding* compose

the covariate vector \mathbf{z} and are used to explain the zero-inflation. The remaining nine variables compose the covariate vector x and explain differences in the individual Poisson parameters.

TABLE 2. Model fitting

RR model	AIC	X^2
Model (1): multinomial model	4425	43.4
Model (2): Poisson model	4422	46.8
Model (3): ZIP model	4376	5.3
Model (4): ZIP model with covariates	4279	0.7

In terms of the Akaike Information Criterion (AIC) Model (2) assuming a Poisson distribution for the units of money earned by working off the books fits slightly better than the multinomial model (1). The X^2 statistics indicate however that the fit of both models is far from satisfactory. Substantial reductions in the AIC are accomplished by the inclusion of a zero-inflation parameter in Model (3), and the inclusion of the 9 β and 2 γ parameters in Model (4). In terms of X^2 both ZIP models seem to exhibit a satisfactory fit.

Model (4) yields an estimate of the zero-inflation parameter of $\hat{\theta} = 0.40$, and an estimated probability vector for the units of money earned by working off the books of $\hat{\pi}_y = (.792, .163, .034, .008, .002, .001)$, with Pearson residuals vector $(.1, -.6, .3, -.1, .5, .0)$. As a comparison, consider the corresponding vectors $\hat{\pi}_y = (.935, .065, 0, 0, 0, 0)$ and $(2.2, .8, .0, -3.3, -3.4, -3.9)$ of the multinomial RR model (1). The vector with the Pearson residuals of this model shows a systematic pattern which is in line with our assumption about self-protective response behavior: in the observed data the zero count is overrepresented and the higher counts are underrepresented.

TABLE 3. Parameter estimates β and γ of Model (4)

covariate	$\hat{\beta}$	se	covariate	$\hat{\gamma}$	se
constant	-2.08	0.45	constant	-0.74	0.32
<i>AIA</i>	-0.23	0.26	<i>trust</i>	-0.33	0.42
<i>DIA</i>	0.04	0.11	<i>understanding</i>	-0.75	0.44
<i>gender</i>	0.15	0.19			
<i>age</i>	0.02	0.02			
<i>education</i>	0.48	0.24			
<i>costs compliance</i>	1.03	0.33			
<i>benefits noncompliance</i>	-2.30	0.34			
<i>norm conformity</i>	0.32	0.43			
<i>informal control</i>	1.25	0.41			

Table 3 shows the estimates for the β and γ parameters. For the β parameters the esti-

mates of *cost of compliance*, *benefits of noncompliance*, *education* and *informal control* are significant, indicating that high costs of compliance, high benefits of noncompliance, higher education and less informal control are associated with higher earnings from working off the books. Judging by the standard errors the estimates of the γ parameters do not seem to be significant. However, the removal of the covariate *understanding* from the model results in a significant likelihood-ratio ($LR = 7.0$, $df=1$, $p < .001$). This results suggests that the zero-inflation in the responses to the sensitive question is negatively related to the extent to which the respondent understands when to answer *yes* and when *no*.

4.1 Conclusions

In this paper we present a ZIP model accounting for self-protective response bias in RR and we illustrate the model with an example from a Dutch social security survey. Obviously, the RR variable in our example is a pseudo count variable and therefore the Poisson assumption is questionable. However, we feel that results of the ZIP model are more realistic than those of the models that assume a multinomial distribution and ignore self-protective response bias.

Table 2 clearly shows that the introduction of the zero-inflation parameter θ substantially improves the fit. Since θ was introduced into the Poisson model to capture the effect of self-protective responses, it seems that the estimate $\hat{\theta} = 0.40$ indicates the proportion of respondents who exhibit self-protective response behavior. However, this interpretation disregards the possibility of true zero-inflation, for example due to respondents who are unable to perform labor or who are fundamentally opposed to working off the books. We therefore interpret the parameter θ in our model as a reflection of multiple sources of zero-inflation, of which self-protective response behavior is only one source. As a consequence, we take the estimate of 40% as overestimate of self-protective bias.

Lastly, an interesting result is the observed relation between the covariate *understanding* and self-protective response behavior. This result suggests that improving the instructions to the respondents may lead to more valid responses to the sensitive question, while improving the respondent's trust in the RR method does not seem to have a significant effect. It would be interesting to see if this result can be reproduced in other studies.

References

- Böckenholt, U., and P.G.M. Van der Heijden (2007). Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. *Psychometrika*. To appear.
- Boruch, R.F. (1971). Assuring confidentiality of responses in social research: a note on strategies, *The American Sociologist* **6**, 308-311.
- Chaudhuri, A., and R. Mukerjee (1988). *Randomized Response: Theory and Techniques*, New York: Marcel Dekker.

- Edgell, S.E., S. Himmelfarb and K.L. Duncan (1982). Validity of forced response in a randomized response model. *Sociological Methods and Research* **11**, 89-110.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14.
- Van der Heijden, P.G.M., G. Van Gils, J. Bouts and J. Hox (2000). A comparison of Randomized Response, Computer-Assisted Self-Interview and Face-To-Face Direct-Questioning. *Sociological Methods and Research* **28**, 505-537.

Modified Profile Likelihood for the Birnbaum-Saunders Distribution

Audrey H.M.A. Cysneiros¹, Francisco Cribari-Neto¹ and Carlos A. Gadelha de Araújo Jr.²

¹ Universidade Federal de Pernambuco - Brazil, audrey@de.ufpe.br

Universidade Federal de Pernambuco - Brazil, cribari@de.ufpe.br

² Ministério Público de Pernambuco - Brazil, caraujo@mp.pe.gov.br

Abstract: This paper obtains adjustments to the Birnbaum-Saunders profile likelihood function. The modified versions of the likelihood function were obtained for both the shape and scale parameters. Modified profile maximum likelihood estimators are obtained by maximizing the corresponding adjusted likelihood functions. We present numerical evidence on the finite-sample behavior of the different associated likelihood ratio tests. The results favor the tests we propose. An empirical application is briefly presented.

Keywords: Birnbaum-Saunders distribution; Likelihood ratio test; Maximum likelihood; Profile likelihood.

1 Introduction

Birnbaum and Saunders (1969a) proposed a two-parameter distribution that can be used to model failure time due to fatigue under cyclic loading when failure follows from the development and growth of a dominant crack. In this paper we shall use the results in Barndorff-Nielsen (1983), Severini (1999) and Cox and Reid (1987) to obtain adjustments to the Birnbaum-Saunders profile likelihood function. Numerical results and an application are presented.

2 Adjusted profile likelihood

Barndorff-Nielsen (1983) proposed an adjusted profile likelihood function which is invariant under reparameterizations of the form $(\tau, \phi) \rightarrow (\tau, \zeta(\tau, \phi))$, where τ is the vector of parameters of interest, ϕ is the vector of nuisance parameters and ζ is a function of τ and ϕ . His proposal follows from the p^* formula, which is an approximation to the conditional density of the maximum likelihood estimator given an ancillary statistic. The proposed adjusted profile likelihood function is $L_{BN}(\tau) = |\partial \hat{\phi}_\tau / \partial \hat{\phi}|^{-1} |j_{\phi\phi}(\tau, \hat{\phi}_\tau)|^{-1/2} L_p(\tau)$, where $j_{\phi\phi}(\tau, \phi) = -\partial^2 \ell / \partial \phi \partial \phi^\top$ is the observed information matrix for ϕ when τ is fixed, $L_p(\tau) = L(\tau, \hat{\phi}_\tau)$ is the profile likelihood function for τ , $L(\cdot)$ being the usual likelihood function and $\hat{\phi}_\tau$ is the restricted maximum likelihood estimator of ϕ given τ . An alternative expression for $L_{BN}(\tau)$ that does not involve $|\partial \hat{\phi}_\tau / \partial \hat{\phi}|$ is available. However, it involves a sample space derivative

and requires an ancillary statistic a such that $(\hat{\tau}, \hat{\phi}, a)$ is minimal sufficient statistic. Some approximations to the sample space derivative of the log-likelihood function have been proposed. An alternative approximation to Barndorff-Nielsen's (1983) adjusted profile likelihood function, say $\check{\ell}_{BN}(\tau)$, was proposed by Severini (1999): $\check{\ell}_{BN}(\tau) = \ell_p(\tau) + \frac{1}{2} \log |j_{\phi\phi}(\hat{\tau}, \hat{\phi}_\tau)| - \log |\check{I}_\phi(\tau, \hat{\phi}_\tau; \hat{\tau}, \hat{\phi})|$, where

$$\check{I}_\phi(\tau, \phi; \tau_0, \phi_0) = \sum_{j=1}^n \ell_\phi^{(j)}(\tau, \phi) \ell_\phi^{(j)}(\tau_0, \phi_0)^\top, \quad \ell_\theta^{(j)}(\theta) = (\ell_\tau^{(j)}(\theta), \ell_\phi^{(j)}(\theta)) \quad (1)$$

being the score function for the j th observation. This approximation can be easily computed and is particularly useful in situations where one is not able to compute expected values of log-likelihood derivatives. We shall now consider an alternative adjustment to the profile likelihood function. Suppose that τ and ϕ be orthogonal, i.e., that the elements of the score vector, $\partial\ell/\partial\tau$ and $\partial\ell/\partial\phi$, be uncorrelated. Cox and Reid (1987) proposed an adjustment that is an approximation to a conditional probability density function of the observations given the nuisance parameter maximum likelihood estimator and can be written as $L_{CR}(\tau) = |j_{\phi\phi}(\tau, \hat{\phi}_\tau)|^{-1/2} L_p(\tau)$. Taking logs we obtain $\ell_{CR}(\tau) = \ell_p(\tau) - \log |j_{\phi\phi}(\tau, \hat{\phi}_\tau)|/2$; the maximizer of $\ell_{CR}(\tau)$ shall be denoted as $\hat{\tau}_{CR}$. It is noteworthy that the score bias is of order $O(n^{-1})$, but the information bias remains $O(1)$. The derivation of $\ell_{CR}(\tau)$ requires that τ and ϕ be orthogonal. However, it is not always possible to find a parameterization that delivers orthogonality. Additionally, unlike $L_{BN}(\tau)$, their adjustment is not invariant under reparameterizations of the form $(\tau, \phi) \rightarrow (\tau, \zeta(\tau, \phi))$.

3 Birnbaum-Saunders adjusted profile likelihoods

At the outset, let α be the parameter of interest and β the nuisance parameter. Also, let $t = (t_1, \dots, t_n)^\top$ denote a random sample of size n from the Birnbaum-Saunders distribution. The profile log-likelihood function, $\ell_p(\alpha)$, is given by

$$-n \log(\alpha \hat{\beta}_\alpha) + \sum_{i=1}^n \log \left[\left(\frac{\hat{\beta}_\alpha}{t_i} \right)^{1/2} + \left(\frac{\hat{\beta}_\alpha}{t_i} \right)^{3/2} \right] - \frac{n}{2\alpha^2} \left(\frac{r}{\hat{\beta}_\alpha} + \frac{\hat{\beta}_\alpha}{s} - 2 \right),$$

where $r = \sum_{i=1}^n t_i/n$, $s = [\sum_{i=1}^n (nt_i)^{-1}]^{-1}$ and $\hat{\beta}_\alpha$, for fixed α is the restricted maximum likelihood estimator of β ; it does not have a closed-form expression, and thus it needs to be obtained using restricted nonlinear optimization methods. In what follows, we shall obtain the adjusted profile likelihoods described in Section 2. We shall omit the derivation details in the interest of space. Note that the interest and nuisance parameters are orthogonal. The observed information $j_{\beta\beta}(\alpha, \beta)$ evaluated at $(\alpha, \hat{\beta}_\alpha)$ can be written as $j_{\beta\beta}(\alpha, \hat{\beta}_\alpha) = -\frac{n}{\hat{\beta}_\alpha^2} + \frac{n}{2} \left(\frac{1}{\hat{\beta}_\alpha^2} + \frac{2K'(\hat{\beta}_\alpha)}{K^2(\hat{\beta}_\alpha)} \right) + \frac{n}{\alpha^2} \frac{r}{\hat{\beta}_\alpha^3}$ and $K(\beta) = \left[\frac{1}{n} \sum_{i=1}^n (\beta + t_i)^{-1} \right]^{-1}$. From (1), we obtain

$$\check{I}(\alpha, \hat{\beta}_\alpha; \hat{\alpha}, \hat{\beta}) = \frac{n}{\hat{\beta}^2} - \frac{1}{\hat{\beta}} \sum_{j=1}^n A_j - \frac{1}{4} \left[\sum_{j=1}^n A_j^2 + \frac{1}{\alpha^2 \hat{\alpha}^2} \sum_{j=1}^n B_j^2 \right]$$

$$- \frac{1}{4} \left[\left(\frac{1}{\alpha^2} + \frac{1}{\hat{\alpha}^2} \right) \left(\sum_{j=1}^n A_j B_j - \frac{2}{\hat{\beta}} \sum_{j=1}^n B_j \right) \right],$$

where $A_j = \left(\frac{t_j^{-1/2} \hat{\beta}^{-1/2} + 3\hat{\beta}^{1/2} t_j^{-3/2}}{t_j^{-1/2} \hat{\beta}^{1/2} + \hat{\beta}^{3/2} t_j^{-3/2}} \right)$ and $B_j = \left(\frac{t_j}{\hat{\beta}^2} - \frac{1}{t_j} \right)$. The likelihood ratio test statistics obtained from the adjusted profile log-likelihood functions given above for the test of $H_0 : \alpha = \alpha_0$ against $H_1 : \alpha \neq \alpha_0$ are given by $LR_{CR}(\alpha) = 2 \{ \ell_{CR}(\hat{\alpha}_{CR}) - \ell_{CR}(\alpha_0) \}$ and $LR_{BN}(\alpha) = 2 \{ \ell_{BN}(\hat{\alpha}_{BN}) - \ell_{BN}(\alpha_0) \}$, where $\hat{\alpha}_{CR}$ and $\hat{\alpha}_{BN}$ are the values of α that maximize $\ell_{CR}(\alpha)$ and $\ell_{BN}(\alpha)$, respectively. These test statistics are asymptotically distributed as χ_1^2 under the null hypothesis.

4 Numerical Evidence

In Table 1 we present the powers of the tests of the null hypothesis $H_0 : \alpha = \alpha_0$. The values of α used ranged from 0.12 to 0.28. Again, the tests were performed using size-corrected critical values (obtained from the size simulations) in order to force all tests to have the same size. The simulations were carried out using $n = 10$, $\alpha = 0.10$ and $\beta = 1.0$. (All entries are percentages.) We note that LR is slightly less powerful than LR_{BN} , which, in turn, is outperformed by LR_{CR} . For example, when $\alpha = 0.20$ and at the 5% nominal level, the nonnull rejection rates of these tests were equal to 76.65%, 82.76% and 82.83%, respectively.

5 Application

We shall now perform profile and adjusted profile likelihood inference using a real data set. We shall assume that observations are random draws from the Birnbaum-Saunders distribution. We consider the data provided by Birnbaum-Saunders (1969b) on the fatigue life of 6061-T6 aluminum coupons cut parallel to the direction of rolling and oscillated at 18 cycles per second (cps). The data set consists of 101 observations with maximum stress per cycle 31,000 psi. Let α be the parameter of interest. Suppose we are interested in testing $H_0 : \alpha = 0.15$ against $H_1 : \alpha \neq 0.15$. The test statistics based on $\ell_p(\alpha)$, $\ell_{CR}(\alpha)$ and $\ell_{BN}(\alpha)$ are, respectively, 3.5771, 3.8421 and 3.8351, with the following corresponding p -values: 0.05858, 0.04998 and 0.05019. Since the sample size is large (101 observations), the values of the three statistics are similar. However, the resulting inference is not the same at the 5% nominal level, since the test based on $\ell_{CR}(\alpha)$, unlike the other two tests, yields rejection of the null hypothesis.

TABLE 1. Nonnull rejection rates, inference on α .

α	5%			1%		
	LR	LR_{CR}	LR_{BN}	LR	LR_{CR}	LR_{BN}
0.12	8.33	12.82	12.72	2.53	4.42	4.40
0.14	24.26	33.25	33.18	11.96	18.00	17.92
0.16	45.28	55.35	55.28	28.68	37.81	37.72
0.18	62.45	70.99	70.83	46.38	55.34	55.26
0.20	76.65	82.83	82.76	63.91	71.50	71.43
0.22	85.54	89.78	89.72	76.08	81.84	81.79
0.24	91.76	94.16	94.14	85.28	89.49	89.44
0.26	94.62	96.24	96.24	89.99	92.81	92.80
0.28	96.60	97.78	97.77	93.47	95.46	95.45

Acknowledgments: Special Thanks to CAPES and CNPq, Brazil.

References

- Barndorff-Nielsen, O.E. (1983). On a formula to the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343-365.
- Birnbaum, Z.W., and Saunders, S.C. (1969a). A new family of life distribution. *Journal of Applied Probability* **6**, 319-327.
- Birnbaum, Z.W., and Saunders, S.C. (1969b). Estimation for a family of life distributions with applications to fatigue. *Journal of Applied Probability* **6**, 328-347.
- Cox, D.R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B* **49**, 1-39.
- Severini, T.A. (1999). An empirical adjustment to the likelihood ratio statistic. *Biometrika* **86**, 235-247.

On the Treatment of Missing Observations in Paired Comparisons Experiments

R. Dittrich¹, B. Francis², R. Hatzinger¹ and W. Katzenbeisser¹

¹ Vienna University of Economics, Department of Statistics and Mathematics, Augasse 2-6, A-1090 Vienna, Austria

² University of Lancaster, Lancaster, UK

Abstract: This paper proposes an extension to a log-linear model for the analysis of dependent paired comparison responses proposed by Dittrich, Hatzinger and Katzenbeisser (2002), which assumes completely observed response pattern vectors. While this model provides a useful method of modelling dependence, it fails to deal with missing data. In this paper we therefore present a log-linear model that can be used to model both incomplete and complete observed response patterns. The model is fitted to an augmented frequency table in which indicators correspond to whether or not a decision is observed or not.

Keywords: Bradley-Terry model; paired comparison data; multiple multinomial responses; log-linear model; nonresponse model

1 Introduction

The method of paired comparisons addresses the problem of determining the scale values of a set of J objects O_1, O_2, \dots, O_J on a preference continuum that is not directly observable. Paired comparisons are judgmental tasks that typically involve repeatedly exposing an individual to all different $\ell = \binom{J}{2}$ pairs of objects from this set and asking which of the pair is preferred.

One of the most prominent and well-known models that covers such situations is due to Bradley and Terry (1952). The (basic) Bradley-Terry model is defined by the equations

$$P\{Y_{ij} = 1|\pi_i, \pi_j\} = \frac{\pi_i}{\pi_i + \pi_j}, \quad P\{Y_{ij} = -1|\pi_i, \pi_j\} = \frac{\pi_j}{\pi_i + \pi_j}, \quad (1)$$

where $\{Y_{ij} = 1\}$ ($\{Y_{ij} = -1\}$) denotes the event that object i (j) is chosen in the comparison of objects i and j . The π 's are unknown non-negative parameters, the so called 'worth' parameters, describing the location of the objects on the preference scale which have to be estimated. The model has been extended by Dittrich et al (2002) who incorporated a set of dependence parameters to allow for lack of independence between responses.

In the tradition of Conaway (1992) the purpose of this paper is therefore to extend the log-linear model for the analysis of dependent paired comparisons to adjust for missing observations.

2 The probability model

2.1 Data structure

The paired comparison experiment can result in two cases. One (case A) is a complete response pattern \mathbf{y} with all $y_{ij} \in \{-1, 1\}$ and the other (case B) is an incomplete response pattern \mathbf{y} with $y_{ij} \in \{-1, 1, 0\}$ where $\{y_{ij} = 0\}$ denotes a missing response in the comparison of objects O_i and O_j . There are 2^ℓ complete response patterns which can be combined into the $(2^\ell \times \ell)$ complete response pattern matrix $\mathbf{Y}^{(A)}$. There are $\sum_{\nu=1}^{\ell-1} \binom{\ell}{\nu} 2^{\ell-\nu} = 3^\ell - 2^\ell - 1$ incomplete response patterns which can be collected into the incomplete response pattern matrix $\mathbf{Y}^{(B)}$. In total there are $3^\ell - 1$ response patterns where the response with only zeros is omitted. Both \mathbf{Y} -matrices can be stacked into the $(3^\ell - 1 \times \ell)$ response pattern matrix $\mathbf{Y} = (\mathbf{Y}^{(A)T}, \mathbf{Y}^{(B)T})^T$. $\mathbf{Y}^{(A)}$ can be seen as the design matrix of a 2^ℓ main effects only design in standard order. $\mathbf{Y}^{(B)}$ is the row-wise concatenation of matrices which are built up by all responses corresponding to a specific nonresponse pattern. Let us denote these submatrices by $\mathbf{Y}_{1;ij}^{(B)}, \mathbf{Y}_{2;ij,kl}^{(B)}, \dots$, where the first index is the number of missing comparisons and the second indices $(ij), (kl), \dots$ represent the missing comparison(s).

The nonresponse indicators can also be arranged in a $(3^\ell - 1 \times \ell)$ matrix \mathbf{R} with the property $r_{ij} = 1$ if the comparison between O_i and O_j is observed, and $r_{ij} = 0$ otherwise. The matrix \mathbf{R} can be represented by $\mathbf{R} = \mathbf{Y} \odot \mathbf{Y}$, the elementwise (Hadamard) product of \mathbf{Y} with itself.

We consider the set of all N respondents as being partitioned into disjoint subsets $N^{(A)}; N_1^{(B)}, \dots, N_{\ell-1}^{(B)}$, where $N^{(A)}$ and $N_k^{(B)}$ denotes also the number of respondents who responded to all ℓ paired comparisons or did not respond to k paired comparisons, respectively, and $N = N^{(A)} + N_1^{(B)} + \dots + N_{\ell-1}^{(B)}$ is the total number of respondents. Thus, e.g. $N_1^{(B)} = N_{1;12}^{(B)} + N_{1;13}^{(B)} + \dots$ is the sum of all those respondents with one given comparison missing.

Example. To illustrate let $J = 3$ and therefore $\ell = 3$. Thus we get the following response pattern matrices: $\mathbf{Y}^{(A)}$ for the complete response pattern and for the incomplete response patterns we obtain the following block matrices $\mathbf{Y}_1^{(B)}, \mathbf{Y}_2^{(B)}$.

$$\mathbf{Y}^{(A)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & -1 \end{pmatrix}, \mathbf{Y}_1^{(B)} = \begin{pmatrix} \mathbf{Y}_{1;23}^{(B)} \\ \mathbf{Y}_{1;13}^{(B)} \\ \mathbf{Y}_{1;12}^{(B)} \end{pmatrix} = \begin{pmatrix} \begin{matrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 0 \end{matrix} \\ \hline \begin{matrix} 1 & 0 & 1 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ -1 & 0 & -1 \end{matrix} \\ \hline \begin{matrix} 0 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & -1 & -1 \end{matrix} \end{pmatrix},$$

$$\mathbf{Y}_2^{(B)} = \begin{pmatrix} \mathbf{Y}_{2;13,23}^{(B)} \\ \mathbf{Y}_{2;12,23}^{(B)} \\ \mathbf{Y}_{2;12,13}^{(B)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & -1 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix}$$

2.2 The probability model

The basic building block is the following representation of the Bradley-Terry specification which is due to Sinclair (1982):

$$p_{ij} = P\{Y_{ij} = y_{ij} | \pi_i, \pi_j\} = \frac{\pi_i}{\pi_i + \pi_j} = \Delta_{ij} \left(\frac{\sqrt{\pi_i}}{\sqrt{\pi_j}} \right)^{y_{ij}} \tag{2}$$

where $\Delta_{ij} = \frac{1}{\sqrt{\pi_i/\pi_j} + \sqrt{\pi_j/\pi_i}}$ is a normalizing constant to assure that the probabilities p_{ij} sum up to unity, and $y_{ij} \in \{-1, 1\}$.

When data are missing, complete responses only can result in biased estimates for the object parameters. With y_{mis} representing the unobserved values which would have been observed ($r_{ij} = 0$) and $y^* = (y_{obs}, y_{mis})$, we can write the joint distribution of the observed data:

$$P\{\mathbf{y}_{obs}, \mathbf{r} | \pi, \psi\} = \int P\{\mathbf{y}^* | \pi\} P\{\mathbf{r} | \mathbf{y}^*, \psi\} dy_{mis}$$

We start by assuming an ignorable missingness process, where, for example, the probability of a missing response can depend on the comparison or on the temporal ordering of the comparison and not on the unobserved response which the subject would have given. In this case, the distribution of a given observed response pattern \mathbf{y}_{obs} with a given nonresponse pattern \mathbf{r} simplifies to:

$$P\{\mathbf{y}_{obs}, \mathbf{r} | \pi, \psi\} = P\{\mathbf{y}_{obs} | \pi\} P\{\mathbf{r} | \mathbf{y}_{obs}, \psi\}$$

Using formula (2) and assuming independence between the paired comparisons we define:

$$\begin{aligned} P\{\mathbf{y}_{obs}\} &= \prod_{i < j} \Delta_{ij}^{|y_{ij}|} \cdot \left(\frac{\sqrt{\pi_i}}{\sqrt{\pi_j}} \right)^{y_{ij}} \\ &= \Delta \cdot \exp \left\{ \sum_{j=1}^J \lambda_j \left(\sum_{\nu=j+1}^J y_{j\nu} - \sum_{\nu=1}^{j-1} y_{\nu j} \right) \right\}, \end{aligned} \tag{3}$$

where the normalizing constant $\Delta = \prod_{i < j} \Delta_{ij}^{|y_{ij}|}$ is different for each block ($\mathbf{Y}^{(A)}, \mathbf{Y}_{1;ij}^{(B)}, \mathbf{Y}_{2;ij,kl}^{(B)} \dots$) and $\lambda_j = \frac{1}{2} \ln \pi_j$. Note that the definition of y_{ij} produces a product over only the observed response values. Note also that the coefficient associated with λ_j is the number of comparisons where object O_j is preferred minus the number of comparisons where object O_j is not preferred.

2.3 Estimation of parameters

Parameter estimation is based on product multinomial sampling. All judges with the same missingness response pattern build a specific response group. The number of judges within a given group will be treated as multinomial distributed with probabilities which are defined in formula (3). Therefore, the expectation of the number of judges with a given response pattern within a given response group can be represented as log-linear. Because the model belongs to the class of Generalized Linear Models, the parameters can be estimated by standard software using a Poisson distribution and a log-link.

The design matrix \mathbf{X} consists of column vectors with suitable entries for the parameters. Because we base our considerations on product multinomial sampling the design matrix is a block matrix where each block corresponds to a given response/nonresponse pattern. In general each block can be written as:

$$(\mathbf{1}, \mathbf{YB}) .$$

The first column corresponds to the different normalizing constants, different for each group. The \mathbf{Y} s are the response pattern matrices defined in section 2.1 and \mathbf{B} is the paired comparison design matrix:

$$\mathbf{B} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 1 & 0 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1. \end{pmatrix}$$

Example: (cont.) The design matrix for our example is given by

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_8 & \mathbf{0} & \mathbf{Y}^{(A)}\mathbf{B} \\ \mathbf{0} & \mathbf{1}_4 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Y}_{1;23}^{(B)}\mathbf{B} \\ \vdots & & & & & & & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_2 & \mathbf{0} & \mathbf{Y}_{2;12,13}^{(B)}\mathbf{B} \end{pmatrix}$$

3 Dependence between comparisons

The model above can easily be extended to include dependence terms between paired comparison responses. There are several approaches discussed in the literature to incorporate within-subject dependencies in paired comparison experiments (e.g. Böckenholt and Dillon, 1997). We focus on the approach of Dittrich, Hatzinger, and Katzenbeisser (2002) who considered a log-linear approach for the analysis of dependent paired comparisons, which was embedded in the analysis of multiple binomial responses of the form $\mathbf{Y} = (Y_{ij})$, $i, j = 1, 2, \dots, J$, $i < j$. The model is extended to

$$P\{\mathbf{y}_{obs}\} = \Delta \cdot \exp \left\{ \sum_{j=1}^J \lambda_j \left(\sum_{\nu=j+1}^J y_{j\nu} - \sum_{\nu=1}^{j-1} y_{\nu j} \right) \right\} \cdot \exp \{ \theta_{1,23} y_{12} y_{13} + \theta_{2,13} y_{12} y_{23} + \theta_{3,12} y_{13} y_{23} + \dots \} .$$

with Δ defined as in formula (3).

4 Non-ignorable missing mechanisms

With non-ignorable missingness, there is dependence of the missingness process on the unobserved y . For example, an individual may choose not to respond to a particular choice if he or she knows that their choice is socially disapproved of (supporting a far right party, liking McDonalds). With y_{mis} representing the unobserved values ($r_{ij} = 0$) and $y^* = (y_{obs}, y_{mis})$, one possible model for missing data could be

$$P\{\mathbf{r} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \psi\} = \prod_{i < j} P\{R_{ij} = r_{ij}\} \quad (4)$$

where, for example,

$$\text{logit } P\{R_{ij} = 1\} = \psi_{ij0} + \psi_{ij1}y_{ij}^* \quad (5)$$

Computational approaches in the literature are mainly based on the EM-algorithm and/or composite link functions. Models with composite link functions are expressed in the form $\mu^* = \mathbf{C}\mu$ where the \mathbf{C} consists of 0's and 1's which define which elements of the unobserved vector μ need to be summed to result in an observed frequency, i.e. $E(\mathbf{Y}) = \mu^* = \mathbf{C}\mu$, $\mu = h(\eta)$, $\eta = \mathbf{X}\beta$ (Rindskopf, 1992, Molenberghs and Goetghebeur, 1997). However in the case of a large number of objects and/or many nonresponse patterns the matrix \mathbf{C} can become very large and computationally intractable. The most prominent approach covering the missing data problem is the EM-algorithm. The estimation step simply estimates the complete data counts given the data to classify counts into the full table according to the current estimates. The maximization step performs maximum likelihood estimation on the filled-in contingency table. Both steps are iterated until convergence, which can be slow.

5 Example

The models proposed in this paper will be applied to a data set where the bidding behaviour in virtual electronic auctions (Ebay auctions) was investigated. The attractiveness of an object was defined by the following characteristics: "number of bids", "quality of the describing pictures", "Rating of the seller" and the "verbal description of the auction object".

References

- Böckenholt, U., and Dillon, W.R. (1997). Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika* **62**, 411-434.
- Bradley, R.A. and Terry, M.E. (1952). Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons. *Biometrika* **39**, 324-345.
- Conaway, M.R. (1992). The Analysis of Repeated Categorical Measurements Subject to Nonignorable Nonresponse. *Journal of the American Statistical Association* **87**, 817-824.

- Dittrich, R. and Hatzinger, R., and Katzenbeisser, W. (2002). Modelling dependencies in paired comparison experiments. *Computational Statistics and Data Analysis* **40**, 39-57.
- Molenberghs, G. and Goetghebeur, E. (1997). Simple Fitting Algorithm for Incomplete Categorical Data. *Journal of the Royal Statistical Society B* **59**, 401-414.
- Rindskopf, D. (1992). A General Approach to Categorical Data Analysis with Missing Data, Using Generalized Linear Models with Composite Links. *Psychometrika* **57**, 29-42.

Factored principal components analysis and likelihood ratio based face recognition

Ian L. Dryden¹, Bai Li² and Linlin Shen²

¹ School of Mathematical Sciences, University of Nottingham, UK

² School of Computer Science and IT, University of Nottingham, UK

Abstract: A dimension reduction technique is proposed for matrix data, with applications to face recognition from images. In particular, we propose a factored covariance model for the data under study, estimate the parameters using maximum likelihood, and then carry out eigendecompositions of the estimated covariance matrix. We also develop a method for classification using a forensic likelihood ratio criterion, and the methodology is illustrated with applications in face recognition.

Keywords: Face recognition; Kernel density estimator; Likelihood ratio; Multivariate normal; Principal components analysis.

1 Introduction

The topic of ‘biometrics’ has attracted enormous interest in recent years, particularly with the requirements to deal with terrorism and other major crimes. An important biometric measurement that is frequently used for identity recognition is a face image. It is essential to develop methods that are reliable and robust given the nature of the applications. However, the task is very difficult.

One key aspect of image analysis is that the data are very high-dimensional. In face recognition and in many other applications it is of interest to summarise high-dimensional data using lower dimensional projections of the data, such as principal components analysis (PCA). In recent years there has been quite a large amount of interest in two-dimensional PCA (2DPCA) and various related methods (Yang et al., 2004; Kong et al., 2005; Ye, 2005). The essential idea of the method is to carry out PCA acting directly on matrices rather than stacking vectors. The method of 2DPCA has been demonstrated to be effective at summarizing the information in image data, where a clear two dimensional structure is present (rows and columns).

2 Factored principal components analysis

2.1 Maximum likelihood estimation

Consider a dataset of n matrices X_1, \dots, X_n which are each of size $r \times c$. It is frequently of interest to summarise the variability in the data by a linear projection to a lower dimensional sub-space.

We shall consider a stochastic model-based approach where the matrices are regarded as identically distributed realizations of a distribution with marginal probability density function $f(X_i), i = 1, \dots, n$, with mean and covariance matrix given by

$$E[X_i] = \int X_i f(X_i) dX_i = \mu, \quad (r \times c \text{ matrix})$$

$$E[\text{vec}(X_i - \mu)\text{vec}(X_i - \mu)^T] = \Sigma,$$

$i = 1, \dots, n$, where $\text{vec}(X)$ denotes the vectorize operator, which involves stacking the first column of X onto the second column of X onto the third column of X etc., to give a vector of length rc , and Σ is a $rc \times rc$ symmetric positive definite matrix. We focus on three particular cases

$$\Sigma = \Sigma_r \otimes I_c, \quad (1)$$

$$\Sigma = I_r \otimes \Sigma_c, \quad (2)$$

$$\Sigma = \Sigma_r \otimes \Sigma_c, \quad (3)$$

where Σ_r and Σ_c are positive definite matrices of size $r \times r$ and $c \times c$ respectively, \otimes denotes the Kronecker product (e.g. see Mardia et al., 1979, p459), and I_k is the $k \times k$ identity matrix. The likelihood of the data if the X_i 's are mutually independent is

$$L(X_1, \dots, X_n | \theta) = \prod_{i=1}^n f(X_i | \theta),$$

where θ is the vector of parameters of the distribution. The maximum likelihood estimator (m.l.e.) of θ is obtained by finding the value of θ which maximizes the likelihood.

2.2 Multivariate Gaussian case

In the case where the X_i are independent and identically distributed (i.i.d.) as multivariate Gaussian the m.l.e.s are given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \text{vec}(X_i - \hat{\mu})\text{vec}(X_i - \hat{\mu})^T.$$

The principal components loadings are given by the eigenvectors of $\hat{\Sigma}$, which has rank $\min(n-1, rc)$. For many practical applications, particularly involving image data, $rc \gg n-1$ and so there will be $n-1$ eigenvectors of $\hat{\Sigma}$ with non-zero eigenvalues. Under the three factored models (1)-(3) the m.l.e. of the mean is again $\hat{\mu}$. If we have the factored matrix of (1) then the m.l.e. of Σ_r is

$$\check{\Sigma}_r = \frac{1}{cn} \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^T,$$

and the eigendecomposition of this matrix gives the principal components loadings of row-based 2DPCA (Yang et al., 2004). If we have the factored matrix of (2) then the m.l.e. of Σ_c is

$$\check{\Sigma}_c = \frac{1}{rn} \sum_{i=1}^n (X_i - \hat{\mu})^T (X_i - \hat{\mu}),$$

and the eigendecomposition of this matrix gives the principal components loadings of column-based 2DPCA (Yang et al., 2004).

If we have the factored matrix of (3) then the m.l.e.s of Σ_r, Σ_c are obtained by solving the system of equations:

$$\hat{\Sigma}_r = \frac{1}{cn} \sum_{i=1}^n (X_i - \hat{\mu}) \hat{\Sigma}_c^{-1} (X_i - \hat{\mu})^T, \tag{4}$$

$$\hat{\Sigma}_c = \frac{1}{rn} \sum_{i=1}^n (X_i - \hat{\mu})^T \hat{\Sigma}_r^{-1} (X_i - \hat{\mu}). \tag{5}$$

Note that we can only estimate Σ_r, Σ_c up to a scale factor, since $a\hat{\Sigma}_r, a^{-1}\hat{\Sigma}_c$ are also solutions, for any $a > 0$. The solutions to equations (4), (5) can be obtained using a simple iterative two stage algorithm.

Consider the spectral decompositions

$$\hat{\Sigma}_r = \Gamma_r \Lambda_r \Gamma_r^T, \quad \hat{\Sigma}_c = \Gamma_c \Lambda_c \Gamma_c^T$$

where Γ_r, Γ_c are orthogonal matrices and Λ_r, Λ_c are diagonal with positive elements (ordered eigenvalues with largest first). The j th column of Γ_r contains the loadings of the j th row PC, $j = 1, \dots, r$. The k th column of Γ_c contains the loadings of the k th column PC, $k = 1, \dots, c$.

In order to reduce the dimension we consider the first q row PCs and s column PCs. Let

$$\begin{aligned} A_q &= [(\Gamma_r)_1, \dots, (\Gamma_r)_q] \quad (r \times q \text{ matrix}) \\ B_s &= [(\Gamma_c)_1, \dots, (\Gamma_c)_s] \quad (c \times s \text{ matrix}) \end{aligned}$$

where $(A)_j$ denotes the j th column of A . The projection onto the first q row and first s column eigenvectors is given by the $q \times s$ matrices of scores

$$S_i = A_q^T (X_i - \hat{\mu}) B_s, \quad i = 1, \dots, n. \tag{6}$$

We call the method of dimension reduction introduced in this section Factored Principal Components Analysis (FPCA), and the matrices S_i are the factored PC scores. The estimated variances of the row PCs are given by the diagonal elements of Λ_r , and the estimated variances of the column PCs are given by the diagonal elements of Λ_c . A low rank PC model for image data can be based on a model for the factored PC scores S_i , for example a multivariate normal model, as used in Section 4.

2.3 Bilateral 2D principal components analysis

An alternative dimension reduction technique for matrices is bilateral 2D principal components analysis (B2DPCA) (Kong et al., 2005) which is equivalent to generalized low rank approximations of matrices (GLRAM) (Ye, 2005). The task is again to find matrices of orthonormal columns A_q, B_s as above, but the estimates are found by least squares, by minimizing

$$\sum_{i=1}^n \|X_i - A_q A_q^T X_i B_s B_s^T\|^2.$$

A similar two-stage iterative algorithm to that of FPCA is used to estimate A_q, B_s . The scores are given again by equation (6).

3 Forensic Likelihood Ratio

In forensic science investigators need to quantify the amount of evidence that data collected from a crime scene (e.g. a CCTV image) really is from a suspect in custody. We investigate the use of forensic identification techniques for classification using the methods of Aitken and Lucy (2004) and Aitken et al. (2007). The strength of evidence is reported as a likelihood ratio (LR) under the two hypotheses that the suspect is that of the CCTV image versus the hypothesis that the person is not. A two stage multilevel random effects model is used to model the population of faces.

Consider a set of n measurements taken from each image (e.g. factored PC scores) $S_{ij}, i = 1, \dots, m, j = 1, \dots, n_i$ in a database of m people. If a new set of measurements is presented based on n_r recovered images then the forensic LR statistic is computed for each of the m individuals in the database. The recovered images are classified as belonging to the person with the highest likelihood ratio. We denote the forensic likelihood classification method as ‘FLR’ when a multivariate normal model is used and ‘FLR-KDE’ when a kernel density estimator model is used (see Aitken et al., 2007, for details).

One of the consistently best performers in classification problems is the support vector machine (SVM), which we consider as an alternative method.

4 Application

We illustrate the utility of the factored PCA approach with an application in face recognition. A total of 200 individuals are selected from the FERET database (Phillips et al., 2000; Shen and Li, 2006) with three frontal photographs of the face per individual with size 128×128 pixels. One of the images is selected at random and used as a test image whereas the other two images are used as training images.

We applied the factored PCA and B2DPCA methods for dimension reduction, and we explored the FLR, FLR-KDE and SVM methods for classification. Some performance rates are given in Table 1. When $sq \geq n$ the estimated covariance matrices in the FLR and FLR-KDE methods are singular. For a pragmatic solution in this case we used $\hat{U} + \tau^2 I, \hat{C} + \tau^2 I$ instead of \hat{U}, \hat{C} respectively, where we take $\tau = 1$ in our application.

TABLE 1. Comparison of classification methods for different choices of q row PCs and s column PCs from the dimension reduction techniques factored PCA (first three rows) and B2DPCA (fourth row). Each row shows the percentage correct classifications for the FLR, FLR-KDE, SVM methods, as indicated. The bold figures are the highest values in each column, and the number of parameters in each model is qs .

	5,10	10,5	10,10	q, s		10,15	15,10	15,15	20,20
				5,20	20,5				
FLR	87.5	82	88	91.5	84	85.5	78	68	87
FLR-KDE	84	80.5	87	89.5	82	83	77	68	87
SVM	44.5	46	56.5	59	56	74.5	66	83	80
FLR	83	87	89	81.5	89.5	84	82.5	71	86

We first discuss the case when factored PCA is used for dimension reduction. Note that in general the performance of the FLR and FLR-KDE methods was much better than SVM for lower numbers of PCs, with FLR usually giving slightly better performance than FLR-KDE (and the FLR-KDE approach is much slower to compute). SVM performs better for higher dimensions than in the lower dimensional cases. For higher numbers of PCs sometimes the SVM method was superior (e.g. $q = 15 = s$) and sometimes the FLR and FLR-KDE methods were better (e.g. $q = 20 = s$). The FLR method gives the best individual performance here, with 91.5% classification accuracy with 100 PC scores ($q = 5, s = 20$). The FLR method is more parsimonious than SVM, and with generally better performance FLR appears to be the best of the classification methods in our application.

It could be argued that the FLR and FLR-KDE methods work well because of the appropriateness of the statistical model, which indirectly accounts for the uncertainties in the observations with regard to expression, illumination, pose etc. by using a two stage model of between and within person variability.

The final row of Table 1 shows the performance of the FLR method when the alternative dimension reduction technique B2DPCA is used. The performance is quite similar to when using factored PCA, although for a fixed number qs of parameters the performance using factored PCA is usually marginally better here.

We have also compared the method with 2DPCA (Yang et al., 2004) and success rates with FLR classification were 29.5%, 43%, 50.8% with 2, 3, 4 row PCs respectively. The success rates with 2, 3, 4 column PCs were 23%, 39.5%, 63%. The numbers of parameters in each case are 256, 384, 512 and so we see here that 2DPCA is clearly inferior to factored PCA and B2DPCA when considering similar numbers of parameters.

References

- Aitken, C. G. G. and Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. Roy. Statist. Soc. Ser. C*, **53**, 109–122.
- Aitken, C. G. G., Zadora, G. and Lucy, D. (2007). A two-level model for evidence evaluation. *Journal of Forensic Science* **52**, 412–419.

- Kong, H., Wang, L., Teoh, E. K., Li, X., Wang, J.-G. and Venkateswarlu, R. (2005). Generalized 2D principal component analysis for face image representation and recognition. *Neural Netw.* **18**, 585–594.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.
- Phillips, P. J., Moon, H., Rauss, P. J. and Rizvi, S. (2000). The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 1090–1104.
- Shen, L. and Bai, L. (2006). Mutualboost learning for selecting gabor features for face recognition. *Pattern Recognition Letters* **27**, 1758–1767.
- Yang, J., Zhang, D., Frangi, A. F. and Yang, J.-Y. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 131–137.
- Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning* **61**, 167–191.

Bayesian modelling of time aggregated water pipe bursts with a zero-inflated, non-homogeneous Poisson process

T. Economou¹, T.C. Bailey¹ and Z. Kapelan¹

¹ School of Engineering, Computer Science and Mathematics, University of Exeter, UK.

Abstract: A commonly used approach to modelling recurrent failures is based on a non-homogeneous Poisson process (NHPP) and requires data on actual failure times. Modelling and predicting bursts in underground water pipes is vital to water companies from both an economic and conservation perspective, but often does not allow for use of a conventional NHPP for two reasons. Firstly, because data is commonly only recorded on numbers of failures over a (relatively long) time period and not on exact failure times. Secondly, because failures are usually only observed in a very small proportion of pipes in the network. This paper proposes a model derived from the conventional NHPP which only makes use of numbers of failures in an observed time period and the age of each pipe at the end of this period, but is still able to capture the age deterioration phase of the reliability curve. The model is then further extended to account for censoring and truncation in the data as well as an excess of zeros. Application of this ‘aggregated’ model and its zero-inflated extension are illustrated on a data set involving a network of 532 cement water pipes in Manukau City, Auckland, New Zealand.

Keywords: NHPP; Zero-inflated; Aggregated; Censoring; Truncation.

1 Introduction

Predicting pipe failures (i.e. bursts or blockages) in water distribution systems is important in terms of scheduling replacements and repairs, and in planning associated budgets. Various modelling approaches have been used for prediction, depending upon the failures of interest, the nature and complexity of the network and the availability, scope and reliability of relevant data (Kleiner, 2001). One approach is to use models drawn from reliability theory which usually treat the occurrence of failures as a non-homogeneous Poisson process (NHPP), i.e. a Poisson process with time varying failure rate. This approach has the advantage of explicitly incorporating and characterizing the non-linear relationship between failure rate and pipe deterioration with age as well as allowing for the inclusion of other covariates. Recently, Watson (2005) has demonstrated how this approach in conjunction with Bayesian Markov Chain Monte Carlo (MCMC) methods can be used effectively to develop a pipe replacement policy.

However, one drawback of NHPP models is that they require detailed data on actual failure times to estimate the shape of the underlying reliability curve. In practice, at least in the UK, this level of detail may not be available since historically many water companies have only recorded total numbers of failures in different parts of the network

over time periods of numbers of years, rather than actual failure times resulting in data being aggregated over the observation period. Moreover the data are most likely limited to only a few years in relation to the age of the pipes in the network (Gat and Eisenbeis, 2000). In other words the data on the number of failures is left truncated, for instance having only 10 years worth of failures for a 100 year old pipe. This added to the fact that bursts are rare events over the lifespan of a pipe, results in data sets having excess zeros in terms of failures. In this article a methodology is developed which extends conventional NHPP models to not only cope with aggregated numbers of failures but also with zero-inflation in the data. The models are implemented within the Bayesian framework using MCMC methods and applied to data involving a network of 532 cement water pipes in Manukau City, Auckland, New Zealand.

2 Model Specification

2.1 Basic Model

Suppose that pipe bursts occur as a NHPP with time-dependent intensity function $\lambda(t)$. An important property of the NHPP is that the number of failures, $N(t)$, in any time interval $[t_1, t_2]$ follow a Poisson distribution with mean $\int_{t_1}^{t_2} \lambda(t) dt = \Lambda([t_1, t_2])$ (Meeker and Escobar, 1998), i.e.

$$Pr(N(t_1) - N(t_2) = n) = \frac{e^{-\Lambda([t_1, t_2])} \Lambda([t_1, t_2])^n}{n!} \quad (1)$$

A variety of models exist that can be used to express the intensity $\lambda(t)$ (Kleiner and Rajani, 2001). Here we adopt a formulation based on the power law (e.g. Landers et al., 2001; Sen, 2002) where:

$$\lambda(t) = \gamma \theta(\mathbf{x}) t^{\theta(\mathbf{x}) - 1}$$

and $\theta(\mathbf{x}) = \beta \mathbf{x}$ is a linear function of suitable pipe covariates $\mathbf{x} = (1, x_1, x_2, \dots, x_k)$ with associated parameters $\beta = (\beta_0, \beta_1, \dots, \beta_k)$. Note that the shape function, $\theta(\mathbf{x})$, can represent both a deteriorating system ($\theta(\mathbf{x}) > 1$) and an improving system ($\theta(\mathbf{x}) < 1$).

2.2 Likelihood Function

Suppose we observe pipe i in the time period $[0, T_i)$ and that n_i denotes the numbers of failures each of which occurred at times: $0 < t_1 < t_2 < \dots < t_{n_i}$. So if $t_{n_i} < T_i$ then the data are time truncated, whereas if $t_{n_i} = T_i$ the data are failure truncated. Then, the conventional NHPP (Rigdon and Basu, 2000) leads to a time truncated likelihood function for the failure times in pipe i as:

$$\begin{aligned} f(t_1, t_2, \dots, t_{n_i}) &= \left[\prod_{j=1}^{n_i} \lambda(t_{ij}) \right] \exp \left\{ - \int_0^{T_i} \lambda(y) dy \right\} \\ &= \left[\prod_{j=1}^{n_i} \lambda(t_{ij}) \right] \exp \{ -\Lambda([0, T_i]) \} \end{aligned}$$

As mentioned previously, this likelihood involves both the number of failures, n_i , and the individual failure times t_{ij} .

Suppose, however, that only n_i and not t_{ij} are available. Then using (1), we see that $n_i \sim \text{Poisson}(\Lambda([0, T_i]))$ which allows direct use of the Poisson likelihood when dealing with aggregated data (i.e. data on number of failures and length of period of observation only). We are making the assumption here that the observation period starts at time zero (i.e. the installation time of the pipe) which is rarely the case since typically we will only have failure information for a few recent years and the pipe will usually have been installed long before the start of that period. So if we suppose that we start observing pipe i at $t_{0i} > 0$, then $n_i \sim \text{Poisson}(\Lambda([t_{0i}, T_i]))$. Assuming we have N pipes and that these pipes are independent, then the overall likelihood for the data on all pipes is:

$$L(\cdot) = \prod_{i=1}^N \frac{e^{-\Lambda([t_{0i}, T_i])} [\Lambda([t_{0i}, T_i])]^{n_i}}{n_i!}$$

and since

$$\Lambda([t_{0i}, T_i]) = \int_{t_{0i}}^{T_i} \lambda(t_{ij}) dt_{ij} = \int_{t_{0i}}^{T_i} \gamma \theta(\mathbf{x}_i) t_{ij}^{\theta(\mathbf{x}_i)-1} dt_{ij} = \gamma [T_i^{\theta(\mathbf{x}_i)} - t_{0i}^{\theta(\mathbf{x}_i)}]$$

we have

$$L(\gamma, \beta) = \prod_{i=1}^N \left(\frac{1}{n_i!} \right) e^{-\gamma [T_i^{\theta(\mathbf{x}_i)} - t_{0i}^{\theta(\mathbf{x}_i)}]} \left(\gamma [T_i^{\theta(\mathbf{x}_i)} - t_{0i}^{\theta(\mathbf{x}_i)}] \right)^{n_i} \tag{2}$$

Note that in this formulation γ is a global parameter, rather than pipe specific. In our case this is appropriate since pipes are of similar material and the ground they are buried in is comparable. However, γ can easily be made pipe-specific if the application demands it.

2.3 Zero Inflated NHPP

A substantial part of the data sets involving water pipe failures are left truncated due to the fact that it is only lately that water companies have started collecting information on failures. Clearly, there are likely to be very few recorded failures over the most recent 5 or 10 years of a pipe installed up to 100 years ago, given that failures over its whole lifespan are relatively rare. This results in data sets where the majority of the pipes appear to have no failures at all. To cope with this zero-inflation in defects of items in manufacturing, Lambert (1992) introduced the Zero Inflated Poisson (ZIP) model which has also been widely used in many other applications (e.g. Gupta et al., 1996; Zorn, 1998; Bohning et al., 1999; Ghosh et al., 2006). The idea in the ZIP model is that it has two states; the first which produces zeros with probability $(1 - p)$ and the second which produces counts from a *Poisson*(μ) distribution with probability p . Assuming a random sample y_1, y_2, \dots, y_m this results in a mixture distribution

$$f(y_k; p_k, \mu_k) = \begin{cases} (1 - p_k) + p_k e^{-\mu_k}, & \text{if } y_k = 0, \\ p_k \frac{e^{-\mu_k} \mu_k^{y_k}}{y_k!}, & \text{if } y_k = 1, 2, \dots \end{cases}$$

whose log-likelihood function is

$$l(\boldsymbol{\mu}, \mathbf{p}; \mathbf{y}) = \sum_{k=1}^m I_{(y_k=0)} \ln [(1 - p_k) + p_k e^{-\mu_k}] + \sum_{k=1}^m I_{(y_k>0)} [\ln(p_k) - \mu_k + y_k \ln(\mu_k) - \ln(y_k!)]$$

where

$$I_{(\text{event})} = \begin{cases} 1, & \text{if event is True,} \\ 0, & \text{if event is False} \end{cases}$$

Using the same idea we can extend our ‘aggregated’ model for pipe burst by assuming that each of the N pipes follows a zero-inflated nonhomogeneous Poisson distribution, so that:

$$f(n_i, T_i, t_{0i}) = \begin{cases} (1 - p_i) + p_i \exp\{-\Lambda([0, T_i])\} & n_i = 0 \\ p_i \frac{\exp\{-\Lambda([0, T_i])\} (\Lambda([0, T_i]))^{n_i}}{n_i!} & n_i = 1, 2, \dots \end{cases}$$

with log-likelihood:

$$l(\gamma, \mathbf{p}, \boldsymbol{\beta}; \mathbf{x}, \mathbf{n}, \mathbf{T}, \mathbf{t}_0) = \sum_{i=1}^N I_{(n_i=0)} \ln [(1 - p_i) + p_i \exp\{-\gamma [T_i^{\theta(\mathbf{x}_i)} - t_{0i}^{\theta(\mathbf{x}_i)}]\}] + I_{(n_i>0)} [\ln(p_i) - \gamma [T_i^{\theta(\mathbf{x}_i)} - t_{0i}^{\theta(\mathbf{x}_i)}] + n_i \ln(\gamma [T_i^{\theta(\mathbf{x}_i)} - t_{0i}^{\theta(\mathbf{x}_i)}]) - \ln(n_i!)] \quad (3)$$

where $\mathbf{n} = (n_1, n_2, \dots, n_N)$.

Following the lines of Lambert (1992) which suggests that the parameter p can itself be parameterized to be a function of covariates, we set

$$\text{logit}(p_i) = \gamma T_i^{\theta(\mathbf{x}_i)}$$

Hence p_i is related to the age of the pipe at the end of the observation period and the characteristics of the reliability curve as determined by $\theta(\mathbf{x}_i)$ and γ .

3 Model Application

Here we consider application of the proposed models to the Howick Pressure Zone data in Manukau city, Auckland, New Zealand (Watson, 2005). These data consist of 532 asbestos cement pipes with 175 recorded failures in the eleven year period 1990-2001. Some pipes experienced multiple failures and in actuality only 81 of the 532 pipes had reported failures, so that 451 pipes had zero failures.

Both the aggregated model (2) and the ZIP model (3) were fitted in WinBUGS (Spiegelhalter et al., 1999) for the first nine years of data and used to derive a posterior predictive distribution for the number of failures in each pipe for the remaining 2 years of the observation period. Covariates used in the model include pipe length, pipe diameter, pressure and absolute pressure.

Two parallel Markov chains were run for each version of the model, using a burn in of 10,000 and then sampling every 25 iterations to collect a total of 10,000 samples from each chain. This was enough to ensure good convergence and rate of mixing for each parameter. A Gaussian prior with zero mean and large variance was assumed for each of the parameters γ , and $\beta_0, \beta_1, \dots, \beta_k$.

Although the data did not contain actual times of failures, both models were still able to capture the ageing process in the vast majority of the pipes through the function $\theta(\mathbf{x}_i)$. However, the ZIP model leads to a substantial reduction in the standard errors of the covariate coefficients involved in $\theta(\mathbf{x}_i)$, suggesting that the deterioration in the pipes is more precisely estimated when zero-inflation is allowed for.

In reality 26 pipe failures were experienced in the 532 pipes in the last two years of the data collection period. The posterior predictive mean of total number of pipe failures from the zero-inflated model for these years based on the previous nine years of data was 21.5, with associated 95% credible interval (10, 29).

4 Conclusions

In this paper we have considered two modifications (time aggregation and zero-inflation) to the conventional NHPP model previously used to predict bursts in underground pipes. In summary, when applied to real data, the ‘aggregated’ zero-inflated model proposed here would appear to be able to adequately capture the ageing process in individual pipes (a key element of the NHPP) and provide usable predictions of numbers of pipe failures, despite the sparsity of failures in the data and the lack of information on actual failure times. On-going work is investigating refinements to the model to incorporate measurement error in covariates, and also to formulate a zero-inflated mixture of both the ‘aggregated’ and conventional NHPP, so as to take advantage of data sets where exact failure times are available for some pipes in the network, but only total failures for others.

References

- Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated poisson and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **162**, 195-209.
- Gat, Y. and Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water* **2**, 173-181.
- Ghosh, S., Mukhopadhyay, P. and Lu, J. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference* **136**, 1360-1375.
- Gupta, P., Gupta, R. and Tripathi, R. (1996). Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis* **23**, 207-218.
- Kleiner, Y. and Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water* **3**, 131-150.
- Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1-14.

- Landers, T., Jiang, S. and Peek, J. (2001). Semi-parametric pwp model robustness for log-linear increasing rates of occurrence of failures. *Reliability Engineering and System Safety* **73**, 145-153.
- Meeker, W. and Escobar, L. (1998). *Statistical methods for reliability data*. New York: John Wiley and Sons Inc.
- Rigdon, S. and Basu, A. (2000). *Statistical methods for the reliability of repairable systems*. New York: John Wiley and Sons Inc.
- Sen, A. (2002). Bayesian estimation and prediction of the intensity of the power law process. *Journal of Statistical Computation and Simulation* **72**, 613-631.
- Spiegelhalter, D., Thomas, A. and Best, N. (1999). *WinBUGS Version 1.2 user manual*. MRC Biostatistics Unit.
- Watson, T. (2005). *A Hierarchical Bayesian Model and Simulation Software for the Maintenance of Pipe Networks*. PhD thesis, Department of Civil and Resource Engineering, University of Auckland.
- Zorn, C. (1998). An analytic and empirical examination of zero-inflated and hurdle poisson specifications. *Sociological Methods Research* **26**, 368-400.

The Smooth Complex Logarithm Model for Quasi-Periodic Signals

Paul H.C. Eilers¹

¹ Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University. P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
P.H.C.Eilers@uu.nl

Abstract: Quasi-periodic signals with variable amplitude can be modelled using smooth curves, constructed with P-splines, for the real and imaginary parts of their complex logarithm. The problem is highly non-linear; good starting estimates are presented. The echo location signal of a bat and sunspots provide illustrations.

Keywords: P-splines, sunspots, time-frequency representation

1 Introduction

Seasonal time series have a fixed period, like a year or a day. Many statistical tools are available for modelling and analyzing them. This can not be said for quasi-periodic time series with, which show changing lengths of the periods and possibly varying amplitude. Examples of such signals can be found in many places in nature and society; examples are ultrasound echo-location by bats, light from variable stars, number of sunspots, business cycles, and EEG recordings. A flexible statistical model for quasi-periodic signals will be useful and interesting. I propose to combine the complex logarithm and P-splines (Eilers and Marx, 1996).

2 The model

A famous formula from complex function theory is due to De Moivre:

$$\exp(\theta + i\phi) = \exp(\theta)(\cos \phi + i \sin \phi). \quad (1)$$

If we let θ and ϕ vary smoothly with time t and keep the real part (or the imaginary part), we have a model for a signal with variable amplitude (e^θ) and variable phase. The instantaneous frequency is given by $(d\phi/dt)/(2\pi)$. The latter might not be directly obvious, but think of a signal with constant frequency f and constant amplitude, $\cos(2\pi ft)$. The period is the reciprocal of the frequency: $p = 1/f$.

Assuming that we have m data points (t_i, y_i) , the modelling goal is to minimize

$$S = \sum_i [y_i - \exp\{\theta(t_i)\} \cos\{\phi(t)\}]^2, \quad (2)$$

under the condition that $\theta(t)$ and $\phi(t)$ are smooth functions. Smoothness can be obtained with P-splines. Write the functions as sums of B-splines: $\theta(t_i) = \sum_j B_j(t_i)\alpha_j$ and $\phi(t_i) = \sum_j B_j(t_i)\beta_j$, plug them into (2), and put roughness penalties on α and β . This is a case of (penalized) non-linear regression, and the established approach works here too: assume that approximations $\tilde{\alpha}$ and $\tilde{\beta}$ are available and linearize the problem using first-order Taylor expansions. Update the coefficients, using (penalized) linear regression, and repeat until convergence.

Unfortunately, the regression problem is highly non-linear and starting values have to be really good, especially for β . Experience has shown that if $\tilde{\phi}$ is off by more than π , this can often not be corrected anymore. It manifests itself as a really bad fit for one or more isolated cycles of the signal. If the initial values are good, rapid convergence will be obtained.

To get good starting values for β , the following procedure was developed: 1) remove noise using a low-pass filter; 2) compute the times of the downward zero crossings, indicate these by u_k , $k = 1 \dots K$; 3) compute $v_k = k + \pi/2$; 4) smoothly interpolate the pairs (u_k, v_k) . To get good starting values for α , the following procedure works well: 1) fit a smooth trend through the data pairs $(t_i, |y_i|)$; 2) compute the log of this trend and fit it with P-splines.

It was silently assumed that the signal has no trend and thus moves almost symmetrically around the zero axis. For some signals one should first fit a trend (again using P-splines) and subtract it.

3 Applications

Figure 1 shows a part of a bat echo location signal and the fit of the model. For better visibility the original data (see Figure 3) were interpolated with a cubic spline, to increase the resolution five-fold. It is easily seen that the frequency gradually goes down, while the strength of the signal increases initially and then stays at a high level. These qualitative impressions are made quantitative by $\exp(\hat{\theta})$ and $d\hat{\phi}/dt$. The former can nicely be portrayed as an “envelope” of the fitted signal.

A second application is shown in Figure 2. These are the yearly sunspot numbers, as obtained from the National Geographic Data Center (see <http://www.ngdc.noaa.gov>). There is a lot of interest in geophysical circles in how period and amplitude of sunspot cycles have been changing over time. The complex logarithm model appears to do a good job. To get a more or less symmetric signal, square roots of the raw numbers were computed from which the trend was removed.

4 Discussion

The smooth complex logarithm model is a powerful tool for analyzing signals with slowly changing frequency and amplitude. The smooth non-parametric function allows more flexibility than the Box-Cox transform proposed by Jiang et al. (2006). Also phase and amplitude are being modelled explicitly, in contrast to the local mean decomposition of Smith (2005).

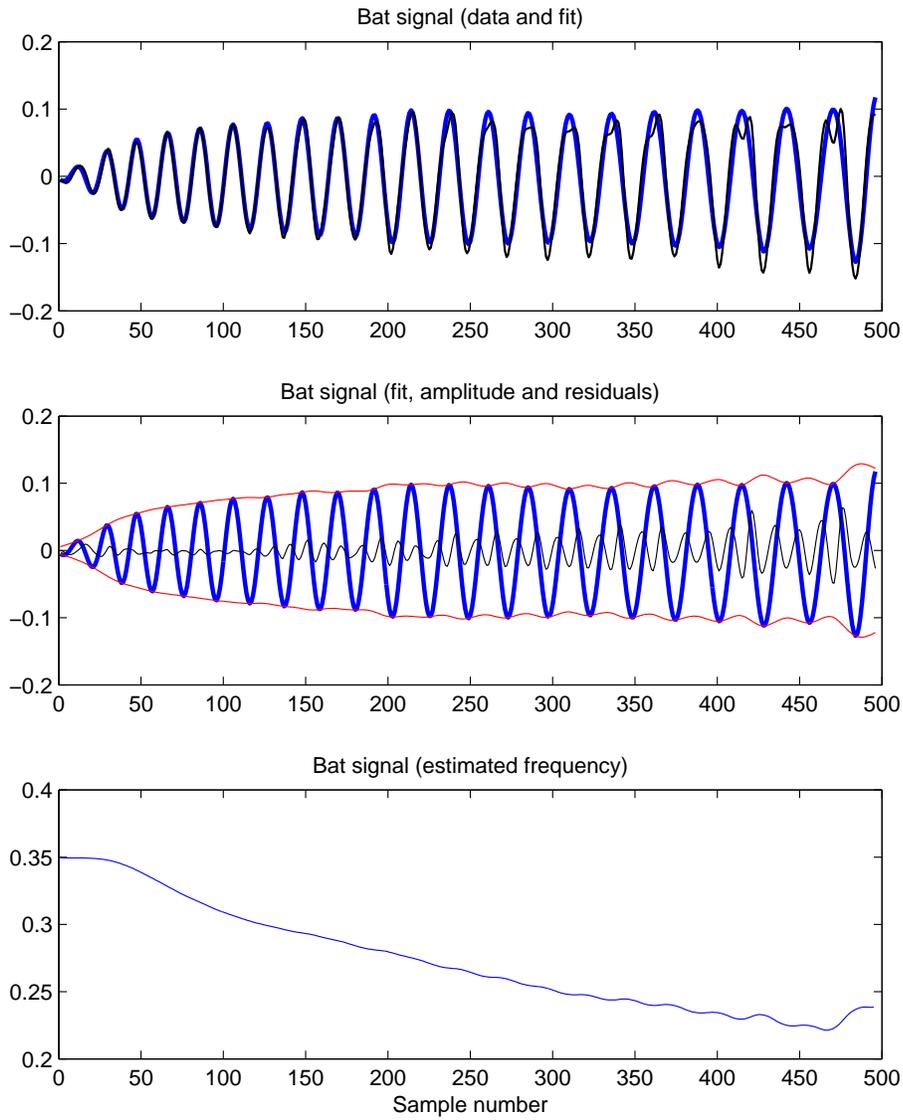


FIGURE 1. Bat echo location signal fitted by the smooth complex logarithm model. Top panel: data (thin line) and model fit (thick line); middle panel: model fit (thick line), residuals (thin line) and amplitude, shown as positive and negative envelope; bottom panel: instantaneous frequency.

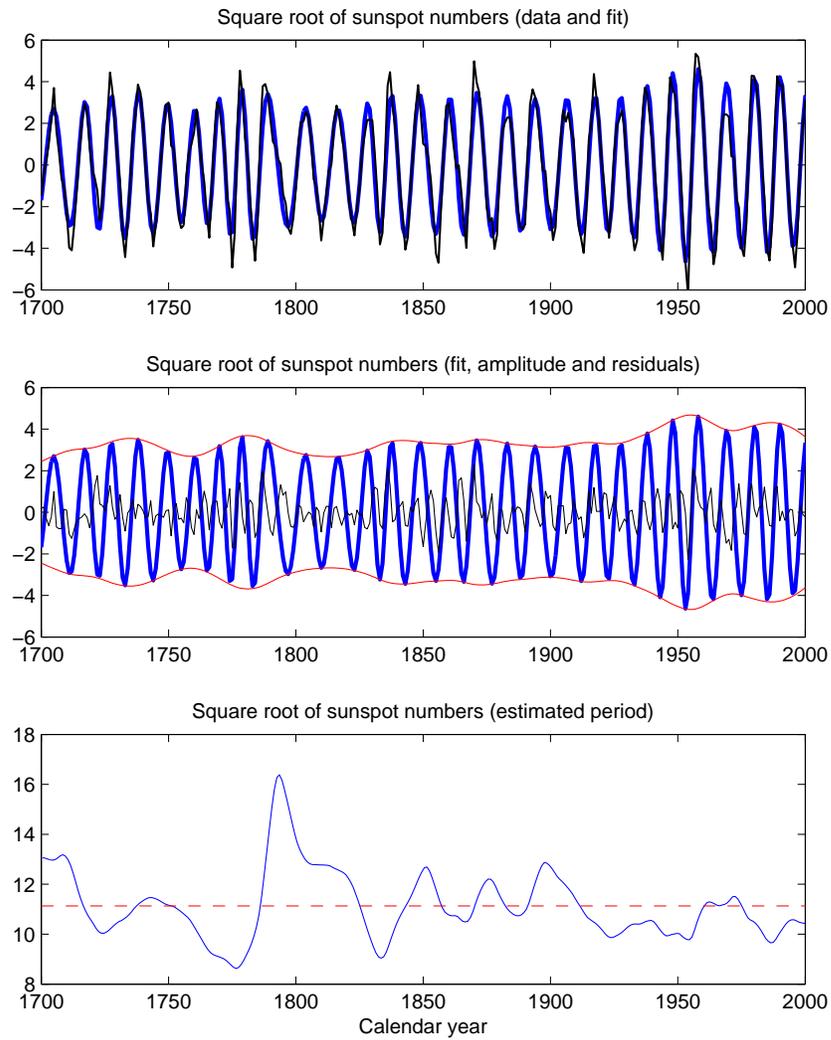


FIGURE 2. Square root of yearly sunspot numbers (detrended) fitted by the smooth complex logarithm model. Top panel: data (thin line) and model fit (thick line); middle panel: model fit (thick line), residuals (thin line) and amplitude, shown as positive and negative envelope; bottom panel: instantaneous period (full line) and average period (11.13) (broken line).

Yet a lot of more work is waiting. Some signals consists of a superposition of several quasi-periodic components; a sum like

$$\mu(t) = \sum_k \exp(\theta_k(t) + i\phi_k(t)) \quad (3)$$

seems a natural choice to model them.

A careful look at the (residuals of the) bat signal shows that a second harmonic of increasing amplitude comes in after a certain time. This suggests a model like (3), but constrained with $\phi_2(t) = 2\phi(t) + \psi$.

Other signals have varying periods and amplitude, with a typical shape of each period which is far from sinusoidal. An example, the light curve of a variable star, is shown by Hall et al. (2000). Two approaches might work. One is to set

$$\mu(t) = \exp\{\theta(t)\}f\{\phi(t) \bmod 2\pi\}, \quad (4)$$

where $f(\cdot)$ is the basic waveform (which has to be estimated too). Another is to use many harmonics, with different relative amplitudes:

$$\mu(t) = \sum_k \exp\{\theta(t) + \gamma_k\} \cos\{\phi(t) + \psi_k\}. \quad (5)$$

This might be useful for musical sound signals (Irizarry, 2001).

However, in all these extended models, it will be much harder to find good starting values, because zero crossings will no longer show a simple pattern. An alternative way to get starting values might be to first compute a two-dimensional time-frequency spectrum, smoothed with a radial gaussian kernel (RGK)(Baraniuk and Jones, 1993). The quasi-periodic components will show up as curved ridges. The footprint of each ridge gives information on $d\phi_k/dt$ and its height on the amplitude. Figure 3 shows the complete bat signal and the RGK spectrum. The existence of two strong and one weak component is clearly visible. With some human help, it will not be difficult to derive good starting values. It will be more challenging to find an automatic procedure.

References

- Baraniuk, R.G. and Jones, D.L. (1993) Signal-Dependent Time-Frequency Analysis Using a Radial Gaussian Kernel, *Signal Processing* **32**, 263–284.
- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science* **11**(2), 89–121.
- Irizarry, R. (2001) Local harmonic estimation in musical sound signals. *JASA* **96** 357–367.
- Jian, H. Gray, H.L and Woodward, W.A. (2006) Time-frequency analysis– $G(\lambda)$ -stationary processes. *CSDA* **51**, 1997–2028.
- Hall, P., Reimann, J. and Rice, J. (2000) Nonparametric estimation of a periodic function. *Biometrika* **87**, 545–557.

Smith J.S. (2005) The local mean decomposition and its application to EEG perception data. *J.R. Soc. Interface* **2**, 443–454.

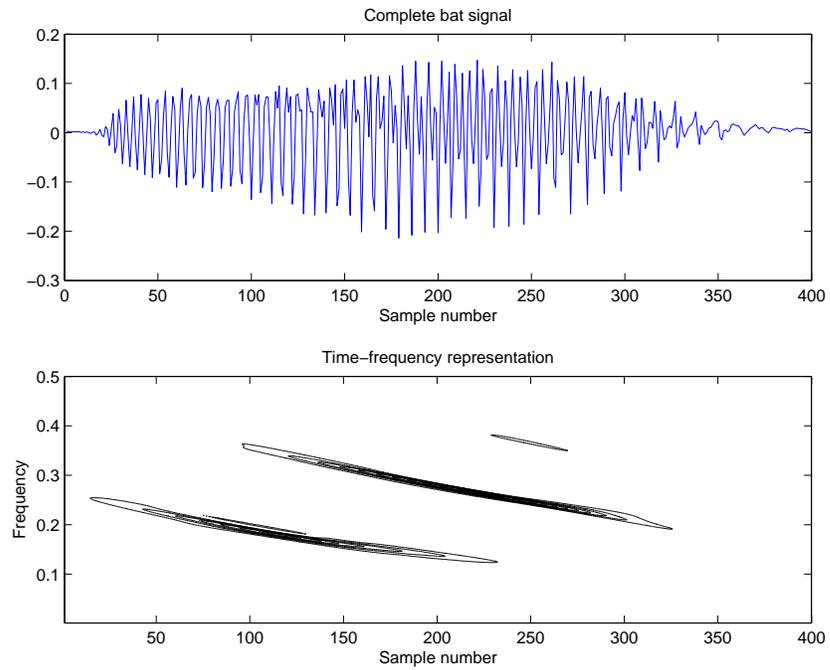


FIGURE 3. Bat signal (complete). Top: time series. Bottom: contour lines of a time-frequency representation.

Modulation Models for Seasonal Incidence Tables

Paul H.C. Eilers, Jutta Gampe, Brian D. Marx, Roland Rau

¹ Department of Methodology and Statistics, Utrecht University, The Netherlands

² Max Planck Institute for Demographic Research, Rostock, Germany

³ Experimental Statistics, Louisiana State University, Baton Rouge, USA

⁴ Terry Stanford Institute of Public Policy, Duke University, Durham, USA

Abstract: We model monthly disease counts on an age time grid using two-dimensional varying coefficient Poisson regression. Since the marginal profile of counts show a very strong and varying annual cyclical behavior over time, sine and cosine regressors model periodicity, but their coefficients are allowed flexibility by assuming smoothness over the age and time plane. The two-dimensional varying coefficient surfaces are estimated using a gridded tensor product B -spline basis of moderate dimension. Further smoothness is ensured using difference penalties on the rows and columns of the tensor product coefficients. The optimal penalty tuning parameters are chosen based on minimization of AIC. The seasonal effects are summarized with two-dimensional amplitude and phase image plots. A motivating and illustrative example is provided using data on monthly deaths due to respiratory diseases, for US females during 1959 – 1999 and for ages 44 – 96.

Keywords: Incidence data; P -splines; seasonality; tensor product.

1 Introduction

Many disease-related events reveal considerable seasonal variation, mostly striking harder in winter than in summer. The size of this seasonal effect may depend on the age and gender of the victims as well as on the type of disease. In a seasonal incidence table, one that presents event counts with a time-precision of months or quarters and age in years, we will see clear ripples with a period of one year. An example is shown in Figure 1 for females of ages 44 – 96 in the United States who died of respiratory diseases during 1959 – 1999. The top panel displays the raw counts and can also be viewed as an image of a flattened large contingency table of monthly death counts with 25,440 cells (53 ages \times 40 years \times 12 months). For fixed year, we generally find that the death counts increase with age, until approximately 85, following the typical density of deaths from human populations. Figure 1 (bottom panel) shows the monthly totals, aggregated over all ages, to visualize marginal seasonal variation, and in fact shows very strong and varying cyclical behavior due to seasonal effects.

It is of interest to quantify the character and the strength of such seasonal patterns. This includes the identification of the peak season and whether it is stable across ages and years, but also whether seasonal amplitudes changed over time or differ among age-groups. Comparisons of different diseases or causes of death may also be of interest. In this paper we present a very general model to give the answers, extending Rau and Gampe (2004).

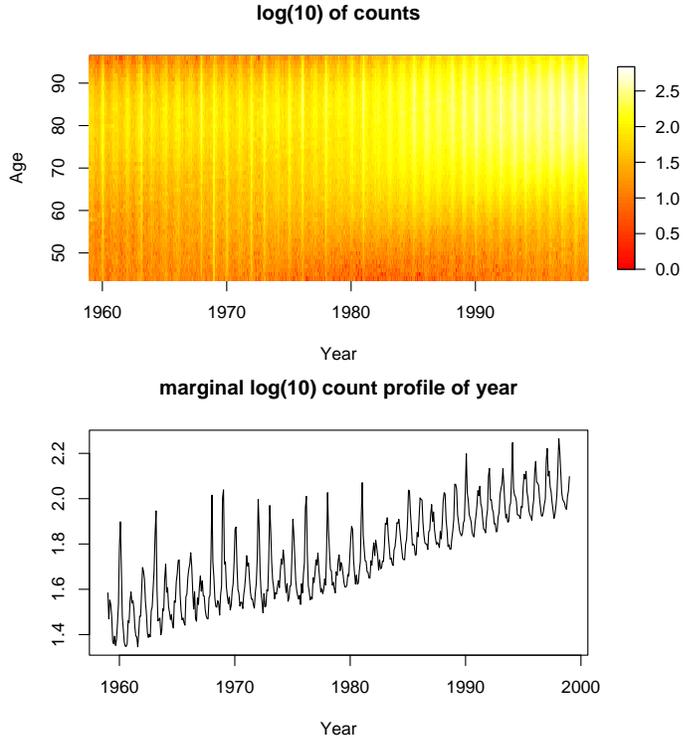


FIGURE 1. Image plot of female respiratory raw counts for ages 44 – 96 during years 1959 – 1999 (top) and the marginal plot of time trend (bottom) reflecting strong and variable seasonality.

2 Modulation models for incidence tables

The P-spline approach has been generalized to two and more dimensions. Tensor products of one-dimensional rich B-spline bases are the building blocks and difference penalties along each dimension allow tuning of smoothness. Extending VCMs and GAMs to two and more dimensions becomes straightforward this way. Instead of, say, a vector f_t , we model a matrix f_{at} as a sum of scaled tensor products of B-splines.

For incidence tables, we consider Poisson regression using a log link function

$$\log(\mu_{at}) = s_{at} + f_{at} \cos(\omega t) + g_{at} \sin(\omega t) = \eta_{at}, \quad (1)$$

with counts Y_{at} and $\mu_{at} = E(Y_{at})$. The index $a = 1, \dots, A$ refers to regressor **age** (44–96), whereas year and month are combined to create a new variable **time**, indexed by $t = 1, \dots, T$ (1 – 480). Annual cyclical behavior in the counts is modelled in the linear predictor η using the periodic sine and cosine regressors, with period 2π approximated at each month. The two regressors are only indexed with t since the

cyclical behavior is only assumed to be associated with **time**. Again we choose $\omega = 2\pi/p$, with $p = 12$. The parameters s , f , g are indexed by both (a, t) and are the smooth (two-dimensional) surfaces for the intercept and slopes for the sine and cosine regressors, respectively. As presented, (1) is over-parameterized with 3×25440 unknown parameters for 25440 cell counts. Clearly some constraint must be placed on the parameters and our choice is to enforce smoothness in estimation.

2.1 Modelling coefficients with tensor product B-splines

We use a cubic tensor product B-spline basis on the indexing variables **age** and **time**. Each tensor product has its own coefficient, which scales the altitude. In general, denote the array of coefficients as $\Theta = [\theta_{kl}]$, $k = 1, \dots, K$ and $l = 1, \dots, L$. Tensor product B-spline basis can produce very general surfaces, however, one difficulty in using such a basis is the choice of the number and the placement of the tensor products in the indexing plane. A P-spline approach takes two steps toward smoothness: 1) A relatively rich (gridded) tensor product basis is used (usually such that $K \times L < 1000$) to purposely overfit the estimated coefficient surfaces, and 2) Penalties are attached onto the rows and columns of Θ such that the influence of each penalty is regulated by its own positive tuning parameter, λ .

To express each of the intercept, sine, and cosine varying coefficients smoothly, it is convenient to work with a vectorized form of Θ denoted as $\theta_u = \text{vec}(\Theta_u)$, $u = 0, 1, 2$. A “flattened” tensor product B-spline basis \mathbf{B} can be formed of dimension $AT \times KL$, such that $\text{vec}(s) = \mathbf{B}\theta_0$, $\text{vec}(f) = \mathbf{B}\theta_1$, and $\text{vec}(g) = \mathbf{B}\theta_2$. In matrix terms, (1) can be reexpressed as

$$\begin{aligned} \text{vec}\{\log(\mu)\} &= \mathbf{B}\theta_0 + \text{diag}[\cos(\omega t)] \mathbf{B}\theta_1 + \text{diag}[\sin(\omega t)] \mathbf{B}\theta_2 \\ &= \mathbf{B}\theta_0 + \mathbf{U}_1\theta_1 + \mathbf{U}_2\theta_2 = \mathbf{M}\theta, \end{aligned} \tag{2}$$

where $\mathbf{M} = [\mathbf{B}|\mathbf{U}_1|\mathbf{U}_2]$ and $\theta' = (\theta'_0, \theta'_1, \theta'_2)$ are the augmented bases and tensor product coefficients, respectively. The diagonalization of the regressors in (2) ensures that the each level of the regressor is weighted by its proper level of the varying coefficient. We now find (2) to be a standard Poisson regression model with effective regressors M of dimension $AT \times KL$ and unknown coefficients θ . The dimension of estimation is now reduced from initially $3 \times AT$ to $3 \times KL$.

2.2 Penalized estimation

We propose to smooth estimation of the coefficient surfaces that is based on penalized maximum likelihood of the unknown tensor coefficients, θ . Assuming that the counts Y_{at} are independent and follow a Poisson distribution with mean μ_{at} , the log-likelihood function (apart from a constant) is given as $l(\theta) = \sum_{at} (-\exp(\eta_{at}) + Y_{at}\eta_{at})$, where $\text{vec}(\mu_{at}) = \exp(\mathbf{M}\theta)$. Rather than directly maximizing $\log(\theta)$, we maximize $l^*(\theta)$, which further discourages roughness among the coefficients in each row and in each column of each tensor field associated with the intercept, cosine term, and sine term. Define the penalized likelihood function as

$$l^*(\theta) = l(\theta) - \sum_{u=0}^2 (\lambda_{Ru} \text{Penalty}_R(\theta_u) + \lambda_{Cu} \text{Penalty}_C(\theta_u)), \tag{3}$$

where the six λ s are the (positive) penalty tuning parameters. The subscripts R and C refer to “rows” and “columns”, respectively. The penalized likelihood function can be written more precisely as

$$l^*(\theta) = l(\theta) - \sum_{u=0}^2 (\lambda_{Ru} \theta'_u P'_R P_R \theta_u + \lambda_{Cu} \theta'_u P'_C P_C \theta_u). \quad (4)$$

The penalty matrices for the vectorized θ can be compactly and conveniently expressed using Kronecker products, $P_R = I_L \otimes D_R$ and $P_C = D_C \otimes I_K$, where I is the identity matrix and D is the matrix formulation of the set of contrasts to construct a difference penalty on a specific row or column. Maximization of (4) results in the iterative solution,

$$\hat{\theta}_{c+1} = \left(M' \hat{W}_c M + \sum_{u=0}^2 (\lambda_{Ru} \theta'_u P'_R P_R \theta_u + \lambda_{Cu} \theta'_u P'_C P_C \theta_u) \right)^{-1} M' \hat{W}_c \hat{z}_c,$$

where $W = \text{diag}\{\text{vec}(\mu)\}$, $z = M\theta + W^{-1}\{\text{vec}(Y) - \text{vec}(\mu)\}$ is the “working” dependent variable and c is the current iterate.

Large values of λ enforce smoothness, whereas small values encourage roughness in either the row or column orientation. We search for the optimal values of λ , and in theory the six tuning parameters in can be free. In practice, we constrain the row (column) λ s constant across sine and cosine terms, yielding four free tuning parameters. Our choice of model performance is measured by Akaike’s information criterion (AIC), which is a compromise between goodness of fit and model complexity, and is defined as $\text{AIC}(\lambda) = \text{deviance}(Y, \theta) + 2\text{trace}(H)$, where $\lambda = (\lambda_1, \dots, \lambda_4)$ and $\text{trace}(H)$ approximates the effective dimension of the model. Upon convergence, the matrix H is the effective “hat” matrix and in this Poisson regression setting its trace can be computed efficiently as

$$\text{tr} \left[M' \hat{W} M \left(M' \hat{W} M + \sum_{u=0}^2 (\lambda_{Ru} \theta'_u P'_R P_R \theta_u + \lambda_{Cu} \theta'_u P'_C P_C \theta_u) \right)^{-1} \right].$$

We search on a linear grid on a log scale for the optimal set λ , monitor AIC, and choose the λ with minimum AIC.

3 Applications

For female respiratory lifetable data, each the intercept, cosine, and sine term, the varying coefficients were constructed on the **age** and **time** grid using a basis with $K \times L = 7 \times 9$ tensor product (cubic) B-splines. The penalty was constructed using a third order difference penalty on each row and column of the tensor product coefficients. The optimal tuning parameters were $\lambda = (0.001, 0.001, 1, 0.01)$, yielding an approximate effective dimension for the model: $\text{trace}(H) = 159$. Note that the initial dimension of the model was 3×25440 , that was reduced to $3 \times 63 = 189$, then finally reduced to 159 after penalization. Figure 2 displays the varying intercept term in the upper,

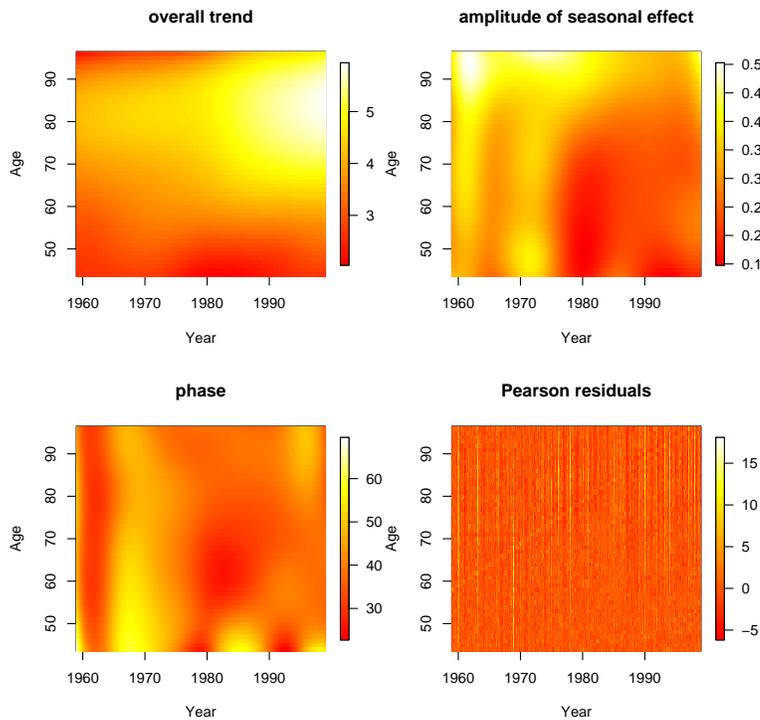


FIGURE 2. Image plot of the varying intercept term or overall trend (upper, left); amplitude and phase (month) resulting from the 2D seasonal effects, (upper, right) and (lower, left) respectively; and the Pearson residuals (lower, right).

right panel. In the two following panels the cosine and sine varying coefficients are transformed into the amplitude and phase surfaces. The phase surface has been further mapped from the interval $(0, 2\pi)$ into $(0, 365)$ and is generally peaked during winter.

4 Discussion

The modulation model for incidence tables allows detailed quantitative description of many aspects of seasonality. The intercept surface captures the overall trend in event counts, as a function of year and age. After conversion to polar coordinates, the modulation surfaces tell us how the relative strength of seasonality changes with year and age, as well as changes in the phase, i.e. the time of the year, in which peaks occur. This can be a powerful tool for detailed epidemiological or demographic analysis. In this paper we studied mortality data, but of course the model can be applied to other types of events, like disease incidence.

Because the estimated surfaces are smooth by design, they allow the computation of derivatives, with respect to age or time, for specialized research in which rates of change are of interest.

The combination of tensor products and discrete penalties is very effective for multidimensional smoothing. One can image extensions of gender models, with shared seasonal components, but differing trends for men and women. One can also envision extension to more dimensions. The US death counts can be subdivided by state. This gives the opportunity to introduce East-West or North-South spatial location for building a three-dimensional incidence table. A further extension is to build four-dimensional models, in which “surfaces” live in the longitude-latitude-time-age space.

References

- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science* **11**, 89-121.
- Eilers, P.H.C. and Marx, B.D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics* **11**(4), 758-783.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.
- Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *Journal of the Royal Statistical Society* **55**(4), 757-796.
- Rau, R. and Gampe, J. (2004). Seasonal variation in death counts: P-spline smoothing in the presence of overdispersion. In: Proceedings of the 19th International Workshop on Statistical Modelling. Florence, Italy.

Smoothing, sampling, and Basu’s elephants

Jochen Einbeck¹, Thomas Augustin² and Julio M. Singer³

¹ Department of Math. Sciences, Durham University, Durham DH1 3LE, UK

² Institut für Statistik, Ludwigstr. 33, 80539 München, Germany

³ Departamento de Estatística, Universidade de São Paulo, CP66281, SP, Brazil

Abstract: We investigate design-weighted local smoothing and show that the optimal (bias-minimizing) weights have similar form and interpretation as the optimal weights given by the Horvitz-Thompson theorem known from sampling theory. We set forth that the hazards in using bias-minimizing weights apply to kernel smoothing, too, suggesting to be cautious with the application of bias-minimizing weights in general.

Keywords: Weighting; Horvitz-Thompson estimator; local polynomials

1 Introduction

A circus owner plans to ship 50 adult elephants and therefore needs a rough estimate of their total weight. As weighing elephants is quite cumbersome, he intends to weigh only one elephant and to multiply the result with 50. However, the circus statistician insists in setting up a proper sampling plan, and to use the Horvitz-Thompson estimator. They agree to assign a selection probability of 99/100 to a previously determined elephant (‘Samba’), which from a previous census is known to have about the average weight of the herd. The probability for all other elephants is 1/4900, including ‘Jumbo’, the biggest elephant in the herd. Naturally, Samba is selected, and the statistician estimates the total weight of the herd by 100/99 times Samba’s weight according to Horvitz-Thompson. If Jumbo were selected, his large weight would even have to be multiplied by 4900 to get the ‘best linear unbiased estimator’ of the total weight! Certainly, after having given these advices, the circus statistician was sacked.

This is a short version of a fable told by Basu (1971), illustrating his reservations against the Horvitz-Thompson (HT) estimator: For a sample of size n drawn from a population Y_1, \dots, Y_N , Horvitz and Thompson (1952) showed that among all linear estimators of the form $\hat{Y} = \sum_{i=1}^N \alpha_i \delta_i Y_i$, the HT estimator $\hat{Y}_{HT} = \sum_{i=1}^N \delta_i Y_i / \pi_i$ (where π_i is the probability that the i -th element is drawn in any of the n draws and δ_i is an indicator taking the value 1 if unit i is selected) is the only unbiased estimator for the population total, Y . Horvitz and Thompson state that if $\pi_i = nY_i/Y$, the estimator \hat{Y}_{HT} has zero variance and sampling will be optimal. Rao (1999) warns that the HT estimator ‘can lead to absurd results if the π_i are unrelated to the Y_i ’, and obviously the probabilities in the fable are far from optimal in this sense. Though HT’s theorem can reduce the bias of an estimate *given* the inclusion probabilities, it may produce useless estimates if they are unfortunately chosen. Nevertheless, HT’s estimator proves to be useful e.g. in the context of ratio estimation, when a second variable X_i is used

to construct selection probabilities which are correlated to the Y_i . In Basu's example, a way out for the unfortunate circus statistician would have been to take the known elephant weights X_i from the previous census, and to set $\pi_i = nX_i/X$, where X was the total weight of the herd measured at that time (Koop, 1971, in the discussion of Basu's essay).

2 Design-weighted local smoothing

One of the statistical fields where weighting is quite common is that of nonparametric smoothing. Given a sample $(x_1, y_1), \dots, (x_n, y_n)$ drawn from a bivariate population $(X, Y) \in \mathbb{R}^2$ with mean function $m(x) = E(Y|X = x)$, we are interested in a smooth estimate $\hat{m}(\cdot)$ of $m(\cdot)$. There are two forms of weighting that have to be distinguished here. Firstly, there are the *kernel weights* $K((x_i - x)/h)$, with a bandwidth h , and secondly, one can use additional *design weights*, $\alpha(\cdot)$, leading to the design-weighted least squares problem

$$\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \alpha(x_i) \left(y_i - \sum_{j=0}^p \beta_j(x)(x_i - x)^j\right)^2. \quad (1)$$

From the vector $(\hat{\beta}_0(x), \dots, \hat{\beta}_p(x))$ minimizing (1), one easily gets estimators of m and its derivatives, $\hat{m}^{(j)}(x) = j!\hat{\beta}_j(x)$, and one has the following

Theorem. *Let $h \rightarrow 0$ and $nh^3 \rightarrow \infty$, and $\mathbb{X} = (x_1, \dots, x_n)$. Under regularity assumptions we get for $p - j$ odd*

$$\text{Bias}(\hat{m}^{(j)}(x)|\mathbb{X}) = e_{j+1}^T S^{-1} c_p \frac{j!}{(p+1)!} m^{(p+1)}(x) h^{p+1-j} + o_P(h^{p+2-j})$$

and for $p - j$ even

$$\begin{aligned} \text{Bias}(\hat{m}^{(j)}(x)|\mathbb{X}) &= e_{j+1}^T \frac{j!}{(p+1)!} \left[\left(\frac{\alpha'(x)}{\alpha(x)} + \frac{f'(x)}{f(x)} \right) s_p m^{(p+1)}(x) + \right. \\ &\quad \left. + S^{-1} \tilde{c}_p \frac{m^{(p+2)}(x)}{p+2} \right] h^{p+2-j} + o_P(h^{p+2-j}), \end{aligned} \quad (2)$$

with $s_p = (S^{-1} \tilde{c}_p - S^{-1} \tilde{S} S^{-1} c_p)$, and kernel moment matrices S , \tilde{S} , and vectors c_p , \tilde{c}_p , for the detailed form of which we refer to Einbeck and Augustin (2005), as well as for the proof, regularity assumptions, and for the asymptotic variance. The more interesting of the two expressions above is the second one, because it shows that in this case the leading term is *not* independent of $\alpha(\cdot)$. This gives the chance to reduce the bias. Note that the augend in the squared bracket in (2) vanishes for $\alpha'(x)/\alpha(x) + f'(x)/f(x) = 0$, and this differential equation is solved for

$$\alpha_{opt}(x) = c \frac{1}{f(x)}, \quad (3)$$

with $c \in \mathbb{R} \setminus \{0\}$. Considering the design density as “selection probability distribution”, this gives a very similar message to that of HT, where we had optimal weights $\alpha_i = 1/\pi_i$. In practice $f(\cdot)$ is mostly unknown, but it may be substituted by a density estimate, $\hat{f}(\cdot)$.

3 A surprising analogy

Formula (3) is exactly the opposite of the recommendation given by Einbeck, André and Singer (2004), who proposed the setting $\alpha(\cdot) = \hat{f}(\cdot)$ in order to robustify against outliers in the design space. It is well known that points near the boundary can have a huge influence on the estimate of the regression function (which is even more true for the derivative estimates, see Newell and Einbeck, 2007). This effect will gain dramatically in power if we even apply weights *inversely* proportional to the design density as suggested by our bias-minimizing criterion above - just as Jumbo had a tremendous influence when selected!

It is at this point worth to take a look into the rejoinder of Basu’s (1971) essay, in which he vehemently denied that the ‘unrealistic sampling plan’ was responsible for the failure of the HT estimator. Basu defended, in contrary, the circus statistician’s sampling plan, as it ensures a *representative* sample, and gave the responsibility for the useless result entirely to the HT estimator itself, ‘being a method that contradicts itself by allotting weights to the selected units that are inversely proportional to their selection probabilities. The smaller the selection probability of a unit, that is, *the greater the desire to avoid selecting the unit*, the larger the weight that it carries when selected.’

Similarly, in the smoothing context, we have derived a bias-minimizing criterion, which may prove useful for large and well-behaved data sets, but may give disastrous results in the presence of outlying predictors. This is exactly the dilemma that Basu was worried about: he did not conform himself to the fact that one has to get the selection probabilities right, and in some sense, he is right. What does one do, for instance, if no auxiliary variable X_i is available to construct a ratio estimator, or if one gets a sample, selected with ‘wrong’ selection probabilities, and has to work with it (we are aware that there exist some techniques to adjust the probabilities ex post, e.g., the Keyfitz (1951) technique, which however have drawbacks as they are actually based on deleting available data). In the smoothing context, the selection probabilities correspond to the design density, which is almost never designed to meet any optimality criterion, and hence there is always a certain potential that things may go wrong.

4 Conclusion

The goal of this paper was to show that there exists an striking analogy between the theories of sampling and smoothing, leading to a similar discrepancy between theoretically optimal and practically useful weighting schemes. We believe that this tells us an important lesson about statistical methods in general: weighting is performed in virtually all statistical disciplines, and a usual way of motivating such weights is to look at theoretical, bias-minimizing criteria. These criteria will often suggest to

choose weights inversely proportional to some kind of selection probability (density). This however makes the estimator extremely sensitive to extreme observations (which correspond to Jumbo in Section 1 and the outlying predictors in Section 2). Hence, we advise to be careful with bias-minimizing estimators if there are any observations which might be labelled by the terms “extreme”, “undesired”, “outlying”, “weak” or “needy”, and the like, and it is likely that this holds far beyond the scope of sampling and smoothing.

References

- Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1 (with discussion), In: Godambe and Sprott (Eds.), *Foundations of Statistical Inference*, 203–242, Holt, Reinhart and Winston, Toronto.
- Einbeck, J., André, C.D.S. and Singer, J.M. (2004). Local smoothing with robustness against outlying predictors. *Environmetrics* **15**, 541–554.
- Einbeck, J. and Augustin, T. (2005). On weighted local fitting and its relation to the Horvitz-Thompson estimator. SFB386 Discussion Paper No. 465, University of Munich.
- Newell, J. and Einbeck, J. (2007). A comparative study of nonparametric derivative estimators. *Proceedings of the 22th IWSM*, Barcelona.
- Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *JASA* **47**, 663–685.
- Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustment for changes in the probabilities. *JASA* **46**, 105–109.
- Rao, J.N.K. (1999). Some current trends in sample survey theory and methods (with discussion). *Sankhyā* **61**, 1–57.

Adaptive Distance-Based Classification

A. Esteve¹, J. Fortiana² and E. Boj³

¹ Centre d'Estudis Epidemiològics sobre l'HIV/sida de Catalunya. Hospital Universitari Germans Trias i Pujol. Ctra. de Canyet, s/n. 08916 Bandanna, Spain.

aeg.ceescat.germanstrias@gencat.net

² Departament de Probabilitat, Lògica i Estadística, Universitat de Barcelona. Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain, fortiana@ub.edu.

³ Departament de Matemàtica Econòmica, Financera i Actuarial, Facultat de Ciències Econòmiques i Empresariales, Universitat de Barcelona, Avinguda Diagonal, 690, 08034 Barcelona, Spain, evaboj@ub.edu

Abstract: We investigate a data-driven version of Distance-Based classification (Cuadras et al. 1997). The distance function is drawn from a family of metrics depending on continuous parameters which, in a sense precised below, are estimated from data.

Keywords: Distance-Based Prediction; Classification; Metrics with Given Marginals; Adaptive Metrics.

1 Introduction

Distance-Based (DB) methods were conceived by Cuadras (1989) with an aim to provide robust nonlinear alternatives to classic multivariate and predictive statistical techniques, capable of sensibly meeting unconventional requirements, such as nonnumerical measurements. Since their inception a stream of contributions has improved on the basic functionality. See, e.g., Boj *et al.* (2007a, 2007b). A cardinal element of DB methods, namely the metric or distance function itself, has so far remained untouched, due to the acceptable performance of standard comprehensive measures such as ℓ^p for continuous variates or the Jaccard coefficient for binary ones. In this paper we investigate metric choice in DB classification, beyond a mere selection from a limited repertory, with metrics depending on parameters which, in a sense precised below, are estimated from data. Section 2 is a concise introduction to the method, Section 3 presents parametric families of metrics with the required properties, Section 4 describes the process of estimating parameters, and Section 5 deals with some implementation details.

2 Distance-Based Classification

The key concept is that of *distance from an individual to a population*, naturally leading to a minimum distance allocation rule (see Cuadras *et al.* 1997). For a training set of n individuals, belonging to g known groups,

$$\Omega = \Omega_1 \sqcup \cdots \sqcup \Omega_g, \quad \#(\Omega_\alpha) = n_\alpha, \quad n = \sum_{\alpha=1}^g n_\alpha,$$

given a new individual ω to be classified on the basis of a certain set of observed variables, the procedure is as follows:

1. From the observed variables, with a metric $d(\cdot, \cdot)$, compute the g within-group matrices \mathbf{D}_α of squared interdistances and the g vectors \mathbf{d}_α of squared distances from ω to each individual in group α .
2. From the above quantities, compute the g proximity functions $f_\alpha(\omega)$. It can be proved that these are equal to the squared distances from ω to the g centroids of the groups in a Euclidean configuration space (see details, formulae and proofs in the above reference).
3. ω is allocated to the group with the nearest centroid.

3 Euclidean Metrics with Given Marginals

Underlying DB classification is the assumption that an abstract configuration space does exist. In turn, this condition depends on $d(\cdot, \cdot)$ being a *Euclidean metric*, in the sense of Multidimensional Scaling (see, e.g., Cox and Cox, 2000). Even though proximity functions derive directly from \mathbf{D}_α and \mathbf{d}_α , without computing actual coordinates, its implicit existence ensures consistent results. As a matter of fact it is an open problem to study the behavior of the DB classification algorithm when supplied with a non-Euclidean metric, as weird effects due to negative “distances” are to be feared. Meanwhile, and to remain on the safe side, we are bound to restrict ourselves to parametric families of Euclidean metrics,

$$\mathcal{D}(\Theta) = \{d_\theta(\cdot, \cdot), \theta \in \Theta\},$$

from which to select an adequate one, adapted to a given problem.

Such families often originate in observations from several information sources (for instance, individuals with some numeric measurements plus some categorical attributes). A natural procedure is to derive separate metrics for each block of variables and, from them, a mixture involving one or more parameters. The result is called a parametric family of *joint metrics* associated with a given collection of *marginal metrics*. This terminological coincidence with the theory of joint probability distributions with given marginals reflects that both descriptions are parallel manifestations of the same reality. For instance, naïve Pythagorean addition (“sum of squared distances”) of metrics implicitly assumes statistical independence of the corresponding sets of variables. Esteve and Fortiana (2002, 2007a, 2007b) have devised strategies for constructing families of metrics, with range restricted to the Euclidean cone, depending on parameters regulating the amount of interaction between sources of information.

4 Tailoring Metrics

Assuming we have chosen a parametric family $\mathcal{D}(\Theta)$ of Euclidean metrics for a given classification problem, the analogue of parameter estimation in the present setting is the calibration procedure for selecting a suitable $\hat{\theta} \in \Theta$, or the corresponding metric $d_{\hat{\theta}}(\cdot, \cdot) \in \mathcal{D}(\Theta)$. There are several possible calibration criteria, all conforming to the

generic description of optimizing some measure of goodness-of-classification for a test sample $\tilde{\Omega}$. If the original training set is large enough, a crossvalidatory scheme may replace an independent testing set (see Section 5).

Since the relative frequency of misclassified individuals in the test sample $\tilde{\Omega}$ is discontinuous as a function of θ , often it is not sensitive enough for optimization purposes. A more smooth criterion is based on the sum of:

$$U(\omega) = f_\gamma(\omega) - \min_{1 \leq \beta \leq g} \{f_\beta(\omega)\}, \quad \omega \in \tilde{\Omega},$$

where $\gamma \equiv \gamma(\omega)$ is the index of the group to which ω is known to belong. Clearly $U(\omega) \geq 0$ and $U(\omega) = 0$ when ω is correctly classified. A similar quantity is the sum of:

$$V(\omega) = \max_{1 \leq \beta \leq g} \{s_\beta(\omega)\} - s_\gamma(\omega), \quad \omega \in \tilde{\Omega},$$

where $(s_\alpha(\omega))$ is the softmin of the vector $(f_\alpha(\omega))$, defined by:

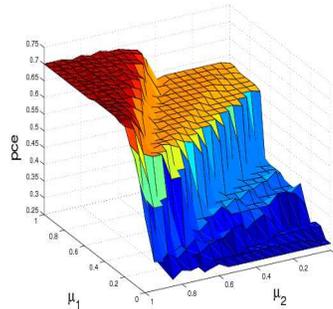
$$s_\alpha(\omega) = \exp(-f_\alpha(\omega)/\sigma^2) / \sum_{\beta=1}^g \exp(-f_\beta(\omega)/\sigma^2), \quad \sigma > 0.$$

5 Implementation and an illustrative example

We have developed a package of MATLAB programs to perform the above optimization process for two of the classes of parametric families described by Esteve and Fortiana (2007b). The current algorithm is based on the *leave-one-out* cross-validatory scheme described in Cuadras *et al.* (1997), which is computationally cheap and easily programmed.

As an illustration of the procedure, Table 1 compares, on the Cancer dataset, already used in Cuadras *et al.* (1997), the number of misclassified samples obtained by DB discrimination with Gower’s distance (δ_G), and two different parametric families, δ_λ and δ_μ , after optimization as described above. Figure 1 shows the estimated probabilities of misclassification as a function of $\mu = (\mu_1, \mu_2)$ for the δ_μ family of metrics.

	Π_1 $n_1=78$	Π_2 $n_2=59$	Total $n=137$
δ_G	18	21	39
δ_λ	36	12	48
δ_μ	16	21	37
LDF	31	27	58
QDF	13	35	48



Acknowledgments: Work supported in part by the Spanish Ministerio de Ciencia y Tecnología and FEDER grant MTM2006-09920.

References

- Boj, E., Claramunt, M. M. and Fortiana, J. (2007a). Selection of Predictors in Distance-Based Regression. *Communications in Statistics—Simulation and Computation* **36**, 87–98.
- Boj, E., Grané, A., Fortiana, J. and Claramunt, M. M. (2007b). Implementing PLS for Distance-Based Regression: Computational Issues. *Computational Statistics*, Special PLS Issue (In press).
- Cox, T. F. and Cox M. A. A. (2000). *Multidimensional Scaling. Second Edition*. London: Chapman & Hall/CRC.
- Cuadras, C. M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In: Y. Dodge (Ed.), *Statistical Data Analysis and Inference*. pp. 459–473. Amsterdam: North-Holland Publishing Co.
- Cuadras, C. M., Fortiana, J. and Oliva, F. (1997). The proximity of an Individual to a Population with application to Discriminant Analysis. *Journal of Classification* **14**, 117–136.
- Esteve, A. and Fortiana, J. (2007a). *Metrics in Product Spaces and their Marginals*. Working paper.
- Esteve, A. and Fortiana, J. (2007b). *Parametric Families of Euclidean Distance Functions*. Working paper.
- Fortiana, J. and Esteve, A. (2002). Euclidean Metrics with Given Marginals. Communication to: *Eleventh International Workshop on Matrices and Statistics*, Lyngby (Denmark).

A High-Dimensional Joint Model for Longitudinal Endpoints of Different Type

Christel Faes¹, Marc Aerts¹, Helena Geys^{1,2}, Geert Molenberghs¹ and Greet Teuns²

¹ Center for Statistics, Hasselt University, Diepenbeek, Belgium.

`christel.faes@uhasselt.be`

² Johnson & Johnson Pharmaceutical Research and Development, Beerse, Belgium

Abstract: In many studies several outcomes are repeatedly measured on the same subject. Often, these outcomes are not all of the same type, but are a mixture of continuous and categorical outcomes. For example, in toxicity studies, many endpoints are measured on the same animal in order to study the toxicity of the compound of interest. While multivariate methods of the analysis of continuous outcomes are well understood, methods for jointly analyzing continuous and discrete outcomes are less familiar. Random effects models can be used in the situation where various outcomes of a different nature are observed (Molenberghs and Verbeke 2005), and these models extend straightforwardly towards inclusion of repeated measures. However, computational problems arise as the number of endpoints increase. Fieuws and Verbeke (2005) propose a pairwise modelling strategy to model high-dimensional longitudinal data of the same type, in which all possible pairwise mixed models are fitted separately, and where inference follows from pseudo-likelihood theory. Here, a similar method is applied for the joint analysis of several binary and continuous endpoints, but where all pairwise bivariate generalized linear mixed models are fitted jointly, making it possible to directly use pseudo-likelihood-based inference. Methods are illustrated using a repeated toxicity study for the evaluation of the neurofunctional effects of a psychotropic drug.

Keywords: Mixed Endpoints, High-Dimensional Joint Model, Longitudinal Data, Pseudo-Likelihood

1 Introduction

When a pharmaceutical company brings a new medicine on the market it must be assured that the product is safe for intended use and in the event of accidental misuse. To properly assess the toxicity of the substance of interest, many endpoints are investigated for possible toxicity and the most appropriate and efficient statistical models should be used. This raises a number of challenges. First, an appropriate statistical model should account for possible correlations among the different endpoints. Since the number of endpoints in these experiments is large, this implies the necessity of a flexible high-dimensional model. Second, one may also deal with outcomes of a mixed continuous/discrete nature. For example, both the malformation status of a live fetus (typically recorded as a binary outcome) or the birth weight (measured on a continuous scale) are important variables in the context of teratogenesis. Perhaps the most common situation is that of a continuous, often normally distributed, and a binary

or ordinal outcome. Third, measurements can be collected repeatedly in time. This induces extra correlation that must be accounted for in the statistical analysis.

While multivariate methods for the analysis of continuous outcomes are well known, methods for mixed continuous and discrete outcomes are less familiar. There broadly are two approaches towards the analysis of combined continuous and discrete endpoints (Aerts et al. 2002). A first modelling approach towards a joint model of a continuous and discrete outcome is to apply a conditioning argument that allows the joint distribution to be factorized in a marginal component and a conditional component, where the conditioning can be done either on the discrete outcome or on the continuous outcome. A drawback of mixed outcome models based on factorization is that they may be difficult to apply for quantitative risk assessment and they do not easily extend to the setting of three or more endpoints. A second modelling approach directly formulates a joint model for both outcomes. Here, a mixed model is used to jointly analyze a continuous and binary longitudinal endpoint. A pseudo-likelihood based approach is used for the extension towards a joint model for many endpoints of mixed type.

2 Longitudinal Continuous-Binary Endpoints

Assume we have two sequences of n outcomes each. We denote the sequence of continuous endpoints for subject i as $\mathbf{Y}_{1i} = (Y_{1i1}, Y_{1i2}, \dots, Y_{1in})$ and the one with binary endpoints as $\mathbf{Y}_{2i} = (Y_{2i1}, Y_{2i2}, \dots, Y_{2in})$. Y_{1ij} and Y_{2ij} represent, respectively, the j th continuous and binary outcome for subject i .

Random effects models are probably the most frequently used models to analyze multivariate data. It is also straightforward to use a mixed model in situations where various outcomes of a different nature are observed (Molenberghs and Verbeke, 2005). For the bivariate response vector $\mathbf{Y}_i = (\mathbf{Y}_{1i}, \mathbf{Y}_{2i})'$ we can assume a general model of the form

$$\mathbf{Y}_i = \mu_i + \epsilon_i, \quad (1)$$

where μ_i is specified in terms of fixed and random effects and ϵ_i is the residual error structure, of which the variance depends on the mean-variance links of the different endpoints. Let $\mu_i = \mu_i(\eta_i) = \mathbf{h}(\mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i)$, in which the components of the inverse link function $\mathbf{h}(\cdot)$ are allowed to change with the nature of the various outcomes in \mathbf{Y}_i . For example, we can choose the identity link for the continuous component, and the logit link for the binary component. \mathbf{X}_i and \mathbf{Z}_i are $(2 \times p)$ and $(2 \times q)$ -dimensional matrices of known covariate values, and β a p -dimensional vector of unknown fixed regression coefficients. Further, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$ are the q -dimensional random effects.

As a result, the correlation among the outcomes can be modelled either using the residual variance of \mathbf{Y}_i or through specification of the random effects structure $\mathbf{Z}_i\mathbf{b}_i$. In what follows, we formulate a possible model to account for the longitudinal structure of joint continuous and binary endpoints, using a conditional independence random-intercepts model with a general variance-covariance matrix \mathbf{D} . This model can be

written in the following form:

$$\begin{pmatrix} Y_{1ij} \\ Y_{2ij} \end{pmatrix} = \begin{pmatrix} \alpha_0 + \alpha_1 X_{ij} + b_{1i} \\ \frac{\exp(\beta_0 + \beta_1 X_{ij} + b_{2i})}{1 + \exp(\beta_0 + \beta_1 X_{ij} + b_{2i})} \end{pmatrix} + \begin{pmatrix} \epsilon_{1ij} \\ \epsilon_{2ij} \end{pmatrix}, \tag{2}$$

where the random effect b_{1i} and b_{2i} are normally distributed as

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix} \right) \tag{3}$$

and where ϵ_{1ij} and ϵ_{2ij} are independent and $v_{2i} = \pi_{2i}[1 - \pi_{2i}]$. The random effects b_{1i} and b_{2i} are used to accommodate for the longitudinal structure in the data. The correlation ρ_{12} among the continuous and binary endpoints is induced by the incorporation of a correlation ρ among the two random effects, and is approximately equal to

$$\rho_{12} = \frac{\rho\tau_1\tau_2v_{2ij}}{\sqrt{\tau_1^2 + \sigma^2} \sqrt{v_{2ij}^2\tau_2^2 + v_{2ij}}}.$$

In the case of conditional independence ($\rho \equiv 0$), the approximate marginal correlation function ρ_{12} also equals zero. In case $\rho \equiv 1$, this model reduces to a shared parameter model, with scale parameter λ equal to τ_1/τ_2 . Standard software can be used to obtain parameter estimates for this bivariate model (e.g. SAS-procedure NLMIXED).

3 Extension to High-Dimensional Data

Assume we have m sequences $\mathbf{Y}_{ki} = (Y_{ki1}, Y_{ki2}, \dots, Y_{kin})$ of n outcomes for individual i . The sequences \mathbf{Y}_{ki} can be either continuous or binary. The m sequences can then be jointly modelled by specifying a joint distribution for the random effects, similar as in (??) and (??), but with an m -dimensional random effects vector \mathbf{b}_i . For each continuous component an identity link function can be used, whereas for each binary component a logit link can be used. However, while these (generalized linear) mixed models are very easily extended to modelling high-dimensional data, computational problems often arise when m increases. In this case, rather than considering the full likelihood contribution for each subject i , i.e., $l_i(\boldsymbol{\Theta}|\mathbf{Y}_{1i}, \mathbf{Y}_{2i}, \dots, \mathbf{Y}_{mi})$, we avoid the computational complexity by using a pseudo-likelihood function, as proposed by Fieuwis and Verbeke (2005). The full likelihood contribution for subject i is replaced by the following pseudo-likelihood function

$$pl_i = \sum_{k=1}^{m-1} \sum_{l=k+1}^m \ln \ell_{ikl}(\boldsymbol{\Theta}|\mathbf{Y}_{ki}, \mathbf{Y}_{li}). \tag{4}$$

Inference for $\boldsymbol{\Theta}$ follows from pseudo-likelihood theory, and is based on a sandwich-type robust variance estimator (Arnold and Strauss 1991). The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\Theta}}$ is given by

$$\sqrt{N}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \sim N(\mathbf{0}, J(\boldsymbol{\Theta})^{-1}K(\boldsymbol{\Theta})J(\boldsymbol{\Theta})^{-1})$$

where $J = J(\Theta)$ is a matrix with elements defined by

$$J_{pq} = - \sum_{k=1}^{m-1} \sum_{l=k+1}^m E \left(\frac{\partial^2 \ln \ell_{ikl}(\Theta | \mathbf{Y}_{ki}, \mathbf{Y}_{li})}{\partial \theta_p \partial \theta_q} \right),$$

and $K = K(\Theta)$ is a symmetric matrix with elements

$$K_{pq} = - \sum_{k=1}^{m-1} \sum_{l=k+1}^m E \left(\frac{\partial \ln \ell_{ikl}(\Theta | \mathbf{Y}_{ki}, \mathbf{Y}_{li})}{\partial \theta_p} \frac{\partial \ln \ell_{ikl}(\Theta | \mathbf{Y}_{ki}, \mathbf{Y}_{li})}{\partial \theta_q} \right).$$

The pseudo-likelihood methodology is very general and flexible. It can be found in many applications and fields of interest. It has been most advantageously used in the spatial data context, where the full likelihood distribution is typically cumbersome. But also in the context where maximum likelihood methods are not feasible, e.g. due to excessive computation requirements, the pseudo-likelihood is an appealing methodology (Aerts et al. 2002, Faes et al. 2004). An important advantage of the pseudo-likelihood approach is the close connection with likelihood, which enables Geys, Molenberghs and Ryan (1999) to construct pseudo-likelihood ratio test statistics that have easy-to-compute expressions and intuitively appealing limiting distributions.

In the pairwise approach as proposed by Fieuws and Verbeke (2005), all pairwise bivariate GLMMs are fitted first, resulting in different parameter estimates for each pair of endpoints $\Theta^* = (\Theta_{1,2}, \Theta_{1,3}, \dots, \Theta_{m-1,m})$. From these models the estimates of the parameters Θ of the joint model are derived by taking averages over estimates obtained from the different bivariate models, or $\hat{\Theta} = A\Theta^*$ (for an appropriate weight matrix A). The covariance matrix of Θ is then given as $A\Sigma(\Theta^*)A'$, where $\Sigma(\Theta^*)$ is the covariance matrix of Θ^* . This method simplifies a computational very challenging problem, making the approach very attractive. A possible disadvantage is that no (pseudo)-likelihood is directly available since all pairwise GLMMs are fitted separately. Also, one might lose some efficiency due to repeatedly estimating of the same parameters. Alternatively, one could estimate all parameters in the pairwise pseudo-likelihood, as defined by (??), simultaneously.

4 Irwin Study

The data considered here come from a 3-day repeated dose toxicity study, and was introduced in Section 2. The purpose of this study is to determine and assess the effects of the chemical on general activity and behavior. For illustration, eight endpoints were selected (4 continuous and 4 binary) and a joint model was estimated. As before, we denote Y_{kij} the j th outcome of the k th response for subject i . The model for each response k is specified as

$$h_k^{-1}(\mu_i) = \eta_{ik} = \beta_{0k} + \beta_{1k}g_i + \beta_{2k}t_i + \beta_{3k}d_i + \beta_{4k}t_id_i + \beta_{5k}g_it_i + \beta_{6k}g_id_i + b_{ik},$$

where h_k is the identity link in case of a continuous endpoint and the logit link in case of a binary endpoint, g_i is an indicator variable taking value 1 for the rats in the vehicle group, t_i is the time after exposure, and d_i is the day of the experiment.

Correlation between the responses at different time points are modelled by inclusion of the random intercepts b_{ik} . Correlation among the different responses are modelled by specifying a joint distribution for the random intercepts

$$\mathbf{b}_i = \begin{pmatrix} b_{i1} \\ b_{i2} \\ \dots \\ b_{i8} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_1^2 & \rho_{12}\tau_1\tau_2 & \dots & \rho_{18}\tau_1\tau_8 \\ \rho_{12}\tau_1\tau_2 & \tau_2^2 & \dots & \rho_{28}\tau_2\tau_8 \\ \dots & \dots & \dots & \dots \\ \rho_{18}\tau_1\tau_8 & \rho_{28}\tau_2\tau_8 & \dots & \tau_8^2 \end{pmatrix} \right\}.$$

Because of the computational complexity, the full likelihood is replaced by a pseudo-likelihood function, as described before.

Table ?? shows the estimated variance-covariance matrix. Based on these results, the correlation matrix among the endpoints can be derived, joint dose-response models are estimated and joint Wald- and pseudo-likelihood ratio tests are performed.

TABLE 1. Estimated correlation matrix for random effects using the pairwise approach. Values on the diagonal are the variances of the random effect.

Grip Strength	0.01							
Pina Reflex	-0.22	1.35						
Pupil Size	-0.21	0.24	3.16					
Sedation	0.04	-0.06	-0.36	2.64				
Temperature	0.11	-0.41	-0.82	0.27	0.12			
Toe Pinch	0.75	-0.56	-0.39	0.07	0.37	2.05		
Vertical Hind	-0.09	0.13	0.39	0.16	-0.54	-0.53	1.03	
Vocalization	-0.61	-0.31	-0.03	0.83	0.24	-0.34	-0.08	35.48

Acknowledgments: We gratefully acknowledge support from the Institute for the Promotion of Innovation by Science and Technology (IWT) in Flanders, Belgium and from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

References

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*, Chapman and Hall.
- Arnold, B.C., and Strauss, D. (1991). Pseudolikelihood estimation: Some examples. *Sankhyā, Series B* **53**, 233-243.
- Faes, C., Geys, H., Aerts, M., Molenberghs, G., and Catalano, P. (2004). Modelling combined continuous and ordinal outcomes in a clustered setting. *Journal of Agricultural, Biological and Environmental Statistics* **11**, 305-322.
- Fieuws, S., and Verbeke, G. (2005). Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics* **62**, 424-431.

Geys, H., Molenberghs, G., and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicity. *Journal of the American Statistical Association* **94**, 734745.

Molenberghs, G., and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. Springer Series in Statistics. New York: Springer.

From Dunkirk to Barcelona with GLIM[®]. A tribute to least-squares

Antoine de Falguerolles¹

¹ Laboratoire de Statistique et Probabilités, Université Paul Sabatier 118, route de Narbonne, F-31062 Toulouse cedex 9.

Abstract: Adrien-Marie Legendre (1752-1833) is the author of the first publication on the method of least-squares, of which he coined the name (1805, 1806). In his paper, Legendre presented an illustrative example in geodetics where he used the famous data set of measures taken in some places located from Dunkirk to Barcelona on the meridian of Paris. Legendre also assumed correlated errors and solved the associated generalised least-squares problem. This paper has two objectives. The first is to present Legendre's data modelling in contemporary notation. The second is to illustrate the early acceptance of the method of least-squares in Spain. An interesting mid-nineteenth century example is found in the work published under the authority of the *Comisión del Mapa de España* (Commission of the map of Spain).

Keywords: Least-squares, generalised least-squares, geodetics.

1 Introduction

In 1805, Adrien Marie Legendre published a book with an appendix presenting the method of least-squares. A second edition followed in 1806. Carl Friedrich Gauss, who has always claimed that he had used least-squares since 1795, published the method (with masterly add-ons) in 1809. The question of priority (see Stigler, 1999, Chapter 17) is not considered in this paper. The focus is on Legendre's statistical modelling. Legendre's appendix on least-squares gives a comprehensive presentation of the method in the context of multiple regression: the criterion is clearly defined for a set of linear equations with unknown coefficients; first and second order optimality conditions are considered; finally, a numerical example is treated. The example considers a famous data set in the history of metrology, namely the measures made along the meridian of Paris in the late eighteenth century by Jean-Baptiste Delambre (1749–1822) et Pierre Méchain (1744–1804).

Legendre's example is not just illustrative. Legendre entertains two models. Both involve two explanatory variables. One is not fitted, possibly because it is non-linear in its parameters. The second model, which is fully treated, is a standard linear regression. However, Legendre notices that, due to the spatial structure of the data, the errors are correlated. Therefore he rejects the use of ordinary least-squares and uses an *ad hoc* method. It turns out that he applies generalised least-squares, which he performs by data augmentation (Falguerolles et Pinchon, 2006).

It is difficult to assess the time span needed for a method to escape the circle of its inventors. One generation? Two? In 1858, the method was used in Spain in the context

of geodesic measurements conducted by the *Comisión del Mapa de España* (Laussedat, 1860, and Soler and Ruíz-Morales, 2006). Again, the example is a masterpiece.

This paper is structured as follows. Section 2 presents the data analysed by Legendre (and earlier by Gauss). In section 3, the multiplicative nature of the parameters in the model partly entertained by Legendre is discussed; a similar model considered earlier by Johann Tobias Mayer (1723–1762) is recalled; an adaptation of the iwls algorithm (Aitkin *et al.*, 2005) is outlined. Section 4 concentrates on the model fully entertained by Legendre, with special interest to its fitting by generalised least-squares. Section 5 recalls the modelling done during the exemplary franco-iberic collaboration involving the French Aimé Laussedat (1819–1907) and the Spaniard Francisco Ibáñez e Ibáñez de Ibero (1821–1891) in the mid-nineteenth century.

In order to focus attention on the modelling, modern notation and language are used. But the reader is reminded that, in Legendre, there is no such term as regression or normal equations, no formal probabilistic structure for the errors, no use of matrix algebra ...

2 The meridian data and problem

For lack of space the data considered by Legendre are not reported here. They can be found in Stigler (1999) and Falguerolles *et Pinchon* (2006). They consist in the names of five places of observations on the meridian of Paris (Dunkirk, the Panthéon in Paris, Évaux, Carcassonne and Montjuïc), their latitudes and the length of the arcs connecting adjacent places. The data were widely circulated even outside France. Published, in the *Allgemeine Geographische Ephemeriden* in 1799, they came to the attention of Gauss (Stigler, 1999, Chapter 17). The notation in Legendre's appendix is as follows: i ($i = 0, \dots, 4$) are the 5 places of observation; L_i , their latitudes; S_i , the lengths of the 4 arcs i ($i = 1, \dots, 4$) between the places of latitudes L_{i-1} et L_i .

Legendre (like Gauss and many scientists earlier) addresses a difficult problem in geodetics, the discipline that deals with the measurement and representation of the earth. The mathematical approximation to the earth, the reference ellipsoid, is described by its equatorial radius (semi-major axis) a and its flattening $f = (a - b)/a$ where b is the polar radius (semi-minor axis). Now, given some measures of the length of arc segments between adjacent points located along the same meridian, how can (functions of) the geodetics coefficients be estimated? This clearly involves the computing of approximations of elliptic integrals.

In his appendix, Legendre entertains two regression models, see section 3 and section 4 below, the linear predictor having the general form:

$$\mu_i(\beta) = [\text{offset}] + [\beta_0] + \beta_1 x_i^1 + \beta_2 x_i^2$$

where $[\star]$ denotes a term which may be absent.

3 Legendre's first formula

The model presented but not fitted by Legendre has:

- Response variable: $y_i = S_i$.

- Linear predictor:

$$\mu_i(\beta) = \beta_1 (L_i - L_{i-1}) + \beta_2 \left(-\frac{270}{\pi}\right) \sin(L_i - L_{i-1}) \cos(L_i + L_{i-1}).$$

Here, β_1 represents the unknown length of a degree at the 45° parallel and $\gamma = \frac{\beta_2}{\beta_1}$ the unknown flattening as defined by Legendre $((a - b)/b)$.

3.1 The multiplicative constraint

One characteristic of this model is that it looks linear but has one regression coefficient which is the product of a coefficient already present by another unknown coefficient: $\mu_i(\beta, \gamma) = [\text{offset}] + [\beta_0] + \beta_1 x_i^1 + (\beta_1 \gamma) x_i^2 = [\text{offset}] + [\beta_0] + \beta_1 (x_i^1 + \gamma x_i^2)$. This multiplicative constraint may explain why Legendre did not pursue it and preferred the linear model presented in section 4.

3.2 Mayer’s earlier example

This multiplicative structure can be found in a model which was fitted by Mayer as early as 1750. The eponym example is addressed to the modelling of lunar data (see Farebrother, 1998, Chapter 1) where Mayers considers the following relationship:

$$\beta - (90 - h_i) \approx \alpha \sin(g_i - k_i) - \alpha \sin(\theta) \cos(g_i - k_i)$$

where g_i et h_i are observed data and the k_i are read in lunar tables computed by Euler. This clearly parallels the structure of Legendre’s multiplicative model for the following choice:

- Response variable: $y_i = h_i - 90$.
- Linear predictor: $\mu_i(\beta, \gamma) = \beta_0 + \beta_1 x_i^1 + \beta_1 \gamma x_i^2 = \beta_0 + \beta_1 (x_i^1 + \gamma x_i^2)$ where $\beta_0 = -\beta$, $\beta_1 = \alpha$, and $\gamma = \sin(\theta)$ are the parameters of interest, the explanatory variables being $x_i^1 = \sin(g_i - k_i)$ and $x_i^2 = \cos(g_i - k_i)$.

Mayer’s method of estimation is skilful: the data being clustered in three groups and averaged within each group, the resulting set of three linear equations can be solved to obtain estimations of the unknown coefficients. Note that this procedure has an invariance property with respect to the selection of the response variable. However, the estimations depend heavily on the determination of a proper partition of the observations: an interesting problem in cluster analysis.

3.3 An adaptation of iwls for multiplicative models

The simplest is to estimate γ from the ratio of two coefficients in an ordinary regression. For an explicit fit, iwls can be considered. The central idea in this widespread method is to use least-squares with a working response and possibly working weights, revised at each iteration, the revision being made in the light of current estimates of the unknown coefficients. Here, denoting by θ the vector of unknown coefficients $([\beta_0,] \beta_1, \gamma)'$, and by \underline{t} its current estimation $([b_0,] b_1, g)'$, it turns out that the minimum of $\sum_i (y_i - \mu_i(\theta))^2$ can be also obtained by iterative least-squares regressions of the working response variate $z_i(\underline{t}) = y_i + b_1 g x_i^2$ onto the working explanatory variables $x_i^1 + g x_i^2$, $b_1 x_i^2$, and, if present, the intercept and the offset. At convergence, a by-product of the fit is the (un)scaled asymptotic covariance matrix of the estimators.

4 Legendre's generalised regression

In this section, the model which Legendre fully entertained is recalled. Special attention is directed to the structure of correlation between errors which he assumed and to the estimation procedure which he used.

4.1 Legendre's second formula

The model is constructed as follows:

- Response variable : $y_i = L_i - L_{i-1}$.
- Linear predictor:

$$\mu_i(\underline{\beta}) = \frac{S_i}{K'} + \beta_1 \frac{S_i}{K'} + \beta_2 K'' \sin(L_i - L_{i-1}) \cos(L_i + L_{i-1})$$

where K' and K'' are known coefficients.

- Moving average dependence: $U_i = E_i - E_{i-1}$ with independent E_i .

Note that there is no intercept in the predictor and that $\frac{S_i}{K'}$ is an "offset".

An *aficionados* of linear regression may object that the latitudes L_i appear on both sides of the model, in the response and in the explanatory variables. But this is the price to pay for a direct interpretation of the regression coefficients: β_2 is the flattening of the earth, as defined by Legendre $((a - b)/b)$, and β_1 a correction term for the approximation of the length of a degree at the 45° parallel ($D = 28500/(1 + \beta_1)$).

4.2 Legendre's generalised regression

Thoroughly revisited, the regression model specified by Legendre is $\underline{Y} = \underline{\mathbf{X}}\underline{\beta} + \underline{U}$ where \underline{Y} and \underline{U} are random vectors of dimension n , where $\underline{\mathbf{X}}$ is a full rank design matrix of dimension (n, p) , and where $\underline{\beta}$ is a coefficients vector of dimension p . Moreover, the serial correlation considered by Legendre is particular case of the following setting: $\underline{U} = \underline{\mathbf{L}}\underline{E}$ where the known matrix $\underline{\mathbf{L}}$ is of dimension $(n, n + k)$ and dimension n and where the random vector \underline{E} of dimension $n + k$ has independent coordinates with null expectation and constant variance.

Thus the following generalised least-squares problem is to be considered:

$$\underline{b}^* = \arg \min \|\underline{y} - \underline{\mathbf{X}}\underline{b}\|_{(\underline{\mathbf{L}}\underline{\mathbf{L}}')^{-1}}^2$$

Now let $\underline{\ell}$ be any matrix of dimension $(k, n + k)$ of rank k such that $\underline{\ell}\underline{\mathbf{L}}' = \mathbf{0}$ and the associated assumption that $\underline{\ell}\underline{e} = \underline{0}$. Consider now the augmented data:

$$\underline{y}_1 = \begin{bmatrix} \underline{\mathbf{L}} \\ \underline{\ell} \end{bmatrix}^{-1} \begin{bmatrix} \underline{y} \\ \underline{0} \end{bmatrix} \quad \underline{\mathbf{X}}_1 = \begin{bmatrix} \underline{\mathbf{L}} \\ \underline{\ell} \end{bmatrix}^{-1} \begin{bmatrix} \underline{\mathbf{X}} \\ \mathbf{0} \end{bmatrix}$$

It is easily seen that ordinary least-squares applied to the augmented data obtained as above is equivalent to generalised least-squares on the observed data:

$$\|\underline{y}_1 - \underline{\mathbf{X}}_1\underline{b}\|^2 = \|\underline{y} - \underline{\mathbf{X}}\underline{b}\|_{(\underline{\mathbf{L}}\underline{\mathbf{L}}')^{-1}}^2$$

In the simple case of the order 1 moving average considered by Legendre, it is enough to fix $\sum_i e_i = 0$ (or $\underline{\ell}' = (1, 1, \dots, 1)$) as he did.

5 Simple linear regression at the Comisión del Mapa de España

Fifty years after Legendre's publication, a report of the *Comisión del Mapa de España*, relates how least-squares were routinely used in 1856 by colonel Ibañez and colonel Saavedra to gauge a measuring device built by Brunner. The report was translated from Spanish to French by Aimé Laussedat, by then a captain of the French military engineers, who had close contacts with his Spanish colleagues (Laussedat, 1860). Brunner and son was a famous instrument maker in Paris and the instrument was to be used for measuring bases in triangulations. An example is the Madrideo base (1858).

The measuring device comprised two matched rules, one in platinum and one in an alloy of zinc and copper (Laussedat, 1860, chapter 1). Colonel Ibañez and colonel Saavedra wanted to assess the dilation of the two rules for varying temperatures and to compare their lengths to the standard meter kept at the *Observatoire Astronomique de Paris*.

The modelling is arranged in two steps. In step I, an ordinary regression involving a pre-processed response variable is performed and the regression coefficients are obtained. In step II, estimates for the parameters of interest are derived from the regression parameters. The pre-processing is critical since it must ensure that the response and explanatory variables can be measured or computed from observable measurements, and that the parameters of interest are computable functions of the regression coefficients.

In passing, the early appearance of the term “normal equations” in the context of least-squares is to be noted.

6 Concluding remarks

The modelling works which have been considered are exemplary in several respects. In all, great care is taken in deriving a working mathematical description of the physical process under consideration including sources of measurement errors. The treatment of errors by a MA(1) process is particularly innovative. Having solved the “normal equations”, all authors compute the fitted values and examine the residuals in details . . . Their authors are to be recommended for honorary membership to the Statistical Modelling Society!

References

- Aitkin, M., Francis B. and Hinde J. (2005). *Statistical modelling in GLIM 4* (2nd edition). Oxford University Press.
- Falguerolles, A. de and Pinchon, D. (2006). Une commémoration du bicentenaire de la publication (1805-1806) de la méthode des moindres carrés par Adrien Marie Legendre. *Journal de la Société Française de Statistique* tome **147**(2), 81–105.
- Farebrother, R. W. (1998). *Fitting linear relationships, a History of the calculus of observations 1750-1900*. New York: Springer.

- Laussedat, A. (translator) (1860). *Expériences faites avec l'appareil à mesurer les bases appartenant à la commission de la carte d'Espagne*. Paris: Librairie militaire.
- Legendre, A. M. (1806). *Nouvelles méthodes pour la détermination des orbites des comètes avec un supplément contenant divers perfectionnemens de ces méthodes et leur application aux deux comètes de 1805*. Paris: Courcier.
- Soler, T. and Ruíz-Morales, M. (2006). Letters from Carlos Ibáñez e Ibáñez de Ibero to Aimé Laussedat: new sources for the history of the nineteenth century geodesy. *Journal of Geodesy* **80**, 313–321.
- Stigler S. M. (1999). *Statistics on the table: the history of statistical concepts and methods*. Cambridge: Harvard University Press.

Modelling and Analysis of Superparasitism Data

John S. Fenlon¹ and Malcolm J. Faddy²

¹ Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

² School of Mathematical Sciences, QUT, Brisbane, QLD 4001, Australia

Abstract: We present a simple solution to the problem of estimating the parameters of a model for superparasitism avoidance that has hitherto been considered quite difficult. The model incorporates both a covariate (host density) and a temporal component to produce a remarkably parsimonious structure with only 7 parameters to describe a data set of over 3000 observations. This model showed that the level of superparasitism declines with increasing host density, and the rates of parasitism decline over time. A refinement using a mixed model to accommodate several large outliers showed that a small number of parasitoids behaved rather differently by parasitising at much higher rates early on but stopping completely after only a few hours.

Keywords: extended Poisson process modelling; superparasitism avoidance; outliers.

1 Introduction

Superparasitism is a phenomenon that occurs when a parasitoid lays more than one egg in a host, but only one egg can mature into an adult within the host. Not surprisingly, there are thought to be biochemical mechanisms that inhibit the parasitoid from laying eggs in a host that already has had eggs oviposited, and there are probabilistic models to describe such avoidance of superparasitism. Bakker *et al.* (1972) considered several models to describe superparasitism, and these were further examined by Rogers (1975) and Griffiths (1977). Daley and Maindonald (1989) set out a very general modelling framework for superparasitism. In this paper use is made of data from Toussidou (2002) on the number of eggs laid by the parasitoid *Aphidus colemani*, a member of the family Aphidiidae (Hymenoptera: Braconidae), which is used in biological control programmes to limit the aphid pest *Aphis gossypii*, many strains of which are pesticide resistant. This dataset is quite extensive, relating to the level of parasitism of individual *Aphidus colemani* foraging for different time periods ($\frac{1}{2}$, 1, 2, 5 and 24 hours) at a range of different aphid densities, and contains some 3000+ entries. The aim of the modelling is to produce a description of the parasitism process over the full 24 hour period that highlights different temporal patterns of parasitism during this time. This is particularly challenging as the basic model for avoidance of superparasitism results in a probability distribution for the number of eggs laid within an individual host aphid that is under-dispersed relative to the Poisson distribution, while the data exhibit varying degrees of dispersion.

2 Modelling

The basic model for superparasitism due to Bakker *et al.* (1972) essentially considers a Markov stochastic process $\{X(t); t \geq 0\}$, with $X(0) = 0$, for the number of eggs laid in an individual host aphid over time, where:

$$P\{X(t + \delta t) = n + 1 | X(t) = n\} = \lambda_n \delta t + o(\delta t)$$

and

$$P\{X(t + \delta t) = n | X(t) = n\} = 1 - \lambda_n \delta t + o(\delta t).$$

If $\lambda_0 > \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots$, then this process results in avoidance of superparasitism (since the probability of an egg being laid in a host with one or more eggs already laid is less than that for a host which is free of eggs), and a probability distribution of $X(t)$ that is under-dispersed relative to the Poisson distribution (*i.e.*, the variance of $X(t)$ is less than the mean). The actual distribution of $X(t)$ can be obtained from the expression (Cox and Miller, 1965, Chapter 4):

$$[P\{X(t) = 0\} \quad P\{X(t) = 1\} \quad \dots \quad P\{X(t) = n\}] = [1 \quad 0 \quad \dots \quad 0] \exp(\mathbf{Q}t)$$

where the matrix \mathbf{Q} takes the form:

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \dots & 0 \\ 0 & -\lambda_1 & \lambda_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_n \end{pmatrix} \quad (1)$$

The simplest sequence $\lambda_0, \lambda_1, \lambda_2, \dots$ has $\lambda_n = \lambda_1$ for all $n \geq 1$, resulting in a two parameter model. A further complication is that the rate parameters λ_0 and λ_1 need to depend on the number of host aphids, h say, available to the parasitoid, as rates of parasitism of individual hosts would decline with increasing h ; the forms used were $\lambda_i = a_i \exp(-b_i h)$ for $i = 0, 1$. And different rate parameters could be expected to apply over the different periods of foraging by the parasitoid: $0 - \frac{1}{2}$ hour, $\frac{1}{2} - 1$ hour, $1 - 2$ hours, $2 - 5$ hours and $5 - 24$ hours, resulting in 5 pairs (λ_0, λ_1) of rate parameters with the above specification, and 5 matrices $\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_5$ defined according to equation (1). With $t_1 = \frac{1}{2}, t_2 = 1, t_3 = 2, t_4 = 5$ and $t_5 = 24$, the probabilities $[P\{X(t_i) = 0\} \quad P\{X(t_i) = 1\} \quad \dots \quad P\{X(t_i) = n\}]$ will then be given by:

$$[1 \quad 0 \quad \dots \quad 0] \exp(\mathbf{Q}_1 t_1) \exp(\mathbf{Q}_2 (t_2 - t_1)) \dots \exp(\mathbf{Q}_i (t_i - t_{i-1})) \quad (2)$$

for $i = 1, 2, \dots, 5$.

3 Results

Each assay (at a particular aphid density and over a specified time period) was carried out independently of all the others so that the likelihood of the data could be obtained simply by multiplying together probabilities derived from equation (2). Maximum

likelihood estimation of all the parameters specifying the model could then be carried out, and possible simplifications considered. One simplification was that the a_0 and a_1 parameters specifying λ_0 and λ_1 (above) could be put equal *within* each time period, and another was that the other parameters b_0 and b_1 could be put equal *across* the different time periods. The resulting parameter estimates (with asymptotic standard errors in brackets) were:

$$\hat{b}_0 = 0.0047(0.00026), \quad \hat{b}_1 = 0.0099(0.00052)$$

$$\hat{a} = \begin{cases} 1.77 (0.110) & \text{for } 0 - \frac{1}{2} \text{ hour} \\ 0.91 (0.162) & \text{for } \frac{1}{2} - 1 \text{ hour} \\ 0.28 (0.093) & \text{for } 1 - 2 \text{ hours} \\ 0.22 (0.036) & \text{for } 2 - 5 \text{ hours} \\ 0.05 (0.007) & \text{for } 5 - 24 \text{ hours.} \end{cases}$$

The ordering of the estimates $\hat{b}_1 > \hat{b}_0$ shows that avoidance of superparasitism is more likely (*i.e.*, $\lambda_0 > \lambda_1$) at higher host densities (h), and the ordering of the \hat{a} estimates over the foraging periods shows that there is an almost halving in the rates of ovipositing after $\frac{1}{2}$ hour followed by a greater reduction for 1 – 5 hours ending with very low rates after 5 hours.

An examination of the fit of this model showed up a number of residuals with rather low exceedance probabilities - 9 (out of 3097 observations) with exceedance probability 0.001. This might be due to some heterogeneity between the different assays - *i.e.*, some parasitoids ovipositing at quite different rates to others. Such a possibility could be tested by fitting a model defined by a mixture of two different processes; this resulted in a much improved fit and a more acceptable pattern of residual behaviour. The estimates from this mixture model showed that some 90% of the parasitoids behaved in a similar manner to the single process model described above, but 10% of them behaved in a markedly different manner with much higher rates of parasitism over the first 5 hours but no parasitism at all after this time.

4 Discussion

This paper has extended the ‘classical’ superparasitism avoidance model of Bakker *et al.*, Rogers, Griffiths and others to incorporate host density dependence and temporal changes in egg laying activity. Most noticeable among the findings were that avoidance of superparasitism becomes more marked at higher host densities, and that as time progresses the rate of oviposition generally declines over the 24 hour period although for a few ‘rogue’ parasitoids more prodigious behaviour is apparent with much greater egg laying activity early on before a premature end after only a few hours’ activity.

References

- Bakker, K, Eijsackers, H.J.P., van Lentern, J.C. and Meelis, E. (1972). Some models describing the distribution of eggs of the parasite *Pseudecoila Bochei* (Hymn., Cynip.) over its hosts, larvae of *Drosophila melanogaster*. *Oecologia* **10**, 29-57.

- Cox, D.R., and Miller, H.D. (1965). *The Theory of Stochastic Processes*. London: Methuen.
- Daley, D.J. and Maindonald, J.H. (1989). A unified view of models describing the avoidance of superparasitism. *IMA J. Maths. Appl. Med. & Biol.* **6**, 161-178.
- Griffiths, D. (1977). Avoidance-modified generalised distributions and their application to studies of superparasitism. *Biometrics* **33**, 103-112.
- Rogers, D. (1975). A model for avoidance of superparasitism by solitary insect parasitoids. *J. Anim. Ecol.* **44**, 623-638.
- Toussidou, M. (2002). Foraging behaviour of *Aphidius colemani* at different spatial scales. *Unpublished PhD thesis, Imperial College, University of London.*

Visualizing hypothesis tests in multivariate linear models

Michael Friendly¹ and John Fox²

¹ Psychology Department, York University, Toronto, Canada.

² Sociology Department, McMaster University, Hamilton, Canada.

Abstract: This paper describes graphical methods for multiple-response data within the framework of the multivariate linear model (MLM), aimed at understanding what is being tested in a multivariate test, and how factor/predictor effects are expressed across multiple response measures.

In particular, we describe HE plots, a new class of visualization methods for the MLM, designed to show the “size” and “shape” of covariation against a multivariate hypothesis (H), relative to covariation due to error (E). For more than two response variables, these relations can be visualized in HE plot matrices or in reduced-rank spaces that correspond to biplots and canonical discriminant spaces.

Keywords: biplot; data ellipse; MANOVA; linear hypotheses;

1 Introduction

This paper describes a new class of visualization methods for linear hypotheses in classical multivariate linear models introduced in Friendly (2006, 2007). The paper begins with a capsule summary of multivariate linear models; it proceeds to explain how dispersion matrices, for data, for variation due to hypotheses (H) and for error (E), can be represented by ellipses or ellipsoids, either in the data space or in a reduced-rank space.

These methods are implemented both in SAS software (Friendly, 2006) and in the `heplots` package for R (Fox et al. 2007).

2 Background

2.1 Multivariate linear models

We consider the multivariate linear model, $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$, where \mathbf{Y} is an $n \times p$ matrix for p response variables, \mathbf{X} is an $n \times q$ full rank model matrix, \mathbf{B} is the $q \times p$ matrix of model coefficients, and \mathbf{E} is the $n \times p$ matrix of errors. Under the assumption that the rows ϵ_i^\top of \mathbf{E} are *iid*, $\epsilon_i^\top \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with common covariance matrix $\mathbf{\Sigma}$, the least squares estimator is $\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$.

All linear hypotheses for subsets of the model effects can be expressed in the form $H_0 : \mathbf{L}\mathbf{B} = \mathbf{0}$, where \mathbf{L} is a $r \times p$ hypothesis matrix of rank r . Any such hypothesis is tested in terms of the $p \times p$ hypothesis sums of squares and products matrix, $\mathbf{SSP}_H =$

$(\mathbf{L}\mathbf{B})^\top [\mathbf{L}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{L}^\top]^{-1} (\mathbf{L}\mathbf{B})$ and a corresponding error matrix $\mathbf{SSP}_E = \widehat{\mathbf{E}}^\top \widehat{\mathbf{E}}$. All multivariate tests are based on the $s = \min(r, p)$ nonzero latent roots $\lambda_1 > \lambda_2 > \dots > \lambda_s$ of the matrix \mathbf{SSP}_H relative to the matrix \mathbf{SSP}_E , or equivalently, the ordinary latent roots of $\mathbf{SSP}_H \mathbf{SSP}_E^{-1}$. The corresponding latent vectors give a set of s orthogonal linear combinations of the responses that produce maximal univariate F statistics for the hypothesis.

We exploit these facts in the visualizations described below. In particular, Roy's maximum root statistic, λ_1 provides a simple visual criterion for judging when \mathbf{SSP}_H is large enough relative to \mathbf{SSP}_E to reject H_0 . The latent vectors of $\mathbf{SSP}_H \mathbf{SSP}_E^{-1}$ help to interpret the results, either in data space or in the reduced-rank canonical discriminant space.

2.2 Data ellipses and ellipsoids

The data ellipse, described by Dempster (1969) and Monette (1990) is a remarkably simple device for visualizing the relationship between two variables, Y_1 and Y_2 in terms of their first and second moments. Let $D^2(\mathbf{y}) = (\mathbf{y} - \bar{\mathbf{y}})^\top \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}})$ be the squared Mahalanobis distance of the point $\mathbf{y} = (y_1, y_2)^\top$ from the centroid, $\bar{\mathbf{y}}$. Then, the data ellipse, \mathcal{E}_c of size c is the set of all points \mathbf{y} with $D^2(\mathbf{y})$ less than or equal to c^2 :

$$\mathcal{E}_c(\mathbf{y}; \mathbf{S}, \bar{\mathbf{y}}) \equiv \{\mathbf{y} : (\mathbf{y} - \bar{\mathbf{y}})^\top \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}}) \leq c^2\}, \quad (1)$$

where \mathbf{S} is the sample variance-covariance matrix.

3 Hypothesis and Error (HE) plots

Hypothesis and Error (HE) plots employ ellipses to represent the sum of squares and product matrices used in all multivariate tests of linear hypotheses. The \mathbf{E} ellipse can be viewed as the data ellipse of residuals, obtained by dividing \mathbf{SSP}_E by its degrees of freedom ($df_e = n - p$), and centered at the grand means, allowing individual factor means to be shown on the same plot to facilitate interpretation.

The \mathbf{H} ellipse for any multivariate linear hypothesis is a similar representation of the \mathbf{SSP}_H matrix, corresponding to a data ellipse of the fitted (predicted) values under that hypothesis, but it is useful to provide two different scalings of the \mathbf{H} ellipse:

significance-sized scaling, where the \mathbf{H} ellipse will protrude beyond the \mathbf{E} ellipse *iff* the corresponding hypothesis can be rejected by the Roy maximum-root test. This is produced by dividing \mathbf{SSP}_H by $\lambda_\alpha(df_e)$, where λ_α is the critical value of Roy's statistic for a test at level α .

effect-sized scaling, where the hypothesis ellipse is put on the same scale as the error ellipse, and approximately represents the data ellipse of the fitted (predicted) values under the alternative hypothesis. Here, \mathbf{SSP}_H is simply divided by df_e .

All of this extends straightforwardly to the 3D case via 3D ellipsoids, and to the nD case via pair-wise HE plots. As well, one can also construct reduced-rank displays by projecting into subspaces corresponding to optimal linear combinations of the responses, including biplot and canonical discriminant versions.

4 Examples

To economize on space, we show just a few simple examples here, using Anderson’s (Anderson, 1935) classic data on four measures of sepal and petal size in three species of iris flowers found in the Gaspé Peninsula.

Figure 1 shows (a) within-species data ellipses for Sepal and Petal length in the iris data, and (b) the corresponding view of the 2×2 portions of the \mathbf{H} and \mathbf{E} matrices for the model $\text{SepalLen SepalWid PetalLen PetalWid} = \text{Species}$, testing whether the species means are equal on all four variables. Here $p = 4$, $r = df_h = 2$ so there are $s = 2$ nonzero latent roots and corresponding (canonical) latent vectors, shown in the plot.

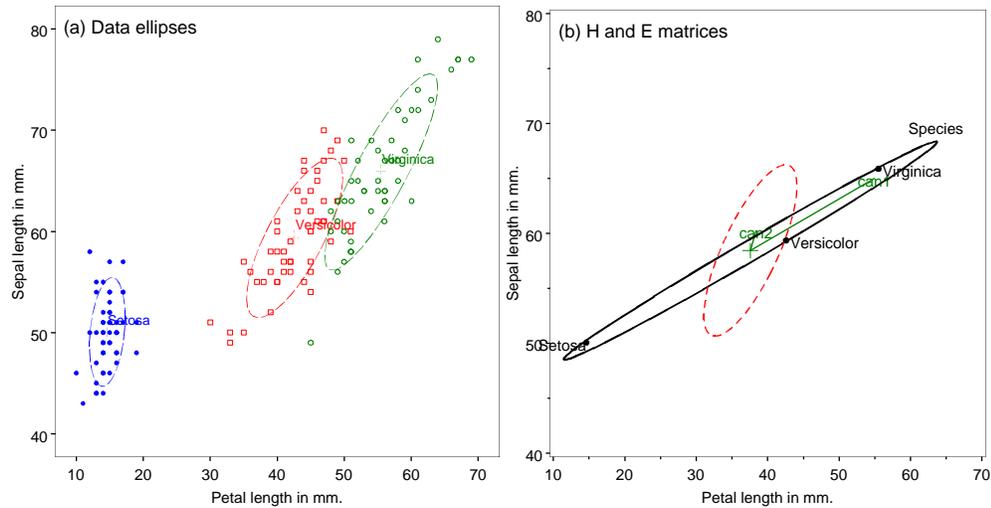


FIGURE 1. Data and HE plots for iris data, showing the relationship between sepal length and petal length in the iris data. (a) data ellipses; (b) \mathbf{H} and \mathbf{E} matrices (effect-size scaling). The green vectors show the projection of the canonical variates in data space.

For more than 2 variables, we can visualize the \mathbf{H} and \mathbf{E} variation in all bivariate views, in the form of an HE plot matrix (not shown for lack of space).

Alternatively, we can imagine projecting the \mathbf{H} and \mathbf{E} ellipses into the space of the canonical variates defined by the latent vectors of $\mathbf{SSP}_H \mathbf{SSP}_E^{-1}$, as shown in Figure 2. This gives a compact 2D view, in which it can be seen that nearly all of the between-species variation is contained in a single dimension.

References

Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society* **35**, 2–5.

Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, MA: Addison-Wesley.

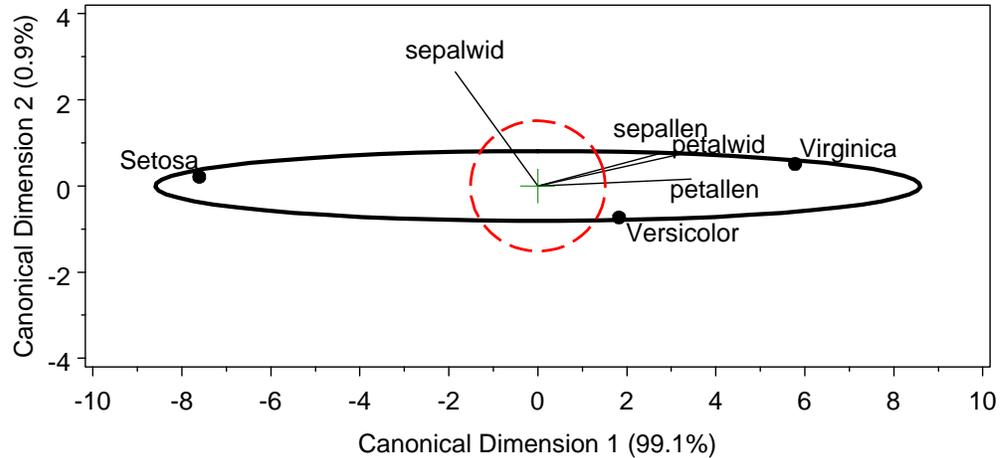


FIGURE 2. Canonical HE plot for the iris data. The lengths of variable vectors are proportional to the correlations of the observed variables with the canonical variates.

Fox, J., Friendly, M. and Monette, G. (2007). Visual hypothesis tests in multivariate linear models: The heplots package for R. In *DSC 2007: Directions in Statistical Computing*. Auckland, NZ.

Friendly, M. (2006). Data ellipses, HE plots and reduced-rank displays for multivariate linear models: SAS software and examples. *Journal of Statistical Software* **17**(6), 1–42.

Friendly, M. (2007). He plots for multivariate general linear models. *Journal of Computational and Graphical Statistics* **16**, In press.

Monette, G. (1990). Geometry of multiple regression and interactive 3-D graphics. In J. Fox and S. Long (Eds.), *Modern Methods of Data Analysis*, chap. 5, (pp. 209–256). Beverly Hills, CA: Sage Publications.

Image estimation from signal-dependent noise observations

M.J. García-Ligero¹, A. Hermoso-Carazo¹, J. Linares-Pérez¹ and S. Nakamori²

¹ Dpto. Estadística e I.O., Universidad de Granada, 18071 Granada, Spain mjgarcia@ugr.es, ahermoso@ugr.es, jlinares@ugr.es

² Department of Technology, Faculty of Education, Kagoshima University, 1-20-6, Kohri-moto, Kagoshima, 890-0065 Japan nakamori@edu.kagoshima-u.ac.jp

Abstract: We propose a recursive filtering algorithm to restore monochromatic images which are corrupted by a signal-dependent additive noise. We assume that the equation which describes the image field is not available and we obtain a filtering algorithm using the information provided by the covariance functions of the signal and noises which affect the measurement equation, and by the fourth-order moments of the signal. The proposed algorithm is applied to restore an image which is affected by signal-dependent additive noise.

Keywords: Signal-dependent noise; Filtering; Covariance information.

1 Introduction

The problem of image restoration consists of estimating the original image from a degraded version of itself. The degradations which affect the original image can be of different kinds although the most often are blur and noise, which is a stochastic phenomenon. Most researches in image restoration have been addressed to consider that the noise process which corrupts the image is additive and signal-independent. In this situation, the restoration problem has been treated by using the techniques of classical filtering theory. Specifically, the techniques of recursive estimation have deserved considerable attention by their advantages of computation. The estimation problem has been treated taking into account the full knowledge of the state-space model by using Kalman filter (Sezan et al. (1990)), as well as when the state-space model is not fully known. In this last case, using covariance information Nakamori et al. (2006) obtain a recursive fixed-interval smoothing algorithm which is applied to the restoration of an image degraded by additive gaussian white noise. However, it is known that a great number of physical processes such as images detected on film including natural scenes as well as many types of medical images respond to the case in which the noise process corrupting the image is signal-dependent. When the state-space model is known, the restoration problem in signal-dependent noise model have been considered by using different approaches.

In this paper, we consider the image restoration problem assuming that the state-space model is not known and the measurement equation responds to an additive signal-dependent noise. Considering this measurement equation, we obtain the filtering

algorithm assuming that the autocovariance function of the signal is known and can be expressed in semi-degenerate kernel form; moreover the second-order moments of the noises and the fourth-order moments of the signal are known. Under these assumptions we transform the additive signal-dependent noise model into a signal-independent noise and from the new measurement equation we obtain a recursive algorithm for image restoration. The recursive filter is obtained by an innovation approach which, as it is known, provides a simple derivation of the estimation algorithms. Since the goodness of the least mean squared estimators is measured by the filtering error variances, recursive formulas for them are also presented. Finally, the proposed filtering algorithm is applied to restore images corrupted by multiplicative and additive noises, as well as only corrupted by purely multiplicative noise.

2 Problem formulation

2.1 Observation model

The degraded image is described by the following measurement equation

$$z(k, l) = u(k, l) + u^\gamma(k, l)w_1(k, l) + w_2(k, l) \quad (1)$$

where $u(k, l)$ for $k, l \geq 1$ represents the gray level of the original image at the location (k, l) , $\gamma \in [0, 1]$ and $\{w_i(k, l); k, l \geq 1\}$, $i = 1, 2$, are white noises. To treat the restoration problem we assume the following hypotheses:

- The image process has zero-mean and its autocovariance function is expressed in a semi-degenerate kernel form; that is,

$$K(k, l, s, \xi) = E[u(k, l)u(s, \xi)] = \begin{cases} \alpha(k, l, \xi)\beta^T(s, l, \xi), & 1 \leq s \leq k, \\ \eta(k, l, \xi)\rho^T(s, l, \xi), & 1 \leq k \leq s, \end{cases}$$

where $\alpha(\cdot, l, \xi)$, $\beta(\cdot, l, \xi)$, $\eta(\cdot, l, \xi)$ and $\rho(\cdot, l, \xi)$ are known matrix functions. Moreover let us suppose that $Var[u^2(k, l)] = \sigma_{u^2}^2(k, l)$ are known $\forall k, l \geq 1$.

- $\{w_i(k, l); k, l \geq 1\}$, $i = 1, 2$, are white sequences with zero mean and $E[w_i(k, l)w_i(s, \xi)] = \sigma_i^2(k)\delta_K(k-s)\delta_K(l-\xi)$, $i = 1, 2$, being δ_K the Kronecker delta function.

- The image process and the noises are mutually independent.

The signal-dependent noise model (1) is rewritten as follows

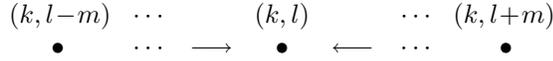
$$z(k, l) = u(k, l) + v(k, l), \quad k, l \geq 1, \quad (2)$$

where $v(k, l) = u^\gamma(k, l)w_1(k, l) + w_2(k, l)$ is a white noise with zero mean and variance function, $E[v^2(k, l)] = E[u^{2\gamma}(k, l)]\sigma_1^2(k) + \sigma_2^2(k)$. Since this variance depends on the unknown moment $E[u^{2\gamma}(k, l)]$, this is approximated using the second-order approximation in Taylor's expansion and so, we obtain the following approximation for the variance

$$E[v^2(k, l)] \simeq \sigma_v^2(k, l)\sigma_1^2(k) \left\{ (\alpha(k, l, l)\beta^T(k, l, l))^\gamma + \frac{\gamma(\gamma-1)}{2} (\alpha(k, l, l)\beta^T(k, l, l))^{\gamma-2} \sigma_{u^2}^2(k, l) \right\} + \sigma_2^2(k).$$

2.2 Restoration problem

We treat the restoration problem of an $N \times M$ discrete monochromatic image from the observations modelled as in (2). To estimate the pixel at the location (k, l) , the observations are made in symmetric and adjacent pixels to this, according to the following sketch



This is only valid for estimating $u(k, l)$ such that $1 \leq k \leq N$ and $m < l \leq M - m$. To estimate the boundary gray levels, we consider m like the maximum of the symmetric pixels around of the pixel which is estimated. So, by noting $z_m(k) = Col(z(k, l - m), \dots, z(k, l + m))$, the available measurements to estimate $u(k, l)$, $1 \leq k \leq N$, $1 \leq l \leq M$, are given by

$$z_m(k) = u_m(k) + v_m(k), \quad k \geq 1$$

where $u_m(k) = Col(u(k, l - m), \dots, u(k, l + m))$ and analogously, $v_m(k) = Col(v(k, l - m), \dots, v(k, l + m))$.

Hypothesis (I) and the properties of the process $\{v(k, l); k, l \geq 1\}$ lead us to the following properties of $\{u_m(k); k \geq 1\}$ and $\{v_m(k); k \geq 1\}$:

- $\{u_m(k); k \geq 1\}$ has zero mean and its autocovariance function can be expressed as $K_{mm}(k, s) = E[u_m(k)u_m^T(s)] = \alpha(k)\beta^T(s)$, $1 \leq s \leq k$, where $\alpha(k) = diag(\alpha(k, l - m), \dots, \alpha(k, l + m))$, $\beta(s) = (\beta(s, l - m), \dots, \beta(s, l + m))$, with the matrix functions $\alpha(k, l) = (\alpha(k, l, l - m), \dots, \alpha(k, l, l + m))$ and $\beta(s, l) = diag(\beta(s, l, l - m), \dots, \beta(s, l, l + m))$.
- $\{v_m(k); k \geq 1\}$ is a zero-mean white sequence with covariance function $E[v_m(k)v_m^T(k)] \simeq R_m(k) = diag(\sigma_v^2(k, l - m), \dots, \sigma_v^2(k, l + m))$.

3 Filtering algorithm

The filter of $u(k, l)$ for $1 \leq l \leq M$ is obtained by

$$\hat{u}(k, k, l) = \alpha(k, l)O(k, l), \quad k \geq 1,$$

where

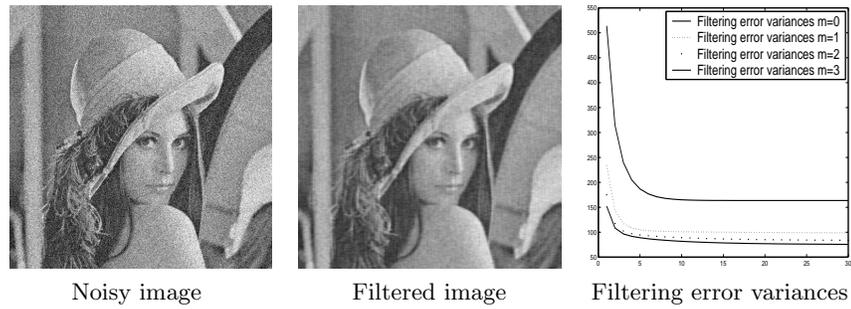
$$\begin{aligned}
 O(k, l) &= O(k-1, l) + J(k, l)\Pi^{-1}(k)\nu(k), \quad k \geq 1; \quad O(0, l) = 0, \\
 J(k, l) &= \beta^T(k, l) - r(k-1, l)\alpha^T(k), \quad k \geq 1, \\
 \nu(k) &= z_m(k) - \alpha(k)O_m(k-1), \quad k \geq 1, \\
 r(k, l) &= r(k-1, l) + J(k, l)\Pi^{-1}(k)J_m^T(k), \quad k \geq 1; \quad r(0, l) = 0, \\
 O_m(k) &= O_m(k-1) + J_m(k)\Pi^{-1}(k)\nu(k), \quad k \geq 1; \quad O_m(0) = 0, \\
 J_m(k) &= \beta^T(k) - r_m(k-1)\alpha^T(k), \quad k \geq 1, \\
 r_m(k) &= r_m(k-1) + J_m(k)\Pi^{-1}(k)J_m^T(k), \quad k \geq 1; \quad r_m(0) = 0, \\
 \Pi(k) &= \alpha(k)J_m(k) + R_m(k), \quad k \geq 1.
 \end{aligned}$$

Recursive expressions for the filtering error variances are

$$\begin{aligned}
 P(k, k, l, \xi) &= \alpha(k, l, \xi)\beta^T(k, l, \xi) - \alpha(k, l)d(k, l, \xi)\alpha^T(k, \xi), \quad k \geq 1, \\
 d(k, l, \xi) &= d(k-1, l, \xi) + J(k, l)\Pi^{-1}(k)J^T(k, \xi), \quad k \geq 1; \quad d(0, l, \xi) = 0.
 \end{aligned}$$

4 Implementation and results

We apply the proposed filtering algorithm to restore “Lenna.tiff” which has been degraded, according to model (1), by two zero-mean Gaussian noises with $\sigma_1^2(k) = 1$ and $\sigma_2^2(k) = 30^2$. We simulate the noisy image for $\gamma = 1/2$ and apply the filtering algorithm fixing $m = 3$. Also, we calculate the filtering error variances for different values of m , noting that like was hope the filtering error variances decrease when the values of m are on the increase.



Acknowledgments: This work was partially supported by “Ministerio de Educación y Ciencia” under contract MTM2005-03601.

References

- Sezan, M.I. and Tekalp, A.M. (1990). Survey of recent developments in digital restoration, *Optics Engineering* **24**, 393-414.
- Nakamori, S., García-Ligero, M.J., Hermoso-Carazo, A. and Linares-Pérez, J. (2006). Derivation of fixed-interval smoothing algorithm using covariance information in distributed parameter systems, *Applied Mathematics and Computation* **176**, 662-672.

Ordinary kriging for functional data

Ramón Giraldo^{1,2}, Pedro Delicado¹ and Jorge Mateu³

¹ Universitat Politècnica de Catalunya

² Universidad Nacional de Colombia en Bogotá

³ Universitat Jaume I de Castelló

Abstract: We present a methodology to carry out geostatistical analysis for functional data. In particular we propose both an estimator of spatial correlation and a kriging predictor. A real data example illustrates the proposals.

Keywords: Functional Data Analysis; Geostatistics; Spatial data.

1 Introduction

The number of problems and the range of disciplines where the collected data are curves are recently increasing. Functional Data Analysis (FDA) is used, since beginning of the nineties, in order to model this type of information (see, Ramsay and Silverman (2005) for a general perspective, and Ferrati and Vieu (2006) for a non-parametric approach). Agronomy, meteorology, ecology and other sciences where the geostatistical analysis (Cressie, 1993) is often used to describe spatial distribution, are not exceptions. However, there are only a few references that use FDA techniques in the spatial context (Yamanishi and Tanaka, 2003; Illian *et al.*, 2006). Consequently the motivation of this work is to offer a solution to the problem of predicting curves on unsampled locations of a region.

In Section 2 we introduce notation and known results. In Section 3 we propose a way to estimate spatial correlation when data are functions. An application in Agronomy is included in Section 4. The paper ends with some conclusions and a list of other topics to be developed.

2 Kriging for functional data

Ferrati and Vieu (2006) define a *functional variable* as a random variable χ taking values in an infinite dimensional space (or functional space). A *functional data* is an observation χ of χ . A *functional data set* χ_1, \dots, χ_n is the observation of n functional variables χ_1, \dots, χ_n distributed as χ . Let $T = [a, b] \subseteq \mathbf{R}$. We work with functional data that are elements of

$$L_2(T) = \{f : T \rightarrow \mathbf{R}, \text{ such that } \int_T f(t)^2 dt < \infty\}.$$

$L_2(T)$ with the inner product $\langle f, g \rangle = \int_T f(t)g(t)dt$ is an Euclidean space.

Let us consider a functional random process $\{\chi_s : s \in D \subseteq \mathbf{R}^d\}$, usually $d = 2$, such that χ_s is a functional variable for any $s \in D$. Let s_1, \dots, s_n be arbitrary points in D . We assume that we can observe a realization of the functional random process χ_s at these sites: $\chi_{s_1}, \dots, \chi_{s_n}$. Our goal is the prediction of χ_{s_0} , the value of the functional random process at s_0 , where s_0 is an unsampled location. Observe that we want to predict a complete function $\chi_{s_0} : T \rightarrow \mathbf{R}$, and not a particular value of a variable, which is the general aim in traditional geostatistics. In this sense our objective is close to multivariable spatial prediction (Ver Hoef and Cressie, 1993). An even more general framework is in Tolosana (2005), where geostatistics in an arbitrary Euclidean space is presented. In fact this section reproduces, with a notation closer to functional data analysis, part of Chapter 4 in Tolosana (2005).

We assume that the random process is second-order stationary and isotropic, in the sense that $E(\chi_s) = m$, for all $s \in D$, where $m : T \rightarrow \mathbf{R}$ is a function, and that

$$\frac{1}{2}V(\chi_r(t) - \chi_s(u)) = \gamma(h; t, u), \text{ for all } t, u \in T, r, s \in D, \text{ where } h = \|s - r\|.$$

The function $\gamma(h; \cdot, \cdot)$, as function of h , is called semivariogram of χ . We consider the family of linear predictors for χ_{s_0} :

$$\hat{\chi}_{s_0} = \sum_{i=1}^n \lambda_i \chi_{s_i}, \quad \lambda_1, \dots, \lambda_n \in \mathbf{R}. \quad (1)$$

The kriging predictor of χ_{s_0} is given by the solution of the following optimization problem:

$$\begin{aligned} \min_{\lambda_1, \dots, \lambda_n} \quad & \int_T E(\hat{\chi}_{s_0}(t) - \chi_{s_0}(t))^2 dt \\ \text{s.t.} \quad & \sum_i \lambda_i = 1. \end{aligned}$$

Working as in the usual geostatistical setting, we obtain that the optimal coefficients are the solution of the linear system

$$\begin{pmatrix} \gamma(s_1 - s_1) & \cdots & \gamma(s_1 - s_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n - s_1) & \cdots & \gamma(s_n - s_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ m \end{pmatrix} = \begin{pmatrix} \gamma(s_0 - s_1) \\ \vdots \\ \gamma(s_0 - s_n) \\ 1 \end{pmatrix}, \quad (2)$$

where $\gamma(s_i - s_j) = \int_T \gamma(\|s_i - s_j\|; t, t) dt$, and m is the Lagrange multiplier used to take into account the unbiasedness restriction ($\sum_i \lambda_i = 1$). The function $\gamma(h)$ can be called trace-semivariogram. The way it is estimated is developed in Section 3.

3 Estimating the trace-semivariogram

An estimator of the trace-semivariogram $\gamma(h) = \int_T \gamma(h; t, t) dt$, where

$$\gamma(h; t, t) = \frac{1}{2}V(\chi_r(t) - \chi_s(t)), \text{ for } r, s \in D \text{ with } h = \|s - r\|.$$

is needed. Given that we are assuming that χ has a constant mean function m over D , $V(\chi_r(t) - \chi_s(t)) = E[(\chi_r(t) - \chi_s(t))^2]$. Observe that

$$\gamma(h) = \frac{1}{2}E \left[\int_T (\chi_r(t) - \chi_s(t))^2 dt \right], \text{ for } r, s \in D \text{ with } h = \|s - r\|$$

by Fubini's Theorem. Then a natural estimator for this quantity is

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt, \tag{3}$$

where $N(h) = \{(s_i, s_j) : \|s_i - s_j\| = h\}$ and $|N(h)|$ is the number of distinct elements in $N(h)$. For irregularly spaced data there are generally not enough observations separated by exactly h . Then $N(h)$ is modified to $\{(s_i, s_j) : \|s_i - s_j\| \in (h - \varepsilon, h + \varepsilon)\}$, with $\varepsilon > 0$ being a small value.

Once we have estimated the trace-semivariogram for a list of K values h_k , we propose to fit a parametric model $\gamma(h; \theta)$ (spherical, Gaussian, exponential or Matérn, for instance) to the points $(h_k, \hat{\gamma}(h_k))$, $k = 1, \dots, K$, as if they were obtained in the classic one-dimensional geostatistical setting. The fit is done by weighted least squares (see, for instance, Cressie, 1993).

Let $\gamma(h; \hat{\theta})$ be the parametric estimated trace-semivariogram. This functional form is used in equation (2) to solve for the kriging coefficients λ_i .

A different procedure, alternative to the parametric fitting, consists in applying smoothing techniques (splines or local linear regression, for instance; see Wasserman (2006) or references therein) to the set of data $(h_k, \hat{\gamma}(h_k))$, $k = 1, \dots, K$, in order to be able to approximately evaluate $\hat{\gamma}(h)$ for any value of $h \in \mathbf{R}^+$. Let $\hat{\gamma}_S(h)$ be this smoothed version of $\hat{\gamma}(h)$. The question of definite-positiveness of $\hat{\gamma}_S(h)$ deserves more attention.

4 Application to penetration resistance curves

In Agronomy, it is usual to measure the soil penetration resistance in the study area before sowing (Chan et al. 2006). Figure 1 shows 32 sampling locations in an experimental plot at the National University of Colombia and some penetration resistance profiles. The complete functional data set with 32 observed functions is shown in Figure 2, left panel.

Smooth curves of observed penetration resistance were obtained by using B-splines (Wasserman, 2006). See Figure 2, right panel. An outlier curve can be detected from previous graphics: that having values over 2.5 for depths in $[30, 40]$. This outlier function was not considered in both estimation and prediction processes. Based on the remain curves and the estimator (3), the trace-semivariogram was calculated for several spatial lags. A spherical model was fitted to the estimated trace-semivariogram (Figure 3, left panel). The fitted model indicate that this data set have a strong spatial autocorrelation, because the maximum distance between sampling points is 190 meters and locations separated 110 meters are correlated. Fitting a reasonable model to the trace-semivariogram is a critical step for subsequent interpolation of functional

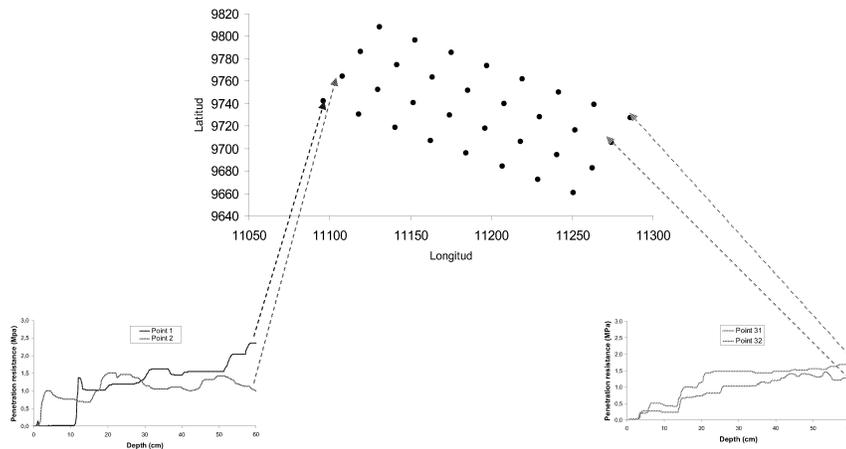


FIGURE 1. Sampling points and some observed penetration resistance curves. Data measured at the Marengo Experimental Station (National University of Colombia) in 2004.

data by kriging. With the sampling scheme considered, there is not possible to have estimations of trace-semivariance near to the origin and it is possible that the nugget parameter has not been estimated well. Consequently, it would be important to include more nearby sampling points in other essays in this experimental plot.

Kriging prediction on an unsampled location, coordinates 11179 (longitude) and 9750 (latitude) (Figure 1), was carried out. λ 's were obtained by solving the system (2) with $\gamma(s_i - s_j)$ estimated by the semivariance model given in Figure 3. Predicted curve (Figure 3, right panel) indicate that in this location there is a good soil compaction level, because the penetration resistance (MPa) estimated is less that 2 MPa, considered the critical limit for root growth, at all range of depth evaluated (Chan et al, 2006).

5 Conclusions and further research

We have introduced a simple kriging predictor when data are functional. More complex procedures can be considered by replacing scalar coefficients λ_i in (1) by functional coefficients $(\lambda_i(t), t \in T)$ or even by double indexed functional coefficients $(\lambda_i(s, t), s, t \in T)$ and using integrals over T as a way to extend the definition of linear combinations. These extension are parallel to regression models with functional responses described in Ramsay and Silverman (2005), Chapters 14 and 16. Other problems requiring further attention are the alternative estimation of the trace-semivariogram, and the automatic detection of outlier functions in the data set.

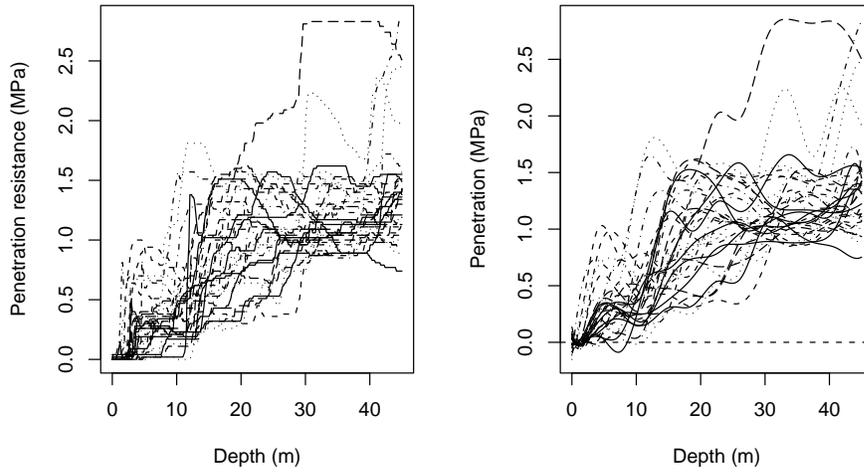


FIGURE 2. Set of 32 penetration resistance functions. Observed (left) and smoothed (right) data.

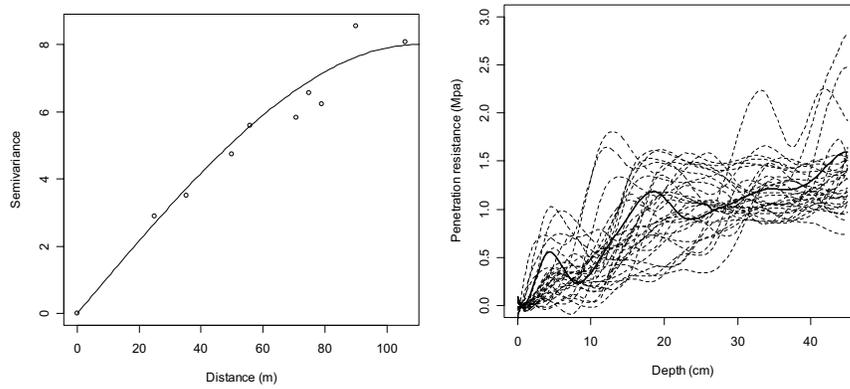


FIGURE 3. *Left panel:* Spherical model fitted to the estimated trace-semivariogram. $\hat{\gamma}(h) = 8(1,5h/110 - 0.5(h/110)^3)$ for $h \leq 110$ and $\hat{\gamma}(h) = 8$ for $h > 110$. *Right Panel:* Measured curves of penetration resistance (dashed lines) and kriging prediction in an unsampled location (solid line).

Acknowledgments: Research partially supported by the Spanish Ministry of Education and Science and FEDER, MTM2004-06231 and MTM2006-09920, and by the EU PASCAL Network of Excellence, IST-2002-506778.

References

- Chan, K., Oates, A., Swan, A., Hayes, R., Dear, B. and Peoples, M. (2006). Agronomic consequences of tractor wheel compaction on a clay soil. *Soil & Tillage Research* **89**, 13-21.
- Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.
- Ferraty, F. and Vieu, P. (2006). *Non Parametric Functional Data Analysis. Theory and Practice*. Springer.
- Illian, J., Benson, E., Crawford, J. and Staines, H. (2006). Principal components analysis for spatial point process data. In: *Case studies in spatial point process modelling*, A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, eds. *Lecture notes in statistics (Springer)* **185**, 135-149.
- Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis. Second edition*. Springer.
- Tolosana-Delgado, R. (2005). *Geostatistics for constrained variables: positive data, compositions and probabilities. Applications to environmental hazard monitoring*. PhD Thesis. University of Girona, Spain.
(http://www.tdx.cesca.es/TDX-0123106-122444/index_an.html)
- Ver Hoef, J. and Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology* **25**, 219-240.
- Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.
- Yamanishi, Y. and Tanaka, Y. (2003). Geographically weighted functional regression. *Journal of Japanese Society of Computational Statistics* **15**, 307-317.

Inverse weighted estimators when there is double censoring

Guadalupe Gómez¹ and Olga Julià²

¹ Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain.

² Departament de Probabilitat, Lògica i Estadística, Universitat de Barcelona, Gran via 585, 08007 Barcelona, Spain.

Abstract: We approach the problem of the estimation of the survival function S_T of a positive random variable T which is doubly-censored, that is, assuming an independent random window of observation $[L, R]$, T is either exactly observed if $L \leq T \leq R$, is left-censored if $T < L$ and right-censored if $T > R$. We propose an inverse-probability-of-censoring estimator for S_T derived from the empirical survival function of T . We apply the proposed methodology to the analysis of the time from starting IV drugs to AIDS in a cohort of drug users recruited in a detoxication program.

Keywords: AIDS data; Doubly-censored; inverse-probability-of-censoring.

1 Introduction

The doubly censoring scheme appears often when both left- and right-censored observations occur for the same data set. An instance of such data is found when analyzing time from HIV-infection to AIDS in a cohort of drug users recruited in a detoxication program, Langohr *et al.* (2004). Some of the patients are AIDS-diagnosed while being in the program (exact observation), some others leave the center AIDS-free (right-censored observation) while a subset of them dies from AIDS without exact AIDS-diagnosis (left-censored observation).

Several nonparametric estimators have been proposed for doubly-censored data. Turnbull's pioneer paper (1974) develops a self-consistent estimator which is proved to be the maximum likelihood estimator assuming a discrete time scale or grouped data. Robins and Rotnitzky (1992) and Robins (1993) take a different perspective to the censoring problems by considering them as a missing data problem introducing the weighting approach. Satten and Datta (2001) show that the Kaplan-Meier estimator for right-censored data can be represented as a weighted average of identically distributed terms.

In this paper we present an inverse-probability-of-censoring representation of the survival estimator when data is doubly-censored. The nonparametric estimation of the survival is our first step to the modelling of a doubly censored response.

2 Notation

Let $\{(T_i, L_i, R_i), i = 1 \dots, n\}$ be a random sample from positive random variables T, L and R , with survival functions S_T, S_R and S_L .

The observable data consists on the pairs $\{(U_i, \delta_i), i = 1 \dots, n\}$ where:

$$\begin{aligned} U_i &= \min\{\max\{L_i, T_i\}, R_i\} = (L_i \vee T_i) \wedge R_i \\ \delta_i &= 1_{\{T_i > R_i\}} - 1_{\{T_i < L_i\}}. \end{aligned}$$

Note that δ_i can take the following three values:

$$\delta_i = \begin{cases} 1 & \text{if } T_i > R_i & \Leftrightarrow U_i = R_i \text{ is a right - censored value} \\ 0 & \text{if } L_i \leq T_i \leq R_i & \Leftrightarrow U_i = T_i \text{ is an exact (uncensored) value} \\ -1 & \text{if } T_i < L_i & \Leftrightarrow U_i = L_i \text{ is a left - censored value.} \end{cases}$$

Let $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ the order statistic corresponding to the observed sample u_1, u_2, \dots, u_n and define $\delta_{(1)}, \delta_{(2)}, \dots, \delta_{(n)}$ as the corresponding censoring values. Whenever tights are present, left observations precede exact observations, which in turn precede right observations. Without loss of generality, we assume $\delta_{(1)} = \delta_{(n)} = 0$ meaning that within the observable data the minimum and maximum values correspond to a death. Let $o_j, j = 1 \dots r$ be the ordered different times among $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ and assume that tights are possible.

3 Estimator for the survival function of T

Our approach starts deriving a first step estimator for each one of the three survival functions S_T, S_L and S_R , assuming that the other two are known. These three first step estimators, which are based on the corresponding empirical survival functions weighted by the inverse of the probability of the corresponding variable being observed, provide a system of three equations. The estimator we propose is the simultaneous solution of the three first step equations. As a by-product of this estimation we obtain estimators for S_L and S_R .

If data were complete, we would observe $u_1 = t_1, \dots, u_n = t_n$ and $\delta_i = 0$, for all i such that $1 \leq i \leq n$ and the empirical survival function for $S_T(t)$ would be given by:

$$S_T^{emp}(t) = \frac{1}{n} \sum_{i=1}^n 1_{\{t_i > t\}} = \frac{1}{n} \sum_{i=1}^n \frac{1_{\{u_i > t\}} 1_{\{\delta_i = 0\}}}{P(T \text{ is observed at time } t_i | T = t_i)}.$$

The probability that the random variables T, L and R are observed within the framework of a doubly censored scheme are as follows:

1. T is observed if and only if $L \leq T \leq R$ and $P(T \text{ is observed at time } t | T = t) = S_R(t^-) - S_L(t)$.
2. L is observed if and only if $T < L$ and $P(L \text{ is observed at time } l | L = l) = 1 - S_T(l^-)$.

3. R is observed if and only if $T > R$ and $P(R \text{ is observed at time } r | R = r) = S_T(r)$.

Following the empirical estimator idea weighted by the inverse of the probability of the corresponding variable being observed, we have as survival estimator for all t :

$$\tilde{S}_T(t) = \frac{1}{n} \sum_{i=1}^n \frac{1}{S_R(u_i^-) - S_L(u_i)} 1_{[\delta_i=0]} 1_{[u_i>t]} \tag{1}$$

which is a valid candidate if both S_L and S_R were known. The same argument is applied to get first step estimators for $S_L(t)$ and for $S_R(t)$ if S_T was known,

$$\tilde{S}_L(t) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 - S_T(u_i^-)} 1_{[\delta_i=-1]} 1_{[u_i>t]} \tag{2}$$

$$\tilde{S}_R(t) = 1 - \frac{1}{n} \sum_{i=1}^n \frac{1}{S_T(u_i)} 1_{[\delta_i=1]} 1_{[u_i \leq t]}. \tag{3}$$

The final estimators $\hat{S}_T(t)$, $\hat{S}_L(t)$ and $\hat{S}_R(t)$ are the solution of the three above equations (1), (2) and (3).

The three functions $\hat{S}_T(t)$, $\hat{S}_R(t)$ and $\hat{S}_L(t)$ are step functions whose jumps are in those points in which the indicator for δ_i is equal to 1. In other words, $\hat{S}_T(t)$ jumps in every $t = u_i$ such that $\delta_i = 0$, $\hat{S}_R(t)$ jumps in every $t = u_i$ such that $\delta_i = 1$ and $\hat{S}_L(t)$ jumps in every $t = u_i$ such that $\delta_i = -1$. We remark that if the sample would not contain left-censored observations, the solution would produce the Kaplan-Meier estimator for right-censored data. The final estimator $\hat{S}_T(t)$ proposed is a self-consistent estimator and it is the maximum likelihood estimator.

4 Illustration

The theoretical development to get an estimator for the survival function when data are doubly censored has been encountered when modelling the elapsed time to AIDS diagnosis from the beginning of starting IV drugs in an observational study conducted at the Hospital Trias i Pujol in Badalona (Spain). A previous analysis can be found in Langohr *et al.* (2004). In that paper we have modelled the time from HIV infection to AIDS diagnosis by means of a log-linear model where the error distribution is assumed from the Weibull family. The ultimate goal of the present analysis is to model T , the time (in days) from the beginning of IV drug use to AIDS diagnosis taking into account that this variable is doubly-censored. As a first step we need the nonparametric estimator of the survival function of T .

Our data set consists on the 266 patients who attended a detoxication program provided by a public Spanish hospital between January 78 and March 97. The cohort has been followed until July 2000.

Among them, 82 developed AIDS during the time of the study (exact observations), 170 were AIDS-free at the last time of follow-up (right-censored observations) and 14 died from AIDS without knowing the exact date of their AIDS-diagnosis (left-censored observations).

References

- Langohr, K. Gómez, G. and Muga, R (2004). A parametric survival model with an interval-censored covariate. *Statistics in Medicine* **23**, pp. 3159-3175.
- Robins, J. (1993). Information recovery and bias adjustment in proportional hazard regression analysis of randomized trials using surrogate markers. *Proceedings of the American Statistical Association – Biopharmaceutical Section*, Alexandria, VA: American Statistical Association, pp. 24–33.
- Robins, J. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology- Methodological Issues*, eds. N. Jewell, K. Dietz, and V. Farewell, Boston: Birkhauser, pp. 297–331.
- Satten, G.A. and Datta, S. (2001). The Kaplan-Meier Estimator as an Inverse-Probability-of-Censoring Weighted Average. *The American Statistician* **55**, No.3, 207-210.

Small Area Estimation using Spanish Labour Force Survey in Canary Islands

E. González-Dávila¹, M.A. González-Sierra¹, R. Dorta-Guerra¹ and A. González-Yanes²

¹ Departamento de Estadística, I.O. y Computación, Universidad de La Laguna, 38205 La Laguna, Tenerife (Canary Islands, Spain), egonzale@ull.es

² Instituto Canario de Estadística, Gobierno de Canarias, 38003 S/C de Tenerife (Canary Islands, Spain)

Abstract: With increasing frequency data at disaggregated level are demanded to National Statistical Agencies and Autonomic Institutes. Thus, the Small Area Estimation is a topic of great interest in Official Statistic researches. There exists a wide range of different methods to provide estimations to small area level. In this work, we introduce a proposal based on indirect estimators that borrow strength information from related areas and we compare it with others design based estimators. We asses them by a Monte Carlo simulation study using the Spanish Labour Force Survey (EPA) in Canary Islands.

Keywords: Small Area Estimation; Design based Estimators; Spanish Labour Force Survey.

1 Introduction

Small Area Estimation is a topic of great interest in many fields, as public statistics, agriculture o disease mapping, because of the increasing demand of data at disaggregated level. There now are a wide range of different methods to provide estimates for small domains. This has led to developments in two directions: design-based methods and model-based methods (see Rao (2003), Jiang and Lahiri (2006)). A typical data file from a complex survey contains survey weights for each individual. Since survey weights contain great information for some population characteristics, they are used in estimation. In particular, the weighted survey design-based estimator for a large area is generally quite robust and is deemed accurate. Typically, an area is regarded as small when the domain sample size is not large enough to yield direct estimates of adequate precision. It is seldom possible to have sample size to support reliable direct estimates for all the domains of interest. In these situations, it is often necessary to use indirect estimates that borrow strength information of the variables of interest from related areas. This information is incorporated into the estimation process through a model that provides a link to related areas through the use of supplementary information related to the variables of interest, such as recent census counts and current administrative records.

The Spanish National Statistical Agency (INE) publishes estimations of activity, unemployed and inactive totals at province for sex level using the Spanish Labour Force Survey (EPA). The domains (small areas) of interest are NUT4 level crossed with sex.

In particular, Canary Islands are constituted for two provinces, Las Palmas and S/C of Tenerife. In the province of Las Palmas exists twelve small areas and in S/C of Tenerife fifteen ones. For each small area, it knows the total population at group age for sex level.

In some research works is used supplementary sample (draw of population) combined with the information provided for mother survey. With respect to the EPA in Canary Islands, the authors of this work assessed it using Monte Carlo simulations in the project supported by Instituto Canario de Estadística titled “Estimadores en Áreas Pequeñas Aplicados a la Estadística Pública Canaria (2005)” in the work group “Áreas Pequeñas” of the INE. The results showed that indirect estimator of the small area that borrow strength information from the province where the small area belongs to, have a behavior acceptable versus direct estimators with both information source.

2 Spanish Labour Force Survey (EPA)

The EPA is a quarterly survey follows a stratified two-stage random sampling design and, for each province a separate sample is extracted. The primary sampling units are Census Sections (geographical areas with a maximum of 500 dwellings) and they are grouping in strata according to the size of municipality. Within each stratum, the primary units are selected without replacement and probabilities proportional to size according to the number of dwellings. In the second stage sampling, the secondary sampling units are dwellings and a without-replacement simple random sampling is applied to draw a fixed number (18 in our case) of secondary units from each selected primary unit.

3 Provided estimators and indicators of yield

In this work, we use the population given for the Census 2001 in the Autonomous Community of Canary, and generate 1000 simulations of EPA’s in this region. Given that we dispose of the population, the real values of the interest variables are also known. We assess the consistency property of usual small area design-based estimators using two indicators of yield, the mean of relative mean square error and the mean of relative bias. Direct and indirect estimators are considered. It is included two synthetic estimators based on implicit linking model. One of them is constructed using information of the recent census counts to form big areas constituted for the joining of homogeneous small area with respect to the interest variables.

We denote the survey weight of observed individual j as w_j , small area crossed with sex as d , group of age as g and the population size of an area l as N_l . The provided estimators, when y denotes the dichotomy variable of interest, are

Direct Estimator:

$$\hat{y}_d^{direct} = \frac{\sum_{j \in EPA} w_j y_j}{\sum_{j \in EPA} w_j} N_d, \quad (1)$$

and its mean $\hat{\hat{y}}_d^{direct} = \hat{y}_d^{direct} / N_d$.

Post-stratify Estimator:

$$\hat{y}_d^{post} = \sum_g \hat{y}_{dg}^{direct} N_{dg}. \tag{2}$$

Synthetic Estimator:

$$\hat{y}_{d,r}^{synth} = \sum_g \hat{y}_{g,r}^{direct} N_{dg}, \tag{3}$$

where r denotes a region formed as a join of small areas. That is, the area d borrows strength information from the region r . In particular, we prove several options to choose the regions. When the region coincides to which the small area belongs we denote such estimator as synth-prov one. In other situations, the regions are constructed using cluster analysis with the information of the variable of interest obtained on recent census counts in the small areas for each group of age. In this case, we denote the estimator as synth- n , with n the number of regions used.

The provided indicators of yield are

Relative Bias:

$$RB_d(\hat{y}) = \frac{1}{Nsim} \sum_{k=1}^{Nsim} \frac{\hat{y}_d(k) - Y_d}{Y_d} 100. \tag{4}$$

Mean of Absolute Relative Bias:

$$MARB(\hat{y}) = \frac{1}{D} \sum_{d=1}^D |RB_d(\hat{y})|. \tag{5}$$

Relative Mean Square Error:

$$RMSE_d(\hat{y}) = \left(\frac{1}{Nsim} \sum_{k=1}^{Nsim} \left(\frac{\hat{y}_d(k) - Y_d}{Y_d} \right)^2 \right)^{1/2} 100. \tag{6}$$

Mean of Relative Mean Square Error:

$$MRMSE(\hat{y}) = \frac{1}{D} \sum_{d=1}^D RMSE_d(\hat{y}), \tag{7}$$

where $Nsim$ and D are the number of simulations and small areas considered, respectively.

4 Results and Conclusions

The obtained results based on the 1000 simulations of EPA's are presented in the table 1. The synth- n estimators present values of MRMSE better than those of the rest of estimators for both sex. While the values of MARB have an intermediate value between the values of the direct estimator and synth-prov ones. The tests realized on real EPA's show that a good candidate estimator would be the synth-2 or synth-3 depending on the variability allowed between consecutive EPA's.

TABLE 1. MRMSE and MARB for occupied and unemployed by sex (M:Male and F:Female) for estimators provided.

	Occupied				Unemployed			
	MRMSE		MARB		MRMSE		MARB	
	M	F	M	F	M	F	M	F
Direct	11, 76	19, 90	1, 47	2, 83	43, 40	47, 02	3, 51	2, 23
Post	10, 50	18, 29	1, 40	2, 39	44, 42	46, 84	3, 85	2, 14
Synth-prov	6, 76	13, 77	6, 07	13, 03	21, 96	20, 08	19, 15	17, 71
Synth-2	4, 82	8, 96	3, 79	7, 39	16, 54	14, 08	12, 66	8, 31
Synth-3	4, 98	8, 37	3, 19	6, 18	17, 97	14, 76	9, 95	8, 31
Synth-4	4, 95	8, 01	2, 80	5, 24	15, 35	15, 88	6, 98	7, 24
Synth-5	5, 67	9, 70	2, 87	5, 51	15, 51	18, 19	5, 53	7, 23

Acknowledgments: This work was supported by the Instituto Canario de Estadística (ISTAC) in the project titled “Estimadores en Áreas Pequeñas Aplicados a la Estadística Pública Canaria (2006)”, and partially supported by the Spanish MEC Proyecto MTM2006-09920.

References

- Jiang, J., and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test* **15**, 1-96.
- Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey: John Wiley & Sons.

Conditional Heteroscedasticity or Stochastic Volatility in Financial Risk Management?

A. Grané¹ and H. Veiga¹

¹ Statistics Department, Universidad Carlos III de Madrid. C/ Madrid, 126, 28903 Getafe, Spain.

Abstract: In this paper we use bootstrap techniques to calculate the minimum capital risk requirement (MCRR) and we compare the MCRRs obtained with the unconditional density to the ones obtained with the conditional approach.

Keywords: Conditional Heteroscedasticity; Stochastic Volatility; MCRR; Bootstrap.

1 Introduction

Risk management modelling has been one of the most fast developing areas of application of statistic and econometric techniques. Consider the following problem: suppose a firm holds a long position of L_t units of a future contract. An important question in risk management is: What is the minimum capital, K_t , needed to cover losses of this long position with, for instance, a 95 percent probability? This minimum capital risk requirement is denoted MCRR (see Brooks 2002).

There are various methods available for calculating the minimum capital risk requirement, including the "delta-normal" method, the historical simulation that involves the estimation of the quantile of the portfolio returns and the structured Monte Carlo simulation. The Monte Carlo approach is clearly powerful and flexible for generating MCRR estimates since any stochastic process for the underlying assets can be specified. However, there are at least two drawbacks with the use of it in this context. First, for a large portfolio, the computational time required to compute the MCRR may be very high. Second, and more important, the calculated MCRR may be inaccurate if the stochastic process that has been assumed for the underlying asset is inappropriate. An alternative approach that could overcome this drawback, would be to use bootstrap rather than Monte Carlo simulation.

The work is organized as follows: In Section 2 we analyze and propose some models for three Index Future Contract series. In Section 3 we compute the MCRRs by bootstrap and we present some results in Section 4.

2 Parametric approach

2.1 Previous study of the series

In this study we calculate the MCRRs for three futures contracts - the FTSE-100 Index Futures Contract, the SPF Standard and Poors Index Futures Contract and the Russell Index Futures Contract. The data were collected from EconWin Financial and spans the period 2 August 1989- 18 May 2005 for the FTSE-100, the period 4 August 1989- 16 October 2006 for the SPF Index and the period 5 February 1993-15

December 2006 for the Russell Index. We have deleted from the data set observations corresponding to non trading days to avoid the incorporations of spurious zero returns, leaving 3980, 4366 and 3421 observations for the FTSE-100, SPF and Russell Indexes, respectively.

In order to test for non-linear dependence of the returns we have used the BDS test of Brock, Dechert, Scheinkman and LeBaron (1996). The null hypothesis of i.i.d. is rejected for all the three return series at a 5 percent level of significance, which is consistent with the results of Hsieh (1993) and Brooks, Clare and Persaud (2000). With the purpose to investigate the cause of this rejection, we estimate the autocorrelation function and the Ljung-box Q statistics. The autocorrelations are not statistically significant which suggests that the rejection of the null i.i.d. is due to non-linear dynamics in the returns. We calculate the bicorrelation coefficients to infer if the non-linearity is in-mean or in-variance and we conclude that the rejection of the null is due to the presence of non-linear dependence in the variance.

2.2 Models proposed

Given the conclusions of Section 2.1, we select several GARCH-type models such as: the traditional GARCH of Bollerslev (1986), the fractional integrated GARCH (FIGARCH) of Baillie, Bollerslev and Mikkelsen (1996) and the hyperbolic GARCH (HYGARCH) of Davidson (2004), with errors that follow a Gaussian or/and a t -distribution. In the context of stochastic volatility, we estimate the autoregressive stochastic volatility model (ARSV) of Taylor (1986) and the autoregressive long memory stochastic volatility model (ARLMSV) of Breidt, Crato and de Lima (1998). The first is a short memory model while the second specifies the volatility process as a fractional integrated process and it is a natural competitor to the FIGARCH and HYGARCH. The estimation methods that have been used are the quasi maximum likelihood (QML) for the conditional heteroscedasticity models and the Whittle estimator for the stochastic volatility models. Moreover, we also report the MCRRs calculated from the unconditional density using moving block bootstrap directly to the observed price changes (see Efron and Tibshirani, 1993, and Lahiri, 2003).

3 Computing the MCCR by bootstrap

In order to evaluate the effectiveness of the previous models in capturing all of the non-linear dependence in the data, we reapply the BDS test to the standardized residuals. If the model has captured all the important features of the data, the standardized residual series should be random. Once these residuals are i.i.d., it is valid to resample from them using the bootstrap technique in order to obtain the MCCR estimates. This is achieved by simulating the future values of the futures price series, using the parameter estimates from the models, and using disturbances obtained by sampling with replacement from the standardized residuals. In this way, 20000 possible future paths of series are simulated (i.e. 20000 replications are used), and in each case, the maximum loss is calculated over holding periods of 1, 5, 10, 30, 90 and 180 days. It is assumed that the futures position is opened on the final day of the sample used to estimate the models. The amount of capital required to cover losses on 95 percent of days is estimated by the 95th percentile of these 20000 maximum losses. To illustrate the methodology we report in Table 3 the obtained results for FTSE-100.

TABLE 1. Capital requirement for 95% coverage probability as a percent of the initial value of the FTSE-100.

Long Position								
No. days	GARCH- gaus.	GARCH- t-Stud	FIGARCH- t-Stud	HYGARCH- gaus.	ARSV	ARLMSV	Uncond	
1	1.00	0.98	0.88	0.89	0.90	0.77	1.26	
5	2.18	2.16	2.05	1.95	2.02	1.73	2.74	
10	3.01	2.99	2.96	2.71	2.88	2.46	3.82	
30	4.77	4.77	5.27	4.40	5.09	4.34	6.41	
90	6.46	6.62	9.28	6.70	9.15	7.74	10.08	
180	6.80	7.10	12.98	8.11	13.32	11.09	12.94	

Short Position								
No. days	GARCH- gaus.	GARCH- t-Stud	FIGARCH- t-Stud	HYGARCH- gaus.	ARSV	ARLMSV	Uncond	
1	1.06	1.05	0.95	0.96	0.94	0.80	1.35	
5	2.38	2.37	2.28	2.17	2.14	1.82	3.14	
10	3.41	3.42	3.42	3.14	3.10	2.64	4.62	
30	6.06	6.16	6.75	5.79	5.79	4.90	8.78	
90	10.37	10.89	14.03	11.06	11.63	9.69	17.55	
180	14.16	15.39	22.80	16.92	19.12	15.48	28.07	

4 Analyzing some results

The results are interesting because, first, the MCCRs derived from bootstrapping the returns themselves (the unconditional approach) are in general higher than those generated from the previous models. This occurs because the level of volatility at the start of the MCCR calculation period was low relatively to its historical level. Therefore, the conditional approach give us forecasts of volatility lower than the historical average. As the holding period increases from 1 to 180 days, the MCCR estimates converge quickly to those of unconditional approach, except for GARCH that presents the lowest values of the MCCR (see Table 3).

We also observe that the MCRRs for short positions are larger than those of comparative long positions. This suggests that on average an upwards move in the futures price is more probable than a fall. Finally, we also compute the bootstrap confidence intervals for the previous MCRRs and an out-of-sample test based upon the selected models.

Acknowledgments: Special Thanks to A.M. Alonso for clarifying some points in the bootstrap computations. Research projects MTM2006-09920 and SEJ2006-03919 from the Spanish Ministry of Education and Science and FEDER.

References

Baillie, R.T., Bollerslev, T. and Mikkelsen, H.O. (1996). Fractionally Integrated Generalized Autoregressive Conditional Heteroscedasticity. *Journal of Econometrics* **74**, 3-30.

- Bollerslev, T. (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* **31**, 307-327.
- Breidt, F., Crato, N. and de Lima, P.J.F. (1998). On The Detection and Estimation of Long Memory in Stochastic Volatility. *Journal of Econometrics* **83**, 325-348.
- Brock, W.A., Dechert, D., Scheinkman, H. and LeBaron, B. (1996). A Test for Independence Based on the Correlation Dimension. *Econometric Reviews* **15**, 197-235.
- Brooks, C. (2002). *Introductory Econometrics for Finance*. Cambridge: University Press.
- Brooks, C., Clare, A.D. and Persaud, G. (2000). A Word of Caution on Calculating Market-Based Minimum Capital Risk Requirements. *Journal of Banking and Finance* **24**, 1557-1574.
- Davidson, J. (2004). Moment and Memory Properties of Linear Conditional Heteroscedasticity Models, and a New Model. *Economic Statistics* **22**, 16-29.
- Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Hsieh, D.A. (1993). Implications of Nonlinear Dynamics for Financial Risk Management. *The Journal of Financial and Quantitative Analysis* **28**, 41-64.
- Lahiri, S.N. (2003). *Resampling Methods for Dependent Data*. New York: Springer-Verlag, Inc.
- Taylor, S. (1986). *Modelling Financial Time Series*. New York: Wiley.

Diagnosing Models from Maps based on Weighted Logratio Analysis

Michael Greenacre¹

¹ Department d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona

Abstract: Weighted logratio analysis leads to graphical displays (also known as spectral maps) that have a lot in common with correspondence analysis, differing only in the initial matrix analyzed. Subsequent steps in the computational algorithm are identical: rows and columns are weighted by the same marginal values, the same (weighted) double-centring and singular value decomposition are applied, and the graphical options are potentially the same. The spectral map has an edge over correspondence analysis as far as theoretical properties are concerned and is more related to the modelling of two-way tables. Correspondence analysis, on the other hand, has the advantage of being able to handle sparse tables with lots of zeros, which is probably why it is so popular for analysing ecological and archeological data.

Keywords: Biplot; Correspondence Analysis; Logratio transformation; Singular value decomposition; Spectral Map.

1 Introduction – Weighted Logratio Analysis

Weighted logratio analysis, also known as the *spectral map* (SM), was developed in the specific context of the analysis of biological activity spectra (Lewi, 1976, 1998). Suppose that we have a matrix of strictly positive data \mathbf{N} ($I \times J$), for example a table of data on the same ratio scale (e.g., measurements all in centimetres, or monetary values all in euros), or a table of positive counts. Suppose that the row and column totals of \mathbf{N} , n_{i+} ($i = 1, \dots, I$) and n_{+j} ($j = 1, \dots, J$), relative to the grand total n , are denoted by \mathbf{r} and \mathbf{c} respectively (so the elements of \mathbf{r} and \mathbf{c} sum to 1 in each case), and that \mathbf{D}_r ($I \times I$) and \mathbf{D}_c ($J \times J$) are the corresponding diagonal matrices of these relative sums. Then weighted logratio analysis (WLR) can be used to visualize these data in the form of a map of points representing the rows and columns, where the rows and columns are weighted by the respective elements in \mathbf{r} and \mathbf{c} . The algorithm for computing the row and column coordinates in WLR can be summarized compactly as follows, using the singular value decomposition (SVD) as the key analytical step, and following almost exactly the algorithm for correspondence analysis (CA):

Step 1: Double-centre the matrix of logarithms of the data with respect to weighted row and column averages:

$$\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\top) \log(\mathbf{N})(\mathbf{I} - \mathbf{c1}^\top)^\top \quad (1)$$

Step 2: Perform a weighted (or generalized) SVD by first pre-transforming \mathbf{A} by:

$$\mathbf{S} = \mathbf{D}_r^{\frac{1}{2}} \mathbf{A} \mathbf{D}_c^{\frac{1}{2}} \quad (2)$$

and then applying the usual (unweighted) SVD:

$$\mathbf{S} = \mathbf{U} \mathbf{D}_\alpha \mathbf{V}^\top \quad \text{where } \mathbf{U}^\top \mathbf{U} = \mathbf{V}^\top \mathbf{V} = \mathbf{I} \quad (3)$$

Step 3: Calculate standard coordinates of the rows and columns as:

$$\text{(row standard) } \mathbf{X} = \mathbf{D}_r^{\frac{1}{2}} \mathbf{U} \quad \text{(column standard) } \mathbf{Y} = \mathbf{D}_c^{\frac{1}{2}} \mathbf{V} \quad (4)$$

and principal coordinates as:

$$\text{(row principal) } \mathbf{F} = \mathbf{D}_r^{\frac{1}{2}} \mathbf{U} \mathbf{D}_\alpha \quad \text{(column principal) } \mathbf{Y} = \mathbf{D}_c^{\frac{1}{2}} \mathbf{V} \mathbf{D}_\alpha \quad (5)$$

As in CA, a *symmetric map* of the data plots the rows and columns in K^* -dimensional space by using the first K^* columns of \mathbf{F} and \mathbf{G} . *Asymmetric maps* (which are also biplots) use the columns of \mathbf{F} and \mathbf{Y} (for row asymmetric – or row principal – map) or \mathbf{G} and \mathbf{X} (for column asymmetric – or column principal – map). Other biplots are possible with interesting graphical interpretations – see Chapter 13 of Greenacre (2007).

2 Connection with Correspondence Analysis

Steps 1–3 above follow exactly the correspondence analysis (CA) algorithm (see, for example, Greenacre, 2007, theoretical and computational appendices), except that the first step (1) involves logarithms of the data, whereas CA starts with the *contingency ratios*, defined as the observed values relative to their “expected” values. When data are counts, the expected values are the usual ones under the hypothesis of independence; for other data these are evaluated in the same way using the margins of the table. So the contingency ratios are $q_{ij} = n_{ij}/(n_{i+}n_{+j}/n)$, which is equal to $(n_{ij}/n)/(r_i c_j)$ in terms of the row and column weights. If we substitute $\mathbf{Q} = [q_{ij}]$ for $\log(\mathbf{N})$ in (1) and continue with Steps 2 and 3, we have the CA solution. Interestingly, if we define the matrix $\mathbf{Q}(\alpha)$ with elements q_{ij}^α , then the above algorithm gives CA for $\mathbf{Q}(1)$ and gives WLR/SM for $\mathbf{Q}(\alpha)$ in the limit as α tends to 0 (this is the same as the Box-Cox transformation $q_{ij}^\alpha - 1$ of the contingency ratios, since the -1 is eliminated by the double-centring). Aitchison & Greenacre (2002) presented the unweighted form of the log-ratio map, but the weighted form inherent in the spectral map is a distinct improvement.

3 Diagnosing models

As shown by Aitchison & Greenacre (2002) and Greenacre & Lewi (2005), specific “equilibrium” models can be diagnosed in the WLR/SM maps when subsets of points

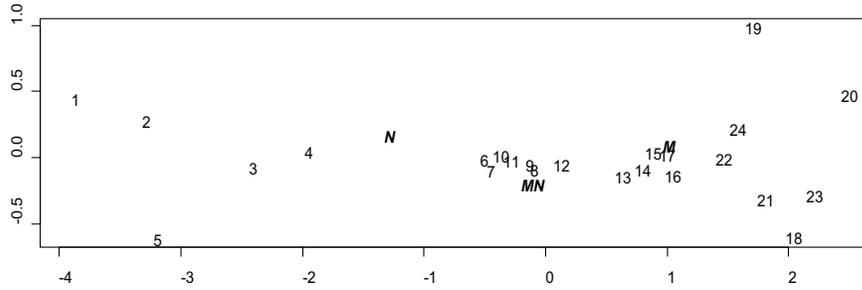


FIGURE 1. Weighted logratio/spectral map of M–N population genetic data, showing the 24 populations (numbers) and the three haplotypes, M (M and M co-occurring), N (N and N) and MN (M and N). The variance explained on the horizontal first axis is 96.8%, showing that the data are highly one-dimensional, indicating an equilibrium relationship.

lie approximately on straight lines; that is, when the difference vectors are approximately collinear. A good illustration of this is the diagnosis of the Hardy-Weinberg equilibrium in population genetics. Figure 1 shows the spectral map of haplotype frequencies of the M, N and M-N blood groups in 24 populations; this is a genetic system with two co-dominant alleles, M and N. The fact that there is 96.8% of the variance on axis 1, i.e., close collinearity of the points, means that the differences (i.e., logratios) $\log(MN) - \log(M) = \log(MN/M)$ and $\log(N) - \log(MN) = \log(N/MN)$ are linearly related. In fact, because the lengths of these difference vectors is almost equal in the map, we can deduce the following linear relationship between the logratios, called an *equilibrium relationship*:

$$\log(MN/M) = \log(N/MN) + C \tag{6}$$

where the constant C can be estimated from a scatterplot of $\log(MN/M)$ vs. $\log(N/MN)$ to be 1.348 (this was calculated by fitting a principal axis to the points in Figure 2 and calculating where it intersected the vertical axis – the principal axis is almost exactly of slope 1). Exponentiating (6) and simplifying we obtain the following diagnosed model:

$$\frac{MN^2}{M \times N} = 3.85 \tag{7}$$

Compared to the Hardy-Weinberg formula for genetic equilibrium of two co-dominant alleles with allele frequencies of p and q ($p + q = 1$), where the frequencies of M-M (M in our notation), N-N (N) and MN (MN) are p^2 , q^2 and $2pq$, respectively, we arrive at the model for perfect equilibrium of:

$$\frac{MN^2}{M \times N} = 4 \tag{8}$$

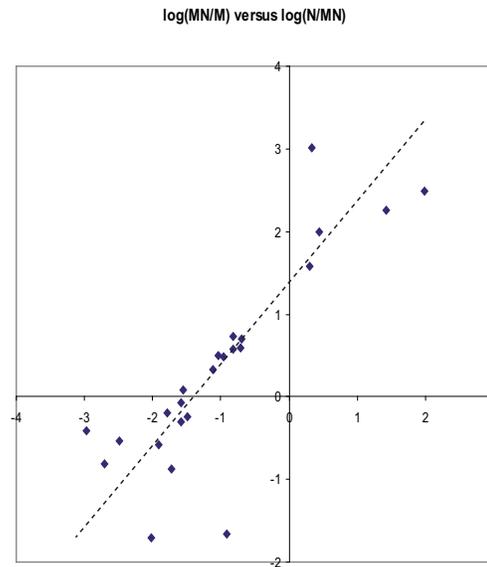


FIGURE 2. Scatterplot of $\log(MN/M)$ versus $\log(N/MN)$, showing the principal axis used to deduce the constant in (6).

which is almost the same as (7).

The CA solution is given in Figure 3, but is not as enlightening to the connection between the haplotypes: the points for the populations follow the curved pattern often found in CA maps called the “arch effect”, owing to the simplex space in which the row points lie (a triangle in this example).

4 Conclusion

We have shown that the weighted logratio map can be used to diagnose a certain class of equilibrium models by identifying straight-line configurations. In the delivered paper we shall show more complex examples, where a subset of points can be identified as lying on an approximate straight line, in which case the corresponding subsets of the data can be diagnosed as following such a class of models. Such diagnosis is not easy in the CA framework, where the equilibrium models turn out to be curves and are not distinguishable from other curves which might not indicate equilibrium models. However, the logratio approach has one drawback: it cannot easily handle lots of zero values. This is the context where CA is extremely useful and is probably the reason why it is the method of choice for ecologists and archeologists worldwide, since their data sets are usually very sparse.

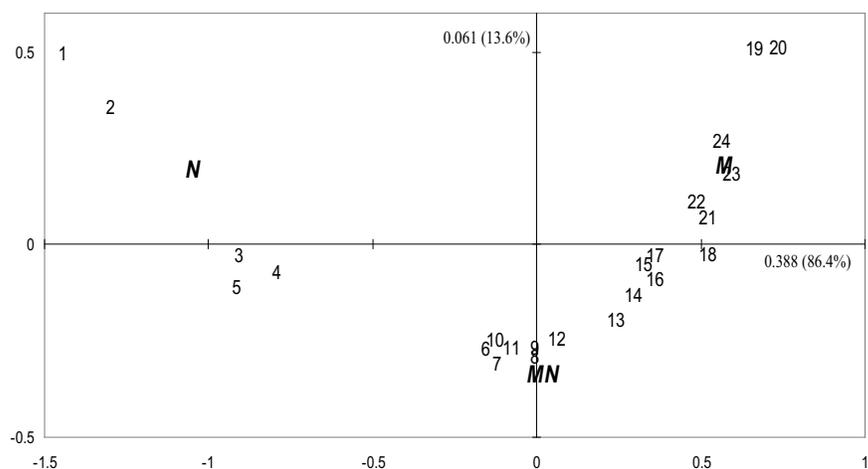


FIGURE 3. CA map of M–N population genetic data, showing the 24 populations following a curve. The variance explained on the first axis is 86.4%, hence the pattern is less one-dimensional than Figure 1.

Acknowledgments: Thanks are due to the Fundación BBVA in Madrid and its director, Prof. Rafael Pardo, for generous research support, and also for partial support from the Spanish Ministry of Education and Science.

References

- Aitchison, J. & Greenacre, M.J. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **51**, 375–392.
- Greenacre, M.J. (2007). *Correspondence Analysis in Practice. Second Edition*. London: Chapman & Hall / CRC Press.
- Greenacre, M.J. & Lewi, P.J. (2005). Distributional equivalence and subcompositional coherence in the analysis of contingency tables, ratio-scale measurements and compositional data. Working report 908, Department of Economics and Business, Universitat Pompeu Fabra. *Submitted for publication*. Available online at www.econ.upf.edu/en/research/onepaper.php?id=908
- Lewi, P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneim. Forsch. (Drug Res.)* **26**, 1295–1300.
- Lewi, P.J. (1998). Analysis of contingency tables. In: *Handbook of Chemometrics and Qualimetrics: Part B*. Chapter 32, pp. 161206. Amsterdam: Elsevier.

Likelihood ratio testing for zero variance components in linear mixed models

Sonja Greven^{1,3}, Ciprian Crainiceanu², Annette Peters³ and Helmut Küchenhoff¹

¹ Department of Statistics, LMU Munich University, Munich, Germany, sonja.greven@stat.uni-muenchen.de,

² Department of Biostatistics, Johns Hopkins University, Baltimore, USA,

³ Institute of Epidemiology, GSF-National Research Center for Environment and Health, Neuherberg, Germany

Abstract: We consider the problem of testing for zero variance components in linear mixed models. Typical applications include testing for a random intercept or testing for linearity of a smooth function. We propose two approximations to the finite sample null distribution of the restricted likelihood ratio test statistic. Our approach applies to a wider variety of mixed models than previous results, including those with moderate numbers of clusters, unbalanced designs, or nonparametric smoothing. Extensive simulations show that both proposed approximations outperform the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation and parametric bootstrap currently used.

Our methods are motivated by and applied to the longitudinal epidemiological study Air-gene, with the aim of assessing non-linearity of dose-response-functions between ambient air pollution concentrations and inflammation.

Keywords: Boundary; Bootstrap; Penalized splines; Nonparametric smoothing; Air pollution.

1 Introduction

Linear mixed models are widely used to model longitudinal or clustered data and, more recently, to estimate smoothing parameters for penalized splines using REML or ML. We focus on linear mixed models of the form

$$Y = X\beta + Z_1b_1 + \dots + Z_Sb_S + \varepsilon, \quad (1)$$

with random effects $b_s \sim N(\mathbf{0}, \sigma_s^2 \mathbf{I}_{K_s})$ pairwise independent and independent of $\varepsilon \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$, K_s columns in Z_s , \mathbf{I}_ν the identity matrix of size ν , and n the sample size. We are interested in testing one of the variance components

$$H_{0,s} : \sigma_s^2 = 0 \quad \text{versus} \quad H_{A,s} : \sigma_s^2 > 0, \quad (2)$$

corresponding, for example, to testing for a zero random intercept or testing linearity against a general alternative. This problem is non-standard due to the parameter on the boundary of the parameter space. Stram and Lee (1994), using results from Self

and Liang (1987), showed that the Likelihood Ratio Test (LRT) statistic for testing (2) has an asymptotic $0.5\chi_0^2 : 0.5\chi_1^2$ null distribution if \mathbf{Y} can be divided into independent and identically distributed (i.i.d.) subvectors. However, for penalized spline smoothing responses are not independent at least under the alternative, and longitudinal studies often have unbalanced data or only moderate numbers of subjects. Crainiceanu and Ruppert (2004) derived the finite sample and asymptotic null distribution of the LRT and restricted LRT (RLRT) for testing (2) in models with one variance component ($S = 1$), and showed that it is generally different from $0.5\chi_0^2 : 0.5\chi_1^2$. For $S > 1$, they recommend a parametric bootstrap, which can be computationally very expensive. As the LRT has been seen to have undesirable properties with a high probability mass at zero, we develop two faster approximations to the finite sample null distribution of the RLRT.

2 Two approximations to the RLRT null distribution

2.1 Fast finite sample approximation

Our first approximation is inspired by pseudo-likelihood estimation (Gong and Samaniego, 1981), where nuisance parameters are replaced by consistent estimators. Liang and Self (1996) showed that under certain regularity conditions the asymptotic distribution of the pseudo LRT is the same as that of the LRT if the nuisance parameters are known. For our problem, we could view the $\mathbf{b}_i, i \neq s$, as nuisance parameters. We assume that under regularity conditions the prediction of $\sum_{i \neq s} \mathbf{Z}_i \mathbf{b}_i$ is good enough to allow the distribution of the RLRT in model (1) to be closely approximated by the RLRT in the reduced model

$$\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_s \mathbf{b}_s + \boldsymbol{\varepsilon}, \tag{3}$$

with $\tilde{\mathbf{Y}} = \mathbf{Y} - \sum_{i \neq s} \mathbf{Z}_i \mathbf{b}_i$ assumed known. As model (3) has only one variance component σ_s^2 , the exact null distribution of the RLRT for testing (2) is known (Crainiceanu and Ruppert, 2004) to be

$$RLRT_n \stackrel{d}{=} \sup_{\lambda \geq 0} \left\{ (n - p) \log \left[1 + \frac{N_n(\lambda)}{D_n(\lambda)} \right] - \sum_{l=1}^{K_s} \log(1 + \lambda \mu_{l,n}) \right\}, \tag{4}$$

where $\stackrel{d}{=}$ denotes equality in distribution, p is the number of columns in \mathbf{X} ,

$$N_n(\lambda) = \sum_{l=1}^{K_s} \frac{\lambda \mu_{l,n}}{1 + \lambda \mu_{l,n}} w_l^2, \quad D_n(\lambda) = \sum_{l=1}^{K_s} \frac{w_l^2}{1 + \lambda \mu_{l,n}} + \sum_{l=K_s+1}^{n-p} w_l^2, \tag{5}$$

$w_l, l = 1, \dots, n - p$, are independent $N(0, 1)$, and $\mu_{l,n}, l = 1, \dots, K_s$, are the eigenvalues of the $K_s \times K_s$ matrix $\mathbf{Z}'_s (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Z}_s$. This distribution can be simulated from very efficiently, as the K_s eigenvalues need to be computed only once, and speed depends only on K_s rather than on the number of observations n .

TABLE 1. Settings for the simulation study.

	Tested Component	Null Hypothesis	Nuisance Component
1	Random Intercept	Equality of Means	-
2	Smooth Function	Linearity	-
3	Random Intercept	Equality of Means	Smooth Function
4	Smooth Function	Linearity	Random Intercept
5	Smooth Function	Linearity	Smooth Function
6	Random Slope	Equality of Slopes	Random Intercept
7	Bivariate Smooth	Additivity and Linearity	Random Intercept
8	Random Intercept	Equality of Means	Bivariate Smooth

2.2 Mixture approximation to the Bootstrap

If a parametric bootstrap is preferred but computationally intensive, we propose the following parametric approximation to the RLRT distribution

$$RLRT \stackrel{d}{\approx} aUD, \quad (6)$$

where $U \sim \text{Bernoulli}(1 - p)$, $D \sim \chi_1^2$, and $\stackrel{d}{\approx}$ denotes approximate equality in distribution. The flexible family of distributions in (6) contains as a particular case the i.i.d. case asymptotic $0.5\chi_0^2 : 0.5\chi_1^2$ distribution with $a = 1$ and $p = 0.5$, and is just as easy to use. p and a can be estimated from a bootstrap sample, while (6) stabilizes estimation of tail quantiles and thus reduces the necessary bootstrap sample size.

Maximum likelihood estimation of p would require the proportion of simulated RLRT values that are exactly zero, and is therefore very sensitive to numerical imprecisions (encountered, for example, with `proc MIXED` in SAS, and the `lme` function in R). We thus propose estimation of p and a using the method of moments, after setting all negative values to zero.

Note that both our proposed approximations are asymptotically identical to the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation when the i.i.d. assumption holds.

3 Simulation Study

We conducted an extensive simulation study, covering a range of important situations with one or two variance components. An overview is given in Table 1. We varied the number of subjects $I = 6, 10$ and observations per subject $J = 5, 25, 50, 100$ for all settings, as well as the value for the respective nuisance variance component $\sigma_1^2 = 0, 0.1, 1, 10, 100$, while all other parameters not restricted to zero under the null were fixed at either 1 or -1 . Covariates were sampled from standard normal distributions, with increasing correlation for the case of two smooth functions. Smooth univariate or bivariate functions were modelled using low-rank thin plate splines with smoothing parameters estimated by REML, and with testing of the corresponding variance component translating to testing for linearity $f(x) = \beta_0 + \beta_1 x$ in the univariate, and

to testing for additivity and linearity $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ in the bivariate case. 10,000 samples each were simulated from the RLRT null distribution and our two approximations compared to a bootstrap and the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation. The fast finite sample approximation produced empirical type I error rates close to the nominal level, comparable to the exact distribution when $S = 1$. The approximation was usually good even for $n = 30$; the necessary sample size increased somewhat when random effects were correlated. The *aUD* approximation reduced the necessary bootstrap sample size by between 10% and 90%, with the reduction more pronounced for smaller α levels or p values. The $0.5\chi_0^2 : 0.5\chi_1^2$ approximation was always very conservative.

4 Testing smooth dose-response-functions

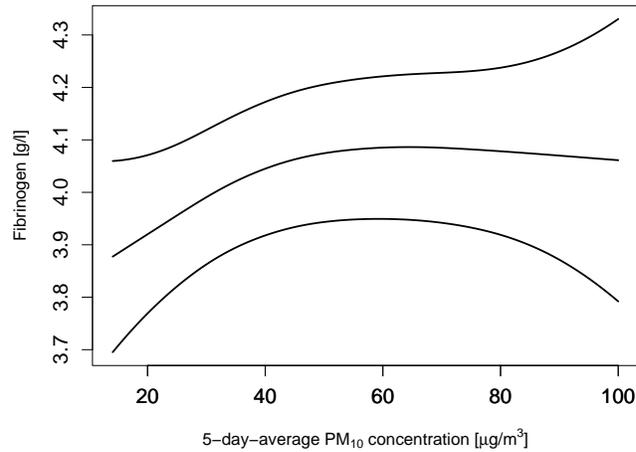
The Airgene study was conducted in six European cities between May 03 and July 04. One of its aims is to assess association between inflammatory responses and ambient air pollution concentrations in myocardial infarction survivors. 3 inflammatory blood markers (CRP, Fibrinogen, IL-6) were measured every month repeatedly up to 8 times in 1,003 patients. Air pollution and weather variables were measured concurrently in each city. Patients were genotyped and additional information collected at baseline. Analyzes had to account for longitudinal data structure and potential nonlinearity of weather and trend variables, with smooth effects estimated in the mixed model framework. As the shape of the air pollution dose-response functions has important policy implications, one aim of the study was to investigate the functional form of the air pollution effects on inflammation. For illustration, we focus on the effect of PM_{10} , particulate matter with diameter less than $10 \mu m$, on Fibrinogen in Barcelona. A total of 1074 valid blood samples and PM_{10} exposures were available for 183 patients. The model used for the PM_{10} -Fibrinogen dose-response function is

$$FIB_{ij} = u_i + f(PM10_{ij}) + \sum_{l=2}^L \beta_l x_{ijl} + \varepsilon_{ij}, \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2), \quad (7)$$

where FIB_{ij} is the j th Fibrinogen value of the i th patient, $u_i \stackrel{iid}{\sim} N(0, \sigma_u^2)$ is a random patient intercept, and $PM10$ indicates the 5-day-average PM_{10} exposure before blood withdrawal. $f(\cdot)$ is a smooth, unspecified, function estimated using penalized cubic B-Splines and penalizing deviations from linearity (Green et al., 2006). Linear covariates x_l are patient's age, asthma diagnosis, time trend, weekday and air temperature (cubic polynomial).

Figure 1 shows the estimated smooth PM_{10} effect on Fibrinogen in Barcelona. An important scientific question is whether the dose-response function is linear. This is equivalent to testing (2) in (7), where σ_s^2 is a variance component controlling the smoothness of $f(\cdot)$. Note that the i.i.d. assumption is violated and that the model includes two variance components.

The test statistic for testing linearity of $f(\cdot)$ against a general alternative takes the value $RLRT = 2.9$. Test results for all four approximations are given in Table 2. The fast finite sample approximation reduces computation time by 4 orders of magnitude,

FIGURE 1. Estimated smooth PM_{10} -Fibrinogen dose-response function in Barcelona.

while results are similar to a bootstrap. The aUD approximation gives results close to the bootstrap even for sample sizes 100 times lower. The $0.5\chi_0^2 : 0.5\chi_1^2$ approximation is clearly conservative. In all cases, results indicate a significant difference from linearity.

TABLE 2. Testing the PM_{10} effect on Fibrinogen in Barcelona for linearity. Computation time was measured on a standard PC using Matlab (f.f.s.) / SAS.

Approximation	Samples	Time	p-value
Fast finite sample	100,000	~35sec	0.023
aUD	100	~4min	0.026
aUD	1,000	~40min	0.031
aUD	10,000	~8h	0.029
$0.5\chi_0^2 : 0.5\chi_1^2$	-	-	0.044
Bootstrap	10,000	~8h	0.025

5 Summary

We have discussed testing for zero variance components in linear mixed models. Possible applications include, but are not limited to, testing for zero random intercepts or slopes and testing for linearity of a smooth function against a general alternative. For models with one variance component, we recommend directly using the exact null distribution of the RLRT statistic derived in Crainiceanu and Ruppert (2004), which can be simulated efficiently. For models with more than one variance component, we have

proposed two approximations to the finite sample null distribution of the RLRT. Extensive simulations showed superiority of both approximations over the $0.5\chi_0^2 : 0.5\chi_1^2$ approximation and parametric bootstrap currently used. Our results extend existing methodology to linear mixed models with more than one variance component and lacking independence assumption. We have illustrated the use in testing for linearity of dose-response-functions for longitudinal data on air pollution health effects.

Acknowledgments: The first author was supported by the German Academic Exchange Service (DAAD). We would like to thank Yu-Jen Cheng and Fabian Scheipl for discussions related to the topic of this paper. Special thanks to the Airgene study group. The Airgene study was funded by EU contract QLK4-CT-2002-O2236.

References

- Crainiceanu, C., and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B* **66**(1), 165-185.
- Gong, G. and Samaniego, F. (1981). Pseudo Maximum Likelihood Estimation: Theory and Applications. *The Annals of Statistics* **9**(4), 861-869.
- Green, S., Küchenhoff, H., and Peters, A. (2006). Additive mixed models with P-Splines. *Proceedings of the 21st International Workshop on Statistical Modelling*, 201-207. Eds.: Hinde J, Einbeck J and Newell J.
- Liang, K.-Y. and Self, S.G. (1996). On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic. *Journal of the Royal Statistical Society: Series B* **58**(4), 785-796.
- Self, S.G. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* **82**(398), 605-610.
- Stram, D.O. and Lee, J.-W. (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics* **50**(3), 1171-1177.

Endogeneity issues in mixed models

Leonardo Grilli¹ and Carla Rampichini¹

¹ Dipartimento di Statistica “Giuseppe Parenti”, Università di Firenze, Viale Morgagni, 59 - 50134 Firenze. {grilli, rampichini}@ds.unifi.it

Abstract: The paper explores some issues related to endogeneity in random effects models, focusing on the case where the random effects are correlated with a level 1 covariate in a linear random intercept model. This type of endogeneity arises from a wrong equality restriction on the between-cluster and within-cluster slopes. A straightforward solution is to allow distinct slopes through the addition of the cluster mean as a further covariate. However, the use of the sample cluster mean instead of the population cluster mean entails a measurement error that yields a biased estimator of the between-cluster slope. In the paper we study the measurement error issue and propose a correction to obtain unbiased estimates. Moreover, we compare alternative estimators and illustrate the main findings through a simulation study.

Keywords: between-within slopes; measurement error; random effects.

1 Introduction

Regression analysis with data from observational studies is often threatened by the problem of endogeneity. Typically, endogeneity arises when there are unobserved covariates affecting the response and correlated with the observed covariates included in the model: in this way, the unobserved covariates are absorbed by the error term, which is thus correlated with the model covariates, causing the standard estimators to be inconsistent.

The topic of endogeneity in mixed models has received some attention in the last years: Ebbes *et al.* (2004), Neuhaus and McCulloch (2006), Snijders and Berkhof (2007) and Kim and Frees (2007).

Mixed models have error terms at every hierarchical level, so the problem of endogeneity can concern error terms at any level. Our contribution focuses on level 2 endogeneity, i.e. the random effects are correlated with level 1 covariates, an issue well known in the setting of panel data due to the famous Hausman test (Ebbes *et al.*, 2004).

We consider a random intercept model with a level 1 covariate X_{ij} :

$$Y_{ij} = \eta + \beta X_{ij} + v_j + e_{ij} \tag{1}$$

where $i = 1, 2, \dots, n_j$ is the level 1 index and $j = 1, 2, \dots, J$ is the level 2 (cluster) index. For example, in a panel setting the level 1 units are waves and the clusters are individuals, while in a cross-sectional framework the level 1 units are individuals and the clusters are entities such as institutions or geographical areas. In equation (1), v_j are level 2 errors, also called random effects, while e_{ij} are level 1 errors.

Level 2 endogeneity arises when $Cov(v_j, X_{ij}) \neq 0$, so $E(v_j | X_{ij}) \neq 0$ and thus the standard estimators are inconsistent for β . In Econometrics it is usual to model $E(v_j | X_{ij})$ as a linear function of the cluster mean \bar{X}_j , so a straightforward remedy to endogeneity is to add \bar{X}_j to the model equation. On the other hand, the statistical literature (Snijders and Berkhof, 2007; Neuhaus and McCulloch, 2006) has pointed out that often the between-cluster and within-cluster effects are conceptually and numerically rather different, and the inclusion of \bar{X}_j as a further regressor is just a way to disentangle the two effects. However, it is usually not recognized that \bar{X}_j is a *sample* cluster mean used as a measurement of a *population* cluster mean: as a consequence, the model including \bar{X}_j is affected by measurement error and thus the estimator of the between-cluster slope is biased. In this work we deal with the measurement error issue, showing that it can be relevant but a straightforward correction is available. The following Sections sketch the issue, while technical details are reported in an extended version.

2 Endogeneity and measurement error

To study endogeneity issues in model (1), the covariate X_{ij} must be treated as random and the hierarchical framework requires to specify how X_{ij} varies between and within clusters. The simplest choice is to assume a variance component model $X_{ij} = X_j^B + X_{ij}^W$, with $E(X_j^B) = \mu_X$, $E(X_{ij}^W) = 0$, $Var(X_j^B) = \tau_X^2 > 0$ and $Var(X_{ij}^W) = \sigma_X^2 > 0$. In the light of this decomposition, it is clear that model (1) implicitly assumes the equality of between-cluster slope β_B and within-cluster slope β_W .

Defining $\delta = \beta_B - \beta_W$, a more general model without such restriction is

$$\begin{aligned} Y_{ij} &= \alpha + \beta_W X_{ij}^W + \beta_B X_j^B + u_j + e_{ij} \\ &= \alpha + \beta_W X_{ij} + \delta X_j^B + u_j + e_{ij} . \end{aligned} \quad (2)$$

By assumption, model (2) is unaffected by endogeneity and is taken as the *true* model. From the second row of (2), it appears that when $\beta_B \neq \beta_W$, model (1) is affected by level 2 endogeneity. This leads to biased estimates of the slope and of the residual cluster variance. In such a case, the slope β of model (1) is a meaningless mixture of β_B and β_W . Level 2 endogeneity can be interpreted as the consequence of the misspecification due to omitting the cluster mean when $\beta_B \neq \beta_W$.

Even if X_{ij} is observable, the components X_j^B and X_{ij}^W are unobservable. Therefore, to fit model (2), X_j^B and X_{ij}^W must be replaced by their observable counterparts, i.e. the cluster mean $\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$ for X_j^B and the deviation from the cluster mean (centered covariate) $X_{ij} - \bar{X}_j$ for X_{ij}^W , giving the *working* model:

$$\begin{aligned} Y_{ij} &= \alpha + \beta_W (X_{ij} - \bar{X}_j) + \beta_B \bar{X}_j + z_j + e_{ij} \\ &= \alpha + \beta_W X_{ij} + \delta \bar{X}_j + z_j + e_{ij} \end{aligned} \quad (3)$$

This model solves the endogeneity issue arising from a wrong equality restriction on the between and within slopes. However, using \bar{X}_j in place of X_j^B causes a measurement error that doesn't affect the within slope β_W , but biases the between slope β_B or, equivalently, the slope difference δ . In particular, for clusters of equal size n , the

estimable between slope is $\lambda_X\beta_B + (1 - \lambda_X)\beta_W$, and the estimable slope difference is $\lambda_X\delta$, where λ_X is the reliability of the sample cluster mean \bar{X}_j as a measure of the population counterpart X_j^B :

$$\lambda_X = \frac{Var(X_j^B)}{Var(\bar{X}_j)} = \frac{\tau_X^2}{\tau_X^2 + \sigma_X^2/n} = \left(1 + \frac{1}{(\tau_X^2/\sigma_X^2)n}\right)^{-1}, \quad \lambda_X \in (0, 1).$$

The estimable δ is attenuated by the reliability λ_X . The attenuation is negligible only when λ_X is about one, e.g. when the cluster size n is large. Moreover, it can be shown that the estimable residual cluster variance of Y is inflated by the quantity $(1 - \lambda_X)\tau_X^2\delta^2$. However, the bias of using \bar{X}_j in place of X_j^B can be corrected quite easily since β_W and λ_X can be unbiasedly estimated and thus also β_B and the residual cluster variance of Y can be unbiasedly estimated.

A simple solution for level 2 endogeneity is the *Fixed Effects* estimator, but several alternative solutions using random effects are available. However, the literature on the topic has two main limitations: it focuses only on the slopes and does not explicitly address the measurement error problem. Here we explore the performances of two estimators with regard to both the slopes and variance components, also considering the consequences of the measurement error.

3 Main results

We assume that the *true* model is (2), and fit models (1) and (3) through Restricted Maximum Likelihood (REML).

When $\delta = 0$, both models lead to unbiased estimators of β_W , but the estimator from model (1) is more efficient. In both models the residual cluster variance is unbiasedly estimated.

When $\delta \neq 0$, both models are affected by endogeneity, though for different reasons: model (1) has an omitted variable problem, while model (3) has a measurement error problem. For the bias in model (3) we derived analytical expressions, while for the bias in model (1) it is necessary to rely on simulations. Table 1 reports the Monte Carlo means of the estimates obtained with the REML method for $\lambda_X = 2/3$ and varying δ . Indeed, the bias, if any, increases with the absolute value of δ . In model (1) the estimated slope is biased for β_W and the estimated residual cluster variance is inflated, as it includes a factor accounting for the between-within slopes difference. In model (3) β_W is unbiasedly estimated, while δ is attenuated by $\lambda_X = 2/3$ and the residual cluster variance is overestimated. However, in contrast with model (1), all the estimates derived from model (3) can be corrected using an estimate of λ_X .

In summary, since level 2 endogeneity is generated by a wrong equality restriction on the within-cluster and between-cluster slopes, an effective solution in the mixed model framework is the inclusion of the sample cluster mean of the covariate under consideration. However, our results show that when the cluster size is small, as in panel or longitudinal data, the measurement error due the use of the sample cluster mean in place of the population mean leads to relevant bias and thus needs a suitable correction.

TABLE 1. MC means on 1000 replications, REML estimates for varying δ .

δ	Model (1): only X_{ij}				Model (3): X_{ij} and \bar{X}_j				
	η	β_W	σ_v^2	σ_e^2	α	β_W	δ	σ_z^2	σ_e^2
-1.0	-0.70	0.70	1.49	1.09	-0.33	1.00	-0.67	1.33	1.00
-0.5	-0.34	0.84	1.11	1.03	-0.17	1.00	-0.33	1.08	1.00
0.0	0.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00
0.5	0.34	1.16	1.12	1.03	0.16	1.00	0.33	1.09	1.00
1.0	0.70	1.30	1.50	1.09	0.33	1.00	0.67	1.34	1.00

True values, model (2): $\mu_X = 1$, $\tau_X^2 = \sigma_X^2 = 1$; $\alpha = 0$, $\beta_W = 1$, $\sigma_u^2 = \sigma_e^2 = 1$

Data structure: $J = 1000$, $n = 2$, $\lambda_X = 2/3$

Bibliography

- Ebbes, P., Bockenholt, U. and Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica* **58**, 161–178.
- Kim, J.S. and Frees, E.W. (2007). Multilevel Modeling with Correlated Effects. *Psychometrika*. To appear.
- Neuhaus, J.M. and McCulloch, C.E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society B* **68**, 859–872.
- Snijders, T.A.B. and Berkhof, J. (2007). Diagnostic checks for multilevel models. In Jan de Leeuw (Ed.), *Handbook of Multilevel Analysis*. New York: Springer. To appear.

A flexible approach to measurement error correction in case-control studies

Annamaria Guolo¹

¹ Department of Statistics, University of Padova, Italy, guolo@stat.unipd.it

Abstract: The problem of measurement error affecting covariates is very common in many scientific areas. Many techniques have been proposed in literature to face this problem. Most of them require several assumptions on the involved variables to be satisfied, otherwise yielding to misleading results. Here, we propose using a flexible parametric model in order to reduce sensitivity to modelling assumptions, mainly on the unobserved and mismeasured phenomenon. The skew normal distribution is used for this purpose. The performance of the method is evaluated through simulation studies, within a case-control setting.

Keywords: case-control data; likelihood; measurement error; skew normal distribution.

1 Introduction

The problem represented by erroneous measurements is very common in many research areas. It has been long recognized that the implications of ignoring measurement errors may be substantial, often turning out in misleading results (Armstrong, 2003). Many correction techniques have been proposed in literature to face this problem during the past 20 years (see Carroll *et al.*, 2006, for a review). Most of them require many assumptions on the involved variables to be satisfied, otherwise yielding to misleading inferential results. However, difficulties in selecting and specifying the relationships among variables may be not neglectable. In the light of this, robustifying measurement error correction techniques is a topic of notable interest. Here we propose to flexibly modelling the distribution of the unobserved phenomenon, with the aim to reduce the sensitivity to distributional assumptions. The skew normal distribution (Azzalini, 1985) is used to this purpose. The performance of the method is evaluated through simulation, within a likelihood-base approach to measurement error correction. Two different distributions for the unobservable variable are considered, specifically a mixture of lognormal distributions and a χ_1^2 distribution. The measurement error is assumed to be nondifferential and multiplicative. Our proposal is compared to the fully parametric approach, in which the distribution of the unobserved variable is correctly specified, and to the *naive* analysis, that is, the one performed by ignoring measurement error. The comparison is in terms of bias, standard error and empirical coverages of confidence intervals of the estimators for the parameters of interest. The focus is on case-control data.

2 Models and methods

Suppose that case-control data are available. Let Y be the case-control status indicator. Let X be the covariates which are not directly observed. In place of X , the mismeasured variables W are observed. Here we focus on uni-dimensional X and W , although extensions to higher dimensions are possible. Let Y be related to X through the logistic model, whose density function is $f_{Y|X}(y|x; \beta_0, \beta_1) = H(\beta_0 + \beta_1 x)^y (1 - H(\beta_0 + \beta_1 x))^{1-y}$, where $H(v) = (1 + v^{-1})^{-1}$. The model does not depend on W , that is, the measurement error is nondifferential. Let $f_{W|X}(w|x; \gamma)$ be the density function of the model relating W to X . Finally, let $f_X(x; \delta)$ be the density function of the assumed model for X . The inferential interest is typically on the parameter β_1 . Within a likelihood-based approach, if n independent observations from (Y, W) are available, the estimates for the parameters are those which maximize the likelihood

$$L(\beta_0, \beta_1, \gamma, \delta) = \prod_{i=1}^n \int_{\mathcal{X}} f_{Y|X}(y|x; \beta_0, \beta_1) f_{W|X}(w|x; \gamma) f_X(x; \delta) dx. \quad (1)$$

Actually, it is often difficult to exactly modelling the distribution of X . Moreover, erroneous specification may lead to inconsistent estimators. A common solution is the use of a semiparametric approach, which avoids the specification of the distribution for X . However, it may be difficult to implement as well as it can lead to considerable loss in efficiency if compared to a parametric approach. Alternatively, the distribution of X may be flexibly modelled through a parametric specification. A mixture of normal distributions is usually adopted to this purpose (Carroll *et al.*, 1999; Richardson *et al.*, 2002). Here we suggest to flexibly modelling the distribution of X through the skew normal distribution, $X \sim SN(\mu, \sigma, \alpha)$, (Azzalini, 1985). Thus, $f_X(x; \delta) = f_X(x; \mu, \sigma, \alpha) = (2/\sigma)\phi((x - \mu)/\sigma)\Phi(\alpha(x - \mu)/\sigma)$, where μ, σ, α are, respectively, the location, the scale and the shape parameter and $\phi(\cdot)$ and $\Phi(\cdot)$ represent the standard normal density and distribution.

3 Simulation studies

We investigate the behaviour of the proposed flexible approach to correct for measurement error (SN), compared to that of the *naive* analysis (NAIVE) and the parametric approach when the distribution of X is exactly specified (LIK). We consider two different distributions for X , which often arise in practice: a mixture of lognormal distributions, $Ln(\mu_1; \sigma_{ln}^2)$ and $Ln(\mu_2; \sigma_{ln}^2)$, with mixing weights 0.8 and 0.2, respectively, and a χ_1^2 distribution. The measurement error model is assumed to be multiplicative, $W = Xe^U$, where $U \sim N(0, \sigma_U^2)$. The number of case-control data is set equal to $n = 600$. In the simulations, the values for the parameters are $(\beta_0, \beta_1)^t = (-1.5, 0.8)^t$ and $(\mu_1, \mu_2, \sigma_{ln})^t = (-2.3, -1.5, 0.9)^t$. Different amounts of measurement error are considered, $\sigma_U = 0.3, 0.5, 0.75$. The number of simulations is 500. Together with the observations from (Y, W) , a validation subsample of data with exactly measures of X is assumed to be available. It is randomly selected as a 10% of the simulated data. The likelihood function (1) is maximized through Nelder and Mead's (1965) optimization algorithm, starting from initial estimates of the parameters provided by *naive* or

TABLE 1. Bias, standard error and coverages of confidence intervals for the *naive* and the corrected estimators of β_1 based on 500 replications, for $\sigma_U = 0.3$. The true value of β_1 is 0.8.

Mixture of lognormals	NAIVE	LIK	SN
BIAS (s.e.)	-0.142 (0.369)	0.020 (0.438)	0.010 (0.436)
90%	0.854	0.914	0.910
95%	0.938	0.964	0.960
χ_1^2	NAIVE	LIK	SN
BIAS (s.e.)	-0.086 (0.088)	0.005 (0.102)	0.011 (0.103)
90%	0.684	0.906	0.906
95%	0.752	0.956	0.954

moment-based results. Integrals are approximated by Gauss-Hermite quadrature. All calculations have been performed using the R programming language (R Development Core Team, 2005).

The results of the simulation studies for the case $\sigma_U = 0.3$ are summarized in Table 1. The bias, the standard error and the empirical coverages of confidence intervals with nominal levels equal to 90% and 95% for the estimator of β_1 are reported. The results indicate that the performance of the flexible approach we suggest is satisfactory for measurement error correction in both the examined scenarios. The method provides estimators for the parameter of interest β_1 whose bias and standard error are comparable to those from a likelihood approach based on the correct specification of the distribution of X . A similar result holds also for the empirical coverages of confidence intervals. As it is expected, the behaviour of the proposed method notably outperforms the one from a *naive* analysis. Moreover, additional investigations show that the satisfactory behaviour of the flexible approach we suggest is preserved under increasing measurement error, $\sigma_U = 0.5, 0.75$. (simulations studies not shown). Discrepancies of the results with respect to those from a correct specification of the distribution of X remain limited.

According to the simulations studies we performed, the use of the proposed flexible approach seems to be promising in order to add a measure of robustness to the analysis. The method is a viable option when the absence of knowledge on the distribution of the unobserved phenomenon X turns out in sensitivity of the results to model specifications. However, further investigation is needed to evaluate the performance of the method, mainly under more complex measurement error structures.

Acknowledgments: This research was partially supported by the Italian Ministry for Education, University and Research.

References

- Armstrong, B. (2003). Exposure measurement error: consequences and design issues. In *Exposure Assessment in Occupational and Environmental Epidemiology*, Oxford University Press, Oxford.
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171-178.
- Carroll, R.J., Roeder, K. and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics* **55**, 44-54.
- Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton: Chapman & Hall, CRC Press.
- Nelder, J.A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal* **7**, 308-313.
- R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Richardson, S., Leblond, L., Jaussent, I. and Green, P.J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society, Series A* **165**, 549-566.

On Likelihood Estimation in Semiparametric Frailty Models

Il Do Ha¹ and Youngjo Lee²

¹ School of Information and Management, Daegu Haany University, South Korea

² Department of Statistics, Seoul National University, South Korea

Abstract: We study likelihood-based approaches, including marginal and hierarchical likelihoods, for frailty models. In semi-parametric frailty models we show that ML (maximum likelihood) estimators can lead to an underestimation of parameters because the number of nuisance parameters in nonparametric baseline hazards increases with sample size. This bias problem can be removed by using adjusted profile-likelihood methods. The proposed methods are demonstrated using a numerical study.

Keywords: Breslow estimator; Frailty models; Hierarchical likelihood; Marginal likelihood; Profile likelihood.

1 Introduction

Semiparametric frailty models have been widely used for analyzing various survival data. For the inferences many authors have worked the ML estimation based on marginal likelihood. In particular, the gamma frailty model has been often studied because it gives an explicit marginal likelihood. The ML estimators are often implemented by the EM algorithm, using the discrete nonparametric Breslow estimates for baseline hazards playing the role of nuisance parameters (Nielsen et al., 1992; Vaida and Xu, 2000). Recently, in semiparametric gamma frailty models Rondeau et al. (2003) and Baker and Henderson (2005) showed numerically that such a ML procedure lead to the underestimation of parameters, particularly for frailty parameter. For the reduction of such biases they proposed the use of the continuous nonparametric estimates, instead of the Breslow estimates, for the baseline hazards. Ha and Lee (2005) showed that an adjustment of profile h-likelihood, even if the Breslow estimates are used, is quite effective in reducing the biases. In this paper we focuss on the estimation of the frailty parameter in semiparametric frailty models and we study how to adjust the profile likelihoods based upon the marginal likelihood and hierarchical or (h-)likelihoods (Lee and Nelder, 1996, 2001), using the Breslow estimates to eliminate nuisance parameters in baseline hazards.

2 Adjusted Profiling Method in Frailty Models

Let T_{ij} ($i = 1, \dots, q$, $j = 1, \dots, n_i$, $n = \sum_i n_i$) be the survival time for the j th observation of the i th subject and C_{ij} be the corresponding censoring time. Let the

observable random variables be $y_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$, where $I(\cdot)$ is the indicator function. Given the unobserved frailty for the i th subject $U_i = u_i$, suppose that the conditional hazard function of T_{ij} is of the form

$$\lambda_{ij}(t|u_i) = \lambda_0(t) \exp(x_{ij}^T \beta) u_i, \tag{1}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, $x_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ is a vector of fixed covariates and $\beta = (\beta_1, \dots, \beta_p)^T$ is a vector of the corresponding regression parameters. The frailties U_i are assumed to be independent and identically distributed with a frailty parameter α . In gamma frailty models the marginal likelihood is explicitly available, whereas in lognormal frailty models it is not. Since the functional form of $\lambda_0(t)$ in (1) is unknown, following Breslow (1972), we consider the baseline cumulative hazard function $\Lambda_0(t)$ to be a step function with jumps at the observed death times, $\Lambda_0(t) = \sum_{k: y_{(k)} \leq t} \lambda_{0k}$, where $y_{(k)}$ is the k th ($k = 1, \dots, r$) earliest distinct death time among the y_{ij} 's, and $\lambda_{0k} = \lambda_0(y_{(k)})$. Note here that $r = r(n)$, the number of nuisance parameters λ_{0k} 's, increases with the sample size n . Let $\omega = (\omega_1, \dots, \omega_r)^T$, where $\omega_k = \log \lambda_{0k}$. Following Lee and Nelder (1996), the h-(log)likelihood for frailty models (1) is defined by

$$h = h(\omega, \beta, \alpha) = \sum_{ij} \ell_{1ij} + \sum_i \ell_{2i},$$

where

$$\sum_{ij} \ell_{1ij} = \sum_k d_{(k)} \omega_k + \sum_{ij} \delta_{ij} \eta_{ij} - \sum_k \exp(\omega_k) \left\{ \sum_{(i,j) \in R_{(k)}} \exp(\eta_{ij}) \right\},$$

$\ell_{1ij} = \ell_{1ij}(\omega, \beta; y_{ij}, \delta_{ij} | u_i)$ is the logarithm of the conditional density function for y_{ij} and δ_{ij} given $U_i = u_i$, $\ell_{2i} = \ell_{2i}(\alpha; v_i)$ is the logarithm of density function for $V_i = \log(U_i)$ with parameter α , $\Lambda_0(t) = \int_{-\infty}^t \lambda_0(k) dk$ is the baseline cumulative hazard function, $\eta_{ij} = x_{ij}^T \beta + v_i$ with $v_i = \log(u_i)$, $d_{(k)}$ is the number of deaths at $y_{(k)}$ and $R_{(k)} = \{(i, j) : y_{ij} \geq y_{(k)}\}$ is the risk set at $y_{(k)}$. Since the dimension of nuisance parameter vector ω increases with sample size, for estimation of $\tau = (\beta, v)$ Ha et al. (2001) proposed the use of profile likelihood

$$h^* = h|_{\omega = \hat{\omega}}$$

with ω being eliminated. Here $\hat{\omega}_k$ are solutions of $\partial h / \partial \omega_k = 0$. For inference about α , Ha et al. (2001) used the adjusted profile likelihood (Lee and Nelder, 2001), $p_\tau(h^*)$, after eliminating τ , defined by

$$p_\tau(h) = [h - \frac{1}{2} \log \det\{D(h, \tau)/(2\pi)\}]|_{\tau = \hat{\tau}}$$

where $D(h, \tau) = -\partial^2 h / \partial \tau^2$ and $\hat{\tau}$ solves $\partial h / \partial \tau = 0$. For gamma frailty models, Ha and Lee (2003) confirmed that the second-order Laplace approximation works better. On the other hand, the profile marginal likelihood becomes

$$m^* = m|_{\omega = \tilde{\omega}},$$

where $m = \log\{\int \exp(h)dv\}$ is the marginal (log-)likelihood, $\tilde{\omega}_k = \log \tilde{\lambda}_{0k}$ and $\tilde{\lambda}_{0k}$ solves $\partial m / \partial \lambda_{0k} = 0$. In particular, the maximization of m^* gives the ML estimators implemented by the EM algorithm using the Breslow estimator; for gamma frailty models see Nielsen et al. (1992) and for lognormal frailty models see Vaida and Xu (2000). There are two independent procedures in the h-likelihood approach, namely (a) the marginalization by integrating out random frailties and (b) the profiling to eliminate the fixed nuisance parameters. If we apply the profiling (b) after the marginalization (a) we have m^* and corresponding approximations based upon the h-likelihood are

$$m^* \simeq p_v^*(h) \equiv p_v(h)|_{\omega=\hat{\omega}} \quad \text{and} \quad m^* \simeq p_v^{*s}(h) \equiv p_v^s(h)|_{\omega=\hat{\omega}},$$

where $p_v^s(h)$ is the second-order Laplace approximation (Lee and Nelder, 2001) to m . On the contrary, if we apply the marginalization (a) after the profiling (b) we have $p_v(h^*)$ and $p_{\beta,v}(h^*)$, recommended by Ha et al. (2001) and Ha and Lee (2005).

Result 1. In the semiparametric frailty models we can show the following approximations:

- (i) $p_\omega(m) \simeq p_{\omega,v}(h) = p_v(h^*) + c,$
- (ii) $p_{\omega,\beta}(m) \simeq p_{\omega,\beta,v}(h) = p_{\beta,v}(h^*) + c,$

where $c = -\frac{1}{2} \sum_k \log\{d_{(k)}/(2\pi)\}$ does not depend on α . Result 1 shows that the two approaches give different estimation procedures. It is interesting to note that $p_v(h^*)$ gives an approximate inference based upon $p_\omega(m)$, not m^* . Note that $p_v(h^*)$ is also the first-order Laplace approximation to a marginal profile likelihood (say, m_P) proposed by Gu, Sun and Huang (2004), which may be computationally intensive. As we shall see in Section 3, the ML procedure m^* (hence $p_v^*(h)$ and $p_v^{*s}(h)$) give severe downward biases for α , while $p_v(h^*)$ and $p_v^s(h^*)$ do not. When m is hard to obtain (e.g. lognormal frailty models) we may use $p_{\omega,v}(h)$ or $p_v(h^*)$. Because the dimension of ω increases with sample size the computation of $p_{\omega,v}(h)$ in the Result (i) is often difficult, so that the use of $p_v(h^*)$ is preferred. For the gamma frailty the second-order approximation, $p_v^s(h^*)$, of $p_\omega(m)$ works better: see the simulation results of Table 1. Note that $p_v^s(h^*)$ is also the second-order Laplace approximation to m_P . Next, following the Result (ii), the use of $p_{\beta,v}(h^*)$ is recommended to obtain REML (restricted maximum likelihood) estimators.

3 Numerical Study

We here consider the model (1) with gamma frailty having $E(U_i) = 1$ and $\text{var}(U_i) = \alpha$. We generate data assuming the exponential baseline hazard $\lambda_0(t) = 1$, one standard normal covariate with $\beta = 1$ and $\alpha = 1.0$. We also consider univariate and bivariate sample cases: $n = \sum_{i=1}^q n_i = 100$ or 200 with $(q, n_i) = (100, 1), (100, 2), (200, 1)$. Note here that we chose fairly extreme cases, with no censoring and small sample size, because these situations yielded the most biased estimates of $\hat{\alpha}$ in the simulation studies by Nielsen et al. (1992) and Barker and Henderson (2005). For the fitting we used the four likelihoods, m^* , $p_v^{*s}(h)$, $p_\omega(m)$ and $p_v^s(h^*)$. From 200 replications of simulated data we compute the mean and mean squared error (MSE) for $\hat{\beta}$ and $\hat{\alpha}$. The

TABLE 1. Simulation results about the estimators $\hat{\alpha}$ and $\hat{\beta}$ under marginal- and h- likelihoods in semiparametric gamma frailty models. The simulation is conducted with 200 replications at true gamma frailty variance $\alpha = 1$ and regression parameter $\beta = 1$ (No censoring).

q	n_i	Method	$\hat{\alpha}$		$\hat{\beta}$	
			Mean	MSE	Mean	MSE
100	1	m^*	0.42	0.469	0.80	0.094
		$p_v^s(h^*)$	0.89	0.393	0.96	0.089
100	2	m^*	0.90	0.067	0.98	0.079
		$p_v^s(h^*)$	0.99	0.066	1.00	0.081
200	1	m^*	0.63	0.231	0.87	0.054
		$p_v^s(h^*)$	0.99	0.192	1.00	0.053

results are summarized in Table 1. The marginal likelihood m^* gives severe downward biases in all cases considered, especially in $n_i = 1$ or frailty parameters α . Moreover, the underestimation of α leads to that of β . Though not reported here, the estimates from m^* and $p_v^{*s}(h)$ ($p_\omega(m)$ and $p_v^s(h^*)$) are very similar. Table 1 demonstrates that the two profile likelihood methods, $p_v^s(h^*)$ and $p_\omega(m)$, reduce such biases effectively.

4 Discussion

Even though unreported we have found that the use of $p_{\lambda_0}(m)$ with $\lambda_0 = (\lambda_{01}, \dots, \lambda_{0r})^T$ does not give such a great bias reduction as $p_\omega(m)$. Thus, the parameterization such as $\omega = \log \lambda_0$ is important to define adjusted profile likelihoods. In this paper we have considered the gamma frailty model which allows an explicit form for the marginal likelihood m , so that we can compute $p_\omega(m)$. However this is not so in general, for example, for models with lognormal distributed frailty, or with nested and/or serially correlated frailty. Thus, adjusted profile likelihoods based on h-likelihood are useful for general frailty models (Ha, Lee and MacKenzie, 2007).

Acknowledgments: This work was partially supported by the Korea Research Foundation Grant (KRF-2006-013-C00092).

References

- Barker, P. and Henderson, R. (2005). Small sample bias in the gamma frailty model for univariate survival. *Lifetime Data Analysis* **11**, 265-284.
- Breslow, N. E. (1972). Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society, Series B* **34**, 216-217.

- Gu, M.G., Sun, L. and Huang, C. (2004). A universal procedure for parametric frailty models. *Journal of Statistical Computation and Simulation* **74**, 1-13.
- Ha, I. D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics* **12**, 663-681.
- Ha, I. D. and Lee, Y. (2005). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models. *Biometrika* **92**, 717-723.
- Ha, I. D., Lee, Y. and MacKenzie, G. (2007). Model selection for multi-component frailty models. *Statistics in Medicine*. To appear.
- Ha, I. D., Lee, Y. and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika* **88**, 233-243.
- Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 619-678.
- Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987-1006.
- Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25-44.
- Rondeau, V., Commenges, D. and Joly, P. (2003) Maximum penalized likelihood estimation in a gamma frailty model. *Lifetime Data Analysis* **9**, 139-153.
- Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine* **19**, 3309-3324.

Generalized Linear Models for Assessing Performance

Karl Heiner¹ and John Hinde²

¹ State University of New York at New Paltz

² Irish National University, Galway

Abstract: This paper investigates the application of generalized linear models, including random effects models, in the area of performance monitoring. We briefly review some highlights from the performance monitoring literature, which is reviewed and summarized by a Royal Statistical Society Working Group in a 2005 JRSS paper. By performance monitoring we are often referring to the oversight and reporting by governmental agencies in a variety of areas, e.g. law enforcement, social services, and health care. In this context, we explore the use of simple general linear models (e.g. logistics regression), random effects logistics models, hierarchical models, and latent class models. Examples will be offered from the areas of social services and health care quality of care.

Keywords: generalized linear models, random effects, latent class models, finite mixture models, total quality management.

1 Background

The Royal Statistical Society Working Group on performance monitoring has developed and extensive report and guidelines for performance monitoring (JRSS, 2005). Among many recommendations, they point out the necessity for developing models that adjust for context to achieve comparability. Including covariates in performance models in an effort to achieve a degree comparability is sometimes known as adjustment for prior status, or in the health care quality literature, case mix adjustment or risk adjustment. The working group point to an example (Bridgewater, et. al., 2003) where individual cardiac surgeons and hospitals were compared with respect to mortality rates. In cases like these, risk adjustment is recommended to achieve some degree of comparability. The context adjustment variables are those outside the influence of the parties being assessed. In addition, the reliance on multiple indicators reported over time is advised.

Goldstein and Spiegelhalter (1996) highlight the need for incorporating uncertainty when reporting performance and enumerate limitations of league table when comparing institutional performance. League tables generally rank entities based on a point estimate of a single measure. Daniels and Gatsonis (1999) recommend using hierarchical generalized linear models when comparing hospitals, individual health care providers and geographic regions. These models allow one to assess contributing factors to outcomes and are thus useful when attempting to achieve comparability and explain variance. We have fit such models and discuss their utility in specific situations.

Mixture models are also recommended when attempting to account for multimodal distributions.

2 Some generalized linear models

Consider an example of a location that has instituted a total quality management (TQM) program. In a TQM program, individuals who are directly involved in providing a service (or, in general producing a product) and their supervisors review information that documents that service for a sample of occasions. The idea is that if the individuals involved in providing the service are also involved its evaluation of performance, quality will improve because those directly involved will have an investment in measuring the quality and will bring knowledge of the process that may be valuable in the quality improvement effort. In addition to this local quality control effort, suppose further that the location is only one of several providing such service and that there is some authority that also seeks to monitor and compare performance at the locations level. It has been the practice of the authority to employ a private independent auditor to assess performance. This independent auditor is initially assumed to be using some gold standard. Here, it would seem natural to fit risk adjusted general linear models for each location, or a model that includes random effects in lieu of fixed effects for locations.

If some of the locations being monitored decide to institute their own total quality management program that seeks to evaluate the location on the same measures as those used by the auditor, an opportunity is created whereby the resources of the oversight authority dedicated to this performance monitoring activity may be reduced. The authority, perhaps naively, may assume that each location is capable of assessing its own program and will report scores on each occasion sampled to the authority as reliably and accurately as would the auditor and that this will be accomplished without bias. In this case, a generalized linear model would allow the authority to score each facility and attempt to achieve comparability by introducing important covariates into the model. Frequently, the performance measure is an indicator variable. These models would be the same as the models that would be used if only auditor data were considered.

A TQM effort may be carried out at several locations, e.g. locations producing the same product within a company; local administrative agencies within a government where each agency is providing the same service, such as issuing food vouchers or medical coverage; or various health care providers providing similar services within a health management organization (HMO), or independently within political region as in hospitals within a state. Thus, we may wish to compare performance among locations using generalized linear models. For example, one might consider a logistic regression model when the performance measure is binary. Here, the covariates and the indicators in the design matrix should be centered so that meaningful scores are derived. Or, performance maybe defined to be time to respond or time until next treatment. In this case, various survival models are considered. Locations may be considered fixed effects or performance scores maybe shrunk by considering locations as random. The use of a more costly independent auditor may be employed to achieve comparability of measurement and to reduce "among location bias".

Another approach would involve reducing the amount of effort of the auditor by having the auditor select a subsample of each locations sampled occasions and apply the gold standard to this subsample. The locations data may then become a predictor of the gold standard and analysis of agreement data within the subsample could be used for training the evaluators at the locations where disagreement occurs. In these models, random effects models may prove more realistic and thus improve fit.

Frequently, the results of obtaining performance scores are made available to the public by a governmental oversight authority or by the news media. It is natural in such instances to compare locations. It is not uncommon for such comparisons to be made by constructing a league table based on point estimates of the performance measure for each location. Of course, this practice does not incorporate the inherent uncertainty associated with such estimates. We have found it useful to construct among location rank distributions. This is accomplished by making a draw from the distribution of the performance measure for each location and then ranking these draws. Now, Repeat this set of draws a large number of times, thus creating a distribution of ranks for each location. Displaying the results using ordered box plots provides a reasonable way of comparing locations. If the boxes for any two locations do not overlap it is unlikely that they have the same rank. Otherwise, the locations may be comparable.

Now suppose that there is some question as to whether the auditors assessment is actually better than the locations self-assessment. After all, the location has more familiarity with its own processes and the occasions that have been selected for evaluation. When the quality measure is an indicator of whether or not the service was provided in according to some definition of best practice, we might consider a latent class model where each occasion belongs to one of two classes, best practiced followed or not.

In an example of a sample of 48 occasions sampled from a particular locations population of occasions, an auditor has subsampled 25. In this example, the location and the auditor agreed 19 times. On the six occasions when they both agreed that the best practice was not followed, the simple latent class model assigned the cases to the not followed class, assigning a probability of 0.77 for each assignment. When there was agreement that best practice was followed, the 13 cases were assigned to the followed class assigning a probability of 0.78 for each assignment. There were six cases of disagreement, three of each type. In all six instances the cases were assigned to the followed class. This is because “best practice followed” was the most frequent assessment. The probability that best practices were followed was 0.61 when the location scored *yes* and 0.60 when the auditor scored *yes*. In the 23 instances when the auditor did not provide an assessment, model assigned the case consistent with the location, six *nos* and 17 *yesses*. Assignment probabilities were 0.63 for *nos* and 0.83 for *yesses*. This simple latent class model seems to produce sensible results. We may now expand this latent class model to include all locations with risk adjustment using fixed effects models and random effects models.

3 Summary

Generalized linear models are a powerful tool for assessing performance. We have outlined how logistics and survival models are used to develop performance scores for

locations and how rank distributions may be derived for comparison purposes. The use of covariates assist in achieving comparability. Gold standards may be modelled using self-assessment as a predictor and latent class models provide a way of combining performance scores from more than one source.

References

- Aitkin, M., Francis, B., and Hinde, J. (2005). *Statistical Modeling in GLIM4, (2nd edition)*. Oxford, UK.
- Daniels, M. and Gatsonis, C.A. (1999). Multilevel Hierarchical Generalized Linear Models with applications to health services research. *Journal of the American Statistical Association* **94**, 29-42.
- Eibeck, E., and Hinde, J. (2006). Nonparametric maximum likelihood estimation for random effect models in R. *Technical Report IRL-GLWY-2006*. National University of Ireland, Galway.
- Goldstein, H., and Spiegelhalter, D.J. (1996). League Tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A* **159**, 385-443.
- Working Party on Performance Monitoring in the Public Services (2005). Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society, Series A* **168**, 1-27.

Randomly Stopped Models

Gillian Heller¹, Mikis Stasinopoulos² and Robert Rigby²

¹ Department of Statistics, Macquarie University, Sydney, Australia.

² STORM Research Centre, London Metropolitan University, U.K.

Abstract: We introduce a method for modelling a continuous response which is the sum of a random number of terms. Examples are total insurance claim sizes (the total of all claims on a policy in a year), or total amount spent by credit card holders in a sector in a month, where there may be multiple spending episodes. The distribution of the number of terms may be, in principle, any discrete distribution defined on the non-negative integers; and each term has a continuous, right-skewed distribution. The resulting marginal distribution of the total amount is a mixed discrete-continuous model, with a probability mass at zero and a continuous component. The model explicitly specifies log-linear models for the four parameters in the total amount distribution. It may be fitted to data using a package written in R. The method is illustrated on an Australian motor vehicle insurance data set.

Keywords: Mixture distribution; mean and dispersion modelling; insurance claims; compound Poisson; discrete mixture.

1 Introduction

We consider models for an outcome variable which is a sum of a random number of non-negative random variables. An example is total insurance claim size in a period, where there can be C claims on a policy within the period, $C = 0, 1, 2, \dots$. In a fixed period, a policy will either experience no claim, in which case the claim amount is identically zero; or one or more claims, which are non-negative amounts typically having extremely right-skewed distributions. The distribution of the total is mixed discrete-continuous: a continuous, right-skewed distribution with a single probability mass at zero. The model explicitly specifies log-linear models for the four parameters in the distribution of the total amount.

We assume that the number of terms (C) in the sum and the total amount (Y) are recorded for each individual, along with the exposure time t ($0 < t \leq 1$) and suitable explanatory variables. The total amount is

$$Y = \begin{cases} 0 & \text{if } C = 0 \\ \sum_{j=1}^C Z_j & \text{if } C = 1, 2, \dots \end{cases} \quad (1)$$

where Z_j is the size of the j th term. Stuart and Ord (1994) call Y a randomly stopped sum. Clearly we have a bivariate dependent variable (Y, C) with joint distribution $f_{Y,C}(y, c) = f_{Y|C}(y) f_C(c)$. We adopt the term *randomly stopped models* (RSMs) to describe models having the above type of likelihood. Specification of the model requires specification of the distributions defining the likelihood function. We approach the

problem by specifying firstly the distribution of the number of terms, $f_C(c)$, and secondly the continuous (conditional) distribution of the total amount, Y , given the number of terms C , i.e. $f_{Y|C}(y)$. We also discuss the marginal distribution of the total amount, $f_Y(y)$, together with problems of estimating model parameters.

1.1 The distribution of C

For a unit exposure time, i.e. $t = 1$, the distribution of the number of terms C can be any discrete non-negative distribution defined on $C = 0, 1, 2, 3, \dots$, with mean μ_c . We use the notation $f_C(c; \mu_c, \sigma_c) = P(C = c; \mu_c, \sigma_c)$ for the distribution of C , where σ_c is a dispersion parameter (if required). The Poisson distribution is an obvious choice for f_C ; with exposure time of t , a reasonable model for the number of terms is $C \sim PO(t\mu_c)$. In order to make the model more flexible, we introduce the random variable W having $E(W) = 1$, which operates multiplicatively on the mean number of terms and reflects individual heterogeneity. The conditional distribution of C given W is assumed to be Poisson, i.e. $C|W \sim PO(t\mu_c W)$, and W has any distribution defined on the non-negative real line with mean one. Hence, given W , C follows a Poisson process over time with mean $\mu_c W$ per unit time. Note that this is effectively a multiplicative random effect (or compound Poisson) model for C , where the value of the random effect W for an individual is assumed constant throughout the exposure time t . The marginal distribution of C has its mean preserved as $t\mu_c$ and exposure can be taken into account by offsetting the mean of the distribution of C .

Example 1. Let W have a gamma distribution: $W \sim GA(1, \sigma_c^{1/2})$, then the marginal distribution of C is $C \sim NBI(t\mu_c, \sigma_c)$, the negative binomial distribution of type I. In the insurance claims example, the explanation of this model is that policyholders have heterogeneous claim rates, which are gamma distributed over the population.

Example 2. Let W be binary with distribution:

$$P(w) = \begin{cases} \sigma_c & w = 0 \\ 1 - \sigma_c & w = 1/(1 - \sigma_c) \end{cases} \quad (2)$$

Then the marginal distribution of C is a zero inflated Poisson distribution. The interpretation of this model in the insurance claims context is that there are two sub-populations: one consisting of individuals who never make a claim (comprising the proportion σ_c of the population) and the other (the remaining proportion $1 - \sigma_c$) who generate claims according to the $PO(t\mu_c/[1 - \sigma_c])$ distribution.

1.2 The conditional distribution of Y given C

In principle the conditional distribution of Y given C can be any distribution defined on the positive real line. Here we take the approach that, since Y is the sum of individual terms Z_j , it is desirable for the distribution of Y to have the same form as the distribution of the individual Z_j . The Z_j are continuous non-negative, iid random variables, each with density $f_Z(z; \mu_z, \sigma_z)$, where μ_z is a location, and σ_z a scale parameter. We assume further that f_Z has the reproductive property :

$$Z_j \stackrel{iid}{\sim} f_Z(z; \mu_z, \sigma_z), \quad j = 1, \dots, c \Rightarrow Y|(C = c) \sim f_Z(y; c\mu_z, \sigma_{y|c}), \quad (3)$$

where $\sigma_{y|c}$ is a simple function of σ_z and c . Hence

$$f_{Y|C=0}(y) = \begin{cases} 1 & y = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{Y|C>0}(y) = \begin{cases} f_Z(y; c\mu_z, \sigma_{y|c}) & y > 0 \\ 0 & \text{otherwise} \end{cases}$$

with $E_{Y|C}(y) = \mu_{y|c} = c\mu_z$. Property (3) holds for f_Z in the exponential family of distributions, which includes the Gaussian, the gamma and the inverse Gaussian distributions.

1.3 The marginal distribution of Y

The unconditional (marginal) distribution of Y , given that condition (3) is satisfied, is the mixture

$$f_Y(y) = \sum_{c=0}^{\infty} f_{Y|C}(y)f_C(c) = \begin{cases} f_C(0) & y = 0 \\ \sum_{c=1}^{\infty} f_Z(y; c\mu_z, \sigma_{y|c})f_C(c) & y > 0 \end{cases} \quad (4)$$

where $\sigma_{y|c} = \sigma_z c^k$ with $k = \frac{1}{2}$ for the Gaussian, $k = -\frac{1}{2}$ for the gamma and $k = -1$ for the inverse Gaussian distributions, respectively. Note that $f_Y(y)$ is a continuous non-negative distribution, with probability mass $P(0; \mu_c, \sigma_c)$ (i.e. the probability of no terms) at zero. Figure 1 displays an example of $f_C(c)$ which is negative binomial, and $f_Z(z)$ which is inverse Gaussian, and the resulting marginal distribution $f_Y(y)$. Note also if $f_C(c)$ is Poisson and $f_Z(z)$ is gamma, then the resulting $f_Y(y)$ is the Tweedie distribution, see Smyth and Jørgensen (2002).

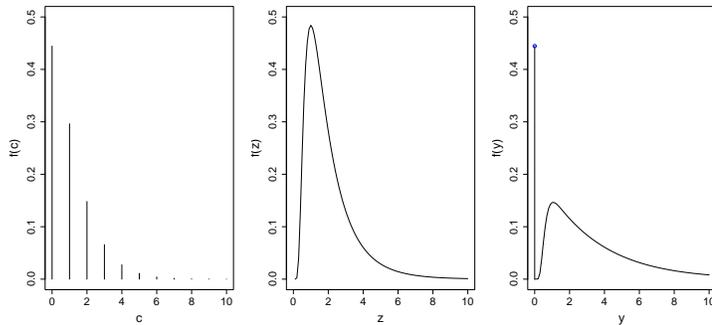


FIGURE 1. Negative binomial $f_C(c)$, inverse Gaussian $f_Z(z)$, and resulting $f_Y(y)$.

1.4 Modelling means and dispersions in terms of covariates

The likelihood function of a RSM will have possibly up to four parameters. Following the philosophy of generalized additive models for location, scale and shape (Rigby and

Stasinopoulos 2005), where all the parameters of the likelihood can be modelled as functions of explanatory variables, we specify the following models on the parameters μ_c , σ_c , $\mu_{y|c}$ and $\sigma_{y|c}$:

$$\ln(\mu_c) = \mathbf{x}'_{11}\beta_1 + h_1(\mathbf{x}_{12}) + \ln(t) \quad (5)$$

$$\ln(\sigma_c) = \mathbf{x}'_{21}\beta_2 + h_2(\mathbf{x}_{22}) \quad (6)$$

$$\ln(\mu_{y|c}) = \mathbf{x}'_{31}\beta_3 + h_3(\mathbf{x}_{32}) + \ln(c) \quad (7)$$

$$\ln(\sigma_{y|c}) = \mathbf{x}'_{41}\beta_4 + h_4(\mathbf{x}_{42}) + k \ln(c) \quad (8)$$

where \mathbf{x}_{1i} , \mathbf{x}_{2i} , \mathbf{x}_{3i} and \mathbf{x}_{4i} , for $i = 1, 2$ are covariate vectors for μ_c , σ_c , $\mu_{y|c}$ and $\sigma_{y|c}$ respectively, which may be different, the same, or may have some but not all elements in common; β_1 , β_2 , β_3 and β_4 are the corresponding parameter vectors; and h_1 , h_2 , h_3 and h_4 are nonparametric functions, typically smoothing splines. Exposure t ($0 < t \leq 1$) is corrected for as an offset in (5) for the mean number of terms. The number of terms c is corrected for by the offsets $\ln(c)$ in (7) and $k \ln(c)$ in (8), where $k = \frac{1}{2}$, $-\frac{1}{2}$ or -1 if the z_j have a Gaussian, gamma or inverse Gaussian distribution respectively. In model equations (5)-(8) logarithmic link functions have been specified. However, in principle any other monotonic, differentiable function may be used as link.

1.5 Estimation

The log-likelihood function of a RSM has the form

$$\ln f_C(c; \mu_c, \sigma_c) + \ln f_{Y|C}(y; \mu_{y|c}, \sigma_{y|c})$$

so maximization can be achieved in two stages: first by maximizing $\ln f_C(c; \mu_c, \sigma_c)$ with respect to the parameters involved in the determination of μ_c and σ_c in equations (5) and (6); and then by maximizing $\ln f_{Y|C}(y; \mu_{y|c}, \sigma_{y|c})$ with respect to the parameters involved in equations (7) and (8) for $\mu_{y|c}$ and $\sigma_{y|c}$ respectively.

The randomly stopped model has been implemented as the `rsm` package in R. In the current version, the Poisson, negative binomial, Poisson inverse Gaussian and zero inflated Poisson are available as options for the discrete distribution of the number of terms $f_C(c)$, and the gamma, inverse Gaussian, and normal for the conditional distribution of total amount $f_{Y|C}(y)$. Maximum (penalized) likelihood estimation is used. The penalized log likelihood function of the each of the two likelihoods above is maximized iteratively using either the RS or CG algorithm of Rigby and Stasinopoulos (2005), which in turn uses a back-fitting algorithm to perform each step of the Fisher scoring procedure. Both RS and CG algorithms use the log likelihood of the data, and its first derivatives (and optionally expected second derivatives) with respect to distributional parameters.

2 Application to motor vehicle insurance

We illustrate the method on a class of motor vehicle insurance policies from an Australian insurance company over a twelve-month period in 2004-05. There were approximately 68,000 policies, of which 93% had no claim in the period of observation. Of

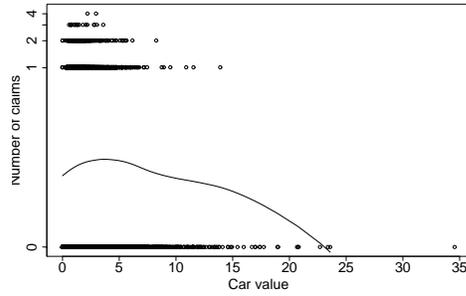


FIGURE 2. Number of claims (log scale) against car value

those that had a claim, 94% had one claim, and the remaining policies between 2 and 4 claims. Table 1 lists potential covariates. The scatterplot of number of claims versus car value is displayed in Figure 2. The figure suggests a nonlinear relationship. Model selection comprises selecting the distributions of the number of claims and the claim size, as well as the sets of covariates for $\mu_c, \sigma_c, \mu_{y|c}$ and $\sigma_{y|c}$. Using the AIC as model selection criterion, the following final model was selected:

$$\log(\mu_c) = \text{age} + \text{area} + \text{body type} + h_1(\text{car value}) + \text{offset}\{\ln(t)\} \tag{9}$$

$$\log(\sigma_c) = \text{car value}$$

$$\ln(\mu_{y|c}) = \text{age} + \text{gender} + \text{area} + \text{offset}\{\ln(c)\} \tag{10}$$

$$\ln(\sigma_{y|c}) = \text{area} + \text{offset}\{k \ln(c)\}$$

where h_1 is the quadratic function, with the negative binomial type I distribution for the number of claims C , and the inverse Gaussian distribution for $Y|C$. Net premium is calculated as the expected claim size $\hat{\mu}_y = \hat{\mu}_c \hat{\mu}_z$, where $\hat{\mu}_c$ is the fitted value from (9) (with $t = 1$), and $\hat{\mu}_z$ the fitted value from (10) (with $c = 1$).

TABLE 1. Covariates for car insurance claims

Covariate	Values
Driver age group	1 (youngest), 2, 3, 4, 5, 6
Gender	male, female
Area of residence	A, B, C, D, E, F
Car value/10,000	0-35
Make	A, B, C, D
Body type	bus, convertible, coupe, hatchback, hardtop, motorized caravan/combi, minibus, panel van, roadster, sedan, station wagon, truck, utility

3 Conclusion

We introduce a method for modelling random sums of continuous random variables using a randomly stopped model, which explicitly specifies log-linear models for the mean number of terms; the dispersion of the number of terms; the mean of a single term; and the dispersion of single term. In principle, any discrete distribution defined on the non-negative integers may be specified for the number of terms; and any continuous, right-skewed distribution for term sizes. Covariates may be in the model equations as parametric forms or as smooth functions. Use of the gamma or inverse Gaussian distributions accommodates the extreme right skewness typically encountered in financial data.

References

- Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics* **54**, 507-554.
- Smyth, G.K. and Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin* **32(1)**, 143-157.
- Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory*, Sixth Edition. London: Edward Arnold.

Recursive estimation of the uncertainty probability in nonlinear systems with uncertain observations

A. Hermoso-Carazo¹ and J. Linares-Pérez¹

¹ Dpto. Estadística e I.O., Universidad de Granada, 18071 Granada, Spain,
ahermoso@ugr.es, jlinares@ugr.es

Abstract: The estimation problem of the uncertainty probability in nonlinear systems with uncertain observations is addressed by a bayesian approach. From an *a priori* Beta distribution, a recursive algorithm to approximate the *a posteriori* mean is obtained. The state and observation estimators, which are required to approximate the *a posteriori* densities, are obtained by applying an extension of the unscented Kalman filter to systems with uncertain observations.

Keywords: Uncertain observations; Bayes estimation; Unscented Kalman filter

1 Introduction

Linear and nonlinear systems with *uncertain observations*, characterized by the fact that the observation at each time can be (randomly) only noise, appear commonly in many practical situations, and the signal estimation problem has been widely investigated when the probability that each observation contains the signal is known. This paper considers the situation in which such uncertainty probability is unknown and, using a bayesian approach, an algorithm to estimate it from the observations of the system is proposed. By assuming an *a priori* distribution for the unknown parameter, the problem is to approximate its conditional mean given the successive observations. This problem connects with that of the state and observation estimation and, since nonlinear equations are considered, approximations of the optimal estimators must be used. The estimation problem in nonlinear systems has received a considerable attention in the last years and different techniques have been used to address it (see Daum 2005). In particular, the so-called *unscented Kalman filter* has been shown to be a good approximation in a variety of application domains. An extension of this filter for nonlinear systems with uncertain observations, developed in Hermoso-Carazo and Linares-Pérez (2007), is applied in this paper. The proposed estimators for the uncertainty probability are obtained by starting from a Beta as *a priori* distribution; after successive approximations of Beta mixtures, a recursive estimation algorithm is derived.

2 System model and problem statement

Consider a nonlinear discrete-time system with uncertain observations:

$$\begin{aligned} x_{k+1} &= f_k(x_k) + w_k, & k \geq 0, \\ y_k &= \gamma_k h_k(x_k) + v_k, & k \geq 1, \end{aligned} \quad (1)$$

where x_k and y_k represent the state and observation vectors, respectively, f_k and h_k are arbitrary functions, and the following hypotheses are satisfied:

- (H1) The initial state x_0 is a Gaussian vector, $\mathcal{N}(0, P_0)$.
- (H2) The noise process $\{w_k; k \geq 0\}$ is a zero-mean white Gaussian sequence with autocovariance function $E[w_k w_k^T] = Q_k$, for all $k \geq 0$.
- (H3) The variables $\{\gamma_k; k \geq 1\}$ describing the uncertainty in the observations are independent Bernoulli random variables with $P[\gamma_k = 1] = p$, for all $k \geq 1$, and the uncertainty probability, p , is unknown.
- (H4) The noise process $\{v_k; k \geq 1\}$ is a zero-mean white Gaussian sequence with autocovariance function $E[v_k v_k^T] = R_k$, for all $k \geq 1$.
- (H5) The initial state and the noise processes are mutually independent.

The problem is to provide a recursive algorithm for the estimation of the unknown parameter p from successive observations of the signal process given by (1). By starting from an *a priori* density, $\pi(p/Y^0)$, and assuming a quadratic loss function, our aim is to approximate the Bayes estimator of p ; that is, given the observations $Y^k = \{y_1, \dots, y_k\}$ and denoting the *a posteriori* density by $\pi(p/Y^k)$, we will approximate the conditional mean, $\bar{p}_k = \int_0^1 p \pi(p/Y^k) dp$, $k \geq 1$, by starting from the initial estimator \bar{p}_0 , the mean of the *a priori* distribution. In pursuit of the recursivity, and noting $g(y_1/p, Y^0) = g(y_1/p)$, the *a posteriori* density is written as

$$\pi(p/Y^k) = \frac{g(y_k/p, Y^{k-1}) \pi(p/Y^{k-1})}{\int_0^1 g(y_k/p, Y^{k-1}) \pi(p/Y^{k-1}) dp}, \quad k \geq 1 \quad (2)$$

and hence, the density $g(y_k/p, Y^{k-1})$ must be computed. For this purpose, denoting $g(y_1/\gamma_1 = j, Y^0) = g(y_1/\gamma_1 = j)$, $j = 0, 1$, and since

$$g(y_k/p, Y^{k-1}) = p g(y_k/\gamma_k = 1, Y^{k-1}) + (1-p) g(y_k/\gamma_k = 0, Y^{k-1}),$$

the computation of $\pi(p/Y^k)$ requires that of the densities $g(y_k/\gamma_k = j, Y^{k-1})$ for $j = 0, 1$. From (H4) and (H5), it is clear that $g(y_k/\gamma_k = 0, Y^{k-1})$ corresponds to the Gaussian distribution $\mathcal{N}(0, R_k)$. However, the computation of $g(y_k/\gamma_k = 1, Y^{k-1})$ requires an ever-growing amount of memory due to the multiplicative noise component in the observations; consequently, it becomes necessary to find approximations of such densities which are more viable from a computational viewpoint.

3 Approximation of the *a posteriori* density

At any time $k \geq 1$ we start with approximations of the mean and covariance of x_{k-1} given Y^{k-1} , $\hat{x}_{k-1/k-1}$ and $P_{k-1/k-1}^x$, and define the σ -points:

$$\begin{aligned} \chi_{k-1/k-1}^0 &= \hat{x}_{k-1/k-1} \\ \chi_{k-1/k-1}^i &= \hat{x}_{k-1/k-1} + \left(\sqrt{(n+\lambda)P_{k-1/k-1}^x} \right)_i, \quad i = 1, \dots, n \\ \chi_{k-1/k-1}^i &= \hat{x}_{k-1/k-1} - \left(\sqrt{(n+\lambda)P_{k-1/k-1}^x} \right)_{i-n}, \quad i = n+1, \dots, 2n \\ W_0^{(m)} &= \lambda/(n+\lambda), \quad W_0^{(c)} = \lambda/(n+\lambda) + (1-\alpha^2 + \beta) \\ W_i^{(m)} &= W_i^{(c)} = 1/2(n+\lambda), \quad i = 1, \dots, 2n, \quad \lambda = \alpha^2(n+\kappa) - n. \end{aligned}$$

From the state equation, the mean and covariance of x_k given Y^{k-1} are approximated by $\hat{x}_{k/k-1}$ and $P_{k/k-1}^x = P_{k/k-1} + Q_{k-1}$, with

$$\begin{aligned} \hat{x}_{k/k-1} &= \sum_{i=0}^{2n} W_i^{(m)} f_{k-1}(\chi_{k-1/k-1}^i), \\ P_{k/k-1} &= \sum_{i=0}^{2n} W_i^{(c)} \left(f_{k-1}(\chi_{k-1/k-1}^i) - \hat{x}_{k/k-1} \right) \left(f_{k-1}(\chi_{k-1/k-1}^i) - \hat{x}_{k/k-1} \right)^T. \end{aligned}$$

Now, we take a set of σ -points $\chi_{k/k-1}^i$ associated to $\hat{x}_{k/k-1}$ and $P_{k/k-1}^x$, and the statistics of $h_k(x_k)$ given Y^{k-1} are approximated by those of the transformed, $h_k(\chi_{k/k-1}^i)$. So the conditional mean and covariance of y_k given $\gamma_k = 1, Y^{k-1}$ are approximated by the following sums:

$$\begin{aligned} \hat{y}_{k/k-1}^1 &= \sum_{i=0}^{2n} W_i^{(m)} h_k(\chi_{k/k-1}^i), \\ P_{k/k-1}^{y1} &= \sum_{i=0}^{2n} W_i^{(c)} \left(h_k(\chi_{k/k-1}^i) - \hat{y}_{k/k-1}^1 \right) \left(h_k(\chi_{k/k-1}^i) - \hat{y}_{k/k-1}^1 \right)^T + R_k. \end{aligned}$$

The knowledge of the first and second-order moments and the Gaussianity of the processes in (1) allow us to take the approximation $\mathcal{N}(\hat{y}_{k/k-1}^1, P_{k/k-1}^{y1})$ for the distribution of y_k given $\gamma_k = 1, Y^{k-1}$ and, thus, an approximation for $\pi(p/Y^k)$ is obtained. For the next step the following expression is used:

$$g(y_k/Y^{k-1}) = \bar{p}_{k-1}g(y_k/\gamma_k = 1, Y^{k-1}) + (1 - \bar{p}_{k-1})g(y_k/\gamma_k = 0, Y^{k-1})$$

and $\hat{x}_{k/k-1}$ and $P_{k/k-1}^x$ are updated from the Kalman filter equations:

$$\begin{aligned} \hat{x}_{k/k} &= \hat{x}_{k/k-1} + \bar{p}_{k-1}P_{k/k-1}^{xy1}\Pi_{k/k-1}^{-1}(y_k - \bar{p}_{k-1}\hat{y}_{k/k-1}^1), \quad k \geq 1; \quad \hat{x}_{0/0} = 0, \\ P_{k/k}^x &= P_{k/k-1}^x - \bar{p}_{k-1}^2P_{k/k-1}^{xy1}; \quad P_{0/0}^x = P_0, \end{aligned}$$

with

$$\begin{aligned} \Pi_{k/k-1} &= \bar{p}_{k-1}P_{k/k-1}^{y1} + R_k + \bar{p}_{k-1}(1 - \bar{p}_{k-1})\hat{y}_{k/k-1}^1\hat{y}_{k/k-1}^{1T}, \\ P_{k/k-1}^{xy1} &= \sum_{i=0}^{2n} W_i^{(c)} \left(\chi_{k/k-1}^i - \hat{x}_{k/k-1} \right) \left(h_k(\chi_{k/k-1}^i) - \hat{y}_{k/k-1}^1 \right)^T. \end{aligned}$$

4 Recursive estimators of the uncertainty probability

The procedure in Section 3 provides a method to approximate the density $\pi(p/Y^k)$ and, from it, the estimator \bar{p}_k . However, the computation grows in complexity with k , depending, in each case, on the selected *a priori* distribution. A new approximation is now proposed by considering a *Beta* as a *a priori* distribution. The *a posteriori* density (2) is rewritten as

$$\pi(p/Y^k) = \delta_k \frac{p\pi(p/Y^{k-1})}{\bar{p}_{k-1}} + (1 - \delta_k) \frac{(1-p)\pi(p/Y^{k-1})}{1 - \bar{p}_{k-1}}, \quad k \geq 1$$

with

$$\delta_k = \frac{\bar{p}_{k-1}g(y_k/\gamma_k = 1, Y^{k-1})}{g(y_k/Y^{k-1})}, \quad k \geq 1.$$

By starting from a $\beta(\alpha_0, \beta_0)$ as a *a priori* distribution, the initial estimator is $\bar{p}_0 = \alpha_0(\alpha_0 + \beta_0)^{-1}$, and $\pi(p/Y^1)$ is the mixture (with parameter δ_1) of the distributions $\beta(\alpha_0 + 1, \beta_0)$ and $\beta(\alpha_0, \beta_0 + 1)$. By approximating this mixture by a single Beta with the same mean, $\beta(\alpha_0 + \delta_1, \beta_0 + 1 - \delta_1)$, and reasoning similarly in the following steps, we find that the *a posteriori* distribution given Y^k and the estimator are approximated by

$$\pi(p/Y^k) \equiv \beta\left(\alpha_0 + \sum_{i=1}^k \delta_i, \beta_0 + \sum_{i=1}^k (1 - \delta_i)\right), \quad \bar{p}_k = \frac{\alpha_0 + \sum_{i=1}^k \delta_i}{\alpha_0 + \beta_0 + k}, \quad k \geq 1.$$

It is easy to see that the estimators satisfy the following recursive relation

$$\bar{p}_k = \bar{p}_{k-1} - \frac{1}{\alpha_0 + \beta_0 + k} [\bar{p}_{k-1} - \delta_k], \quad k \geq 1; \quad \bar{p}_0 = \frac{\alpha_0}{\alpha_0 + \beta_0}.$$

Remark: The replacement of \bar{p}_k in the estimation algorithms proposed in Hermoso-Carazo and Linares-Pérez (2007), provides adaptive algorithms when the probability p is unknown.

Acknowledgments: This work is partially supported by the ‘Ministerio de Ciencia y Tecnología’ through the project MTM2005-03601.

References

- Daum, F. (2005). Nonlinear filters: Beyond the Kalman filter. *IEEE Aerospace and Electronic Systems Magazine* **20(8)**, 57-69.
- Hermoso-Carazo, A., and Linares-Pérez, J. (2007). Different approaches for state filtering in nonlinear systems with uncertain observations, *Applied Mathematics and Computation*, doi:10.1016/j.amc.2006.08.083.

Change Point Detection for Panel Data Models

Johannes Hofrichter¹, Herwig Friedl²

¹ Institute of Applied Statistics, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Steyrergasse 25a, 8010 Graz, Austria

² Institute of Statistics, Graz University of Technology, Steyrergasse 17, 8010 Graz, Austria

Abstract: A new method for estimating continuous change points for panel data is presented. For each panel a generalized linear model with two change points is considered. It is further assumed that the fitted values at the change point of any two consecutive segments are the same. Moreover, for all these panels it is especially assumed that the slope parameter in the last segments is the same. The performance of a new estimation method is investigated by a Monte Carlo simulation study. Finally this method is applied on hydrological runoff data.

Keywords: Hydrological Runoff Model; Segmented Generalized Linear Model; Continuous Change Points; Monte Carlo Simulation.

1 Introduction

This work is motivated by a real problem in hydrology. The general question is to find a suitable statistical model that allows to describe specific properties of the catchment of a river. One of the properties of interest is the volume of the groundwater storage. Information about this volume can be obtained by analyzing the runoff of a river after a heavy rainstorm. Under certain circumstances, we will be able to recognize two changes in the time dependent behavior of the runoff. A main part of the hydrological analysis is to determine these two points of change which divide the data into three segments. As shown in Hofrichter (2007) the runoff in each segment can be described by a generalized linear model (GLM), the challenge is to detect the change points in such a recession model. Usually there exists data from more than one runoff due to the fact that more than one rainstorm occurs over the entire observation period. Thus, such a data set can be interpreted as a panel data set where each panel belongs to the time after a rainstorm. According to hydrological considerations a common slope for all last segments is mandatory. Thus, we like to estimate both change points in such a GLM for panel data where the slope parameter in the last segments is the same for all panels.

2 Theory

In what follows we only consider GLMs with changes in the mean structure. This means that we do not allow for different link or variance functions. We further assume that the dispersion parameter is the same for all segments. The changes are determined by some specific data points, usually called the *change points*. In particular, we especially

utilize GLMs with *continuous* changes at all change points. Consequently, the means in the change point of two consecutive models have to be the same.

2.1 GLMs with two continuous change points

Let (x_i, y_i) , $i = 1, \dots, n$, denote independent pairs of observations, where y_i is the response and x_i some explanatory variable. Let us further assume that these n pairs can be arranged in some natural ordering of the predictor variable, i.e. $x_i \leq x_{i+1}$. The two continuous change points, say γ_1 and γ_2 , partition the data into three segments. The parameters of interest are the segment specific coefficients $\beta_k = (\beta_{k0}, \beta_{k1})^T$, $k = 1, 2, 3$, and the change points $\gamma = (\gamma_1, \gamma_2)^T$ of the model

$$g(\mu_i) = \begin{cases} \beta_{10} + \beta_{11}x_i & a \leq x_i \leq \gamma_1 \\ \beta_{20} + \beta_{21}x_i & \gamma_1 < x_i \leq \gamma_2 \\ \beta_{30} + \beta_{31}x_i & \gamma_2 < x_i \leq b, \end{cases}$$

where $\mu_i = E(y_i|x_i)$, $a = \min_i x_i$, and $b = \max_i x_i$. The continuity constraint at e.g. change point γ_1 is

$$g(E[y_i|\gamma_1]) = g(\beta_{10} + \beta_{11}\gamma_1) = g(\beta_{20} + \beta_{21}\gamma_1).$$

As the link function is monotonic differentiable, the continuity constraint can be simplified here to

$$\beta_{10} + \beta_{11}\gamma_1 = \beta_{20} + \beta_{21}\gamma_1.$$

If the change points are unknown, no closed form solution of the estimates of the linear parameters and the change points exists. Stasinopoulos and Rigby (1992) suggested to use a grid search in order to find the estimate of the change points. Notice that the change points are not restricted to any observed values of the explanatory variable. Hence, the grid can be arbitrarily chosen. Küchenhoff (1997) first proposed an exact method to determine the estimate of the change points. His method consists of two steps. First, assume that both change points γ_1 and γ_2 lie in arbitrary open intervals $(x_r; x_{r+1})$ and $(x_s; x_{s+1})$, respectively, where $2 < r < s < n - 2$. Then calculate a candidate of the maximum likelihood estimate (MLE) of the change points. If this candidate lies in the assumed interval, it is the MLE. In the case where this candidate is not an element of the corresponding interval, the boundaries of the rectangle $[x_r; x_{r+1}] \times [x_s; x_{s+1}]$ are investigated. These two steps are done for all feasible rectangles and the MLEs of the change points and the linear parameters are those values, which yield the global maximum of these locally maximized likelihoods.

2.2 Panel data with a common slope in the last segment

Now we extend these change point models to panel data with $j = 1, \dots, n$ panels, where each panel consists of $i = 1, \dots, n_j$ observations. This notation allows either balanced panels, $n_1 = n_2 = \dots = n_n$, or unbalanced panels with $n_{j_1} \neq n_{j_2}$ for at least one $j_1 \neq j_2$ with $j_1, j_2 \in \{1, \dots, n\}$. As before we assume that the response distribution is the same for all panels and segments within each panel. Furthermore, we again do not allow for different link and variance functions nor for different dispersions.

Let $\mathbf{y}_j = (y_{j1}, \dots, y_{jn_j})^T$ be the response vector of panel j , where y_{ji} follows a distribution from the linear exponential family, and assume that the corresponding explanatory variable x_{ji} has a natural ordering, i.e. $x_{ji} \leq x_{j,i+1}$. The two continuous change points $\gamma_{j1}, \gamma_{j2} \in [x_{j1}; x_{jn_j}]$, with $\gamma_{j1} < \gamma_{j2}$, partition the data into three segments. In addition, we assume that the slope parameter in the last segment is the same for all panels and write the model as

$$g(\mu_{ji}) = \begin{cases} \beta_{j10} + \beta_{j11}x_{ji} & a_j \leq x_{ji} \leq \gamma_{j1} \\ \beta_{j20} + \beta_{j21}x_{ji} & \gamma_{j1} < x_{ji} \leq \gamma_{j2} \\ \beta_{j30} + \delta x_{ji} & \gamma_{j2} < x_{ji} \leq b_j \end{cases}$$

with continuity constraints

$$\begin{aligned} \beta_{10} + \beta_{11}\gamma_{j1} &= \beta_{20} + \beta_{21}\gamma_{j1} \\ \beta_{20} + \beta_{21}\gamma_{j2} &= \beta_{30} + \delta\gamma_{j2}. \end{aligned}$$

Let \mathbf{X}_j be the design matrix of panel j belonging to β_j . Furthermore, let \mathbf{T}_j the column vector of the explanatory variable which corresponds to the last segment. Notice that the structure of \mathbf{X}_j and \mathbf{T}_j depends on the location of the change points. The mean vector of panel j is $\boldsymbol{\mu}_j = (\mu_{j1}, \dots, \mu_{jn_j})^T$ and the global model for all panels can be written as

$$\begin{pmatrix} g(\boldsymbol{\mu}_1) \\ g(\boldsymbol{\mu}_2) \\ \vdots \\ g(\boldsymbol{\mu}_n) \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{T}_1 \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} & \mathbf{T}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_n & \mathbf{T}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \\ \delta \end{pmatrix}. \tag{1}$$

The problem here is that estimating all unknowns by utilizing either a grid search or the exact method is not really rational within feasible time. Thus, the idea is to divide the fitting procedure into two steps (for details see Hofrichter, 2007). In the first step we estimate the common slope for the last segments. In the second step, given this estimate of the common slope, we estimate the change points. The advantage of this separation is, that for a given last slope δ , the design matrix of the global model can be divided into two terms. The linear terms corresponding to the common slope can be interpreted as an offset in these models. Thus, model (1) can be written as

$$\begin{pmatrix} g(\boldsymbol{\mu}_1) \\ g(\boldsymbol{\mu}_2) \\ \vdots \\ g(\boldsymbol{\mu}_n) \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} + \delta \begin{pmatrix} \mathbf{T}_1 \\ \mathbf{T}_2 \\ \vdots \\ \mathbf{T}_n \end{pmatrix}.$$

The consequence is that the remaining design matrix is block diagonal. Therefore, each panel can be analyzed separately using either a grid search or the exact method by Küchenhoff (1997).

Briefly, the two tasks of this algorithm are:

1. Estimate a common slope δ for the last (right most) segments for all panels.

TABLE 1. Empirical (MC) means and standard deviations of the common slope estimate and the three estimated change points, based on 1000 replications.

parameter	mean	std.dev.	true value
$\hat{\delta}$	-0.101	0.009	-0.10
$\hat{\gamma}_1$	7.749	0.363	7.75
$\hat{\gamma}_2$	8.200	0.445	8.20
$\hat{\gamma}_3$	5.741	0.329	5.75

- Given an estimate of δ , we estimate the parameters in the GLM with continuous change points for each panel separately.

These two steps are iterated until the algorithm has converged. The performance of this algorithm is investigated by the following Monte Carlo (MC) simulation study.

2.3 Monte Carlo simulation

We especially consider some panel data consisting of three different panels with the common slope $\delta = -0.1$. For each panel, a GLM with one continuous change point is assumed. The response variable in all six segments follows a Poisson distribution with canonical link function, i.e. $g(\mu) = \log(\mu)$, for its mean μ . Furthermore, the parameters of the three left segments and the location of the respective continuous change points are different as also the domains of the explanatory variables. For this MC simulation 1000 panel data were generated. The empirical mean and the empirical standard deviation of these 1000 estimates of the common slope and the change points γ_1 , γ_2 , and γ_3 and their corresponding true values are listed in Table 1. There is a negligible bias for all four parameter estimates and it seems that this new algorithm is appropriate to analyze such kind of data.

3 Application

This new method is applied on 13 runoff curves monitored at the river Sulm at Leibnitz (Austria) during 1999. These curves are plotted in Figure 1. As already argued in Hofrichter (2007), hydrological considerations result in a GLM for Gaussian responses, where $\mu^{-1/2}$ is described by the linear predictor $\beta_0 + \beta_1 t$, t denoting time in days after a rainstorm. In addition, there is a strong evidence that the recession process changes twice over time. The direct flow is followed by a so called surface flow, which then finally changes to the base flow. The time points of these changes are of course unknown and should be simultaneously estimated together with all other parameters. One aspect of this analysis is to get an appropriate estimate for the common slope for the base flow, which is a parameter describing some properties of such a catchment.

We also compared the results when modelling each recession curve separately or together by utilizing the model described before. Figure 2 shows the fits of these two models for one specific recession curve. It can be seen that the respective estimates of

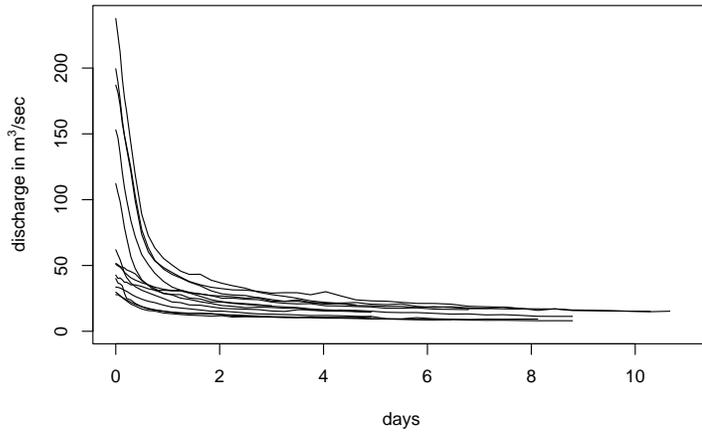


FIGURE 1. 13 runoff curves observed at the river Sulm at Leibnitz (Austria).

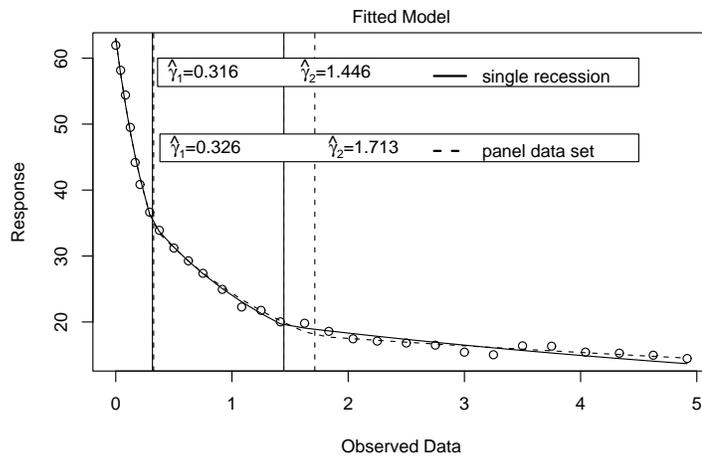


FIGURE 2. Fitted models of one single recession (solid) and the panel data model assuming a common slope for the last segment (dashed).

the first change point are almost the same, whereas those for the second one slightly differ. Of course this might be caused by the fact, that the estimates of the slopes in the last segments are different.

4 Discussion

Our MC simulation studies as well as the application of this model onto real data have shown, that this new method is appropriate to analyze such kind of hydrological data. This algorithm besides other functions for detecting change points in GLMs are implemented in R and available on request.

References

- Hofrichter, J. (2007). *Change Point Detection in Generalized Linear Models*. Unpublished PhD Thesis, University of Technology Graz, Austria.
- Küchenhoff, H. (1997). An exact algorithm for estimating breakpoints in segmented generalized linear models. *Computational Statistics* **12**, 235-247.
- Stasinopoulos, D., and Rigby, R. (1992). Detecting break points in generalized linear models. *Computational Statistics & Data Analysis* **13**, 461-471.

Mixture-Regression Cluster Model applied to Longitudinal Microarray Experiments

Emma Holian¹ and John Hinde²

¹ Department of Mathematics and Statistics, University of Limerick, Ireland.

`emma.holian@ul.ie`

² Department of Mathematics, National University of Ireland, Galway, Ireland.

`john.hinde@nuigalway.ie`

Abstract: The aim of this work is to explore various statistical techniques to identify genes which contribute to some change in phenotype level. For example, the response of fish kept under stressful conditions for various lengths of time. We aim to assess the level of *differential* expression of each gene in the tissue samples and also attempt to model the expression patterns of genes over time, not only to classify genes by similarities in expression patterns, but also to model these patterns as specified functions.

Keywords: Microarray; Longitudinal; Mixtures; Regression; Random effects.

1 Introduction and Background

Microarray technology measures genetic expression in the cells of a tissue sample and is implemented to identify the function of genes in an organism. A cDNA microarray can measure the genetic expression exhibited in two tissue samples. The animal sources from which these tissue samples are taken are often chosen because they differ in phenotype for some particular trait, for example, trout fish displaying symptoms of stress versus unstressed trout. The source with the phenotype trait is often labelled as the *treatment* and the source not displaying the trait labelled the *control*.

Of course the genetic makeup of any one phenotype consists of many thousands of genes and so detecting which genes are relevant to that particular trait is no menial task. The microarray facilitates detection of the presence and abundance of the expression exhibited in the tissue samples of thousands of genes simultaneously since an array consists of thousands of probes of different genetic material spotted at key locations on a glass slide. The level of expression of a gene at a spot is measured by recording the levels of intensity of two fluorescent dye molecules, Cyan 5 and Cyan 3, when the array is excited by a laser. Say the Cyan 5 molecule is attached to the treatment genetic expression and the Cyan 3 attached to the control genetic expression then the ultimate aim is to examine the level of *differential expression* estimated between these two tissue sources.

In order to estimate and remove the effects of experimental and biological variation, the arrays are repeated with various combinations of technical and biological replicate samples. In this simple two tissue variety experiment, the analysis is theoretically a two sample comparison test for each gene, *filtering* for those genes that display

significant differential expression. In practice however, in the context of microarrays, the researcher often finds numerous difficulties in this task. To name a few for example, the relatively low number of biological replicates available versus the high degree of experimental error seen in these experiments and the need for corrections for multiple testing.

If the level of differential expression for a gene is not significant the gene is said to be *inactive* with respect to the phenotype difference, that is, in the example given, the gene is inactive with respect to exposure to stress. If the level of measured expression for the gene is significantly higher in the treatment tissue than in the control tissue then the gene is said to be *over-expressed* meaning the gene becomes more active when the treatment phenotype trait is seen. Conversely, if the level of measured expression for the gene is significantly higher in the control tissue than in the treatment tissue, *under-expression*, then the gene becomes less active when the treatment phenotype trait is seen. Thus there is a natural classification of gene expression into groups or *clusters* of inactive, over-expressed and under-expressed genes.

As microarray experiments are increasing in popularity, geneticists have become more adventurous in the genetic questions they aim to answer. Thus experiments are increasing in complexity, requiring statistical consultation in first attaining the most efficient design for the experiment and later to analyse the data since these more adventurous experiments leads to unique computational and statistical problems in the fields of gene filtering and gene clustering.

In particular gene expression profiles obtained from time-course microarray experiments exhibit a unique opportunity to model the trends and correlations between longitudinal genetic expressions. In the following we outline details of the longitudinal trout fish stress experiment, and explore how the differential expression profiles can be clustered under a flexible parametric framework.

2 Outline of the Data Provided

Samples of liver tissue were extracted from rainbow trout fish exposed to confinement stress for varying lengths of time, at times 2,6,24,168 and 504 hours of stress, these times represented by $t, t \in 1 : 5$, and labelled as tissue variety *treatment*. Samples of liver tissue from unstressed fish left for the same period of time in a neighbouring tank are also included in the analysis, labelled as *control* samples.

Although a comparison is required between the expressions of treatment and control tissue samples at each time-point, these are not measured together directly on the same array. Instead each sample is paired with a common reference sample on an array.

Let the measured expression for a probe at spot $s \in [1 : 21168]$ be denoted by y_{sijkt} where $i \in [1 : 80]$ indicates on which of the 80 arrays the expression was measured on, j is the dye indicator where $j = 1$ for Cyan 5 and $j = 2$ for Cyan 3, k denotes tissue variety, $k = 0$ for reference, $k = 1$ for treatment and $k = 2$ for control.

In order to remove some experimental variation, a series of corrections are applied to the measured expressions, a process referred to as *normalisation*. As part of this process we assume the treatment and control intensities are calibrated for array effects using

the reference intensities. The following probe or spot-specific model is then applied to normalised measured expressions y'_{sijkt} ,

$$\log_2 y'_{sijkt} = \mu_s + Ref_s + D_{sj} + V_{sk} + T_{st} + VT_{skt} + \epsilon_{sijkt}$$

where μ_s is the overall mean and D_{sj} is the spot specific dye effect. Note that the reference variety, $k = 0$ is parameterized separately as Ref_s and since the calibration applied earlier assumes the reference values to be uniform over time, T_{st} represents the spot-specific time effect common to both varieties treatment and control. V_{sk} is the spot specific treatment and control effects, $k \in [1 : 2]$, while VT_{skt} estimates the changes in treatment and control expressions over the time-course of the experiment. Using these estimates, the differential expression profile between treatment and control for each probe s can be calculated as \mathbf{Y}_s , with elements

$$Y_{st} = (\widehat{V}_{s1} - \widehat{V}_{s2}) + (\widehat{VT}_{s1t} - \widehat{VT}_{s2t}) \text{ for } t = [1, 5].$$

Clustering these probes into groups of similarity may give some indication as to genes that co-regulate in the production of proteins in response to exposure to stress.

3 Formulation of Cluster Model and Estimation

The proposed Mixture-Regression Cluster Model is developed to model *and* cluster the genes into groups according to their expressions measured over time. This model is similar to that of the multivariate normal mixture model in that clusters are identified by the EM algorithm but is adapted to incorporate the flexibility of regression curves to fit the trends. In this way, additional features such as covariates, random effects and correlation structures can be incorporated into the model while potentially offering a considerable saving on the number of parameters required to model the trends.

For a particular cluster $i \in [1 : c]$ let the differential response vector be modelled by $\mathbf{Y}_s = X_s \boldsymbol{\beta}_i + Z_s \mathbf{b}_{is} + \epsilon_{is}$. for fixed effects $\boldsymbol{\beta}_i$, usually the regression curve in time, specified by the design matrix X_s and any optional random effects \mathbf{b}_{is} specified by design matrix Z_s . Where the errors have a normal density $\epsilon_{is} \sim N(0, \Sigma_i)$, and the random effects \mathbf{b}_{is} have a normal density function $f_i(\mathbf{b}_{is}) = \phi(0, D_i)$, then the marginal model for \mathbf{Y}_s is normally distributed, $f_i(\mathbf{Y}_s) = \phi(X_s \boldsymbol{\beta}_i, V_{is})$ with $V_{is} = Z_s D_i Z_s' + \Sigma_i$. The full distribution $f(\mathbf{Y}_s; \Psi)$ is then a mixture of the clusters so that

$$\mathbf{Y}_s \sim \sum_{i=1}^c \pi_i N(X_s \boldsymbol{\beta}_i, V_{is}).$$

Let the set of parameters for each of the component densities be denoted by $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, V_{is})$ then $\Psi = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_c, \pi_1, \dots, \pi_{(c-1)})'$ is the vector containing all unknown parameters.

Estimation of these parameters, for a pre-specified number of components, c , can be done via the Expectation-Maximization (EM) algorithm, an iterative procedure which is initiated by a random allocation of probes into the clusters. The M-step estimates the parameters, $\Psi^{(1)}$, using this initial allocation, by a weighted regression using R-subroutines GLS for a model with no random effects and LME fitting a model with

a random intercept or slope. The E-step then updates the allocation ratios using the estimated set of parameters $\Psi^{(1)}$ from the M-step. The iterations continue until there is little difference in the observed log likelihood as calculated in the E-step.

4 Results and Remarks

Simulations have shown that the mixture-regression model can recover clusters successfully and for each resulting cluster can provide a parametric model for the longitudinal trend followed by probes in the same cluster.

To find the model specification of optimal fit to the data, certain features of the model can be varied and the model refitted. For example re-specifying, the number of clusters c , or re-specifying $X_s\beta_i$ to be polynomials of varying degrees in time, re-specifying $Z_s\mathbf{b}_{si}$ to include a random intercept or slope and varying the correlation structure within each cluster Σ_i .

The optimal model is selected so that the log-likelihood is maximized, or if a penalization for the number of parameters is more desirable aim to minimize Akaike's Information Criterion AIC , or the more prudent Bayesian Information Criterion BIC . We show how these procedures were applied to a filtered subset of the fish stress probes resulting in a 17-component mixture. Some discussion will also follow as to how a number of these interesting clusters have proven to be a very useful source of information in understanding the molecular processes tested in these experiments.

Acknowledgments: Special Thanks to colleagues working at the National Diagnostics Centre, and in the Mathematics Department of National University of Ireland, Galway, where the work was carried out.

References

- Bowtell, D. and Sambrook, J. (2002). *DNA Microarrays*. Cold Spring Harbor Laboratory Press.
- Gentleman, Rossini, Dudoit and Hornik (2003). The Bioconductor FAQ, <http://www.bioconductor.org>
- McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons.
- McLachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.
- Goldstein, H. (1995). *Multilevel statistical models*. London : E.Arnold; New York : Halsted Press.

- Diggle, P.J., Heagerty, P., Liang, K. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.
- Cui, X., Kerr, K.K. and Churchill, G.A. (2003). *Transformations for cDNA Microarray Data*. *Statistical Applications in Genetics and Molecular Biology* **2(1)**, Article 4.
- Wit, E., and McClure, J. (2004). *Statistics for Microarrays: design, analysis and inference*. John Wiley & Sons.
- Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S (2001). *Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models*. *Journal of Computational Biology* **8(6)**, 625-637.

A hidden illness-death model to estimate life expectancies

Ardo van den Hout¹ and Fiona E. Matthews¹

¹ MRC Biostatistics Unit, Institute of Public Health, Cambridge, U.K.

Abstract: Longitudinal data can be used to estimate transitions between healthy and unhealthy states prior to death. A general continuous-time hidden three-state model is presented where transition intensities are allowed to change over time. When health is defined with respect to cognitive ability during old age, the trajectory of performance is either static or downward. The three-state model is used to describe the underlying categorized cognitive decline, where observed improvement of cognitive ability is modelled as misclassification. The methodology is extended to estimate life expectancy with and without cognitive impairment.

Keywords: Life expectancy; Markov model; misclassification; survival.

1 Introduction

The burden of disability in the population can be estimated using longitudinal follow-up and mortality information. In studies of the older population the measurement of cognitive ability is essential as it is an important predictor of health, survival, and need for care. This paper presents an illness-death model, where the three states are defined as the healthy state, an illness state, and the death state. Intensities of transitions between the states are estimated and used to compute life expectancies. If the states are defined with respect to cognitive ability, the model can help to understand how cognitive decline develops over time and which factors play a role.

The basis of the model is a continuous-time hidden Markov model where time-dependent intensities are dealt with in a piecewise-constant fashion. Measurement error is taken into account by logistic regression models for the misclassification and for the distribution of the states at baseline.

The model is an extension to the hidden Markov models in Satten and Longini (1996) and Jackson and Sharples (2002), the difference being our modelling of the latent distribution at baseline and the piecewise-constant approach to the intensities. Piecewise-constant continuous-time models are also presented by Chen and Sen (1999), but the number of parameters in their model depend on the imposed time grid. In our model, loglinear modelling of the time effect keeps the number of parameters manageable and allows us to extrapolate over time to estimate life expectancies.

2 Model and life expectancies

At time $t \geq 0$, the true state of an individual is $S_t \in \{1, 2, 3\}$ whereas the observed state is $S_t^* \in \{1, 2, 3\}$. Death state 3 is measured without error, but state 1 and state

2 are allowed to be misclassified.

First, we assume that transition intensities do not depend on time. A transition from state r to state $s \neq r$, occurs with intensity q_{rs} , where $q_{rs} \geq 0$ for $(r, s) \in \{(1, 2), (1, 3), (2, 1), (2, 3)\}$, and $q_{rs} = 0$ for $(r, s) \in \{(3, 1), (3, 2)\}$. The intensity matrix \mathbf{Q} is given by

$$\mathbf{Q} = \begin{pmatrix} -q_{12} - q_{13} & q_{12} & q_{13} \\ q_{21} & -q_{21} - q_{23} & q_{23} \\ 0 & 0 & 0 \end{pmatrix}.$$

For time interval $(t_1, t_2]$, the transition probability matrix is $\mathbf{P}(t_1, t_2) = \exp((t_2 - t_1)\mathbf{Q})$, with entries $p_{rs} = \Pr(S_{t_2} = s | S_{t_1} = r)$, for $r, s \in \{1, 2, 3\}$, see, e.g., Norris (1997). Intensities are allowed to depend on covariates and random effects by the log-linear model

$$q_{rs}(\mathbf{x}) = \lambda_{rs} \exp(\mathbf{a}'_{rs}\mathbf{x} + \tau_{rs}), \tag{1}$$

where $\mathbf{a}_{rs} = (a_{rs,1}, \dots, a_{rs,p})'$, $\mathbf{x} = (x_1, \dots, x_p)'$, and each random effect τ_{rs} is normally distributed with mean zero and unknown variance σ_{rs} . The model easily accommodates a more complex random-effect structure. If $\tau_{12} = \tau_{13} = \tau_{23} = \tau$, then τ is a shared random effect (Salazar et al., 2005).

Assume that an individual i has observations at times t_1, \dots, t_M and that we model a shared random effect. In shorthand notation and using the Markov assumption, the contribution of this individual to the likelihood is

$$\begin{aligned} L_i &= \int_{-\infty}^{\infty} \Pr(S_{t_1}^*, S_{t_2}^*, \dots, S_{t_M}^*) \phi(\tau) d\tau \\ &= \int_{-\infty}^{\infty} \left[\sum_{S_{t_1}, S_{t_2}, \dots, S_{t_M}} \Pr(S_{t_1}^* | S_{t_1}) \Pr(S_{t_1}) \Pr(S_{t_2}^* | S_{t_2}) \Pr(S_{t_2} | S_{t_1}) \right. \\ &\quad \left. \times \dots \times \Pr(S_{t_{M-1}}^* | S_{t_{M-1}}) \Pr(S_{t_{M-1}} | S_{t_{M-2}}) \times \Lambda \right] \phi(\tau) d\tau, \end{aligned}$$

where the sum is over all possible paths of latent states. In case of death, i.e., $S_{t_M}^* = 3$, we define

$$\Lambda = \Pr(S_{t_M} = 1 | S_{t_{M-1}}) q_{13} + \Pr(S_{t_M} = 2 | S_{t_{M-1}}) q_{23}.$$

So we assume an unknown latent state at time t_M and then an instant death. In case of censoring, it is known that the respondent is alive at the end of the study, but the state is unknown. In this case,

$$\Lambda = \Pr(S_{t_M} = 1 | S_{t_{M-1}}) + \Pr(S_{t_M} = 2 | S_{t_{M-1}}).$$

Next, intensities are allowed to depend on time. We impose a time grid independently of the observations times. Define $t_1^o = 0$, and $t_{g+1}^o = t_g^o + h$, for $g = 1, 2, \dots, G - 1$ such

that t_G° is the end time of the study. Parameter h is set by the researcher. For time since entry to the study t , piecewise-constant intensity matrices are defined by

$$\mathbf{Q}(t) = \begin{cases} \mathbf{Q}(t_1^\circ), & t_1^\circ \leq t < t_2^\circ \\ \mathbf{Q}(t_2^\circ), & t_2^\circ \leq t < t_3^\circ \\ \vdots & \\ \mathbf{Q}(t_{G-1}^\circ), & t_{G-1}^\circ \leq t < t_G^\circ, \end{cases}$$

where

$$q_{rs}(t_g^\circ) = \lambda_{rs} \exp(\nu_{rs} t_g^\circ) \exp(\mathbf{a}'_{rs} \mathbf{x} + \tau_{rs}), \quad g \in \{1, \dots, G-1\}. \quad (2)$$

Compared to model (1) there is an extra multiplicative term in (2) that describes the effect of time. Heterogeneity between individuals is modelled via the covariate vector \mathbf{x} .

An observed time interval $(t_m, t_{m+1}]$ is embedded in the time grid as follows. Determine the largest g_1 such that $t_{g_1}^\circ \leq t_m$, and determine the largest g_2 such that $t_{g_2}^\circ < t_{m+1}$, where $g_1, g_2 \in \{1, \dots, G-1\}$. The subintervals for $(t_m, t_{m+1}]$ are given by $(t_m, t_{g_1+h}^\circ]$, $(t_{g_1+h}^\circ, t_{g_1+2h}^\circ]$, \dots , $(t_{g_2}^\circ, t_{m+1}]$. We approximate the transition probability matrix for $(t_m, t_{m+1}]$ by multiplying the piecewise-constant transition probability matrices as follows

$$\begin{aligned} \mathbf{P}(t_m, t_{m+1}) &= \mathbf{P}(t_m, t_{g_1+h}^\circ) \times \dots \times \mathbf{P}(t_{g_2}^\circ, t_{m+1}) \\ &\approx \exp[(t_{g_1+h}^\circ - t_m) \mathbf{Q}(t_{g_1}^\circ)] \times \dots \times \exp[(t_{m+1} - t_{g_2}^\circ) \mathbf{Q}(t_{g_2}^\circ)]. \end{aligned}$$

To model misclassification, we use logistic regression models given by

$$\begin{aligned} \Pr(S_t^* = 2 | S_t = 1, \mathbf{y}(t)) &= \exp(\mathbf{b}'_1 \mathbf{y}(t)) / [1 + \exp(\mathbf{b}'_1 \mathbf{y}(t))] \\ \Pr(S_t^* = 1 | S_t = 2, \mathbf{y}(t)) &= \exp(\mathbf{b}'_2 \mathbf{y}(t)) / [1 + \exp(\mathbf{b}'_2 \mathbf{y}(t))]. \end{aligned}$$

Note that the vector $\mathbf{y}(t)$ is only important at the observation times: The models for the misclassification concern the observed states, not the intensities. For the baseline distribution of the latent states (i.e., state 1 or 2 at $t_1 = 0$), we use a logistic regression model given by

$$\Pr(S_{t_1} = 1 | \mathbf{z}(t_1)) = \exp(\mathbf{c}' \mathbf{z}(t_1)) / [1 + \exp(\mathbf{c}' \mathbf{z}(t_1))].$$

Expected life expectancy in state $s \in \{1, 2\}$ given initial state $r \in \{1, 2\}$ and covariate vector \mathbf{x} is given by

$$e_{rs}(\mathbf{x}) = \int_0^\infty \Pr(S_t = s | S_0 = r, \mathbf{x}) dt.$$

Note that we only need the parameters from the hidden Markov model to estimate the life expectancies (cf. Izmirlian et al., 2000).

TABLE 1. Maximum likelihood estimates for the hidden Markov model in the application. Estimated standard errors in parentheses.

Parameter		Parameter		Parameter	
λ_{12}	0.018 (0.004)	$\alpha_{12.1}$	0.196 (0.019)	$b_{1.1}$	-0.017 (0.178)
λ_{13}	0.064 (0.004)	$\alpha_{13.1}$	0.072 (0.007)	$b_{2.1}$	-0.192 (0.298)
λ_{23}	0.195 (0.024)	$\alpha_{23.1}$	0.040 (0.010)	$b_{1.2}$	-1.204 (0.113)
ν_{12}	0.056 (0.046)	c_1	2.386 (0.107)	$b_{2.2}$	-0.546 (0.119)
ν_{13}	0.064 (0.010)	c_2	-0.166 (0.015)		
ν_{23}	0.051 (0.019)				

3 Application

We analyzed a random subset of the Medical Research Council Cognitive Function and Ageing Study where individuals have had up to eight interviews in the period 1991 to 2004. Cognitive impairment was measured using the Mini-Mental State Examination (MISE), which is a common cognitive ability scale from 0 up to 30. The research question is about cognitive impaired life expectancy, where the MISE score 21 is used as cut-point: Individuals who score less than 22 are impaired. Over time, transitions occur between the not-impaired state 1, the impaired state 2, and the death state 3. The last observed state is either death or censored. The subset consists of 2015 men who are aged 65 years or older. For the model, age is centered by subtracting 77.

For individuals in state s , $s \in \{1, 2\}$, missing MMSE scores were imputed by the mean of the MMSE scores in state s . In total, 139 scores were imputed in the presence of 5025 observed scores. Given the relative small number of missings, we assume the underestimation of the variance due to the mean imputation to be very small.

The trajectory of cognitive performance is assumed to be either static or downward. Hence, we restrict $q_{21}(t)$ to zero and assume that observed improvement is due to misclassification of the underlying true states. The probability of misclassification might be larger for individuals with an MMSE score close to the cut-point. We take this into account by regressing the misclassification probabilities on the absolute distance between the score and the cut-point:

$$\begin{aligned} \text{logit} [\Pr(S_t^* = 2 | S_t = 1, \mathbf{y}(t))] &= b_{1.1} + b_{1.2}y(t) \\ \text{logit} [\Pr(S_t^* = 1 | S_t = 2, \mathbf{y}(t))] &= b_{2.1} + b_{2.2}y(t), \end{aligned}$$

where $y(t) = |\text{MMSE score at time } t - 21.5|$.

For the logistic regression of the initial distribution, we use centered age at baseline as a covariate and estimate intercept c_1 and regression coefficient c_2 . Centered age at baseline was also used as a covariate in the model (2) for the intensities. Including a shared random effect did not have a significant effect ($\hat{\sigma} = 0.196$ with estimated standard error 0.210) and was therefore not included in the final model.

For the piecewise approximation of the intensities we use a grid with time interval equal to 1 year. The likelihood was maximized using the general-purpose optimizer `optim` in R. Table 1 presents the maximum likelihood estimates and the estimated

TABLE 2. Estimated life expectancies (LEs) for men, given their age and state at baseline 1991. In parentheses estimates standard errors.

Age at baseline	Not-impaired LE	Impaired LE	Impaired LE
	given baseline state = 1	given baseline state = 1	given baseline state = 2
65	15.59 (0.53)	0.19 (0.11)	6.36 (0.92)
75	9.21 (0.22)	0.54 (0.11)	4.61 (0.40)
85	4.11 (0.21)	1.14 (0.15)	3.29 (0.22)

standard errors. The latter were estimated by evaluating the observed information matrix.

We obtain $\hat{b}_{1.1}, \hat{b}_{2.1}, \hat{b}_{1.2}, \hat{b}_{2.2} < 0$. Thus a smaller $y(t)$ means a bigger logit indicating a higher probability of misclassification. For example, estimated misclassification matrices for MMSE scores 15 and 20 are given by

$$\mathbf{C}_{\text{MMSE}=15} = \begin{pmatrix} 1.000 & 0 & 0 \\ 0.023 & 0.977 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathbf{C}_{\text{MMSE}=20} = \begin{pmatrix} 0.861 & 0.139 & 0 \\ 0.267 & 0.733 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where entry (k, l) denotes $\Pr(S_t^* = l | S_t = k)$. The estimated misclassification concerns mainly the misclassification of the true state 2 as an observed state 1. This means that measurement error mainly concerns not detecting an impaired state. It is also for this situation that the distance of the MMSE to the cut-of point is important. As was to be expected, both the effect of time since entry to the study and the effect of age at baseline are positive.

Table 2 presents estimated life expectancies (LEs). Standard errors are estimated by simulating the variation in the estimation of the LEs. That is, we consider the multinomial distribution with expectation equal to the maximum likelihood estimate of the parameter vector and the covariance matrix equal to the estimated covariance matrix at the optimum. By simulating parameters values from this distribution and computing the LEs for each of the simulated values, the sample variation in the estimation of the LEs will be reflected (cf. Aalen *et al.*, 1997). Due to the extrapolation over time, the standard errors are relatively larger for the younger ages.

As was to be expected, not-impaired LE decreases with increasing age. LE given baseline state 1 is smaller than impaired LE given baseline state 2 because a large proportion of the people in state 1 will die before they will develop cognitive problems. The performance of the model was assessed visually by checking survival curves: given an age and a baseline state, model-based latent survival is translated to fitted survival and compared to the survival estimated from the semi-parametric Cox regression model. Although this comparison does not address all the aspects of the multi-state model, it is a insightful way to get a first idea of goodness of fit. Results (not reported) were checked for various ages and seemed reasonable.

References

- Aalen O.O., Farewell V.T., De Angelis D., Day N.E. and Gill O.N. (1997) A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine* **16**, 2191-2210.
- Izmirlan, G, Brock, D., Ferrucci, L. and Phillips, C. (2000) Active life expectancy from annual follow-up data with missing responses. *Biometrics* **56**, 244-248.
- Jackson, C.H., and Sharples, L.D. (2002) Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine* **21**, 113-128.
- Norris, J.R. (1997) *Markov Chains*. Cambridge University Press: Cambridge.
- Salazar, J.C., Schmitt, F.A., Yu, L., and Mendiondo, M.M. (2005) Shared random effects analysis of multi-state Markov models: application to a longitudinal study of transitions to dementia. *Statistics in Medicine* (to appear).
- Satten, G.A., and Longini, I.M. (1996) Markov chains with measurement error: estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease (with discussion). *Applied Statistics* **45**, 275-309.
- Chen, P.L., and Sen, P.K. (1999) A piecewise transition model for analyzing multi-state life history data. *Journal of Statistical Planning and Inference* **78**, 385-400.

A Weighted Kaplan-Meier Approach for Estimation of Recurrence of Colorectal Adenomas

Chic-Hises Hsu¹, Jeremy M. G. Taylor², Qi Long³ and David S. Alberts¹

¹ University of Arizona; phsu@azcc.arizona.edu

² University of Michigan

³ Emory University

Abstract: The treatment effect of a colorectal polyp prevention trial is often evaluated on the colorectal adenoma recurrence status at the end of the trial. Due to early colonoscopy from some participants, the data can be considered as current status data. The early colonoscopy could be informative of status of recurrence and induce informative differential follow-up into the data. In this paper we use mid-point imputation to handle interval censored observations and then perform a weighted Kaplan-Meier method on the early colonoscopy status to the imputed data to adjust for potential informative differential follow-up. In a simulation study, we show that the weighted Kaplan-Meier method can produce reasonable estimates of recurrence rate under an informative early colonoscopy situation with prognostic covariates compared to conventional logistic regression, weighted logistic regression, and Kaplan-Meier estimator. The method described here is illustrated with an example from a colon cancer study.

Keywords: mid-point imputation; weighted Kaplan-Meier estimator.

1 Introduction

Most of colorectal cancer prevention trials use colorectal adenomas to study preventive agents. The treatment effect is evaluated on recurrent adenomas by performing colonoscopy at follow-up. The follow-up colonoscopy is scheduled to be performed at the end of the trial (e.g. 3 years) to evaluate the status of recurrence. The actual event time for each participant is then only known as occurring either before or after three years. Some participants could have their follow-up colonoscopy before the schedule of examinations because of their family history of colorectal cancer and health conditions (considered as “early colonoscopy”) and then create differential follow-up. Due to differential follow-up, the recurrent adenoma data can be considered as current status data (“case 1” interval censored data). In a colorectal polyp prevention trial often over 50% of participants had their only follow-up colonoscopy performed at the end of trial. When participants had their only follow-up colonoscopy performed at the end of trial, there was little information with regard to the actual time of recurrence, which would contribute to estimating the survival function. In this paper, we are interested in estimating recurrence rate at the end of the study and propose using mid-point imputation to handle interval censored observations and a weighted Kaplan-Meier (WKM) method to adjust for potential informative early colonoscopy.

2 Methods

For an interval-censored recurrence time, mid-point imputation is used to impute time to recurrence by the mid-point of the interval. For right-censored recurrence time, time to recurrence is treated as right censored. The KM and WKM methods are then conducted on the imputed data. Often there are more than one covariate associated with early colonoscopy or risk of recurrence in a colorectal polyp prevention trial. In this paper we adapt and generalize the ideas in Hsu et al. (2006) to incorporate multiple covariates into the WKM method. We propose to fit two logistic regression models, one for status of early colonoscopy and one for risk of recurrence to reduce the covariates to two risk scores, which provides indicators of an individual's early colonoscopy probability and risk of recurrence. The two risk scores will be continuous and can be categorized into groups based on dichotomization or quartiles. The WKM can then be easily derived based on the categorized groups.

3 Application to UDCA Data

In 1996, the Arizona Cancer Center initiated a multi-center trial to determine whether ursodeoxycholic acid (UDCA) can prevent the recurrence of colorectal adenomas (Alberts et al., 2005). A total of 1192 subjects underwent at least one follow-up colonoscopy and were thus considered for the endpoint analysis, 579 in the placebo group and 613 in the UDCA group. There were 282 (49%) having early colonoscopy in the placebo group and 307 (50%) in the UDCA group. Early colonoscopy is highly associated with age and risk of recurrence and marginally associated with previous polyp history. This indicates informative early colonoscopy for the UDCA study.

The primary interest of the study is in estimating recurrence rate of adenomas at the end of the study for both the placebo and UDCA groups. A logistic regression model (Logit), WKM and KM methods are performed to the data. In Table 1, Logit produces a lower recurrence rate compared to the KM and WKM methods for both placebo and UDCA groups. The WKM method, which incorporates age and previous polyp history into analysis to adjust for informative early colonoscopy, produces a lower recurrence rate for both placebo and UDCA groups compared to the KM method. The WKM method, which incorporates early colonoscopy directly into analysis to adjust for informative early colonoscopy, produces a slightly higher recurrence rate for the placebo group and an almost identical recurrence rate for the UDCA group compared to the KM method.

4 Discussion

The research in this paper uses mid-point imputation to handle interval censored observations and uses a weighted Kaplan-Meier approach to adjusting for potential informative early colonoscopy through the use of prognostic covariates while estimating recurrence rate for a colorectal polyp prevention trial. This approach can handle a situation with multiple prognostic covariates by deriving risk scores from a logistic

regression model. Although the idea of this approach might appear simple, the simulation results (Table 2) do show that the weighted Kaplan-Meier approach can provide a reasonable recurrence rate estimate under both non- and informative early colonoscopy without losing efficiency or introducing bias into estimation. Whereas, weighted logistic regression (WLogit), which simply uses a weight function of follow-up length to adjust for differential follow-up in logistic regression, tends to under-estimate recurrence rate and the performance of the KM and WLogit methods rely on the assumption of non-informative early colonoscopy. Simply using mid-point imputation to handle interval censored observations might produce bias survival estimates and misleading results. In this paper, we focus on estimating recurrence rate at the end of the study using the WKM method. Under an independent censoring assumption, mid-point imputation will not contribute bias to the estimate of recurrence rate at the end of the study.

TABLE 1. UDCA: Estimation of recurrence rate for placebo and UDCA groups at the end of the study.

Method	Placebo		UDCA		Odds Ratio
	estimate	standard error	estimate	standard error	
KM	0.469	0.023	0.439	0.022	0.886
WKM ^a	0.476	0.025	0.437	0.021	0.854
WKM ^b	0.479	0.028	0.431	0.022	0.824
Logit	0.439	0.021	0.409	0.020	0.884

^aprognostic covariate: early colonoscopy.

^bprognostic covariate: early colonoscopy, age and previous polyp history.

References

- Alberts D. S., Martínez M. E., Hess L. M., et al. (2005) Phase III Trial of Ursodeoxycholic Acid To Prevent Colorectal Adenoma Recurrence. *Journal of the National Cancer Institute* **97**, 846-853.
- Hsu C.-H., Taylor J. M. G., Murray S. and Commenges D. (2006). Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine* **25**, 3503-3517.

TABLE 2. Monte Carlo Results: Estimation of recurrence of adenomas at 3 years, where sample size is 100, maximum follow-up is 3 years, M (early colonoscopy indicator) is from $Bernoulli(1, p)$ and censoring time is from $Exponential(1)$ for participants with early colonoscopy, X_1 and X_2 are from a $Uniform(0, 1)$

Method	$F(t) \sim Exp(\lambda^a)$				$F(t) \sim Lognormal(\mu^b, 1)$			
	est	SD ^c	SE ^d	CR ^e	est	SD	SE	CR
$p = 0.5$	Recurrence rate: $F(3) = 0.607$				$F(3) = 0.691$			
KM	0.498	0.0588	0.0588	55.0	0.600	0.0558	0.0577	65.7
WKM	0.599	0.0666	0.0629	91.8	0.690	0.0569	0.0588	94.3
WLogit	0.501	0.0587	0.0593	58.0	0.599	0.0558	0.0583	66.6
$p = 0.6$	Recurrence rate: $F(3) = 0.656$				$F(3) = 0.734$			
KM	0.537	0.0612	0.0617	52.7	0.637	0.0599	0.0593	64.0
WKM	0.650	0.0686	0.0661	92.0	0.737	0.0596	0.0602	94.0
WLogit	0.546	0.0610	0.0623	59.6	0.640	0.0593	0.0600	66.9

^a $\lambda = [0.1 + 0.05X_1 + 0.05X_2 + (1 - M)(0.1 + 0.45X_1 + 0.45X_2)]$

^b $\mu = [1.0 + 0.2X_1 + 0.1X_2 + (1 - M)(-1.5 + 0.2X_2)]$

^c empirical standard deviation of 1000 point estimates.

^d average of 1000 estimated standard errors.

^e fraction of 95% confidence intervals containing the true value.

Stationary versus Non-stationary Correlation Models for Familial Longitudinal Count Data

Vandna Jowaheer and Brajendra Sutradhar

¹ Department of Mathematics, University of Mauritius, Reduit, Mauritius

² Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, Canada, A1C5S7

Abstract: In familial longitudinal set up, the responses from all members collected over a period of time exhibit familial as well as longitudinal correlations. In the context of count data, Sutradhar and Jowaheer (2003, JMVA, 398-412) used a stationary longitudinal correlation structure based familial longitudinal correlation model for the construction of the generalized quaslikelihood (GQL) estimating equations for the regression effects, and variance component of the random family effects. The longitudinal correlation parameter was estimated by using the method of moments. In this paper we use a non-stationary correlation model appropriate for the time varying covariates based data analysis and exploit the combined GQL and moment approach for the inferences about the parameters of the model. The performances of the non-stationary versus stationary correlation structure based inferences are examined through a limited simulation study.

Keywords: Discrete data; Familial and longitudinal correlations; Generalized quaslikelihood estimation; Time dependent covariates

1 Introduction

Let y_{ijt} denote the count response for the j th ($j = 1, \dots, n_i$) individual on the i th ($i = 1, \dots, I$) family/cluster at a given time t ($t = 1, \dots, T$). Also, let $x_{ijt} = (x_{ijt1}, \dots, x_{ijtu}, \dots, x_{ijtp})'$ denote the p covariates associated with the response y_{ijt} . For example, in health care utilization data, the number of visits to the physician by the members of a large number of independent families may be recorded over a period of several years. Also the information on the covariates-gender, number of chronic conditions, education level and age- may be recorded for the members of each family. Note that as the members of the i th ($i = 1, \dots, I$) family are likely to be influenced by a common family effect say γ_i , the count responses of any two members of the same family at a given year are likely to be correlated. This correlation is usually referred to as the familial correlation. Furthermore, conditional on the unobservable family effect γ_i , the repeated count data collected from the same member of the i th family are also likely to be correlated. This correlation is usually referred to as the longitudinal correlation. Note that if the covariates, such as education level and age, collected from the same individual over a period of time are time dependent, then the longitudinal lag correlations for the same individual will be non-stationary. It is of scientific interest to find the effects of the covariates on the count response of an individual after taking the familial and non-stationary longitudinal correlations into account.

Suppose that conditional on the random family effect γ_i , the response of the j th ($j = 1, \dots, n_i$) member of the i th ($i = 1, \dots, I$) family at a given year t ($t = 1, \dots, T$), i.e., y_{ijt} , marginally follows the Poisson density with the mean and the variance given by

$$E(Y_{ijt}|\gamma_i) = \text{var}(Y_{ijt}|\gamma_i) = \mu_{ijt}^* = e^{(x'_{ijt}\beta + \gamma_i)}. \tag{1}$$

As far as the correlation structure is concerned, Sutradhar and Jowaheer (2003) used a stationary correlation structure given by

$$\text{corr}(Y_{iju}, Y_{ikt}|\gamma_i) = \begin{cases} \rho^{|t-u|} & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases} \tag{2}$$

[see also Woolridge (1999)]. This, under the assumption that $\gamma_i \stackrel{iid}{\sim} N(0, \sigma^2)$, yields the unconditional mean, variance of y_{ijt} as

$$E(Y_{ijt}) = e^{(x'_{ijt}\beta + \sigma^2/2)} = \mu_{ijt}, \text{ var}(Y_{ijt}) = \mu_{ijt} + (e^{\sigma^2} - 1)\mu_{ijt}^2, \tag{3}$$

and the covariance between y_{iju} and y_{ikt} as

$$\text{cov}(Y_{iju}, Y_{ikt}) = \begin{cases} \rho^{|t-u|} \{\mu_{iju}\mu_{ikt}\}^{\frac{1}{2}} + \mu_{iju}\mu_{ikt}[e^{\sigma^2} - 1] & \text{for } j = k \\ \mu_{iju}\mu_{ikt}[e^{\sigma^2} - 1] & \text{for } j \neq k \end{cases} \tag{4}$$

Sutradhar and Jowaheer (2003) then exploited the mean, variance and the covariance structures in (3)-(4) to construct the GQL estimating equations for β and σ^2 , and the ρ parameter was estimated by the method of moments. This combined GQL and moment approach produces consistent estimates for all three parameters, whereas the β estimator can be highly efficient too. Note however that it may not be possible to construct a valid stationary correlation structure such as (2) when the data are non-stationary, especially when the covariates are time dependent. This raises a concern to develop a model that allows non-stationary correlation structure. In the next section we develop such a model.

2 Proposed Correlation Model

As a generalization to the stationary model of McKenzie (1988), we now consider

$$y_{ijt}|\gamma_i = \rho \circ [y_{ij,t-1}|\gamma_i] + d_{ijt}|\gamma_i, \tag{5}$$

$$\rho \circ y_{ij,t-1} = \sum_{s=1}^{y_{ij,t-1}} b_s(\rho), \tag{6}$$

where $Pr[b_s(\rho) = 1] = \rho$ and $Pr[b_s(\rho) = 0] = 1 - \rho$. Furthermore, suppose that d_{ijt} for $t = 2, \dots, T$ follow the Poisson distribution with mean parameter $\mu_{ijt}^* - \rho\mu_{ij,t-1}^*$, i.e.,

$d_{ijt}|\gamma_i \sim P(\mu_{ijt}^* - \rho\mu_{ij,t-1}^*)$. Also suppose that d_{ijt} is independent of $z_{ij,t-1} = \rho \circ y_{ij,t-1}$. The above model produces the longitudinal correlations conditional on γ_i as

$$\text{corr}(Y_{iju}, Y_{ijt}|\gamma_i) = \rho^{|t-u|} \left[\frac{\mu_{iju}^*}{\mu_{ijt}^*} \right]^{\frac{1}{2}} = \rho^{|t-u|} \left[\frac{\mu_{iju}}{\mu_{ijt}} \right]^{\frac{1}{2}} = \rho^{|t-u|} r_{ijut}, \tag{7}$$

with $r_{ijut} = \exp\{-\frac{1}{2}(x_{ijt} - x_{iju})'\beta\}$. These correlations depend on the time dependent covariates and hence are non-stationary. Furthermore, at any two time points, the responses of any two members are conditionally uncorrelated. Thus, $\text{corr}[Y_{iju}, Y_{ikt}] = 0$ for $j \neq k$. Consequently, one obtains the non-stationary covariances as

$$\text{cov}(Y_{iju}, Y_{ikt}) = \begin{cases} \mu_{ijt} + (e^{\sigma^2} - 1)\mu_{ijt}^2 & \text{for } k = j; u = t \\ \rho^{t-u}\mu_{iju} + (e^{\sigma^2} - 1)\mu_{iju}\mu_{ijt} & \text{for } k = j; u < t \\ (e^{\sigma^2} - 1)\mu_{iju}\mu_{ikt} & \text{for } k \neq j; u \leq t \end{cases} \tag{8}$$

The mean and the variance of y_{ijt} under the model (5) remain the same as in (3).

Note that the above correlation structure in (7) becomes stationary only when the covariates are stationary, i.e., $x_{ijt} = x_{iju}$ for all t and u . In practice this will however happen very rarely. The purpose of the next section is to exploit the non-stationary covariance structure in (8) and develop the GQL estimating equations for β and σ^2 and moment equation for the ρ parameter.

3 Estimation of Parameters

3.1 The GQL Estimation of β and σ^2

Let $\psi_i = (\psi'_{i1}, \dots, \psi'_{ij}, \dots, \psi'_{in_i})'$ be the $n_i T$ -dimensional mean vector of $y_i = (y'_{i1}, \dots, y'_{ij}, \dots, y'_{in_i})'$. Further, denote the covariance matrix of y_i by $\Sigma_i(\beta, \sigma^2, \rho) = (\sigma_{ijkut})$. One may then obtain the generalized quaslikelihood (GQL) estimate of β by solving the GQL estimating equation for β given by

$$\sum_{i=1}^I \frac{\partial \psi'_i}{\partial \beta} \Sigma_i^{-1}(\beta, \sigma^2, \rho)(y_i - \psi_i) = 0, \tag{9}$$

where $\partial \psi'_i / \partial \beta$ is the $p \times n_i T$ first derivative matrix. Note that for given σ^2 and ρ , the estimating equation (9) produces both consistent and efficient estimate for β . This is because, the estimating function in the left hand side of (9) is an unbiased estimating function for zero which assures the consistency of the solution. Further, the weight matrix used in (9) is the exact covariance matrix of y_i which leads the estimator of β as a highly efficient estimator among the moments based estimators.

Next, to obtain a GQL estimate of σ^2 for known β and ρ , we exploit the vector of all first and second order responses as a basic statistic (see also Sutradhar (2004), Jiang

(1998)). Let $g_{ij(s)} = (y_{ij1}^2, \dots, y_{ijt}^2, \dots, y_{ijT}^2)'$ be the T -dimensional vector of squares and $g_{ij(p)} = (y_{ij1}y_{ij2}, \dots, y_{ij(T-1)}y_{ijT})'$ be the $T(T-1)/2$ -dimensional vector of distinct pair-wise products of the elements of y_{ij} vector. Further, let $g_{ij} = (g'_{ij(s)}, g'_{ij(p)})'$ be the $T(T+1)/2$ dimensional combined vector of squares and pair-wise products for the j th ($j = 1, \dots, n_i$) member of the i th family. Next we write

$$g_i = (g'_{i1}, \dots, g'_{ij}, \dots, g'_{in_i})', \text{ and } E(g_i) = \lambda_i, \tag{10}$$

where g_i is the $n_i T(T+1)/2$ -dimensional vector of squares and distinct products for all n_i individuals of the i th ($i = 1, \dots, I$) family, and $\lambda_i = (\lambda'_{i1}, \dots, \lambda'_{ij}, \dots, \lambda'_{in_i})'$, where $\lambda_{ij} = (\lambda'_{ij(s)}, \lambda'_{ij(p)})'$. One may then solve a ‘working’ independence assumption based GQL estimating equation

$$\sum_{i=1}^I \frac{\partial \lambda_i}{\partial \sigma^2} \Omega_i^{-1}(I; \rho = 0)(g_i - \lambda_i) = 0, \tag{11}$$

to obtain the GQL estimate of σ^2 . Note that as the GQL estimating equation (11) is unbiased, it produces consistent estimator for σ^2 , even if $\rho = 0$ is used. By the same token, if the ρ parameter is large and $\rho = 0$ is used, the estimating equation (11) may not produce highly efficient estimate.

3.2 Moment Estimation of ρ

Finally, the ρ parameter may be consistently estimated by solving a moment equation which can be constructed by equating the population covariance

$$\text{cov}(Y_{ijt}, Y_{ij,t+1}) = \rho \mu_{ijt} + (e^{\sigma^2} - 1) \mu_{ijt} \mu_{ij,t+1},$$

to its sample counterpart. See Sutradhar and Jowaheer (2003) for details. To be specific, the moment estimating equation may be written as

$$\begin{aligned} \sum_{i=1}^I \sum_{j=1}^{n_i} \sum_{t=1}^{T-1} [(y_{ijt} - \mu_{ijt})(y_{ij,t+1} - \mu_{ij,t+1})] &= \sum_{i=1}^I \sum_{j=1}^{n_i} \sum_{t=1}^{T-1} [\rho \mu_{ijt} \\ &+ (e^{\sigma^2} - 1) \mu_{ijt} \mu_{ij,t+1}], \end{aligned} \tag{12}$$

which produces a consistent estimate for the ρ parameter. Note that the moment estimate of the same ρ parameter would be different if the moment equation is constructed based on the stationary correlation structure (2), whereas the moment equation (12) is constructed by using the non-stationary correlation structure (7).

4 A Simulation Study

The purpose of the simulation study is to examine the performance of the GQL estimates for β and σ^2 , and the moment estimate of ρ , obtained from (9), (11), and (12) respectively. Note that these equations are constructed based on the non-stationary

TABLE 1. Simulated mean (SM), standard error (SE), and relative bias (RB) of the non-stationary correlation structure (NSCS) versus stationary correlation structure (SCS) based GQL estimates for parameters of the non-stationary familial longitudinal model with one covariate for selected values of σ^2 and ρ ; $K = 100$; $n = 2$; $T = 4$; $\beta = 1.0$; 500 simulations.

Working correlation structure	Correlation parameter(ρ)	Quantity	Estimates		
			$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\rho}$
NSCS	0.25	SM	0.968	0.488	0.279
		SE	0.077	0.169	0.264
		RB	42	7	11
	0.50	SM	0.976	0.480	0.487
		SE	0.060	0.152	0.206
		RB	40	13	6
	0.75	SM	0.980	0.494	0.715
		SE	0.047	0.134	0.161
		RB	43	4	22
SCS	0.25	SM	0.968	0.485	0.216
		SE	0.078	0.171	0.206
		RB	41	8	17
	0.50	SM	0.977	0.474	0.381
		SE	0.060	0.154	0.159
		RB	38	17	75
	0.75	SM	0.981	0.482	0.565
		SE	0.048	0.139	0.123
		RB	40	13	150

correlation structure (7) which also was used to generate the data. For simplicity, we consider $n_i = n = 2$ members in each of $I = 100$ families. Also we consider one dimensional egression effect with covariate for the first member defined as

$$x_{i1t1} = \begin{cases} (t^2 - 2.5)/8 & \text{for } i = 1, \dots, I/2; t = 1, \dots, 4 \\ t^2/8 & \text{for } i = I/2 + 1, \dots, I; t = 1, \dots, 4, \end{cases}$$

whereas for the second member the time dependent covariate is taken as:

$$x_{i2t1} = \begin{cases} 0.1 + (t - 1) \times 0.25 & \text{for } i = 1, \dots, I/4; t = 1, \dots, 4 \\ (1 + t + t^2)/12 & \text{for } i = I/4 + 1, \dots, 3I/4; t = 1, \dots, 4 \\ (t^2 - 2.5)/8 & \text{for } i = 3I/4 + 1, \dots, I; t = 1, \dots, 4. \end{cases}$$

The simulated mean (SM), standard error (SE), and relative bias (RB) of all these three estimates based on 500 simulations are reported in the top half of the Table 1. Similar results for the situation where data were generated based on correlation structure (7) but estimating equations were constructed based on (2), are given in the

bottom half of the same Table 1. The results of the table show that the RB's in the bottom half are generally larger than those displayed in the top half. This shows that using a simple stationary correlation structure can be detrimental in estimating the parameters.

Acknowledgments: This research is supported partially by a grant from the Natural Sciences and Engineering Research Council of Canada

References

- Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *J. of American Statistical Association* **93**, 720-729.
- McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Adv. in Appl. Probb.* **20**, 822-35.
- Sutradhar, B. C. (2004). On exact quaslikelihood inferences in generalized linear mixed models. *Sankhya: The Indian Journal of Statistics* **66**, 261-89.
- Sutradhar, B. C. and Jowaheer, V. (2003). On familial longitudinal Poisson mixed models with gamma random effects. *J. Mult. Var. Anal.* **87**, 398-412.
- Woolridge, J. (1999). Distribution-free estimation of some non-linear panel data models. *Journal of Econometrics* **90**, 77-97.

Some regression methods in evaluation of genotypes in series of experiments

Zygmunt Kaczmarek¹, Elżbieta Adamska¹ and Teresa Cegielska-Taras²

¹ Institute of Plant Genetics, Polish Academy of Sciences,
60-479 Poznań, URL. Strzeszyński 34, Poland. zkac@igr.poznan.pl

² Plant Breeding and Acclimatization Institute, Poznań, Poland

Abstract: The paper gives main theoretical and methodological results concerning the analysis of genotypes by environment (*GE*) interaction studied in of a series of experiments conducted with the same genotypes in different environments. In the analysis based on a mixed model a special attention is given to estimation and hypothesis testing problems concerning the studying the *GE* interaction. A proper analyses of the regression are used in drawing inferences on the genotypes and their behaviour in various environments. Estimates of the relevant regression coefficients are given and testing the hypotheses concerning the linear regression of the *GE* interactions on the environment main effects and a multiple linear regression of *GE* interaction effects on chosen concomitant variables pertaining to environments are discussed. The application of the regression methods presented in the paper is illustrated by a practical example. The example is related to some genetic and plant breeding research project concerning the contents of oleic fatty acids in doubled haploid population of winter oilseed rape.

Keywords: MANOVA; *GE* interaction; multiple regression; oilseed rape; fatty acids.

1 Introduction

Statistical methods of the analysis of a series of plant breeding experiments developed by several authors, e.g., by Caliński et al. (1997), allow evaluation individual genotypes in various environmental conditions including assessment of stability and adaptability. The linear regression of genotype by environment interaction effects on the environment main effect may provide a practical method of evaluating trends in response of a given trait to various environmental conditions. The paper presents also some proposition of multiple linear regression to evaluation of genotypes when additional information about environments are available. Practical application of this approach will be shown on the real data concerning the fatty acid composition of seed oil in a population of winter rape.

2 The data

The data include results from 6 experiments with the same 37 genotypes of winter oilseed rape. The experiment was carried out in a completely randomized block design with three replications. The contents of five fatty acids: palmitic, stearic, deic, linoleic

and linolenic were estimated. Moreover, two traits: seed yield and oil content treated as concomitant variables given additional information about environments were measured. The analysis described here is confirmed to data concerning oleic fatty acid only. The fact that there might be several more fatty acids estimated during the experiment will be ignored here. Certainly, the same type of analysis could be applied to the rest fatty acids, also after a suitable transformation of the data. The calculations were made for all 37 genotypes. It would not be possible to present them here in full. Only some results of the interaction and regression analysis of 9 randomly chosen *DH* lines will be shown. In fact these *DH* lines form only a subset of a much larger set of genotypes compared in all the experiments.

3 Methods

Suppose that I genotypes are compared in an experiment repeated at J randomly chosen environments. A simple model for the observed mean of the studied trait of genotype i at environment j is

$$y_{ij} = \mu + \alpha_i^G + a^E(j) + a_i^{GE}(j) + e_{ij},$$

where μ denotes the "true" overall mean, α_i^G denotes the main effect of genotype i ($= 1, 2, \dots, I$), $a^E(j)$ denotes the main effect of environment j ($= 1, 2, \dots, J$), $a_i^{GE}(j)$ denotes the effect due to the interaction of genotype i and environment j , and e_{ij} denotes the mean random error from L replications. Taking the assumptions as described in the paper by Caliński et al. (1997), the analysis of variance can be performed.

The presence of the *GE* interactions give rise to questions concerning their causes. One question is whether these interactions can be explained by their dependence on the overall mean studied trait in the environment.

It follows that the interaction effect $a_i^{GE}(j)$ can be correlated with $a^E(j)$, the main effect of environment j . This correlation, and the corresponding regression, may have in many cases an important contribution to *GE* interactions. For practical application it may be useful to note that the correlation between $a_i^{GE}(j)$ and $a^E(j)$ is different from zero if and only if the correlation between $a_i^{GE}(j)$ and observable values $y_{.j}$ is nonzero. It means that the regression of $a_i^{GE}(j)$ on $y_{.j}$ is of particular interest.

The minimum variance unbiased estimate of the vector coefficients of the regression $\beta = [\beta_1, \beta_2, \dots, \beta_I]'$ is $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_I]'$ = $I(\mathbf{1}'_I \mathbf{S}_E \mathbf{1}_I)^{-1} \mathbf{G} \mathbf{S}_E \mathbf{1}_I$, where \mathbf{S}_E is the covariance matrix and $\mathbf{G} = \mathbf{I}_I - I^{-1} \mathbf{1}_I \mathbf{1}'_I$, with \mathbf{I}_I being the $I \times I$ identity matrix and $\mathbf{1}_I$ denoting the $I \times 1$ vector of unit elements. Testing hypothesis $H_\beta : \beta = 0$, concerning the regression of the interaction effects on the environment main effect (or the mean $y_{.j}$) is described by Caliński et al. (1997). The rejection of the hypothesis H_β implies that the significant part of the *GE* interactions can be explained by their statistical dependence on the environment main effect. In this case testing the hypotheses $H_{\beta_i} : \beta_i = 0$ ($i = 1, 2, \dots, I$) is of interest.

In present paper a multiple linear regression of interaction effects of genotypes on concomitant variables pertaining to environments is also discussed.

The analysis by a multiple linear regression of *GE* interaction effects on chosen concomitant variables pertaining to environments is made in accordance with the method used by Caliński et al. (1997).

4 Results

The analysis of variance for the data described in the paragraph "The data" was made. Because the general hypothesis of no GE interaction was rejected ($F_{GE} = 1.45$ at $F_{0.05} = 1.22$), the individual analysis for each genotype was justifiable. Information concerning the performance of randomly chosen 9 genotypes (DH lines) can be read from Tables 1 and 2. Table 1 contains the results of individual analysis for genotypes, i.e. the main effect of estimate, the F -statistic for it, the F -statistic for GE interaction of genotypes with environments and the results concerning regression of interaction of particular genotypes on environment main effect. In this case the environment is characterized by the environmental deviations, i.e. estimates of their main effects. Only one genotype (No. 28) indicates some regression of their interaction on environment effect. In Table 2 the results of the multiple linear regression effects of genotypes on two canonical variables concerning seed yield and oil content, are given. The GE interaction for a few genotypes is closely depended with one or two concomitant variables.

TABLE 1. Estimates and results of testing the hypotheses concerning genotypes, their GE and interactions and hypothesis of no regression between the interactions and environment main effects.

Genotype	Estimated main effect	F -statistic for the main effect	GE interaction	Coefficient of determination (%)	regression	regression	F -statistic for deviation from regression
G2	2.38	23.06*	2.55*	55.0	-0.41	4.88	1.44
G6	-1.70	32.50*	0.93	29.3	0.18	1.66	0.82
G11	4.58	150.50*	1.45	17.3	-0.17	0.84	1.50
G16	1.93	2.44	15.93*	14.1	-0.52	0.65	17.13*
G18	0.32	0.89	1.19	47.8	-0.26	3.66	0.78
G24	-2.61	23.95*	2.98*	36.2	0.36	2.27	2.37
G25	-2.25	76.87*	0.69	11.6	0.10	0.53	0.76
G28	-1.43	28.22*	0.76	91.7	0.29	44.39*	0.08
G31	-2.18	27.99*	1.87	28.7	-0.25	1.61	1.58

* - significant at the 5% level

TABLE 2. Estimates and results of testing the hypotheses concerning genotypes, their GE and interactions and hypotheses concerning multiple linear regression of GE interaction effects on two canonical variables pertaining to environments.

Genotype	Coefficient of determination (%)	F -statistic for regression	Norm. seed yield	coeff. of reg. oil contents	After elim. nonsignificant variable	Coefficient of determination (%)	F -statistic for regression
G2	59.9	2.24	-0.20	-0.84*	56.5	5.20	
G6	47.2	1.34	0.16	0.74*	45.2	3.29	
G11	14.0	0.24	-0.40	-0.10	13.2	0.61	
G16	45.2	2.03	0.73*	0.23	40.7	2.74	
G18	98.0	73.79*	0.59*	-0.59*			
G24	52.5	1.66	-0.62*	-0.19	49.5	3.92	
G25	74.9	14.47*	0.92*	-0.15	73.0	10.82*	
G28	70.6	3.60	0.01	0.84*	70.6	9.60*	
G31	89.3	12.50*	0.85*	-0.19	86.2	25.05*	

* - significant at the 5% level

5 Final remarks

Both presented methods have a potential to provide important results concerning explanation of the GE interactions observed in particular experiments. Comparing the results from Table 1 and Table 2 one can say that at least part of the GE interactions can be explained by their statistical dependence on the concomitant variables.

References

Caliński T., Czajka S. and Matchmark Z. (1997). A multivariate approach to analysing genotype-environment interactions. In: *Advances in Biometrical Genetics*. Poznań, 1997, P. Krajewski and Z. Matchmark (eds.), 3–14.

Discrete valued time series models for examining weather effects in daily accident counts

Dimitris Karlis¹, George J. Sermaidis² and Tom Brijs³

¹ Department of Statistics, Athens University of Economics and Business, 76 Patission str., 10434 Athens, Greece, e-mail: karlis@aub.gr

² Department of Statistics, University of Warwick, Coventry CV4 7AL, England

³ Transportation Research Institute, Hasselt University, Wetenschapspark, Gebouw 5, B-3590 Diepenbeek, Belgium

Abstract: In this paper we aim at examining the effect of weather conditions in daily accident counts. In order to account for the serial correlation and the overdispersion present to the data, we make use of two models for discrete valued time series using covariate information. The models considered are the model of Zeger and the Integer Autoregressive model including covariates. Estimation procedures and possible extensions of the models are discussed. Data from 27 major cities roads in the Netherlands are examined. We make use of a meta-analysis approach in order to combine the effects retrieved for each site with site-specific covariate information.

Keywords: INAR model, Zeger's model, accidents statistics

1 Introduction

The last few years, road accidents statistics are the subject of increased interest both on the part of policy makers and academia. The objective is to better understand the complexity of factors that are related to road accidents in order to take corrective actions to remedy this situation. In this context, the modelling of accidents over time has obtained considerable attention by researchers in the past. In this paper, we study the effects of weather conditions on daily accidents for 27 major cities in the Netherlands. The use of weather conditions is motivated by earlier research where significant influences of weather conditions on accidents have been found. To do so we propose and apply two models adequate for modelling discrete valued time series models, namely the Integer Autoregressive model proposed by McKenzie (1985) and Al-Osh and Al-Zaid (1987) and the model proposed by Zeger (1988). The first belongs to the category of observation driven models, since we relate directly the observations themselves, while the second one to the category of the parameter driven models as the dependence structure comes from a time dependent process in the parameters of the model.

In order to capture the effect of weather covariates in the accident counts we use covariate information related to weather conditions. This includes covariate information about the rainfall, the wind, the temperature and other weather characteristics in the area of examination.

Furthermore in order to account for the different characteristics of the 27 roads we proceed with meta-analysis of the derived results. This approach allows to combine the results from the different sites and to examine site-specific effects.

2 The models

2.1 INAR model

McKenzie (1985) and Al-Osh and Al-Zaid (1987) defined a process for discrete data which mimics the standard autoregressive model for continuous data, called the Integer-valued autoregressive (INAR) process as follows:

A sequence of random variables $\{Y_t\}$ is an INAR(1) process if it satisfies a difference equation of the form

$$Y_t = \alpha \circ Y_{t-1} + R_t, \quad t = 1, 2, \dots, \quad (1)$$

where R_t is the innovation term, which is a discrete random variable. According to the choice of the distribution of the innovations certain marginal properties can be deduced for the process. The operator " \circ " denotes the binomial thinning operator defined by $\alpha \circ Y = \sum_{t=1}^Y Z_t$, where Z_t are independent Bernoulli random variables with $P(Z_t = 1) = \alpha = 1 - P(Z_t = 0)$, $\alpha \in [0, 1]$. Thus, conditional on Y_t , $\alpha \circ Y_t$ is a binomial random variable where Y_t denotes the number of trials and α the probability of success in each trial.

The basic ingredient of the INAR model is that it assumes that the realization of the process at time t is composed by two parts, the first one clearly relates to the previous observation, while the second one is independent from it and depends only on the current time point. Thus, the first part represents the influence of previous time periods while the innovation term captures the effects of the present time point. Although it is possible to incorporate higher-order lags into the model, we do not pursue them since their interpretation is not straightforward. More detail on such models can be found in Jung and Tremayne (2006).

Assuming that R_t follows a Poisson distribution the Poisson INAR model arises, which assumes that the marginal distribution of Y_t is a Poisson distribution. The simple Poisson INAR model can be extended to a Poisson INAR regression model by adding covariates to both the innovation term and/or the autocorrelation parameter. The model then takes the form

$$\begin{aligned} Y_t &= \alpha_t \circ Y_{t-1} + R_t \\ R_t &\sim \text{Poisson}(\lambda_t) \\ \log \lambda_t &= \mathbf{z}_t' \boldsymbol{\beta} \\ \log \left(\frac{\alpha_t}{1 - \alpha_t} \right) &= \mathbf{w}_t' \boldsymbol{\gamma} \end{aligned}$$

where \mathbf{z}_t and \mathbf{w}_t are vectors of covariates at time t for the innovation term and the autocorrelation parameter respectively while $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vector of the associated regression coefficients. Note that the covariate information for the two parts of the

model are not necessarily the same. We have developed an EM type algorithm for fitting this model to real data making use of the convolution representation of the process. Details of the algorithm are omitted.

Extensions of the model to allow for overdispersion can be made by assuming an overdispersed innovation distribution.

2.2 Zeger's Model

We describe the model proposed by Zeger (1988). Let's suppose we have observed a time series of counts y_t , $t = 1, 2, \dots, T$, as well as a vector of covariates \mathbf{x}_t . Our goal is to describe $\mu_t = E(Y_t)$ as a function of the $p \times 1$ vector of covariates. Furthermore, assuming that the distribution of y_t is Poisson, that is $y_t \sim \text{Poisson}(\mu_t)$, where $\mu_t = \exp(\mathbf{x}_t' \mathbf{b})$, maximum likelihood method can be used to estimate the unknown vector of coefficients \mathbf{b} . In practice, quite often the sample variance exceeds the sample mean, providing evidence that an overdispersed relative to the Poisson distribution must be used. In this case quasi-likelihood methods which allow a variety of variance-mean relation is more appropriate.

Extensions of log-linear models which account for dependence are necessary to obtain valid inference about the relationship of y_t and \mathbf{x}_t . Zeger suggested that if ϵ_t is an unobservable noise process then the conditional distribution of y_t on ϵ_t is Poisson with mean equal to the product of the latent process value and the predictor as in a simple log-linear model. Therefore

$$Y_t | \epsilon_t \sim \text{Poisson}(\epsilon_t \exp(\mathbf{x}_t' \mathbf{b})) \quad (2)$$

Assume that ϵ_t is a non-negative time series with mean 1, autocovariance function $\gamma_\epsilon(h)$ and variance σ_ϵ^2 . Letting $\delta_t = \log \epsilon_t$, then the conditional mean of Y_t on ϵ_t can be written as

$$u_t = \exp(\mathbf{x}_t' \mathbf{b} + \delta_t) \quad (3)$$

We assume $E(\exp(\delta_t)) = 1$. Unless the δ_t is a stationary Gaussian process, there is not an explicit relationship between the autocovariance functions of ϵ_t and δ_t .

For this model, the marginal variance of Y_t is greater than its marginal mean providing this way a degree of overdispersion which depends on the variance of the latent process σ_ϵ^2 . Another interesting property of this model is that the form of the autocorrelation of the observed counts inherits its structure from that of the latent process. It is also true that even if there is no significant autocorrelation in y_t , it does not necessarily mean that autocorrelation is not present in ϵ_t either. This implies that the autocorrelation function of the observed count process will tend to underestimate that of the latent process, even in the simplest case where no regressors are present. Therefore, the latent process introduces both autocorrelation and overdispersion in Y_t . The interpretation of any element of the vector of coefficients \mathbf{b} in the above model is the same as in a simple Poisson regression model.

Estimation of this model is not easy. The full likelihood of the model cannot be written easily as it is defined recursively. A GEE approach has been proposed by Zeger (1988) in order to estimate the parameters of the model. We have followed this approach.

Concluding this section, the two model, despite their different generation mechanism implied, have some more differences in the sense that the model of Zeger allows for overdispersion. In the sequel we applied both model to our data.

3 Meta-Analysis

Meta-analysis can be defined as the quantitative review and synthesis of the results of related but independent studies. By combing information over different studies, an integrated analysis will have more statistical power to detect a specific effect than an analysis based on only one study. When several studies have conflicting conclusions, a meta-analysis can be used to estimate an average effect. For an excellent review on meta-analysis the reader can refer to Normand (1999). In this paper we aim at combining results from different sites in order to synthesize a general effect.

A fixed-effects model assumes that each study summary statistic Y_i (in our case a regression coefficient summarizing the effect of a weather variable) is a realization from a population of study estimates with common mean θ . Let α be the central parameter of interest and assume there are $i = 1, 2, \dots, k$ independent studies. Assume that Y_i is such that $E(Y_i) = \theta$ and let $Var(Y_i) = s_i^2$ be the variance of the summary statistic in the i th study. For moderately large study sizes, each Y_i should be normally distributed (by the central limit theorem) and approximately unbiased. Thus

$$Y_i \sim N(\alpha, s_i^2) \text{ for } i = 1, 2, \dots, k \quad (4)$$

and s_i^2 assumed known. The central parameter of interest is α which quantifies the average effect.

The random-effects model assumes that each study summary statistic Y_i is drawn from a distribution with a study-specific mean, α_i , and variance s_i^2 .

$$Y_i | \alpha_i, s_i^2 \sim N(\alpha_i, s_i^2) \text{ for } i = 1, 2, \dots, k \quad (5)$$

Furthermore, each study-specific mean α_i is assumed to have been drawn from some superpopulation of effects with mean α and variance τ^2 with

$$\alpha_i | \alpha, \tau^2 \sim N(\alpha, \tau^2) \quad (6)$$

The parameters α and τ^2 are to referred as hyperparameters and represent, respectively, the average effect and inter-study variation. Thus we introduce one more level of variability.

4 Application

This study is based on the daily accident counts that were obtained from the major roads covered by the surface of 27 big cities in the Netherlands in the year 2001. The cities were selected based on two criteria: a) their proximity to some national weather stations in order to obtain accurate daily weather conditions and b) they were far enough apart in order to prevent that weather conditions would be identical for the different sites for too many of the observations.

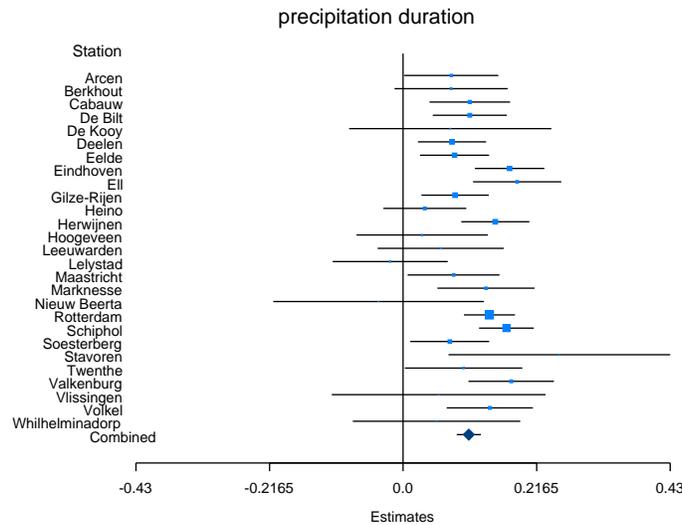


FIGURE 1. Weighted forest plot for the precipitation duration

For each site, the number of daily counts on accidents together with detailed weather information was collected. The day was used as a proxy of the different traffic volumes, i.e. as a proxy to the exposure. The data have several distinct features since for some roads the autocorrelation and the overdispersion varied considerably.

We make use of both the fixed and the random effects meta-analysis models. A typical forest plot can be seen in Figure 1 for the precipitation duration. One can see the different effects for each site and the combined estimator. This combined estimator shows that there is a positive effect of the precipitation duration.

The meta-regression models identified variables which influence the effect of the covariates. A summary of the main finding is that an increase of one unit of the maximum temperature decreases the effect of the mean temperature on the accidents and a decrease of a unit of the minimum temperature increases the temperature below zero effect. The effect of humidity covariate becomes stronger when combined with lower minimum temperatures. The rainfall intensity effect was related to the increase of the rainfall duration.

However, one must interpret the findings with care since weather variables can also have some influence on the exposure. Hence, the effects found are not necessarily explicit on the accidents but they can be implicit through the increase/decrease of the exposure.

References

- Al-Osh M.A. and Al-Zaid A.A. (1987). First Order Integer Valued Autoregressive Process, *Journal of Time Series Analysis* **8**, 261-275.

- Jung, R.C. and Tremayne, A.R. (2006), Binomial thinning models for integer time series, *Statistical Modelling* **6**, 81-96
- McKenzie E. (1985). Some Simple Models for Discrete Variable Time Series. *Water Resources Bulletin* **21**, 645-650.
- Normand S.L. (1999) Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 321-359.
- Zeger S.L. (1988). A Regression Model for Time Series of Counts. *Biometrika* **75(4)**, 621-9.

Smooth models of mortality with period shocks

James Kirkby¹ and Iain Currie¹

¹ Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland, I.D.Currie@hw.ac.uk

Abstract: We suppose that we have mortality data arranged in two-way tables of deaths and exposures classified by age at death and year of death. It is natural to suppose that there is a smooth underlying force of mortality, the mortality surface, that varies with age and year (or period). However, observed mortality is subject to more than stochastic deviation from this smooth surface; for example, flu epidemics, hot summers or cold winters can disproportionately effect the mortality of certain age groups in particular years. We call such an effect a period shock. We describe the mortality surface with an additive model with two components: the underlying smooth surface is modelled with 2-dimensional P -splines; the period shocks are modelled with a 1-dimensional P -spline in the age direction for each year. This is a large regression model but array methods (Currie *et al.*, 2006) enable the computations to be performed. We illustrate our methods with Swedish mortality data taken from the Human Mortality Database.

Keywords: Generalized linear array model; mortality; P -splines; period shock; smoothing.

1 A smooth model of mortality with period shocks

We suppose that we have mortality data arranged in two-way tables of deaths and exposures classified by age at death and year of death. It is natural to suppose that there is a smooth underlying force of mortality, the mortality surface, that varies with age and year (or period). However, observed mortality is subject to more than stochastic deviation from this smooth surface; for example, flu epidemics, hot summers or cold winters can disproportionately effect the mortality of certain age groups in particular years. We call such an effect a period shock.

An example of a period shock is the Spanish flu epidemic of 1918 which affected the mortality of those under the age of 60. We illustrate the extent of this effect with data on Swedish males from the Human Mortality Database; the data runs from 1900 to 2003 and from age 10 to 90. The upper right panel of Figure 1 shows the differences between the logarithms of observed mortality for 1918 and 1919. The remaining panels show the corresponding differences for other years. The 1918/1919 experience is extreme but each panel indicates a systematic age dependent departure from a smoothly changing underlying mortality.

We propose an additive model with two components for the mortality surface : the first component describes a smooth two-dimensional surface and the second describes the period shocks. We describe each component in turn.

We suppose that the deaths and exposures are arranged in $n_a \times n_y$ matrices \mathbf{Y} and \mathbf{E} , that $\mathbf{y} = \text{vec}(\mathbf{Y})$ and $\mathbf{e} = \text{vec}(\mathbf{E})$ are their vector equivalents, and that the rows and columns of \mathbf{Y} and \mathbf{E} are classified respectively by ages \mathbf{x}_a and years \mathbf{x}_y each arranged in ascending order. The first component of our model uses two-dimensional P -splines

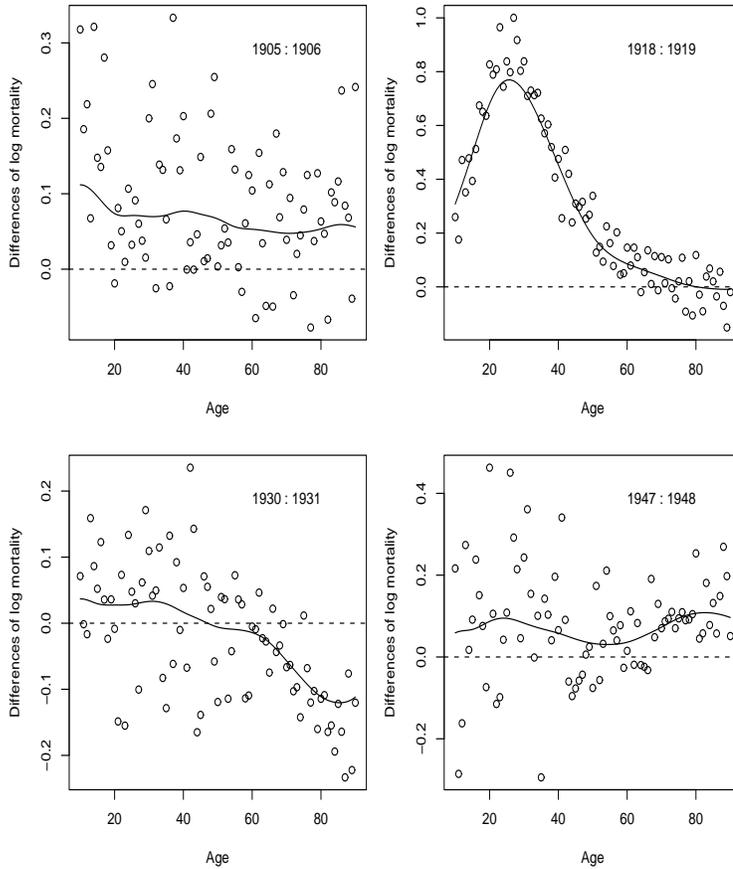


FIGURE 1. Differences of the logarithms of observed and fitted mortalities for Swedish males for selected successive years

(Eilers and Marx, 1996, Currie *et al.*, 2004) to model the smooth underlying mortality surface. Let $\mathbf{B}_a = \mathbf{B}(\mathbf{x}_a)$, $n_a \times c_a$, and $\mathbf{B}_y = \mathbf{B}(\mathbf{x}_y)$, $n_y \times c_y$, be one-dimensional regression matrices of B -splines evaluated at age and year respectively; \mathbf{B}_a and \mathbf{B}_y are known as marginal regression matrices. The Kronecker product $\mathbf{B}_y \otimes \mathbf{B}_a$ creates a two-dimensional regression basis. We suppose that the number of deaths y_{ij} at age i and year j follows a Poisson distribution with mean $\mu_{ij} = e_{ij}\theta_{ij}$ where θ_{ij} is the force of mortality. We define a generalized linear model (GLM) for \mathbf{y} with regression matrix $\mathbf{B}_y \otimes \mathbf{B}_a$, offset $\log \mathbf{e}$, log link and Poisson error. If we suppose that we have a rich basis of B -splines for age and year then a smooth surface is obtained by marginal penalization. We define the penalty matrix

$$\mathbf{P} = \lambda_a \mathbf{I}_{c_y} \otimes \mathbf{D}'_a \mathbf{D}_a + \lambda_y \mathbf{D}'_y \mathbf{D}_y \otimes \mathbf{I}_{c_a} \tag{1}$$

where \mathbf{D}_a , λ_a , \mathbf{D}_y and λ_y are the difference matrices and smoothing parameters for

age and year respectively. We have defined a two-dimensional P -spline model; see Currie *et al.* (2004, 2006) and elsewhere for details of fitting these models.

We require the second component to have two properties. First, it must be versatile and capable of modelling different patterns in different years and second, the underlying patterns in different years must be smooth. Figure 1 illustrates both these features: the patterns in different years are distinct and follow separate underlying smooth curves. We define a second regression matrix by $\mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a$ where $\check{\mathbf{B}}_a = \check{\mathbf{B}}_a(\mathbf{x}_a)$, $n_a \times c$, is a marginal regression matrix of B -splines. We suppose that c is small so that the modelling by age for each year is quite crude. With c regression coefficients for each of n_y years $\mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a$ is a large matrix, even with small c . The underlying smooth mortality surface is modelled by the first component of our model so we force smoothness on the second component by applying a ridge penalty to the age coefficients in each year. We define the penalty matrix

$$\check{\mathbf{P}} = \lambda_s \mathbf{I}_{n_y} \otimes \mathbf{I}_c = \lambda_s \mathbf{I}_{n_y c} \quad (2)$$

where λ_s is the smoothing parameter for the shocks. Our two component model of mortality is thus a penalized generalized linear model with two additive components, linear predictor

$$\log \mu = \log e + \mathbf{B}_y \otimes \mathbf{B}_a \mathbf{a} + \mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a \check{\mathbf{a}} \quad (3)$$

and block diagonal penalty blockdiag[$\mathbf{P} : \check{\mathbf{P}}$]. This is a computationally demanding problem since three smoothing parameters must be chosen within the framework of a large GLM with $n_a n_y$ observations and $c_a c_y + c n_y$ regression variables. We describe how we deal with the computational problem in the next section.

2 Generalized linear array models

Generalized linear array models or GLAMs, introduced by Currie *et al.* (2006), provide a structure and a computational procedure for fitting GLMs whose model matrix can be written as a Kronecker product and whose data can be written as an array. The GLAM form of the 2-dimensional smooth model with regression matrix $\mathbf{B}_y \otimes \mathbf{B}_a$ is

$$\log \mathbf{M} = \log \mathbf{E} + \mathbf{B}_a \mathbf{A} \mathbf{B}_y' \quad (4)$$

where $\mathbf{M} = E(\mathbf{Y})$ and \mathbf{A} , $c_a \times c_y$, is the matrix of coefficients. In a large problem the GLAM approach gives very substantial savings in both storage and computational time over the usual GLM algorithm. The method can be extended to GLMs with additive components each of which has the GLAM form. The GLAM form of (3) is

$$\log \mathbf{M} = \log \mathbf{E} + \mathbf{B}_a \mathbf{A} \mathbf{B}_y' + \check{\mathbf{B}}_a \check{\mathbf{A}} \quad (5)$$

where $\check{\mathbf{A}}$, $c \times n_y$, is a further matrix of coefficients. We use the GLAM procedure to fit model (3). Efficient computation of the linear predictor in (3) is provided by (5). The fitting of a GLM also requires the computation of a weighted inner product; for the additive model (3) we require

$$\begin{bmatrix} (\mathbf{B}_y \otimes \mathbf{B}_a)' \mathbf{W} (\mathbf{B}_y \otimes \mathbf{B}_a) & (\mathbf{B}_y \otimes \mathbf{B}_a)' \mathbf{W} (\mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a) \\ (\mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a)' \mathbf{W} (\mathbf{B}_y \otimes \mathbf{B}_a) & (\mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a)' \mathbf{W} (\mathbf{I}_{n_y} \otimes \check{\mathbf{B}}_a) \end{bmatrix} \quad (6)$$

where \mathbf{W} is the diagonal matrix of weights. Details of the efficient computation of this matrix are given in Currie *et al.* (2006), in particular the example in section 7.1. The GLM is now fitted with the usual scoring algorithm with the linear predictor, weighted inner product and working variable all computed using array computations.

A simpler method of modelling shocks is the mean shock, by which we mean a constant shock for all ages within a year. The GLAM form of the linear predictor for this model is

$$\log M = \log E + \mathbf{B}_a \mathbf{A} \mathbf{B}'_y + \mathbf{1}_{n_a} \mathbf{h}' \tag{7}$$

where $\mathbf{1}_{n_a}$ is a vector of 1's of length n_a and $\mathbf{h}' = (h_1, \dots, h_{n_y})$. Mean shocks are used in some well-known models of mortality, such as the Lee-Carter and the Age-Period-Cohort models; neither of these models is capable of modelling the kind of effects we see in Figure 1 but we consider model (7) for comparison.

3 An application to Swedish data

We use the Bayesian Information Criterion (BIC) for model selection and fit our three models to the Swedish mortality data used in section 1. We have $n_a = 81$ ages and $n_y = 104$ years and used cubic B -splines with $c_a = 19$, $c_y = 24$ and $c = 9$. The full regression matrix in (3) is 8424×1392 , a large regression matrix. The fitted values have been added to Figure 1; it appears that our model has successfully modelled the systematic increases and decreases seen for our selected years. The extent of the period shocks is shown in Figure 2; the age structure of these shocks is evident.

Table 1 gives various summary statistics for all three models: the basic 2-dimensional smooth model (4), the mean shock model (7) and the age dependent shock model (5). It is clear that the period shock model is superior to the mean shock model which is, in turn, superior to the 2-dimensional model.

TABLE 1. Summary statistics

Model	$(\lambda_a, \lambda_y, \lambda_s)$	Trace	Deviance	BIC
2- d smooth	(10, 7, -)	293	21226	23871
Mean shock	(0.05, 30, 2000)	367	15538	18852
Period shock	(0.01, 1900, 850)	489	9670	14089

4 Concluding remarks

The 1918 observations are extreme. An alternative modelling strategy is to regard the 1918 data as missing. This has the effect of increasing the value of λ_s (since the remaining data are much less variable by year). The estimates of the remaining period shocks are smoother compared to the shocks from the full data set.

It is possible to express all our models in mixed model form. One consequence of this approach is that the period shocks can be regarded as random effects which is

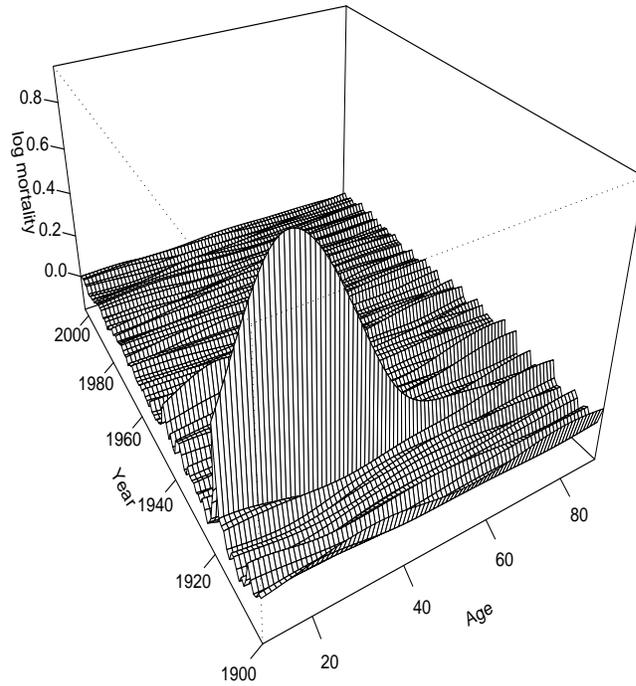


FIGURE 2. Age dependent shocks to mortality surface by year

appealing given their nature. The GLAM methodology is available for mixed models (Currie *et al.*, 2006).

The models in this paper and model (3) in particular are computationally intensive. Further computational savings over and above those provided by GLAM can be made by taking advantage of the form of the regression matrices of the period components in (3) and (7).

In conclusion, the period shock model was successful in modelling age-dependent departures within years for the Swedish data considered in this paper. Experience with other data sets has shown that the model is more widely applicable. We see two main uses of the model: first, it can detect age effects within single years and this will often be of interest in its own right; second, the two component model enables the underlying smooth surface to be more successfully identified.

References

- Currie, I.D., Durban, M., and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* **4**, 279-98.
- Currie, I.D., Durban, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B* **68**, 259-80.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. *Statistical Science* **11**, 89-121.
- Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org.

Generalized linear mixed model with a flexible random-effects distribution

Arnošt Komárek¹ and Emmanuel Lesaffre¹

¹ Katholieke Universiteit Leuven, Biostatistical Centre, Kapucijnenvoer 35, Leuven, Belgium.
E-mail: Arnost.Komarek@med.kuleuven.be

Abstract: It is known that misspecification of the distribution of random-effects in a generalized linear mixed model (GLMM) may lead to biased estimation of the fixed-effects. In this paper, we will show how the distribution of the random-effects can be specified in a flexible way using a penalized Gaussian mixture. The methodology will be illustrated on a longitudinal study from a clinical trial in dermatology, where the outcome is binary. For practical analyses an R package has been written.

Keywords: Clustered data; Logistic regression; Longitudinal study; Markov chain Monte Carlo.

1 Generalized linear mixed model

The generalized linear mixed model (GLMM) is a popular tool to regress a discrete response when the measurements are clustered (e.g., multicenter clinical trials and longitudinal studies). Let $Y_{i,l}$ ($i = 1, \dots, N, l = 1, \dots, n_i$) be the l -th response in the i -th cluster or the l -th longitudinal response of the i -th unit. In the remainder of the paper, we will restrict ourselves to longitudinal studies. However, the whole methodology can equally be used for clustered data.

In the GLMM, the effect of the covariates on the response is modelled as

$$h\{E(Y_{i,l} | \mathbf{b}_i)\} = \mathbf{b}'_i \mathbf{z}_{i,l} + \boldsymbol{\beta}' \mathbf{x}_{i,l} \quad (i = 1, \dots, N, l = 1, \dots, n_i), \quad (1)$$

where h is a known link function, $\mathbf{x}_{i,l}$ is a vector of covariates and $\mathbf{z}_{i,l}$ represents a subset of covariates for which the effect may vary randomly across units. Further, $\boldsymbol{\beta}$ is the vector of regression coefficients (fixed-effects) and $\mathbf{b}_1, \dots, \mathbf{b}_N$ are unit specific zero-mean vectors of random-effects.

2 Distribution of random effects

It is conventionally assumed that the random effects are normally distributed, i.e., that $\mathbf{b}_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{D})$. However, it has been shown that, in contrast to the linear mixed model (LMM), misspecification of the random-effects distribution in GLMM can lead to biased estimates of the fixed-effects (e.g., the treatment effect) that are usually of primary interest. See Molenberghs and Verbeke (2005, Chap. 23) for more

details and references. For this reason, there is a need for GLMMs with a more flexible random-effects distribution.

Recently, a flexible model for the random effects distribution, called ‘penalized Gaussian mixture (PGM)’, has been suggested, based on the idea of penalized smoothing promoted by Eilers and Marx (1996). Ghidry et al. (2004) used the PGM to model the random-effects distribution in a LMM. Komárek et al. (2005) used it as a flexible specification of the distribution of a logarithm of the baseline event time in a survival regression model. Further, Komárek and Lesaffre (2006) used the bivariate PGM in a survival regression model for paired data. Finally, Komárek and Lesaffre (2007) exploited the PGM to model all distributional parts in a random-effects survival regression model.

In this paper, we will show how the PGM can be used as a model for the random-effects distribution in a GLMM. The methodology will be illustrated on longitudinal binary data arising from a dermatological clinical trial using a random-intercept logistic model. An R (R Development Core Team, 2006) package has been written and is available upon request from the first author.

3 Penalized Gaussian mixture

For the purpose of this paper, only a univariate PGM will be described which can, e.g., serve as a model for the distribution of the random intercept in GLMM. Namely, we will assume that univariate random effects b_1, \dots, b_N are i.i.d. distributed with a density $g(b)$. Let φ_{μ, σ^2} denote the density of the normal distribution $\mathcal{N}(\mu, \sigma^2)$. The density $g(b)$ is modelled by a PGM in the following way

$$g(b) = \tau^{-1} \sum_{j=-K}^K w_j \varphi_{\mu_j, \sigma^2}(\tau^{-1}b), \quad (2)$$

where τ is an unknown scale parameter and $\mathbf{w} = (w_{-K}, \dots, w_K)'$ is a vector of unknown PGM weights. We consider the model (2) as a spline-like smoothing of the unknown function $g(b)$ where $\varphi_{\mu_{-K}, \sigma^2}, \dots, \varphi_{\mu_K, \sigma^2}$ form a basis specified over a grid of knots μ_{-K}, \dots, μ_K .

In the philosophy of penalized smoothing (Eilers and Marx, 1996), we will use a relatively high number ($2K + 1$ equal to 30–40) of *equidistant* knots. In the remainder we denote the distance between two consecutive knots as δ . Since the random effects should have mean zero, the knots are centered around zero, i.e. $\mu_0 = 0$ and $\mu_j = j\delta$ ($j = -K, \dots, K$).

To obtain a reasonable model for the unknown distribution the knots should be put in the area where the true random intercept density has a significant amount of the probability mass. Since a scale parameter τ is included in the model (2), the knots should cover an area where a zero-mean and unit-variance distribution has most of its probability mass. To this end, the choice of the boundary knots $\mu_{-K} \approx -5$, $\mu_K \approx 5$ usually suffice.

The choice of the basis standard deviation σ corresponds to some extent to the choice of the degree of a B-spline, if the B-splines were used instead of normal densities.

The value of $\sigma = (2/3)\delta$ is motivated by the correspondence to cubic B-splines, as explained in Komárek et al. (2005).

Finally, satisfactory conditions for the PGM weights to ensure that (2) is a density are $w_j > 0$ ($j = -K, \dots, K$) and $\sum_{j=-K}^K w_j = 1$. To avoid constrained estimation, the vector of transformed weights $\mathbf{a} = (a_{-K}, \dots, a_K)'$ given by

$$a_j = \log\left(\frac{w_j}{w_0}\right), \quad w_j = \frac{\exp(a_j)}{\sum_{k=-K}^K \exp(a_k)} \quad (j = -K, \dots, K), \quad (3)$$

is used. For identifiability reasons $a_0 = 0$.

To avoid overfitting and identifiability problems caused by the relatively high number of parameters, Eilers and Marx (1996) suggested to use the method of penalized maximum-likelihood for estimation purposes. In our paper, we suggest to use a Bayesian specification of the model with uninformative prior distributions for all the model parameters except for the transformed weights \mathbf{a} . The prior $p(\mathbf{a})$ for \mathbf{a} is specified as an intrinsic Gaussian Markov random field (IGMRF) which corresponds closely to the penalty term used by Eilers and Marx (1996), see Lang and Brezger (2004). Thus

$$p(\mathbf{a}) \propto \exp\left\{-\frac{\lambda}{2} \sum_{j=-K+m}^K (\Delta^m a_j)^2\right\}, \quad (4)$$

where Δ^m denotes the backward difference operator of order m ($m = 3$ is used in our applications). Further, the smoothing hyperparameter λ can be interpreted as an inverse variance of the IGMRF and can be given a Gamma prior $G(\xi_1, \xi_2)$ with small values of ξ_1 and ξ_2 . Inference is based on a sample from the posterior distribution obtained using Markov chain Monte Carlo methods.

TABLE 1. Toenail data. Posterior median and 95% credible interval for model parameters.

Effect	PGM model		Normal model	
Intercept	-1.588	(-2.945, -0.705)	-1.609	(-2.610, -0.758)
Treatment	0.431	(-0.448, 1.298)	-0.175	(-1.340, 0.996)
Time	-0.383	(-0.480, -0.300)	-0.394	(-0.486, -0.312)
Time:Treatment	-0.134	(-0.277, 0.002)	-0.137	(-0.276, -0.004)
sd(b)	3.446	(2.733, 4.783)	4.034	(3.361, 4.887)

4 Practical example: Toenail data

We will apply the developed methodology on a longitudinal clinical trial in dermatology which was set up to compare the efficacy of two oral treatments for toenail infection (De Backer et al., 1998). One of the end points of the study was the degree of onycholysis which expresses the degree of separation of the nail plate from the nail-bed (0, absent; 1, mild; 2, moderate; 3, severe) and was evaluated at seven visits (on weeks 0, 4, 8, 12, 24, 36 and 48). In total, 1 908 measurements on 294 patients are available.

The effect of the treatment on the dichotomized onycholysis (0, absent or mild; 1, moderate or severe) has already been analyzed by Lesaffre and Spiessens (2001) with a logistic random-effects model assuming normally distributed random intercepts. The methodology of our paper allows us, among other things, to evaluate whether the assumption of normality of the random effects was reasonable.

Let $Y_{i,l}$ represent the dichotomized onycholysis of the i -th subject at the l -th visit. We will model it using the following random-effect logit model:

$$\text{logit}\{P(Y_{i,l} = 1 | b_i, \boldsymbol{\beta})\} = b_i + \beta_0 + \beta_1 \text{trt}_i + \beta_2 t_{i,l} + \beta_3 t_{i,l} \cdot \text{trt}_i, \quad (5)$$

where trt_i denotes the binary treatment indicator of the i -th subject and $t_{i,l}$ the time of the l -th visit for the i -th subject in months. Further, b_i is the subject-specific random intercept and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ is the vector of regression coefficients.

In the PGM Model we assume that the random intercepts b_i ($i = 1, \dots, N$) are i.i.d. with a density given by (2). For comparison purposes, we also fitted the Normal Model where we assumed that b_i ($i = 1, \dots, N$) are i.i.d. with a normal distribution $\mathcal{N}(0, \tau^2)$. Posterior summary statistics are given in Table 1, pointwise posterior mean of the random intercept density is shown in Figure 1. It is seen that the fitted random intercept distribution is quite distinct from normality.

In their simulation study, Litière et al. (2007) observed a bias in the estimation of fixed effects ($\boldsymbol{\beta}$ coefficients) when the random-effect distribution in the GLMM is misspecified. Moreover, situation with a bimodal normal mixture (the most similar to our fitted random-effect distribution) as a true random-effect distribution showed one of the highest biases. Their findings are confirmed by our example, see the difference in the posterior medians for main treatment effect between the PGM Model and the Normal Model.

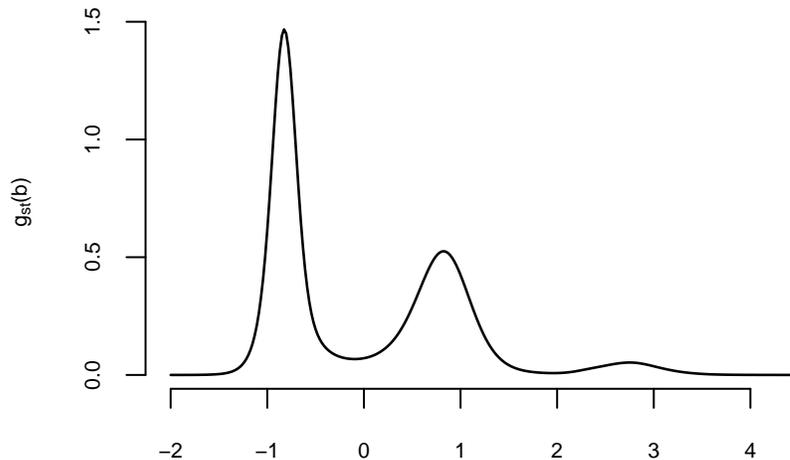


FIGURE 1. Posterior pointwise mean of the random-effect density (standardized to have zero-mean and unit-variance) in the PGM model.

5 Conclusions

We have suggested a method for smooth estimation of the random-effects distribution in the GLMM which allows to relax a conventional assumption of normality. We confirm findings of others, i.e. that a misspecified random-effect distribution can lead to biased results for the fixed-effects. For practical computations, a software package has been written in R.

Acknowledgements

The research of the first author was performed in the framework of a postdoctoral mandate PDM/06/242 financed by the Research Funds, Katholieke Universiteit Leuven. Support from the Interuniversity Attraction Poles Program P6/03 - Belgian State - Federal Office for Scientific, Technical and Cultural Affairs is also acknowledged. The authors thank Novartis, Belgium, for permission to use their dermatological data for statistical research.

References

- De Backer, M., De Vroey, C., Lesaffre, E., Scheys, I., and De Keyser, P. (1998). Twelve weeks of continuous onychomycosis caused by dermatophytes: A double blind comparative trial of terbafine 250 mg/day versus itraconazole 200 mg/day. *Journal of the American Academy of Dermatology* **38**, S57–S63.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with Discussion). *Statistical Science* **11**, 89–121.
- Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics* **60**, 945–953.
- Komárek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics* **14**, 726–745.
- Komárek, A. and Lesaffre, E. (2006). Bayesian semiparametric accelerated failure time model for paired doubly-interval-censored data. *Statistical Modelling* **6**, 3–22.
- Komárek, A. and Lesaffre, E. (2007). Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association*. To appear.
- Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.
- Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics* **50**, 325–335.

- Litière, S., Alonso, A., Molenberghs, G., and Geys, H. (2007). The impact of a misspecified random-effects distribution on maximum likelihood estimation in generalized linear mixed models. Submitted.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.
- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL: <http://www.R-project.org>.

Penalized likelihood for a three-parameter Rasch Model

Ioannis Kosmidis¹

¹ University of Warwick, Dept. of Statistics, Coventry, CV4 7AL, UK,
i.kosmidis@warwick.ac.uk

Abstract: We apply a bias reduction method for ML estimation (Firth 1993) to a three-parameter Rasch model. For its application the scores and thus the likelihood function are appropriately modified so that the resultant estimator has bias of order $O(n^{-2})$. The form of the modified scores is given and a pseudo-response to be used for fitting procedures is derived. Kosmidis (2007) shows that application of the method to logistic regression models corresponds to an improved alternative of ordinary maximum likelihood mainly due to the shrinkage properties of the bias-reduced estimator. It is illustrated that these properties extend to the case of the 3-parameter Rasch model.

Keywords: Penalized likelihood; Bias reduction; Logistic regression

1 A three-parameter Rasch model

We focus on inference for a three-parameter Rasch model that has the form

$$\log \frac{\pi_{rs}}{1 - \pi_{rs}} = \eta_{rs} = \alpha_r + \beta_r \gamma_s \quad (r = 1, \dots, N), (s = 1, \dots, n), \quad (1)$$

where π_{rs} is, for example, the probability that the person s answers correctly to question r and α_r , β_r , γ_s are unknown parameters. In contrast to an ordinary logistic regression model, note that the log-odds η_{rs} are a non-linear function of the parameters. Specifically, η_{rs} is a 'partially linear' combination of parameters; if we fix either β_r or γ_s we end up with a linear model on the logistic scale.

2 Bias reduction

2.1 Modified score functions

Firth (1993) proposed appropriate modifications to the efficient scores such that the roots of the modified scores result in a first-order $O(n^{-1})$ unbiased estimator. Bull et al. (2002) and Kosmidis (2006), (2007) studied the case of ordinary logistic regression showing that the bias-reduction method constitutes an improvement over traditional maximum likelihood (ML) mainly due to the shrinkage properties of the penalized maximum likelihood (PML) estimator. This motivated the study of the behaviour of the PML estimator for models such as (1), where we depart from the assumption of a linear predictor.

Consider realizations y_{rs} of independent binomial random variables Y_{rs} with totals m_{rs} . Also, let $\delta = (\alpha_1, \dots, \alpha_N, \beta_1, \dots, \beta_N, \gamma_1, \dots, \gamma_n)$ be the vector of parameters of the Rasch model. Using the modifications based on the expected information (see Firth (1993)) and after some algebra, the t -th component of the modified score vector for δ takes the form

$$U_t^* = \sum_{r=1}^N \sum_{s=1}^n \left(y_{rs} + \frac{1}{2}h_{rs} - (m_{rs} + h_{rs})\pi_{rs} + c_{rs}v_{rs} \right) z_{rst}, \tag{2}$$

where h_{rs} is the s -th diagonal element of the $n \times n$ projection matrix $H_r = Z_r F^{-1} Z_r^T \Sigma_r$, with Z_r the $n \times 2N + n$ matrix of first derivatives of $\eta_r = (\eta_{r1}, \dots, \eta_{rn})$ with respect to δ , F is the Fisher information on δ and Σ_r is a diagonal matrix, with s -th diagonal element $Var(Y_{rs}) = v_{rs} = m_{rs}\pi_{rs}(1 - \pi_{rs})$. Also, z_{rst} denotes the (s, t) -th element of Z_r and c_{rs} is the asymptotic covariance of the estimator of β_r and the estimator of γ_s and can be obtained by the appropriate block of F^{-1} .

The term $c_{rs}v_{rs}$ in (2) reflects the aforementioned 'partial linearity' of the predictor; fixing either β 's or γ 's the model reduces to an ordinary binary logistic regression model and $c_{rs}v_{rs} = 0$, retrieving the form of the modified scores for binary logistic regression (see, for example, Heinze & Schemper (2002)).

2.2 Implementation of the bias-reduction method

By the form of the modified scores (2) and considering H_r as if they were matrices of known constants, the PML is formally equivalent to the use of ML after making the following adjustments to the response frequencies:

Counts of "successes"	$y_{rs} + \frac{1}{2}h_{rs} + c_{rs}\pi_{rs}I(c_{rs} \geq 0)$.
Counts of "failures"	$m_{rs} - y_{rs} + \frac{1}{2}h_{rs} - c_{rs}(1 - \pi_{rs})I(c_{rs} < 0)$.
Binomial totals	$m_{rs} + h_{rs} + c_{rs}\pi_{rs} - c_{rs}I(c_{rs} < 0)$,

with $I(e > 0)$ taking value 1 if $e > 0$ and 0 else. This fact can be directly used for obtaining the PML estimates. Estimated standard errors for the PML estimator can be obtained by using the square roots of the diagonal of F^{-1} evaluated at the PML estimate.

2.3 Properties of the PML estimator

Kosmidis (2007) proved that the PML estimates for binary logistic regression models have always finite values and that they are shrunk towards the origin relative to the ML estimates. In the case of the Rasch model, empirical work not reported here indicates that the PML estimators of the predictors η_{rs} enjoy the same properties. The importance of the shrinkage effect is recognized in terms of smaller variance and smaller estimated standard errors for the PML estimator, as well as reduced bias.

3 Illustration. 2001 U.S. House of Representatives roll call data

The data consists of 20 roll calls selected by 'Americans for Democratic Action' (ADA) that were voted by 439 members of the U.S. House of representatives on 2001. The s -th member votes 'with ADA' or 'not with ADA' for the r -th roll call and the outcome is considered as the realization of a Bernoulli random variable Y_{rs} that takes value either 1 or 0, respectively. Missing values that appear due to either denial of voting or speaker's privilege or unavailability of voter are excluded from our analysis. Model (1) is fitted to the data with $\pi_{rs} = P(Y_{rs} = 1)$. We apply the bias-reduction method and obtain the PML estimates and the corresponding estimated standard errors for every η_{rs} with $r = 1, \dots, 20$, $s = 1, \dots, 439$. To illustrate the shrinkage effect, we plot the maximum likelihood estimates (MLEs) against the bias-reduced estimates and the estimated standard errors for the maximum likelihood estimator against the ones for the bias reduced estimator (Figure 1). The line crossing both plots is a 45° line passing through the origin. Because of shrinkage the points in the first plot fall above the line when both estimates are positive and below it when both are negative. Also, note that the higher the absolute value of the MLE, the more apparent is the shrinkage effect. Further, since the bias-reduced estimates are smaller in absolute value than the MLEs, their estimated standard errors are also reduced, as noted in the second plot.

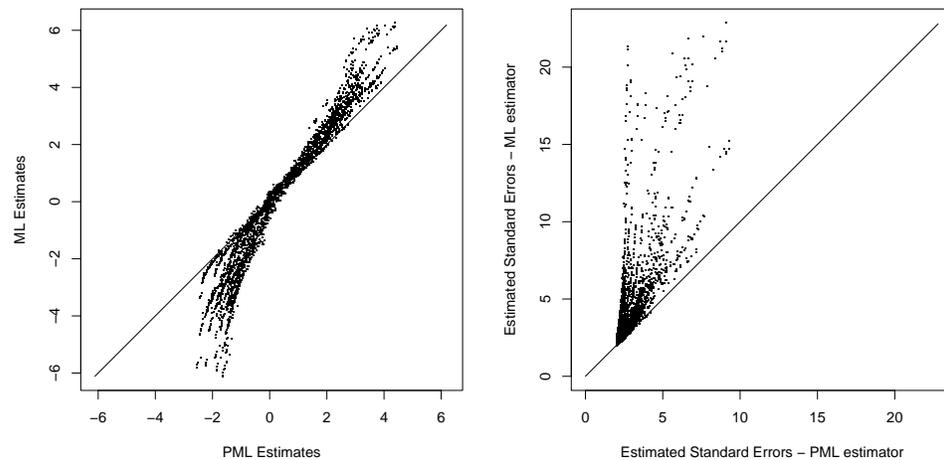


FIGURE 1. Plots of ML estimates against PML estimates and of estimated SEs for the ML estimator and for the PML estimator

Acknowledgments: The author was supported partially by an EPSRC research studentship.

References

- Americans for Democratic Action, ADA (2002). 2001 voting record: Shattered promise of liberal progress. *ADA Today* **57**, 1–17.
- Bull, S. B., Mak, C. and Greenwood, C. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics and Data Analysis* **39**, 57–74.
- de Leeuw, J (2006) Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis* **50**, 21–39.
- Firth D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.
- Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine* **21**, 2409–2419.
- Clinton, J., Jackman, S. and Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review* **98**, 355–370.
- Kosmidis, I. (2006) Bias reduction and shrinkage in multinomial logit models. In *21st International Workshop on Statistical Modelling*, Eds. J. Hinde, J. Einbeck & J. Newell, pp. 294–302. Galway, Ireland. ISBN 1-86220-180-3.
- Kosmidis, I. (2007) Bias reduction and shrinkage in logistic regression. *Unpublished report*

Cochran's Q -test for variance stabilized effect size estimates and a random effect size model

Elena Kulinskaya¹ and Robert G. Staudte²

¹ Statistical Advisory Service, Imperial College, Sir Alexander Fleming Building, South Kensington Campus, London, SW7 2AZ, UK,

e.kulinskaya@imperial.ac.uk

² Department of Statistical Science, La Trobe University, Melbourne, Australia 3086, r.staudte@latrobe.edu.au

Abstract: The standard method of testing for homogeneity (equal fixed effects) in several studies is to compare Cochran's Q statistic with a critical point of the Chi-squared distribution, and if homogeneity is rejected, a random effects model is then often assumed. Unfortunately, when the weights needed for Cochran's Q need to be estimated, the null distribution of Q is no longer Chi-squared, rendering the procedure suspect. Here we propose to stabilize the variance of estimated effect sizes via suitable transformations before applying Cochran's Q which then has known weights proportional to the sample sizes in the respective studies. If homogeneity of effect sizes is rejected, a variance component can be introduced on the space of the transformed effect sizes to allow for heterogeneity in the various studies. This allows one to find t -confidence intervals for the overall mean transformed effect size, and by back-transformation, an overall effect size.

Keywords: meta-analysis; standardized effects; variance component

1 Cochran's Q test applied to transformed effect sizes

1.1 The traditional test of homogeneity

Let μ_k , $k = 1, \dots, K$ be unknown fixed effects in K related studies. If it were known that all μ_k were equal (the homogeneous case) then one could proceed to estimate the common effect. Otherwise, a more complicated model needs to be adopted, so a test of homogeneity is often carried out. Assume there are independent estimated effects $\hat{\mu}_k$, $k = 1, \dots, K$, for these studies that satisfy $\sqrt{w_k}(\hat{\mu}_k - \mu_k) \rightarrow N(0, 1)$ for some constants w_k . Cochran's Q (1954) is defined by $Q = \sum_k \hat{w}_k (\hat{\mu}_k - \hat{\mu}_{\hat{w}})^2$, where $\bar{\mu}_w = \sum w_k \mu_k / \sum_k w_k$ is the weighted effect, and \hat{w}_k^{-1} is an estimator of the unknown asymptotic variance w_k^{-1} of $\hat{\mu}_k$. We restrict attention to situations where for each k there are n_k observations in the k th group and $w_k^{-1} = \sigma_k^2/n_k$ for a fixed, but usually unknown $\sigma_k^2 > 0$. If all $n_k \rightarrow \infty$ at the same rate, the limiting distribution of Q is the Chi-squared distribution with $\nu = K - 1$ degrees of freedom, so it is traditional to reject the hypothesis of homogeneity at level α when $Q \geq \chi_{\nu, 1-\alpha}^2$. However this procedure is unreliable for small sample sizes, see Table 1.

1.2 A test of homogeneity of effect sizes

Sometimes one can find a (variance stabilizing) transformation of the estimator $\hat{\mu}$, or a Studentized version of it, that is approximately normal with constant variance. For example, with the normal model $N(\mu, \sigma^2)$ and null hypothesis $\mu = \mu_0$ the effect size, (often called the standardized effect), $\delta = (\mu - \mu_0)/\sigma$ can be estimated by $\hat{\delta} = (\hat{\mu} - \mu_0)/\hat{\sigma} = t_{n-1}/\sqrt{n}$, where t_{n-1} is the Student- t statistic with $n - 1$ degrees of freedom. The transformation $\kappa = \kappa(\delta) = \sqrt{2} \sinh^{-1}(\delta/\sqrt{2})$ was shown by Azorin (1953) to stabilize the variance of $\hat{\kappa} = \kappa(\hat{\delta})$ and further to a good approximation $\hat{\kappa} \sim N(\kappa, 1/n)$. Suppose now there are K studies, each having an effect size δ_k of interest, and one wants to know if these effect sizes are equal or not. Cochran's Q statistic can be calculated for the $\hat{\kappa}_k$'s with known weights n_k : $Q^* = \sum_k n_k (\hat{\kappa}_k - \hat{\kappa}_n)^2$. This $Q^* \sim \chi_{K-1}^2$ under homogeneity, exactly, but only if the variance stabilization succeeds and the resulting distribution is normal; nevertheless it works approximately, see Table 1. Note that Q^* and Q do not test for the same type of homogeneity, the former testing equality of effects and the latter equality of effect sizes. But if the σ_k 's are equal, the homogeneity of the μ_k 's is equivalent to homogeneity of the δ_k 's.

2 Simulation studies

When one models the observations in the k th study by the normal distribution $N(\mu_k, \sigma_k^2)$, one estimates μ_k by the sample mean $\hat{\mu}_k = \bar{X}_k$ and w_k by $\hat{w}_k = n_k/s_k^2$, where s_k^2 is the sample variance. In order to compare Q with Q^* a modest empirical study based on 40,000 simulations of $K = 4$ normal samples of equal size was carried out. The results are displayed in the upper left half of Table 1. For example, when all sample sizes $n_k = 5$, the actual size of Q is close to 0.15 rather than the nominal 0.05; as n_k grows to 80, the empirical size gradually approaches 0.05. Similarly, the empirical mean \bar{Q} and standard deviation s_Q gradually approach the mean and standard deviation of the χ_3^2 distribution, respectively 3 and $\sqrt{6} = 2.45$. The performance of Q^* for the same samples reveals that it more quickly approaches the limiting χ_3^2 distribution. The experiment was repeated 3 times, each time using Q , Q^* as though they were based on data generated by a normal distribution, but now unknowingly, the data are generated by other distributions. Somewhat surprisingly for Student- t_3 generated data, both statistics performed slightly better than they did for normal data; but again Q^* outperforms Q , see Columns 6–8 of the upper half of the Table 1.

3 Random transformed effect size model

Assume the map $\delta \rightarrow \kappa(\delta)$ is strictly monotone increasing and continuous, and define the *random transformed effect size model* by assuming, for $\kappa_k = \kappa(\delta_k)$, that $\kappa_1, \dots, \kappa_K$ are a sample from the $N(\kappa, \gamma^2)$ model with unknown mean $\kappa = \kappa(\delta)$ and unknown variance γ^2 . Further assume $\hat{\kappa}_k = \kappa(\hat{\delta}_k)$ has a conditional distribution, given κ_k , which is $N(\kappa_k, 1/n_k)$. To obtain the unconditional properties of such estimators one must average over the distribution $N(\kappa, \gamma^2)$. By using the conditioning formulae for expectations and variances, one finds for all k

$$E[\hat{\kappa}_k] = \kappa \quad \text{and} \quad \text{Var}[\hat{\kappa}_k] = \gamma^2 + 1/n_k. \quad (1)$$

TABLE 1. The empirical size of the nominal level 0.05 Q -test under the hypothesis of homogeneity, as well as its mean and standard deviation. Also shown is the performance of Q^* based on the same simulated samples. The results in the upper half of the table in Columns 3–5 are generated from the $N(0, 1)$ model; while those in Columns 6–8 are generated from the Student t_3 model. The lower half of the table shows results when samples are generated from the double exponential model (Columns 3–5) and from an asymmetric model in which 80% of the sample is standard normal while 20% is a standardized χ^2_2 distribution. (Columns 6–8).

Statistic	n_k	size	\bar{Q}	s_Q	size	\bar{Q}	s_Q
Q	5	0.155	4.515	5.259	0.127	4.197	4.353
	10	0.095	3.575	3.350	0.078	3.468	2.920
	20	0.069	3.239	2.793	0.058	3.201	2.554
	40	0.059	3.132	2.641	0.052	3.099	2.473
	80	0.054	3.066	2.533	0.052	3.058	2.455
Q^*	5	0.115	3.810	3.979	0.089	3.541	3.434
	10	0.073	3.261	2.953	0.058	3.154	2.629
	20	0.058	3.086	2.612	0.049	3.048	2.429
	40	0.053	3.055	2.551	0.047	3.025	2.419
	80	0.051	3.027	2.488	0.050	3.022	2.430
Statistic	n_k	size	\bar{Q}	s_Q	size	\bar{Q}	s_Q
Q	5	0.128	4.164	4.200	0.150	4.476	5.270
	10	0.082	3.487	2.987	0.095	3.566	3.295
	20	0.063	3.223	2.668	0.070	3.250	2.817
	40	0.058	3.128	2.558	0.059	3.129	2.607
	80	0.053	3.045	2.503	0.055	3.055	2.522
Q^*	5	0.083	3.471	3.331	0.116	3.839	4.133
	10	0.060	3.160	2.630	0.076	3.279	2.970
	20	0.053	3.067	2.507	0.060	3.114	2.667
	40	0.053	3.050	2.474	0.054	3.061	2.536
	80	0.050	3.007	2.462	0.052	3.024	2.487

Let $\bar{\kappa} = (\sum_k \hat{\kappa}_k)/K$ and $s_{\kappa}^2 = \sum_k (\hat{\kappa}_k - \bar{\kappa})^2/(K - 1)$ denote the sample mean and variance of the $\hat{\kappa}_k$'s. Then s_{κ}^2/K is an unbiased estimate of the variance of $\bar{\kappa}$, and $E[s_{\kappa}^2] = \gamma^2 + (\sum_k 1/n_k)/K$, which leads to an estimate of γ^2 , namely $\hat{\gamma}^2 = \max\{0, s_{\kappa}^2 - (\sum_k 1/n_k)/K\}$.

3.1 Confidence intervals for κ and δ

If all n_k 's were equal, or if all $1/n_k$'s are negligible compared to γ^2 , then the $\hat{\kappa}_k$'s are just a sample from a normal population; hence a $1 - \alpha$ confidence interval for κ is the usual Student- t interval $\bar{\kappa} \pm c s_{\kappa}/\sqrt{K}$, where $c = t_{K-1, 1-\alpha/2}$. The small sample interval for δ is then $[\kappa^{-1}(L), \kappa^{-1}(U)]$.

3.2 Example: Drop in systolic blood pressure

The results of 6 studies showing the drop in systolic blood pressure following a weight reducing diet are summarized in Table 2; they are selected from the review by Mulrow *et. al* (2004). The standardized effects $\hat{\delta}_k$ and their transformed values $\hat{\kappa}_k$ are also tabled, where $\kappa = \kappa(\delta) = \sqrt{2} \sinh^{-1}(\delta/\sqrt{2})$. The $\hat{\kappa}_k$'s have weighted mean $\hat{\bar{\kappa}}_n = 0.6168$ and $Q^* = 58.67$ suggesting a random transformed effects model. Further, the sample mean $\bar{\kappa} = 0.6567$ and variance $s_{\kappa}^2 = (0.735)^2 = 0.54$, so all the $1/n_k$'s are small compared to $\hat{\gamma}^2 = 0.54 - 0.05 = 0.49$. Thus a 95% Student-*t* interval for κ is given by $[L, U] = [-0.115, 1.429]$; and for δ is $[\kappa^{-1}(L), \kappa^{-1}(U)] = [-0.115, 1.685]$.

TABLE 2. For each of six studies drop in systolic blood pressure for patients undergoing a weight-loss regime, summarized by n , \bar{y}_k , s_k .

Study k	n_k	\bar{y}_k	s_k	$\hat{\delta}_k$	$\hat{\kappa}_k$
Haynes (1984)	27	-4.8	13.8	-0.35	-0.34
MacMahon (1985)	20	13.3	8.1	1.64	1.40
Croft (1986)	66	11.0	17.1	0.64	0.62
Andersson (1991)	10	4.0	15.3	0.26	0.26
Jalkanen (1991)	24	8.0	20.4	0.39	0.39
Gordon (1997)	19	12.5	6.3	1.98	1.61

References

- Azorin, P.F. (1953). Sobre la distribución t no central I,II. *Estadística* **4**, 173–198 & 307–337.
- Mulrow, C.D., Chiquette, E., Angel, L., Cornell, J., Summerbell, C. et al. (2004). Dieting to reduce body weight for controlling hypertension in adults (Cochran Review). In *The Cochran Library*, Issues 3. Chichester: John Wiley & Sons.

Bayesian Density Estimation from Grouped Observations

Philippe Lambert¹ and Paul H.C. Eilers²

¹ Institut des sciences humaines et sociales, Université de Liège, Boulevard du Rectorat 7 (B31), B-4000 Liège (Belgium), p.lambert@ulg.ac.be

² Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University. P.O. Box 80140, 3508 TC Utrecht, The Netherlands.
p.h.c.eilers@uu.nl

Abstract: Grouped data occur frequently in observed data, either because of limited resolution of instruments, or because data have been summarized in relatively wide bins. Combining the composite link model with roughness penalties or equivalent priors, effective penalized likelihood and Bayesian procedures are presented to estimate smooth densities from such data.

Keywords: Coarsening, Langevin-Hastings algorithm, P-splines

1 Introduction

Descriptive statistics are easily computed when data are available with high precision. Unfortunately, this is not always the case. Measured concentrations may be below the detection limit, or an instrument may have poor resolution. It may also be the case that precise data have been summarized with a small number of wide intervals. An obvious quick-and-dirty solution replaces the observations in each interval by its midpoint and handles them as if they were actually observed. This can work quite well when computing the mean and the variance, but it fails for the estimation of quantiles. An improvement is to assume a parametric model for the underlying distribution, and fit it to the grouped data with the EM (estimation-maximization) algorithm. If an approximation to the distribution is available, one distributes the counts in each wide interval to pseudo-counts on a grid of narrow intervals, proportionally to the current approximation (the E step). The midpoints of the narrow intervals are used as “precise data”, with weights proportional to their pseudo-counts, to estimate distribution parameters by, say, maximum likelihood (the M step). This process is repeated till convergence.

When the observed coarse distribution has a simple shape, the parametric method can work quite well. But when it is skewed or multi-modal, or both, a lot of trial-and-error may be needed to find the right model. A non-parametric model will be more attractive then.

Histosplines are a popular non-parametric method. The idea is to compute a smooth spline under the condition that integrals over the given wide intervals are equal to the observed counts. Early work in this area was reported by Boneva et al. (1971).

Although the computations are not complicated, histosplines are not without disadvantages. There is no guaranty that the estimated distribution will be non-negative everywhere. Especially when there are wide intervals with zero observations next to intervals with positive counts, the spline, maintaining its smoothness, may undershoot the horizontal axis. A more fundamental objection is that the sampling variation in the observed counts is ignored. Generally one sees that a histospline shows a number of unrealistic wiggles.

We present here a non-parametric spline-based model: 1) the logarithm of a smooth latent distribution is modelled with P-splines, 2) expected counts are obtained from integrals over the wide intervals, and 3) observed counts are modelled with multinomial distribution, with the given expectations. The results is a penalized composite link model (Eilers, 2007).

We introduce a Bayesian variant of the model, allowing uncertainties to be quantified for model components and derived quantities, like quantiles. As is common for complex Bayesian models, analytic results are not obtainable, so we use a simulation algorithm. A combination of the Langevin-Hastings algorithm and rotation of the P-spline parameters allows fast computation, making the model a practical tool for everyday use.

2 The model

In the setting of grouped data with a multinomial distribution, the composite link model (CLM) of Thompson and Baker (1981) has the following structure. Let π be the latent distribution, defined on a grid of narrow intervals. We model these probabilities using polytomous logistic regression $\pi_j = e^{\eta_j} / (e^{\eta_1} + \dots + e^{\eta_J})$ with $\eta = B\phi$. We cannot observe π itself, but only sums over certain intervals. The expected values for these intervals are $\gamma = C\phi$, where C is an indicator (0/1) matrix, such that $c_{ij} = 1$ if narrow interval j contributes to wide interval i . In other words: row i of C shows which elements of π are added to form γ_i . The observed counts are in the vector \mathbf{y} drawn from a Multinomial distribution with cell probabilities γ . In summary:

$$(Y_1, \dots, Y_I) \sim \text{Mult}(y_+; \gamma_1, \dots, \gamma_I) \text{ with } \gamma = C\pi. \tag{1}$$

See Eilers (2007) for a more detailed description of the model and applications.

3 Inference

3.1 Estimation in a frequentist setting

Thompson and Baker (1981) showed how to estimate the parameters of the CLM in a frequentist setting. It boils down to a polytomous logistic regression of \mathbf{y} on a “working” matrix $X = W^{-1}CHB$, with $H = \text{diag}(y_+\pi(1-\pi))$ and $W = \text{diag}(y_+\gamma(1-\gamma))$. The well-known scoring algorithm then leads to iteration of

$$\check{X}'\check{W}\check{X}\check{\phi} = \check{X}'(y - y_+\gamma + \check{W}\check{X}\check{\phi}), \tag{2}$$

where the “breve” symbol, as in $\breve{\phi}$, indicates the current approximation. In the spirit of P-splines (Eilers and Marx, 1996), we put a roughness penalty on ϕ , to force the estimated distribution to be smooth. The penalized log-likelihood is

$$L^* = \sum y_i \log(\gamma_i) - \frac{\lambda}{2} |D\phi|^2, \quad (3)$$

where D is the r th order differencing matrix such that $D\phi = \Delta^r \phi$. The penalty modifies the scoring algorithm slightly:

$$(\breve{X}'\breve{W}\breve{X} + \lambda D'D)\phi = \breve{X}'(y - y_+ \gamma + \breve{W}\breve{X}\breve{\phi}), \quad (4)$$

3.2 Inference in a Bayesian setting

Following Lambert and Eilers (2005, 2006), the roughness penalty translates into a smoothness prior for the spline coefficients

$$p(\phi|\tau) \propto \tau^{\mathcal{R}(P)/2} \exp\left\{-\frac{\tau}{2} \phi'P\phi\right\} \quad (5)$$

where $P = D'D$ and $\mathcal{R}(P)$ is the rank of P . A gamma prior with large variance for τ is a possible choice, although more robust results can be obtained with a mixture of gammas (Jullion and Lambert, 2007).

Closed forms for the log-posterior of (ϕ, τ) and its gradient can be obtained. Markov chain Monte Carlo (McMC) methods can be used to draw a sample, $\{(\phi^{(m)}, \tau^{(m)}) : m = 1, \dots, M\}$, from the posterior. Here, we use the Langevin-Hastings algorithm on a carefully reparametrized posterior (Lambert and Eilers, 2006), $p(\psi, \tau|\mathbf{y})$, with $\phi = L\psi + \hat{\phi}_{\tau_0}$ where $\hat{\phi}_{\tau_0}$ and L denote the MLE of ϕ and the lower triangular matrix resulting from the Choleski decomposition of the asymptotic variance-covariance matrix of the MLE for a value of the penalty parameter $\lambda = \tau_0$ selected using the BIC criteria.

To each element of the sample, $\phi^{(m)}$, corresponds a density $f^{(m)}(y)$ for which any summary measure $\xi^{(m)}$ of interest such as the mean, the standard deviation or quantiles can be computed. Point estimates and credible intervals for ξ can be derived from the so-obtained sample $\{\xi^{(m)} : m = 1, \dots, M\}$.

Specific properties such as unimodality or log-concavity can be imposed on the estimated density by excluding, through the prior, the configurations of ϕ corresponding to non-desirable densities.

4 Simulation

A small simulation study was performed to assess the performances of the Bayesian CLM for varying degrees of coarsening when the mean, the standard deviation and selected quantiles are estimated using the obtained fitted density. The data were simulated using a gamma distribution with mean 5 and variance $\sigma^2 = 2.5$. The compact interval (0,15) was taken as the support of the target gamma distribution: this is a reasonable approximation as it contains 99.999% of the probability mass. That interval

was subdivided into $I = 64$ (small) bins of equal width ($= 0.15\sigma$). $J (= 4, 8, 16, 32, 64)$ wide bins of equal width ($= 2.37\sigma, 1.19\sigma, 0.59\sigma, 0.30\sigma, 0.15\sigma$) were obtained by grouping consecutive small bins, smaller values of J corresponding to coarser data. $S = 100$ datasets of size $n = 1,000$ were simulated. After a single burn-in of length 10,000, a chain of length $M = 2,000$ was built for each dataset to explore the posterior distribution of the spline parameters in the Bayesian CLM. The fitted density $f^{(s)}$ for the s th dataset corresponds to the MCMC estimate $\frac{1}{M} \sum_{m=1}^M \phi^{(m)}$ of the posterior mean of the spline parameters ϕ . It can be used to derive a point estimate for the mean, standard deviation and selected quantiles of the unknown density. The boxplot of these point estimates are given in the first two rows of Figure 1: good performances arise when $J \geq 8$. Interestingly, a larger precision is obtained than with the usual sample estimators computed independently from the raw (ungrouped) data. The third row of Figure 1 depicts the grouped data, the target gamma density and the fitted density for one of the $S = 100$ simulated datasets for decreasing degrees of coarsening.

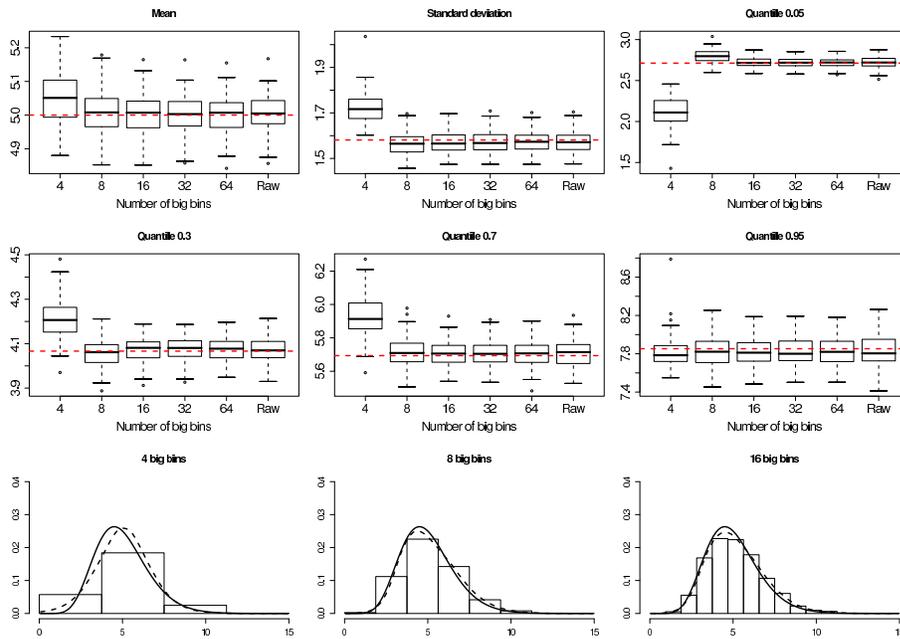


FIGURE 1. Rows 1-2: boxplot of the posterior point estimates for the mean, standard deviation and selected quantiles for the simulation; the dashed horizontal line indicates the true value and the 'Raw' entry gives the usual sample estimates computed from the raw (ungrouped) data. Row 3: the true density (solid line) and the fitted density (dashed line) for one of the $S = 100$ simulated datasets and a given number of wide bins.

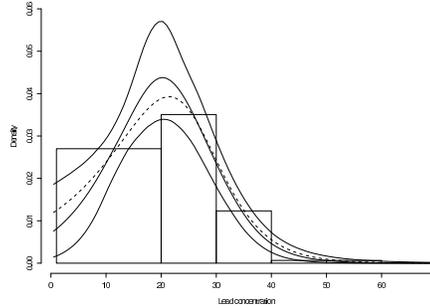


FIGURE 2. Histogram of the grouped data and the corresponding frequentist (dashed line) and Bayesian (thick solid line) estimates of the latent density. The thin solid lines delimit the (Bayesian) 90% pointwise credible intervals for the density.

5 Illustration

Consider the following grouped data corresponding concentrations of lead in the air in New York (Hasselblad et al., 1980):

	Lead concentration (in $\mu\text{g}/\text{dl}$)					
Wide interval	(1,20)	(20,30)	(30,40)	(40,50)	(50,60)	(60,70)
Freq. y_i	79	54	19	1	1	0

We propose to estimate the latent lead concentration density using the above described strategy. A grid of narrow intervals of width $\Delta = 1$ on $(0.5,70.5)$ is proposed. The probability to have a measurement in the j th narrow interval is $\pi_j = \int_{\text{bin } j} f(y) dy \approx f(u_j)\Delta$ where u_j denotes the interval midpoint.

A frequentist estimate of f can be derived from Section 3.1 with λ selected in the grid $\{2^k : k = 0, \dots, 20\}$ to minimize the BIC. When $\lambda = 4096$, this yields the dashed curve on Figure 2. The Bayesian estimate for the latent density (thick solid line) is also shown, with the 90% pointwise credible interval computed from a chain of length $M = 50,000$. Unimodality was forced in the Bayesian approach.

Point estimates and 90% credible intervals for the mean, the standard deviation and two quantiles were also derived:

	μ_Y	σ_Y	Quant. 0.20	Quant. 0.80
Posterior mean	20.3	9.6	11.4	27.8
90% credible intervals	(18.6,21.8)	(8.3,10.9)	(8.8,14.1)	(26.0,29.6)

This can be done for any function of the density.

References

Boneva, L.I., Kendall, D.G. and Stefano, I. (1971). Spline transformations. Three new diagnostic aids for the statistical data-analyst. *J. R. Statist. Soc. B* **33**(1), 1–71.

- Eilers, P.H.C. (2007) Ill-posed Problems with Counts, the Composite Link Model, and Penalized Likelihood. *Statistical Modelling*. To appear.
- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science* **11**(2), 89–121.
- Hasselblad, V., Stead, A. G. and Galke, W. (1980) Analysis of coarsely grouped data from the lognormal distribution. *JASA* **75**, 771–778.
- Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis* **51**, 2542–2558.
- Lambert, P. and Eilers, P.H.C. (2005). Bayesian proportional hazards model with time-varying coefficients: a penalized Poisson regression approach. *Statistics in Medicine* **24**, 3977–3989.
- Lambert, P. and Eilers, P.H.C. (2006). Bayesian multidimensional density smoothing. Proc. of the *21st International Workshop on Statistical Modelling*, 313–320, Galway.
- Thompson, R. and Baker, R.J. (1981). Composite link functions in generalized linear models. *Applied Statistics* **30**(2), 125–131.

Smoothing mixed models for overdispersed spatial count data

Dae-Jin Lee¹ and María Durbán¹

¹ Department of Statistics, Universidad Carlos III de Madrid, Spain,
dae-jin.lee@uc3m.es, mdurban@est-econ.uc3m.es

Abstract: We propose the use of Penalized splines (P -splines) in a mixed model framework to handle spatial effects and overdispersion in spatial count data. Spatial effects are modelled by a two-dimensional smooth function of latitude and longitude, and two approaches are presented to model overdispersion. We apply the methodology proposed to the analysis of cancer incidence rates

Keywords: Mixed Models; P -splines; Overdispersion; Spatial Count Data; *Scottish Lip Cancer Data*

1 Introduction

Penalized splines (Eilers and Marx, 1996) are a well established method for smoothing in generalized lineal models, in one or more dimensions. Currie et al. (2006) introduced a mixed model representation of array regression methods, which gave a fast and compact methodology for $2-d$ smoothing when data have an array structure. In this paper we use the mixed model approach to model overdispersed spatial counts when data are scattered instead of being in a regular grid.

Suppose we have Normal spatial data, (x_{1i}, x_{2i}, y_i) where \mathbf{x}_1 and \mathbf{x}_2 are latitude and longitude and \mathbf{y} is the response variable. A smooth model for the data would be given by:

$$\mathbf{y} = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (1)$$

where $\boldsymbol{\theta}$ is the vector of coefficients, and \mathbf{B} is a regression basis proposed by Eilers et al. (2006) and defined as the “row-wise” Kronecker product of the individual basis, $\mathbf{B} = \mathbf{B}_2 \square \mathbf{B}_1$, where each row of this matrix is the Kronecker product of the corresponding rows of \mathbf{B}_2 and \mathbf{B}_1 . The smoothness is imposed via the penalty matrix \mathbf{P} , based on second order difference matrices \mathbf{D} :

$$\mathbf{P} = \lambda_2 \mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_1} + \lambda_1 \mathbf{I}_{c_2} \otimes \mathbf{D}'_1 \mathbf{D}_1 \quad (2)$$

The aim is to set a new basis which allows the representation of (1) and its associated penalty into a mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (3)$$

where \mathbf{G} is a diagonal matrix which depends on the smoothing parameters λ_1 and λ_2 . Following a similar approach to Currie et al. (2006), and using the properties of

row-wise Kronecker product (Liu, 1999) it is easy to show that

$$\mathbf{X} = [\mathbf{X}_2 \square \mathbf{X}_1] \quad \text{and} \quad \mathbf{Z} = [\mathbf{Z}_2 \square \mathbf{X}_1 : \mathbf{X}_2 \square \mathbf{Z}_1 : \mathbf{Z}_2 \square \mathbf{Z}_1], \quad (4)$$

where $\mathbf{X}_2 = [\mathbf{1}_n : \mathbf{x}_2]$, $\mathbf{X}_1 = [\mathbf{1}_n : \mathbf{x}_1]$, $\mathbf{Z}_2 = \mathbf{B}_2 \mathbf{U}_{2s}$ and $\mathbf{Z}_1 = \mathbf{B}_1 \mathbf{U}_{1s}$, where \mathbf{U}_{2s} and \mathbf{U}_{1s} correspond to the non-zero eigenvalues of the singular value decomposition of $\mathbf{D}'_1 \mathbf{D}_1$ and $\mathbf{D}'_2 \mathbf{D}_2$. The estimation of the coefficients follow from standard mixed model theory:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \quad \text{and} \quad \hat{\boldsymbol{\alpha}} = \mathbf{G} \mathbf{Z}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}), \quad (5)$$

where $\mathbf{V} = \sigma^2 \mathbf{I} + \mathbf{Z} \mathbf{G} \mathbf{Z}'$. Smoothing parameters may be selected by maximizing the residual log-likelihood (REML). In the generalized linear mixed model (GLMM) case we use the penalized quasi-likelihood (PQL) of Breslow and Clayton (1993).

2 Modelling Overdispersion

Count data often present extra Poisson variation; this phenomenon is known as *overdispersion*. The most popular approach for analyzing overdispersed spatial count data takes account of the spatial heterogeneity by a conditional autoregressive (CAR) model and add individual random effects for overdispersion. We model the spatial variation by a smooth surface and take two approaches to account for overdispersion: (i) introduce individual random effects and (ii) use of negative binomial distribution. Using the mixed model representation of *P*-splines we can estimate efficiently the spatial effects and overdispersion.

2.1 The PRIDE approach

Perperoglou and Eilers (2005) gave an approach based on penalized likelihood, using individual random effects which adds extra parameters ($\boldsymbol{\gamma}$) to the linear predictor of a Poisson GLM (with log-link) for each observation.

$$\boldsymbol{\eta} = \mathbf{B} \boldsymbol{\theta} + \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \kappa^{-1} \mathbf{I})$$

This model is called PRIDE (“*Penalized Random Individual Dispersion Effects*”). The linear predictor can be reparameterized as a mixed model

$$\boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \boldsymbol{\alpha} + \boldsymbol{\gamma}, \quad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}), \quad \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \kappa^{-1} \mathbf{I})$$

Using PQL we obtain the following set of equations:

$$\begin{bmatrix} \mathbf{X}' \mathbf{W} \mathbf{X} & \mathbf{X}' \mathbf{W} \mathbf{Z} & \mathbf{X}' \mathbf{W} \\ \mathbf{Z}' \mathbf{W} \mathbf{X} & \mathbf{G}^{-1} + \mathbf{Z}' \mathbf{W} \mathbf{Z} & \mathbf{Z}' \mathbf{W} \\ \mathbf{W} \mathbf{X} & \mathbf{W} \mathbf{Z} & \kappa \mathbf{I} + \mathbf{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \mathbf{W} \mathbf{y} \\ \mathbf{Z}' \mathbf{W} \mathbf{y} \\ \mathbf{W} \mathbf{y} \end{bmatrix}, \quad (6)$$

where \mathbf{y} is the *working vector*, $\mathbf{y} = \boldsymbol{\eta} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, and \mathbf{W} is the diagonal matrix of weights $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \exp(\mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \boldsymbol{\alpha} + \boldsymbol{\gamma})$. This yields a weighted linear Gaussian model with error variance $\sigma^2 = 1$, the coefficients are estimated as in (5). Then, conditional on the estimates of the coefficients, we estimate λ_1 , λ_2 and κ .

2.2 GLMM for Negative Binomial distribution

The negative binomial distribution is derived by letting the mean of the Poisson distribution vary according to a fixed parameter given by the Gamma distribution. The marginal distribution of \mathbf{y}_i is negative binomial with mean μ_i and variance $(\mu_i + \mu_i^2)/\kappa$, where κ is a dispersion or shape parameter. The fact that the negative binomial has two parameters and is not in the exponential family, makes it more difficult to extend the methodology developed for Poisson data. However, it can be shown that the estimation of fixed and random effects, and variance components is done as in the case of Poisson data but the matrix of weights is given by

$$\mathbf{W} = \kappa \operatorname{diag} \left(\frac{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})}{\kappa \mathbf{1} + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha})} \right)$$

Note that the weights are very similar to the ones in the PRIDE approach, but in PRIDE model, we can estimate the individual random effect $\boldsymbol{\gamma}$.

3 Application to Scottish Lip Cancer data

The data set consists on the observed (\boldsymbol{o}), and expected (\boldsymbol{e}) number of cases of lip cancer registered in 56 counties in Scotland, together with the proportion of the population in each county working in agriculture, forestry or fishing industries (\mathbf{aff}). In our approach, we include latitude (\mathbf{lat}) and longitude (\mathbf{lon}) of the centroid of each county as covariates and use a 2- d P -spline in order to account for the spatial variation. The basic spatial model for these data is:

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = \log(\boldsymbol{e}) + \mathbf{aff} + f(\mathbf{lat}, \mathbf{lon})$$

where f is estimated by a two-dimensional surface. Three models were fitted to the data: a Poisson model with spatial trend ignoring overdispersion (as above), and the two models proposed in the paper. The models were compared in terms of the Akaike Information Criteria and the Bayesian Information Criteria. The PRIDE model was chosen by both criteria. A CAR model was also fitted but the effective dimension of this model was much larger than the model proposed. Figure 1 shows the fitted values (in the scale of the linear predictor) for the model chosen.

Concluding Remarks

We have presented a new approach to the analysis of spatial count data in the presence of overdispersion. The spatial variation is accounted for by a 2- d smooth function of the spatial coordinates, and two methods are proposed to deal with the overdispersion. The mixed model representation of multidimensional P -splines for scattered data yields a fully parametric model which can be easily implemented in standard software.

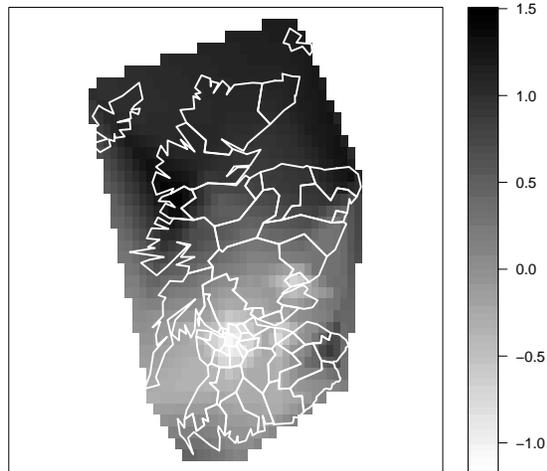


FIGURE 1. Plot of fitted spatial trend plus individual random effects $f(\text{lat}, \text{lon}) + \gamma$ for PRIDE model

References

- Breslow, N. E. and Clayton, D. G. (1993). Approximated Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* **88**(421), 9-25.
- Currie, I. D., Durbán, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B* **68**, 1-22.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B -Splines and Penalties. *Statistical Science* **11**, 89-121.
- Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis* **50**(1), 61-76.
- Liu, S. (1999). Matrix results on the Khatri-Rao and Tracy-Singh products. *Linear Algebra and its Applications* **289**, 267-277.
- Perperoglou, A. and Eilers, P. H. C. (2005). Overdispersion modelling with individual random effects and penalized likelihood. *Submitted*.

Assessing Surrogacy in the Counterfactual Framework Using Bayesian Models

Yun Li¹, Jeremy M.G. Taylor¹ and Michael R. Elliott¹

¹ Department of Biostatistics, University of Michigan, yunlisph@umich.edu

Abstract: The usage of a valid surrogate in predicting treatment effect can substantially reduce trial duration and cost. To assess the degree to which the treatment effect occurs through a surrogate marker, Frangakis and Rubin (2002) proposed to evaluate surrogacy through the association between potential outcomes of the true endpoint and the principal strata defined based on potential outcomes of the surrogate marker. In contrast to the Prentice and Freedman criteria, these surrogacy measures always have causal interpretation. We carry out the evaluation using Bayesian methods in situations when the surrogate marker, true endpoint and treatment assignment are binary. To address nonidentifiability, we incorporate assumptions that are plausible in the surrogate context into prior distributions. These measures are used to predict the causal treatment effect. We extend the approach to multiple trial settings using hierarchical modelling.

Keywords: Causal Models; Principal Stratification; Bayesian; Clinical Trials; Surrogate.

1 Introduction

Surrogate markers (S) in randomized clinical trials are often intermediate physical or laboratory indicators in a disease progression process that can be measured earlier than the true endpoint (T). When T is rare, later-occurring or costly to obtain, the usage of a valid surrogate in extracting information on the treatment (Z) effect on T can substantially reduce trial duration and size, lower the expense and lead to earlier decision making. Examples of potential surrogate markers include CD4 counts and viral load for HIV infection, blood pressure and serum cholesterol level for cardiovascular disease and prostate-specific antigen for prostate cancer.

Prentice (1998) proposed a formal definition of perfect surrogacy that requires the changes in S fully capture the effect of the treatment on T. Less stringent measures for surrogacy using the proportion of the treatment effect explained by S have been proposed by Freedman (1992). These surrogacy measures and the prediction of the treatment effect on T using S often utilize models for the distribution of $T|S, Z$. However, since S is a post-treatment variable, the measures and the predicted treatment effect do not have the causal interpretation defined by counterfactual models.

An alternative approach is to measure surrogacy in a causal inference framework. The general idea of causal inference hypothesizes the setting wherein each individual has two potential outcomes, corresponding to the two possible treatment regimes (e.g., $Z = 1$ for treatment and $Z = 0$ for placebo). In reality, we only observe one of the outcomes since either the treatment or placebo (not both) is assigned to a patient. The

TABLE 1. Surrogacy Measures: Probabilities from Counterfactual Models.

Principal Stratum	Potential Outcomes	$(T(0), T(1))$		
	$(s(0), s(1))$	$(0, 0)$	$(0, 1)$	$(1, 1)$
1	$(0, 0)$	p_{11}	p_{12}	p_{13}
2	$(0, 1)$	p_{21}	p_{22}	p_{23}
3	$(1, 1)$	p_{31}	p_{32}	p_{33}

causal treatment effect would be the comparison between these two potential outcomes for the same set of the individuals. The framework has been used to model noncompliance. An effective surrogate occurs along the causal pathway between the treatment and the true endpoint. To evaluate the degree to which the effect of treatment on T occurs only if the treatment has affected S, Frangakis and Rubin (2002) proposed to measure surrogacy of S through the association between the potential outcomes of T and the principal strata defined by potential outcomes of S. This approach has been investigated by Taylor et al (2005) to study the causal interpretation of the Freedman's criteria. When both S and T are binary, the potential outcomes for S and T are denoted by $(S(Z) = 0, 1)$ and $(T(Z) = 0, 1)$ with respect to Z . Table 1 displays the probabilities as surrogacy measures associated with the cross-tabulations of principal strata and potential outcomes. They partition the population into 9 groups. In contrast to the Prentice and Freedman criteria, these surrogacy measures and the causal effect within each stratum always have a causal interpretation because the comparisons of the potential outcomes are made on the same set of subjects. As far as we know, no literature has provided practical implementation of the evaluation and applied to the real data. In this paper, first, we use Bayesian methods to carry out estimation in a single trial setting and then prediction of the treatment effect using S. We further extend the methods to a multiple-trial setting. The methods are applied to the data from the collaborative initial glaucoma treatment study (CIGTS).

2 Bayesian Estimation

Since we only observe one of the potential outcomes, the counterfactual model contains more parameters than the number of independent observations, some assumptions are required for the model to be identifiable. We first exclude the values $(S(0) = 1, S(1) = 0)$ and $(T(0) = 1, T(1) = 0)$ because when the treatment is effective, neither are scientifically sensible. Based on the biological mechanism, we believe that the values of the potential outcomes $(S(0), S(1))$ of a potential good surrogate are more likely to agree with those of $(T(0), T(1))$. Thus, in terms of the probabilities in Table 1, $p_{11} > p_{12} > p_{13}$, $p_{22} > p_{21}, p_{23}$, and $p_{33} > p_{32} > p_{31}$. Via prior distributions, we encourage the ordering restriction of these probability parameters. We treat the unobserved potential outcomes as missing data and estimate them via imputation. Multinomial distributions with Dirichlet priors are assumed for the complete data. The approach extends naturally to multiple trials. Prediction of the causal treatment effect using S when T is partially observed can be obtained using these estimated

probabilities and integrating over all possible unobserved potential outcomes of S.

3 Application to the Glaucoma data

The methods are illustrated using the data from CIGTS study. It is a randomized multi-center trial conducted to compare the effects of two treatments, surgery ($Z = 1$) and medicine ($Z = 0$), on reducing the intraocular pressure among glaucoma patients. Glaucoma is a leading cause for blindness and the intraocular pressure (IOP) is a major risk factor. The true endpoint is the IOP measurement at the 96th month and we use the IOP measure at the 12th month as the surrogate marker. Both S and T are defined as 1 if the IOP values are less than 17.5mmHg and as 0 if otherwise.

4 Results and Discussion

We combine all the glaucoma data where both S and T are observed to illustrate our methods and the data summary is in Table 2. The probabilities from Table 1 are estimated and summarized in Table 3. All probabilities are quantities of interest and help interpret how the treatment affects the surrogate and the true endpoint and their relationships. For example, p_{22} is defined as the associative effect because the effect on T is accompanied by the change on S; conversely, p_{12} and p_{32} are dissociative effects. The ratio $p_{22}/(p_{12} + p_{22} + p_{32})$ measures the extent of causal treatment effect explained by the surrogate and is estimated as 0.29 with its 95% credible interval (CI) (0.11, 0.52). These measures are used to predict the causal treatment effect from the surrogate. We conduct sensitivity analysis to examine the impact of the prior assumptions. Due to non-identifiability, we are unable to identify the causal parameters using maximum likelihood, but Bayesian approach incorporating prior belief allows us to obtain relatively narrower ranges of posterior distributions. However, even with the incorporation of the prior knowledge, the posterior distributions of some estimates obtained in a single trial setting are still relatively flat and not informative about the relative likelihood of values within the ranges. When multiple trials are available, with the hierarchical modelling approach, we make distributional assumptions and allow the sharing of information on the properties of the surrogate across trials, much more precise and informative inferences are obtained with even less informative prior parameters. In practice, we collect data in a scientific context and with a considerable amount of a priori knowledge. This research explores the possibility of utilizing biological assumptions in the surrogate context to increase the precision of the estimated causal parameters.

TABLE 2. Number of Patients and Observed Proportions by Treatment Groups and IOP Values at the 12th and 96th Months from the Glaucoma Data.

		T		
		S	0	1
$Z = 0$	0	71(56.3%)	16(12.7%)	
	1	32(25.4%)	7(5.6%)	
$Z = 1$	0	24(23.5%)	21(20.6%)	
	1	14(13.7%)	43(42.2%)	

TABLE 3. Summary Statistics for Probabilities Defined by Principal Strata and Potential Outcomes based on their Posterior Distributions for the Glaucoma data.

	Mean	95%CI		Mean	95%CI		Mean	95%CI
p_{11}	0.24	(0.16, 0.32)	p_{12}	0.14	(0.056, 0.23)	p_{13}	0.065	(0.0098, 0.13)
p_{21}	0.061	(0.0096, 0.14)	p_{22}	0.12	(0.040, 0.23)	p_{23}	0.061	(0.011, 0.13)
p_{31}	0.075	(0.012, 0.15)	p_{32}	0.16	(0.072, 0.25)	p_{33}	0.073	(0.037, 0.12)

References

Frangakis, C.E., and Rubin, D.B. (2002). Principal stratification in casual inference. *Biometrics* **58**, 2129.

Freedman, L.S., Graubard, B.I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic disease. *Statistics in Medicine* **11**, 167178.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials, definition and operational criteria. *Statistics in Medicine* **8**, 431-440.

Taylor, J.M.G., Wang, Y., and Thiébaud, R. (2005). Counterfactual links to the proportion of treatment effect explained by a surrogate marker. *Biometrics* **61**, 1102-1111.

Analysis of Breast Cancer Survival in Local Health Authorities

Joseph Lynch ¹ and Gilbert MacKenzie ¹

¹ Centre of Biostatistics, University of Limerick, Ireland

Abstract: Kaplan-Meier analysis of a large breast cancer dataset is carried out under all-cause and cause-specific survival. The results are compared with a variety of model-based analyses, including Cox's Proportional Hazard (PH) model and its Gamma frailty variants, along with the non-PH Generalised Time-Dependent Logistic Model (GTDL) and its Gamma frailty variants.

Keywords: Kaplan-Meier; cause-specific; all-cause; PH/non-PH; Gamma frailty

1 Introduction

Coleman (1999) reported that North Staffordshire Local Health Authority (LHA) was ranked last of 99 LHAs in England & Wales with respect to breast cancer survival. His report was based on a relative survival approach which did not take account of *case-mix* factors and he analysed incident cases diagnosed between 1991-1993. We re-analyse an augmented dataset from the West Midlands of England, including North Staffordshire, by more traditional methods and report on the resulting *case-mix* adjusted league table.

The population data analysed comprise 15,516 incident cases of cancer of the female breast diagnosed in the West Midlands, UK, between 1991-1995 and followed-up to the end of 2001. Survival time is defined as the time in years from diagnosis to death or censoring. Both cause-specific and all-cause definitions are used. The cause-specific definition refers to deaths in which breast cancer is registered as the primary cause - other outcomes being regarded as censored (at their time of occurrence). In contrast the all-cause refers to death from all causes including breast cancer.

2 Model Definitions

We consider several models of increasing complexity; the proportional hazard (PH) model of Cox (1972), the Gamma frailty variant (Hougaard, 1994), the Generalised Time-Dependent Logistic Model (MacKenzie 1996, 1997) and the Gamma frailty variant discussed by Blagojevic, MacKenzie & Ha (2003). The models are defined in order by:

$$\begin{aligned}\lambda(t|x) &= \lambda_0(t)\exp(x\beta) \\ \lambda(t|u, x) &= u\lambda_0(t)\exp(x\beta)\end{aligned}\tag{1}$$

$$\begin{aligned}\lambda(t|u, x, z) &= \lambda_0(t)\exp(x\beta + zu) \\ \lambda(t|x) &= \lambda p(t|x) \\ \lambda(t|u, x) &= u\lambda p(t|x)\end{aligned}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, β is a $p \times 1$ vector of regression parameters associated with fixed covariates, x and $U \sim \text{Gamma}$ with $E(U) = 1$ and $V(U) = \sigma^2$. Because some of the covariates studied do not obey the PH assumption we also adopted the non-PH GTDL model with $p(t_i|x_i) = \exp(t_i\alpha + x'_i\beta) / 1 + \exp(t_i\alpha + x'_i\beta)$ for $i = 1, \dots, n$ subjects. In the third model, u is a $q \times 1$ vector of coefficients associated with the q Health Authorities entered as random covariates (z).

3 Results

Below we report the major findings from the PH model with which the Epidemiologists involved in the study are most familiar. Further results from the more sophisticated models will be presented in the poster.

The data were obtained from the West Midlands Cancer Intelligence Unit (the Cancer registry) in Birmingham, UK. There are 10 major covariates of interest including: age, diagnosis basis, stage, grade, morphology, whether or not the cancer was screen-detected, Townsend score (a measure of deprivation), year of diagnosis, Local Health Authority (14 including one unknown category) and treatment. All covariates were categorical.

Figure 1 shows the overall KM survival rates, the upper curve refers to cause-specific survival (mean = 8.33 years) and the lower curve to all-cause survival (mean = 7.43 years), so that (typically) all-cause mortality is higher, leading to shorter survival.

From the KM analysis, using LHA as a factor, we were able, somewhat surprisingly, to produce a result analogous to Coleman's in this extended data set, which had more cases recruited and a longer follow-up period. North Staffordshire was again bottom of the league - this time the West Midland's league. As with Coleman's analysis, no covariates were used at this stage.

In the Cox analysis, all ten of the covariates were statistically significant. The major factors influencing survival were stage treatment, grade, age and tumour morphology. We addressed the underlying variable selection issue by generating 100 bootstrap samples from the original data and re-fitting the model using a variety of stepwise algorithms and cut-off points including different values of $t = \hat{\beta}/\text{se}(\hat{\beta})$ and various Odds Ratios. The stability of variable selection was re-assuringly good in the cause-specific analysis. Based on the bootstrap analysis, we elected to adjust the LHA rates for the remaining 9 covariates.

Table 1 shows the resulting 5-year survival league tables for all-cause (unadjusted and adjusted.). There is wide variation between the unadjusted and adjusted ranking results. However these findings should be interpreted cautiously as: (a) the quantitative differences are small and (b) when North Staffordshire is regarded as the reference category in the regression analysis only Coventry and Birmingham have significantly better survival.

Thus we see that Coleman's claim - that North Staffordshire is at the bottom of the League is not quite justified when key *case-mix* factors are taken into account.

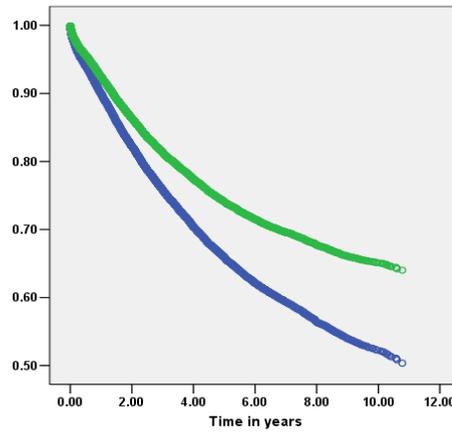


FIGURE 1. KM Breast Cancer Survival: cause specific=upper curve, all cause=lower curve

TABLE 1. All-cause 5-year Survival League tables

LHA	$\hat{S}_{KM}(t = 5)$	LHA	$\hat{S}_{PH}^*(t = 5 \bar{x})$	LHA	$\hat{S}_{PH}^{**}(t = 5 \bar{x})$
Solihull	0.71	Unkn	0.783	Unkn	0.779
Worcester	0.68	Birm	0.763	Cov	0.765
Hereford	0.68	Her	0.753	Birm	0.762
Shropshire	0.68	Cov	0.752	Her	0.752
Warwick	0.68	War	0.732	War	0.735
Wolver	0.67	Sand	0.727	Wolv	0.735
Coventry	0.67	Shrop	0.727	Shrop	0.723
SStaff	0.65	Sol	0.727	Sand	0.722
Walsall	0.65	Wolv	0.722	Dud	0.720
Unknown	0.65	Worc	0.716	Sol	0.717
Birm	0.65	Dud	0.706	Wal	0.714
Dudley	0.65	Wal	0.706	NStaff	0.711
Sandwell	0.63	NStaff	0.696	Worc	0.711
NStaff	0.58	SStaff	0.686	SStaff	0.689

* $\hat{S}_{PH}^{**}(t = 5|\bar{x})$ is the stratified Cox model employing a separate baseline hazard function for each Health Authority and \bar{x} is the West Midlands mean.

However, clearly, breast cancer survival in North Staffordshire was not optimal over the study period.

4 Discussion

This is a large study and we have only just begun to scratch the surface. The preliminary findings suggest that covariate adjustment is required for valid interpretation. However, here, the PH model is merely an approximation to the truth since in the course of the analysis we found that several important covariates did not obey the PH assumption. There is some evidence of Gaussian frailty, but not much evidence of gamma frailty. Fitting frailty models in R is a slow process and the standard errors of the frailty variance are not produced in standard R output. In the poster we shall compare the findings obtained by fitting the non-PH models, discuss the value of frailty in this context and comment on the process of obtaining adjusted survival curves.

References

- Blagojevic M., MacKenzie G. and Ha I.D. (2003) A Comparison of non-PH & PH - Gamma frailty models. *IWSM 2003*, 39-43.
- Coleman, M.P., de Stavola, B., Harris, S., Sloggett, A. Quinn, M. and Babb P. (1999) Cancer Survival in the Health Authorities of England up to 1998; A report prepared for the National Health Services Executive .
- Cox DR (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34**, 187-220.
- Hougaard, P. (1994). Heterogeneity Models of Disease Susceptibility, with Applications to Diabetic Nephropathy. populations. *Biometrics* **50**, 1178-1188.
- MacKenzie, G. (1996) Regression models for survival data: the generalised time dependent logistic family. *JRSS Series D* **45**, 21-34.
- MacKenzie, G. (1997) On a non-proportional hazards regression model for repeated medical random counts. *Statistics in Medicine* **16**, 1831-1843.

An analysis of deprivation in Portugal based on Bayesian latent class models

Carla Machado¹, Carlos Daniel Paulino² and Francisco Nunes³

¹ Direcção-Geral de Estudos, Estatística e Planeamento, Ministério do Trabalho e da Solidariedade Social, Lisboa

² Departamento de Matemática e Centro de Matemática e Aplicações, Instituto Superior Técnico, Universidade Técnica de Lisboa

³ Departamento de Economia, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa

Abstract: The aim of this work is to analyse household deprivation in Portugal during the second half of the 1990s. The concept of household deprivation is based on the multidimensional aspects of well-being and as such, from the *European Community Household Panel* data, we make a previous selection of deprivation categorical indicators clustered in the following well-being dimensions: housing, durable goods, economic strain and social relationships. In a first stage, we adopt a Bayesian latent class model (LCM) for each dimension in order to identify groups with a differential degree of deprivation and each household's deprivation profile. From a generation of values for the corresponding four latent variables for the sample households, we then apply a new latent class model with the purpose of obtaining an estimated risk of overall deprivation and the most vulnerable dimensions of household's living standard. The analyses of the several LCM are carried out via Markov Chain Monte Carlo methods.

Keywords: Deprivation; Latent Class Model; Bayesian methods; Label-switching.

1 Introduction

In the framework of the new European Social Agenda, social inclusion policies are essential to prevent and combat the multiple facets of poverty. The success of the renewed European strategy requires an improved and more up-to-date knowledge of living conditions. In Portugal, the research on poverty measurement has been limited when compared with other areas of sociology and economy. Although poverty is often referred to as being a multidimensional phenomenon, the most common approach focuses on the monetary dimension of poverty, and so poverty has been assessed using income-based measures. However, using income as a proxy for analysing household's well-being may misrepresent what is actually available to a household for the purpose of meeting its needs. The approach followed in this article focuses exactly on including multidimensional measures into poverty measurement, providing a contribution to an innovative analysis of poverty in Portugal. Following Townsend (1979) we use a relative deprivation concept focused on the household's lack of needs (material and immaterial) that results from a lack of their resources, i.e., the available resources are below the

minimum acceptable standards of living which are defined in relation to social norms or expectations.

Information on household deprivation is based on national data from the *European Community Household Panel* (ECHP) carried out by the Portuguese National Statistical Institute. The ECHP is a harmonised longitudinal survey focusing on income, social conditions and lifestyle (Eurostat, 2003). It was implemented in every European Union Member State under Eurostat co-ordination and nowadays provides the official data to analyse poverty. In this first approach to the problem we make a static analysis of deprivation by examining the second and eighth waves (that correspond to the years of 1995 and 2001). This micro data set contains information on living conditions where the main unit of observation is the household (4614 households).

2 Methodology of analysis

The multiple dimensions of well-being identified in the data justify a methodology based on a few steps to analyse household's deprivation. The starting point was to select a set of deprivation categorical indicators from the ECHP data, which are clustered into the following dimensions of well-being: housing, durable goods, economic strain and social relationships. In this point, it is important to select the indicators of well-being that reflect the household's vulnerability in the society where they live; to use the same unit of observation (household) for all variables and to pick up the sample units provided with full information; to make sure that all indicators are categorized; and to conduct an appropriate descriptive analysis of the variables.

The second step consisted of making a separate analysis by well-being dimensions in order to construct a composite deprivation indicator for each dimension. To this purpose, we considered a latent class model that describes the association among the manifest variables through a conditional independence structure on an augmented contingency table with a categorical latent variable (see, e.g., Bartholomew and Knott, 1999; Paulino and Singer, 2006). This latent variable intends to identify a deprivation indicator for each well-being dimension represented by the observed variables of the corresponding table. This analysis allowed us to obtain an estimated risk of deprivation (*i.e.* the probability of the most deprived class) and identify household's profiles with a greater or lesser deprivation per dimension.

At last, the third step provided an overall approach to the deprivation problem by bringing together the partial outcomes of the previous step. This is achieved by applying a latent class model to "data" formed from the values for the latent variables associated with the four dimensions that are generated by a sensible criterion from the estimated conditional probabilities of the latent classes given every household's observed profile. This analysis allowed us to create an indicator of overall deprivation relied upon three levels, non-deprived, partly deprived and fully deprived, as well as to identify the most vulnerable dimensions of household's living standard.

The underlying statistical model on which the LCM structure is imposed is a multinomial model. In contrast to most analyses of the deprivation problem (e.g., Moisisio, 2005; Pérez-Mayo, 2005), Bayesian methods were used for drawing inferences about the parameters of interest, due to parametric complexity of the model. Given that information on the LCM parameters elicitable from experts in the area of application

was not clear-cut enough, we used a non-informative Dirichlet prior distribution for them (Paulino *et al.*, 2003). The analysis of the model proceeded via Markov Chain Monte Carlo (MCMC) methods.

Due to the size of the observed data sets and the need of achieving interpretability criteria, we opted by an analytical scheme based on specifying a fixed number of latent classes (K). The models for several values of K were compared in terms of goodness of fit, complexity and predictive ability measured by DIC, BIC and Pseudo-Bayes factors. Examination of the results obtained by these criteria, together with the ease of interpretation of the latent classes, allowed us to select one of them. One of the main challenges in LCM applications is the nonidentifiability related to the invariance of the likelihood under relabelling of the model parameters - the so-called label switching problem (Stephens, 2000; Jasra *et al.*, 2005). To overcome this problem, after a judicious choice of the observed variables in the first step, we applied either simple inequality constraints on the parameter space to remove the symmetry in the likelihood or a relabelling algorithm (Kullback-Leibler algorithm) proposed by Stephens (2000).

3 Brief Conclusions

This approach identified four classes for the deprivation latent indicator on housing and economic strain dimensions, and three classes for the deprivation indicator on durable goods and social relationship dimensions. Overall, we found three relevant classes for the analysis of global deprivation. It is also interesting to note that we achieved a sharp distinction among the situations of non-deprived, partly deprived and fully deprived in all applications of Latent Class modelling. Due to space limitations we only present some primary outcomes in terms of posterior point estimates.

Over the period analysed, the posterior mean of overall deprivation risk decreased substantially. The third step of the applied methodology showed that the estimated risk of deprivation decreased from 27% (1995) to 13% (2001). The intermediate situation of partial deprivation for both years encompassed half of the households in posterior belief terms, and hence the posterior estimate of the non-deprived class probability was higher in 2001 (39%) when compared to 26% in 1995.

With reference to analyses per dimension, the estimated risk of deprivation for both years is higher for the economic strain dimension, even though it had strongly decreased from 1995 (47%) to 2001 (30%). Although the posterior estimate of the deprivation risk in social relations appeared to be nearly constant, this is the second dimension in which it is higher. The probability of the most deprived class was estimated *a posteriori* to be around 12%. In turn, the posterior mean of deprivation risk in durable goods decreased largely over the period of analysis (18% to 9%). Regarding housing dimension, the corresponding deprivation risk estimate was 8% (2001) and 10% (1995). At last, it is worth mentioning that this static analysis deserves to be supplemented with a longitudinal analysis of the data for the available waves. This study on the dynamic evolution of the deprivation in Portugal is currently in progress.

References

- Bartholomew, D.J. and Knott, M. (1999). *Latent variable models and factor analysis*. 2nd ed. London: Arnold.
- Eurostat (2003). *ECHP UDB Description of variables*. Doc. PAN 166/03. Luxembourg: European Commission.
- Jasra, A., Holmes, C. and Stephens, D.A. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science* **20**(1), 50-67.
- Paulino, C.D., Turkman, A. and Murteira, B. (2003). *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian.
- Paulino, C.D. and Singer, J.M. (2006). *Análise de dados categorizados*. São Paulo: Edgard Blücher.
- Moisio, P. (2005). *A latent class application to the measurement of poverty*. IRISS-C/I Working paper. Florence: European University Institute.
- Pérez-Mayo, J. (2005). Identifying deprivation profiles in Spain: a new approach. *Applied Economics* **37**, 943-955.
- Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society: Series B (Methodological)* **62**(4), 795-809.
- Townsend, P. (1979). *Poverty in the United Kingdom*. Middlesex: Penguin Books.

Modelling Survival Data with Crossing Hazards

Gilbert MacKenzie¹ and Il Do Ha²

¹ Centre of Biostatistics, Department of Mathematics, University of Limerick, Ireland

² School of Information & Management, Daegu Hanaay University, South Korea

Abstract: We revisit the crossing hazards problem in survival analysis and compare the use of Cox's semi-parametric model with a parametric non-PH model from the generalised time-dependent logistic family (GTDL). A set of gastric cancer data is analysed and a GTDL gamma-frailty model is shown to explain the observed data well. The role of heterogeneity in the crossing hazards problem is discussed.

Keywords: Crossing Hazards, GTDL family, non-PH survival modelling

1 Introduction

Despite the ubiquity of Cox's proportional hazards (PH) model it is being realised increasingly that not all survival data obey the PH assumption. In multi-factor studies the effect of one or more covariates may be noticeably non-PH. A clear signal is that of crossing hazards. A classical example is the well-known data set of the Gastrointestinal Tumor Study Group (GTSG)(1982), reporting the effects of chemotherapy and combined chemotherapy and radiotherapy on the survival times of gastric cancer patients (Figure 1). The question then arises as to how best to model these effects. Sometimes, in practice, non-PH covariates are ignored and they are analysed as being PH in a larger model, but the optimality of this expediency is unclear.

An alternative approach is to adopt a model which can cope with non-PH and PH effects. The Generalised Time-Dependent Logistic family of survival models contains two non-PH parametric models which are potential competitors for Cox's model, namely, the GTDL model (MacKenzie, 1996) and the logistic accelerated life model, the LAL (Al-tawarah & MacKenzie, 2003). Recently, the family has been extended to incorporate frailty (Blagojevic, MacKenzie & Ha, 2003) and to more general multivariate forms (Blagojevic & MacKenzie, 2007).

In relation to tests and models developed specifically for crossing hazards situations *per se* we refer the reader to Stablein & Koutrouvelis (1985), Aalen(1994), Hseish (2001) and Bagdonavicius *et al* (2005).

2 Models & Interpretations

Here we take a rather simpler approach when comparing some PH and non-PH models in the GTDL family in the analysis of crossing hazards data. In particular, we consider

fitting the following set of models to the gastric cancer data:

$$\lambda(t|x) = \lambda_0(t)\exp(x\beta) \tag{1}$$

$$\lambda(t|u, x) = u\lambda_0(t)\exp(x\beta) \tag{2}$$

$$\lambda(t|x) = \lambda p(t|x) \tag{3}$$

$$\lambda(t|u, x) = u\lambda p(t|x) \tag{4}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function, β is a $p \times 1$ vector of regression parameters associated with fixed covariates, x , $\lambda > 0$ is a scalar, $U \sim \text{Gamma}$ with $E(U) = 1$, $V(U) = \sigma^2$ and $p(t|x) = \exp(t\alpha + x'\beta) / (1 + \exp(t\alpha + x'\beta))$.

2.1 Interpreting $\lambda_0(t)$

Consider the two group case with a single binary covariate, x . First we note that in the PH model (1), $\lambda_0(t)$ emerges when $x = 0$, as the baseline hazard function. However, it also emerges when $\beta = 0$, whence there is no PH regression. Then the subscript '0' is redundant and $\lambda_0(t)$ should be denoted $\lambda(t)$, since the hazard is now arbitrary. When $\beta \neq 0$ we may proceed to estimate $\lambda_0(t)$, eg, via Breslow's method, and compare the resulting (marginal) survival function with the KM estimator as a check on the goodness of fit of the model. However, when we are dealing with one covariate indexing two-groups, the comparison with KM is not available when $\beta = 0$.

By contrast, the interpretation of parametric models is much clearer, as when $x = 0$ or $\beta = 0$ the hazard typically reduces to a specific function of time eg, $\lambda \exp(t\alpha) / (1 + \exp(t\alpha))$ in model (3) above, and the corresponding parametric survival function can always be tested against the KM estimator as a check on fit.

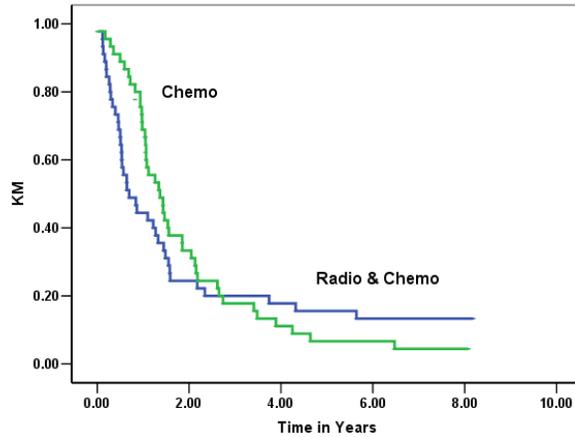


FIGURE 1. KM Survival Functions in Treatment Groups

3 Results

We re-analysed the survival times of the 90 patients with gastric cancer. Figure 1 shows the KM survival functions in the chemotherapy and combined therapy groups.

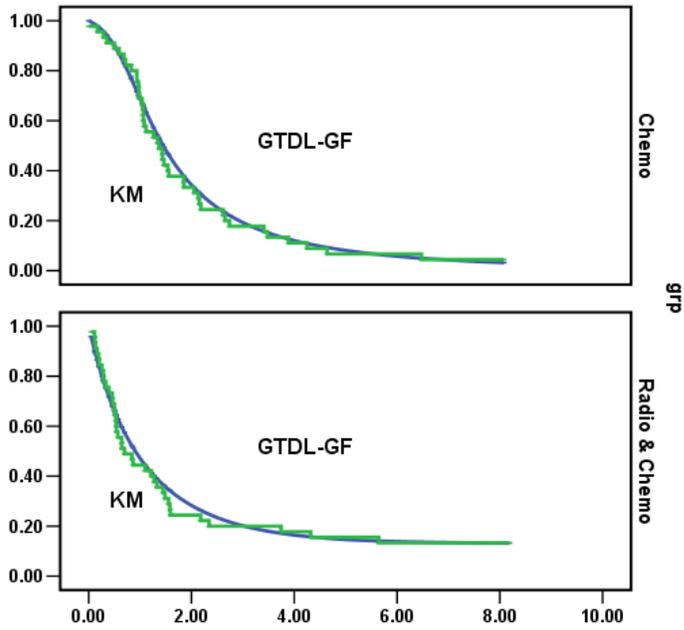
TABLE 1. Models fitted and their marginal mles & (s.e.)

Model	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$\hat{\ell}$
Cox	-	-	-	-0.106	-	-307.47
				(0.223)		
Cox GF	-	-	-	-1.146	1.717	-306.50
				(0.675)	(1.024)	
GTDL	-0.832	-0.094	1.494	-1.380	-	-132.55
	(0.242)	(0.192)	(0.666)	(0.822)		
GTDL GF	-0.789	3.499	2.380	-4.612	0.400	-127.89
	(0.326)	(1.408)	(1.413)	(1.676)	(0.176)	

The crossing survival functions are a clear sign of non-proportionality and survival is lower in patients receiving combined therapy for the first three years and thereafter it is better. We fitted the sequence of models (1)-(4) presented above using (marginal) maximum likelihood estimation and the results are shown in Table 1. As expected the PH model cannot cope with this situation - the log-rank test is non-significant - as indicated by $\hat{\beta}_1 / \text{se}(\hat{\beta}_1)$. On the other hand, the generalised Wilcoxon statistic is ($\chi^2 = 3.96, \text{df} = 1, p < 0.05$). In a two group comparison with no explanatory covariates we should examine the role of heterogeneity, via frailty. Here again the semi-parametric PH model is uninformative. The GTDL model with separate time parameters for each group (α_0 & α_1) is equally unhelpful, but when the GTDL model is extended to Gamma frailty, all of the resulting parameters are statistically significant, suggesting that the GTDL frailty model provides a satisfactory explanation of the data. The resulting fit, confirming this, is shown in Figure 2, while Figure 3 shows where the fitted GTDL-GF model survivor functions cross.

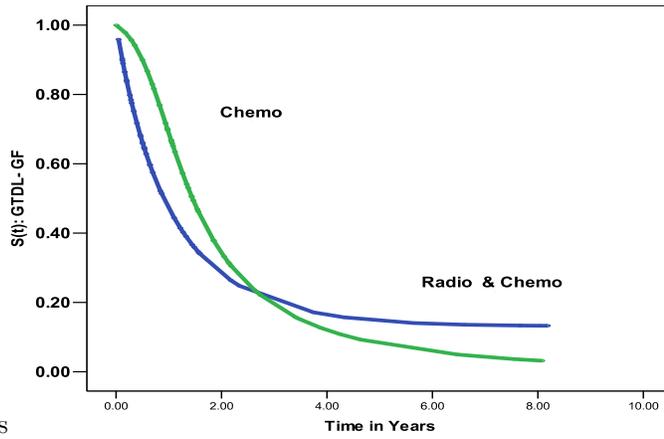
4 Discussion

The use of parametric models is convenient when dealing with crossing hazards data. It is natural, in these circumstances, to consider a non-PH family such as the GTDL. It might be have been thought that the use of separate time parameters, which is always a viable strategy with parametric models, would have been sufficient to capture the structure of the data. However, our analysis shows clearly the important role of heterogeneity modelled via the Gamma distribution.



SS

FIGURE 2. KM Survival & Fitted GTDL Gamma Frailty Model



SS

FIGURE 3. GTDL-GF Survival - Fitted GTDL Gamma Frailty Model

References

Cox DR (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. B* **34**, 187-220.

Hougaard, P. (1994). Heterogeneity Models of Disease Susceptibility, with Applications to Diabetic Nephropathy. populations. *Biometrics* **50**, 1178-1188.

MacKenzie, G. (1996) Regression models for survival data: the generalised time dependent logistic family. *JRSS Series D* **45**, 21-34.

Stablein DM I. & Koutrouvelis, IA (1985). A Two-Sample Test Sensitive to Crossing Hazards in Uncensored and Singly Censored Data *Biometrics* **41**(3), 643-652.

Bayesian multivariate disease mapping and ecological models with errors-in-covariates: Mapping disability adjusted life years

Ying C MacNab¹

¹ Division of Epidemiology and Biostatistics, Department of Health Care and Epidemiology, University of British Columbia, Canada.

Abstract: We present a Bayesian disability adjusted life year (DALY) methodology for spatial analysis of disease and/or injury burden. A Bayesian disease mapping model framework, which blends together multivariate spatial modelling, shared-component modelling, and ecological modelling with errors-in-covariates, is developed for small area DALY estimation and inference. In particular, we develop a model framework that enables shared-component modelling as well as multivariate CAR modelling of non-fatal and fatal disease or injury rates and associated risk factors. Fully Bayesian estimation and inference, with Markov chain Monte Carlo implementation, will be illustrated. We present a methodological framework within which the DALY and the Bayesian disease mapping methodologies interface and intersect. Its development brings the relative importance of premature mortality and disability into the assessment of community health and health needs in order to provide reliable information and evidence for community-based public health surveillance and evaluation, disease and injury prevention, and resource provision. The methods are motivated and illustrated using road traffic accident injury data from 1991 to 2000 in British Columbia, Canada.

Keywords: Bayesian disease mapping; Multivariate CAR; Shared-component model; Ecological regression with errors-in-covariates, Disability-adjusted life years.

1 Disability adjusted life year methodology

The disability adjusted life year (DALY) methodology (commonly known as burden of disease methodology) was initially developed, through a World Bank and World Health Organization collaboration, to provide information to support global health policy and priority setting. The methodological framework has been explored and debated for its potential to inform and facilitate health-sector planning. In particular, relevant literature explored DALY's potential as a population health indicator for public health assessment and as a 'currency' for cost-effectiveness analysis in order to set priority for health interventions. A key element of a burden of disease study is the estimation and assessment of DALYs, quantified by the sum of years of life lost to premature mortality (YLLs) and years of life lost to disability (YLDs): $DALYs = YLLs + YLDs$. The hallmark of the DALY methodology is a single measure of disease or injury burden designed to capture the impact of both premature death and disability. The methodological development, showcased in a comprehensive publication of the Global Burden of Disease and Injury Series (Murray and Lopez, 1996), marked

a new transition to population health and health needs measures that reflect both the *quantity* and *quality* of life.

Specifically, the basic DALY framework allows the combined impact of mortality and morbidity to be incorporated and assessed simultaneously, in order to provide information and evidence for population health and health care management. The methodology is based on analysis of incidence- or prevalence-based epidemiological data and assessment of disease-based health-related quality of life, with a set of disability weights being developed using expert assessment (a person trade-off method) for disease and injury conditions. In the case of incidence-based DALY estimation, a set of duration of disability estimates can be derived within the DALY framework presented in Murray and Lopez(1996). The Global Burden of Disease and Injury Study also incorporates the following key social preferences or values into the development of the indicator: (1) the use of gender specific ‘standard’ expected years of life lost, derived from the life expectancy at birth of 82.5 years for women, and 80 years for men, the highest national life expectancies, those of the Japanese; (2) the use of age-weighting $Cxe^{-\beta x}$, where x denotes age and C and β are parameters indicating relative importance of healthy life at different ages (or age groups); and (3) the use of time preference, discounting health gains in the future at an annual rate of , say, 3% (commonly used in DALY studies).

Since the age-weighting and future health discounting are two controversial value choices that are seen considerable debates in the DALY literature, in this study and without loss of generality we illustrate the methodology assuming 3% discounting rate but no age-weighting. Let U_{ij} and V_{ij} represent the numbers of fatal and non-fatal cases for a particular disease or injury to an age- and gender-specific population, where i indexes age or age group and j indexes region; let W_i be the corresponding disability weight, D_i the duration of disability, L_i the standard life expectancy at a particular age or age group, N_{ij} the ‘at risk’ population, and r the discount rate. One derives the simplified version (Murray and Lopez, 1996) of $DALY_{s_{ij}}$: $DALY_{s_{ij}} = YLLs_{ij} + YLDs_{ij}$, where $YLLs_{ij} = U_{ij}(1 - \exp(-rL_i))/r$ and $YLDs_{ij} = V_{ij}W_i(1 - \exp(-rD_i))/r$, with $YLLs_{ij}/N_{ij}$, $YLDs_{ij}/N_{ij}$, and $DALY_{s_{ij}}/N_{ij}$ being the corresponding YLLs, YLDs and DALYs per capita.

2 Bayesian DALYs

In this study, we bring DALYs into the context of spatiotemporal disease and injury surveillance and develop Bayesian DALY measures for small area burden of disease and injury analysis and mapping. To facilitate joint modelling of non-fatal and fatal disease or injury outcomes for small-area DALY estimation and inference, we utilize the current state-of-the-art for the disease mapping methodology to build a considerably rich Bayesian hierarchical model framework that enables ‘borrowing strength’, accommodate shared latent risks, quantify cross-component correlation and observed covariate effects, and account for overdispersion, spatial correlation, and measurement errors. We develop a Bayesian DALY methodology that makes it possible to smooth, describe and draw inference on small area burden of disease or injury. A Bayesian DALY analysis will be presented using road traffic accident injury fatal and nonfatal data from 1991 to 2000 in 84 local health areas in British Columbia (BC), Canada.

In particular, the main feature of the Bayesian DALY methodology is that it incorporates DALY estimation into the statistical modelling of the observed epidemiological data in order to facilitate exploration of spatial clustering of DALYs, to quantify DALY uncertainty, and to offer DALY inference. Estimation of the YLLs, YLDs, and DALYs, at the provincial and the local areal levels, requires the input of adequate epidemiological statistics quantifying area-, age-, gender-, time-, and type-specific rates of non-fatal and fatal outcomes. We develop a Bayesian DALY framework within which estimation of YLLs, YLDs, and DALYs becomes part of the analysis of spatial and spatiotemporal non-fatal and fatal disease or injury rates, with convergence monitoring and posterior sample inference on the YLLs, YLDs, and DALYs.

2.1 Spatial models for DALYs: Shared component models

Taking the main ideas from Held and Best (2001), we introduce the following shared component formulation, with Poisson likelihood, for joint spatial modelling of the non-fatal and fatal injury outcomes:

$$\log(\mu_{1j}|b) = \log(n_{1j}) + a_1 + b_{0j}\gamma + b_{1j}, \quad (1)$$

$$\log(\mu_{2j}|b) = \log(n_{2j}) + a_2 + b_{0j}/\gamma + b_{2j}, \quad (2)$$

where $\mu_{ij}|b$ is the expectation of Y_{ij} conditioning on the random effects $b = (b_0^\top, b_1^\top, b_2^\top)^\top$; a_i is the intercept with respect to non-fatal ($i = 1$) and fatal ($i = 2$) outcomes; $m_i = \exp(a_i)$ is the corresponding global mean rate; $b_0 = (b_{01}, b_{02}, \dots, b_{0J})^\top$ is a random effects ensemble representing shared effects/component common to both non-fatal and fatal outcomes; $b_i = (b_{i1}, b_{i2}, \dots, b_{iJ})^\top$ is a random effects ensemble that only has relevance to the non-fatal ($i = 1$) or fatal ($i = 2$) outcome risk; $\gamma > 0$ is an unknown scale parameter allowing for different ‘risk gradients’ for non-fatal and fatal outcomes respectively.

The primary motivation for the model formulation (1)+(2) is that for a specific type of injury such as road traffic accident, the non-fatal and fatal outcomes often share similar risk factor(s). The method also enables us to examine and discover similarities and dissimilarities in the geographical distributions of the non-fatal and fatal outcome risks. In addition, the shared component modelling facilitates ‘borrowing strength’ between the two rates. This has particular appeal and can potentially improve YLLs (and therefore DALY) estimation and inference, for to most injuries fatal outcome rates are often much smaller and have higher chance variation in comparison with the corresponding non-fatal rates. Note also that the model can be readily extended to jointly model non-fatal and fatal outcomes of multiple diseases and injury categories, the associated covariates effects, and to model data under Binomial likelihood (say, for non-rare disease or health outcomes). Modifications to the model formulation (1)+(2) and spatial prior formulations for the random effects bs will be presented and discussed.

2.2 Spatial models for DALYs: Bivariate CAR models

As an alternative, joint modelling of the non-fatal and fatal rates, as well as multivariate disease or injury outcomes, may be carried out via bivariate (or multivariate) CAR

modelling that models responses correlation and enables ‘borrowing strength’. In particular, recently developed multivariate CAR models (MacNab, 2007, and references within), which enable us to account for and estimate cross-component correlation, can be considered. For example, for a predefined spatial/neighborhood configurations of the areal units (say, area adjacency), the following bivariate model may be considered for bivariate modelling of the non-fatal and fatal injury outcomes:

$$\log(\mu_{ij}|b) = \log(n_{ij}) + a_i + b_{ij}, \quad i = 1, 2 \quad (3)$$

with the second-level prior

$$b \sim \text{MVN}(0, \Omega_b), \quad (4)$$

where $b = (b_1^\top, \dots, b_J^\top)^\top$, $b_j = (b_{1j}, b_{2j})^\top$. In this study, we present a MCAR analysis for the following MCAR formulations (MacNab, 2007):

$$\Omega_b = (L - \lambda W) \otimes \Gamma; \quad (5)$$

$$\Omega_b = (M - \alpha W) \otimes \Gamma; \quad (6)$$

$$\Omega_b = M \otimes \Gamma - W \otimes \Gamma R \Gamma; \quad (7)$$

where $\lambda \in (0, 1)$, $\alpha \in (0, 1)$, $\alpha_i \in (0, 1)$, $i = 1, 2$, $L = \text{diag}(1 - \lambda + \lambda m_1, \dots, 1 - \lambda + \lambda m_J)$, $M = \text{diag}(m_1, \dots, m_J)$, $R = \text{diag}(\alpha_1, \alpha_2)$; I_2 is an identity matrix of dimension 2; Γ is a unstructured 2 by 2 symmetric positive definite matrix; W is a J by J (neighbourhood) matrix with elements $w_{ll} = 0$ and $w_{lm} = 1$ when the l th and m th regions are neighbours and zero otherwise. Extension to multiple disease or injury outcomes is straightforward and will be presented and discussed.

2.3 Separate modelling of spatial rates

It is worth noting that assuming the following second-level prior for model (3) enables separate modelling of the non-fatal and fatal rates:

$$b_i \sim \text{MVN}(0, \Omega_{b_i}), \quad i = 1, 2, \quad b_1 \perp b_2 \quad (8)$$

where $b_i = (b_{i1}, \dots, b_{iJ})^\top$, $i = 1, 2$. Spatial or non-spatial (IID) priors may be considered for b_1 and b_2 respectively.

2.4 Ecological regression with errors in covariates

Epidemiological investigations in which associations between disease occurrence and potential risk factors are studied over aggregated groups (e.g. areas) have gained increased popularity and recognition in recent years. This is mainly due to the recent developments of Bayesian statistical methodologies that enable us to explore and address important issues such as risk estimation, unmeasured confounding, spatial dependence, measurement errors, and ecological bias. In this study, we will illustrate how DALYs may be attributed to shared risks and shared risk factors via ecological regression analysis. One potential utility of the resulting DALY investigation, for example, is the evaluation of DALYs that attributable to regional characteristics and

identification of regions to which injury prevention resources may be directed to reduce DALYs. To this end, the aforementioned spatial models were extended to include covariates. Ecological models with errors-in-covariates were developed such that attribution of regional characteristics to DALY variations and changes may be explored and evaluated taking into consideration potential influence of measurement errors on risk attribution.

In particular, our study has identified five ecological covariates (educational attainment, neighbourhood socio-economic status, seatbelt violations, speeding charges, and recent immigrants) that were associated with the road traffic accident (RTA) injury rates to boys and girls aged 15 to 24 in British Columbia, Canada; and these regional characteristics explained the majority of spatial and regional RTA rate variations. A Bayesian hierarchical model framework with Binomial and Poisson measurement error submodels and spatial and non-spatial submodel priors have been developed for regional variables of proportions and rare events.

3 Illustration and Discussion

To illustrate spatial DALY estimation and inference, we present a 10-year aggregated regional analysis of non-fatal and fatal RTA occurrences to boys and girls aged 15-24 in BC's 84 local health areas. Full Bayesian analyses were carried using WinBUGS. Model assessment and selection were carefully carried out by comparing the DIC, deviance, and the effective number of parameters for the fitted models. We also assessed goodness-of-fit and model adequacy by visual inspection of the Bayesian cross-validation residual plots.

Recent developments of the Bayesian disease mapping methodology, the DALY framework, and geographic information system (GIS) technology present opportunities to develop model- and GIS-based spatiotemporal disease and injury surveillance systems for public health monitoring and evidence-based priority setting and health planning. As exemplified herein, the Bayesian DALY analysis and mapping facilitate exploration of patterns of DALYs, quantify DALY uncertainty and provide DALY inference. We will illustrate a methodological framework within which the DALY and the Bayesian disease mapping methodologies interface and intersect. The methodology presents a new approach to DALY assessment for evidence-based disease and injury prevention and surveillance, policy initiative, and service provision. The methods can also be extended to facilitate cost-effectiveness analysis of health care interventions.

The Bayesian hierarchical spatial models presented herein have been extended to spatiotemporal context for spatiotemporal DALY analysis, estimation and inference (MacNab, 2007). A brief outline of the spatiotemporal DALY methods will be presented.

Acknowledgments: The author thanks the BC Ministry of Health, the BC Vital Statistics Agency, the BC Stats, and the University of BC's Centre for Health Services and Policy Research for provision of the data. The work was partially funded by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Health Research, the Michael Smith Foundation for Health Research, and the British Columbia Child and Family Research Institute.

References

- Murray, C.J.L. and Lopez A.D. (eds) (1996). *The Global Burden of Disease*. Cambridge, Massachusetts: Harvard University Press.
- Knorr-Held, L. and Best, N.G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A* **164**, 73-85.
- MacNab, Y.C. (2007). Mapping disability-adjusted life years: A Bayesian hierarchical model framework for burden of disease and injury assessment. *Statistics in Medicine*. To appear.

Out-of-Sample Bootstrap Tests for Non-Nested Models

Patrick Mair¹ and Achim Zeileis¹

¹ Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Augasse 2-6, A-1090 Vienna, Austria, patrick.mair@wu-wien.ac.at

Abstract: When testing non-nested models, the asymptotic distribution theory of the ordinary likelihood ratio statistic is not valid anymore. Several test statistics, some of them based on information criteria, have been proposed in order to test such non-nested hypotheses. Concerning bootstrap approaches to simulate goodness-of-fit measures such as the likelihood ratio value, have been elaborated as well. Based on these methods, we extend existing bootstrap simulations towards out-of-sample bootstrap evaluation. As an application, a parametric bootstrap on simulated regression data is provided.

Keywords: Non-nested models; Out-of-sample bootstrap.

1 Introduction

In standard statistical theory, nested models are usually compared by using a likelihood ratio (LR) statistic which is asymptotically χ^2 -distributed with the corresponding difference in the degrees of freedom. However, in order to test hypotheses driven by subject-matter knowledge, decisions between models which are non-nested can be desirable. In this case, typically AIC or BIC are used to compare the models on a “descriptive” level: The model which minimizes the corresponding IC is chosen. However, this comparison does not allow for a statement that one model fits significantly better than the other one.

Basically, statistical models can be non-nested in the likelihoods, in the regressors, and in the link-function (e.g., linear against log-linear regressions). The estimated parameters optimize a certain target function, typically the likelihood. An in-sample evaluation of the target function may be too “optimistic” with respect to the generalizability of the results; an out-of-sample evaluation is more feasible. In addition, by using the classical IC approach, a certain risk or goodness-of-fit functional is approximated by in-sample calculations. A more modern approach is to estimate that risk functional by resampling or bootstrap techniques. Thus, we present corresponding out-of-sample bootstrap (OOB) evaluations on the differences in the log-likelihoods as target function which is applicable for nested as well as for non-nested model hypotheses. For such non-nested models, a broad application spectrum has been shown, ranging from econometrics to psychometrics.

2 General Problem Formulation

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$ denote a random vector with the corresponding observations $\mathbf{y} = (y_1, y_2, \dots, y_n)$. When testing non-nested hypotheses, it is not obvious, which one should be considered as “null hypothesis” H_0 and “alternative” H_1 , respectively. Thus, for model M_f we state the hypothesis H_f that M_f fits the data, and that model M_g fits the data is imposed in H_g . Pertaining to parameter notation, $\boldsymbol{\theta}_f \in \Theta_f$ is the parameter vector for M_f and $\boldsymbol{\theta}_g \in \Theta_g$ is the parameter vector for M_g . It follows that the LR statistic which tests the appropriateness of M_f (Cox, 1962) can be expressed by

$$T_f = \left(L_f(\mathbf{y}|\hat{\boldsymbol{\theta}}_f) - L_g(\mathbf{y}|\hat{\boldsymbol{\theta}}_g) \right) - E_{\boldsymbol{\theta}_f} \left(L_f(\mathbf{y}|\hat{\boldsymbol{\theta}}_f) - L_g(\mathbf{y}|\hat{\boldsymbol{\theta}}_g) \right). \quad (1)$$

In order to test the fit of M_g , T_g can be established straightforwardly by switching the hypotheses. However, this way to state the hypotheses can lead to ambiguous results. Hence, within the context of our bootstrap approach, we test that “ M_f and M_g do the same” (H_0) vs. “one is better than the other” (H_1). A formal representation is given in the next section.

Further developments and generalizations emanating from Cox’s test statistic are numerous. A recent overview of existing testing approaches can be found in Watnik, Johnson, and Bedrick (2001).

3 Out-of-Sample Bootstrap Approach

3.1 Simulation Setting and DGP

Hinde (1992) accomplished a bootstrap simulation based on the LR -criterion in order to make a decision between two non-nested GLM’s M_f and M_g . We extend this approach with respect to a parametric OOB simulation following Hothorn, Leisch, Zeileis, and Hornik (2005). The bootstrap sampling is carried out by repeatedly drawing (i.e., from $b = 1, \dots, B$) with replacement samples of size n from the Monte Carlo simulated data. When using subsampling, samples of size $n/2$ are drawn without replacement. Based on the resulting observations, each bootstrap sample is splitted into a training and a test data set and the differences in the log-likelihoods, i.e., $\Delta L = L_f(\mathbf{y}|\hat{\boldsymbol{\theta}}_f) - L_g(\mathbf{y}|\hat{\boldsymbol{\theta}}_g)$, are computed using the training data and evaluated on the test data.

As an application, we simulate data which are consistent with the following regression model M_f (see also Watnik et al., 2001):

$$\mathbf{Y} = \mathbf{V}\boldsymbol{\beta}_v + \mathbf{X}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}_x \quad (2)$$

\mathbf{Y} is the $n \times 1$ vector of the dependent variables pertaining to the simulated observations, \mathbf{V} is an $n \times 2$ matrix with the intercept vector and one regressor. The remaining k regressors are defined in the $n \times k$ matrix \mathbf{X} . The joint vector of regression parameters is $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_v, \boldsymbol{\beta}_x)$.

The alternative model M_g is expressed as

$$\mathbf{Y} = \mathbf{V}\boldsymbol{\beta}_v + \mathbf{Z}\boldsymbol{\beta}_z + \boldsymbol{\epsilon}_z. \quad (3)$$

\mathbf{Z} are k different regressors with respect to \mathbf{X} . Thus, \mathbf{V} is the vector of common regressors, ϵ_x and ϵ_z are i.i.d. with mean $\mu = 0$ and variance $\sigma^2 = 1$, whereas $\beta_z = \beta_x$. Pertaining to the data generation process (DGP), the regressor matrices are drawn from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ and the similarity of M_f and M_g is controlled by the correlation $r(\mathbf{X}, \mathbf{Z})$ in the VC-matrix Σ . The lower $r(\mathbf{X}, \mathbf{Z})$, the larger the difference in model fit between M_f and M_g . As a consequence, $\Delta L = L_f(\mathbf{y}|\beta_v, \beta_x) - L_g(\mathbf{y}|\beta_v, \beta_z)$ (i.e. the deviance without χ^2 correction) becomes large and the OOB test should detect this deviation.

In the simulation, for each bootstrap sample b on a particular data set, the observation-wise out-of-sample likelihoods are computed and correspondingly the mean is taken. Thus, the likelihood values $L_f^{(b)}$ and $L_g^{(b)}$ result. In order to obtain one deviance value for each data set, again the mean is taken, i.e. $\overline{\Delta L} = 1/B \sum_{b=1}^B (L_f^{(b)} - L_g^{(b)})$. Thus, a z -test can be carried out straightforwardly and, consequently, p -values and rejection rates for each data set can be computed.

3.2 Simulation Conditions and Bootstrap Results

The correlation parameters for model similarity vary from .20 to .99, the sample size n from 20 to 100. In order to examine effects with respect to the variations in goodness-of-fit of M_f , different β -constellations from .20 up to 2.00 are provided. Within each combination of these parameters, 500 data sets for $k = 2$ x - and z -predictors are simulated. On each of these data sets $B = 250$ bootstrap samples are drawn and evaluated out-of-sample. Finally the rejection rates of the hypothesis that $\Delta L = 0$ is determined.

By inspecting the resulting rejection rates in Table 1, several issues of test behavior of the OOB test become obvious: For low β -values, which implies that neither of the models fits the data, the test is not able to detect differences even for low model correlations. With growing β -values, it is capable to discriminate between M_f and M_g with respect to the goodness-of-fit. For large β 's and large n -values, the OOB test becomes significant also for negligible model differences (i.e., high correlations). By inspecting the sample size n for medium ranged regression coefficients, the decrease in test power over the model correlations is striking. For $n = 20$, the rejection rate is considerably low, whereas an increasing n leads to a noticeable gain in testing power. However, further inspection of this bootstrap testing approach is needed, on the one hand, from a formal point of view with respect to a bias due to finite samples and, on the other hand, from an application point of view to compare the results to other competing test statistics with respect to different model types.

TABLE 1. Rejection rates for non-nested regression models

β	n	$r(\mathbf{X}, \mathbf{Y})$								
		.20	.50	.70	.80	.85	.90	.95	.97	.99
.2	20	.31	.35	.32	.27	.27	.27	.34	.26	.28
.2	30	.39	.44	.30	.37	.35	.33	.35	.34	.34
.2	50	.52	.46	.47	.46	.46	.46	.48	.45	.44
.2	100	.71	.67	.70	.70	.57	.65	.60	.67	.54
.5	20	.40	.33	.36	.34	.32	.36	.31	.27	.36
.5	30	.67	.69	.59	.52	.53	.46	.46	.46	.34
.5	50	.96	.92	.82	.79	.73	.72	.67	.62	.56
.5	100	1.00	.99	.99	.95	.94	.91	.80	.78	.73
1.0	20	.67	.57	.54	.49	.47	.45	.49	.36	.41
1.0	30	.95	.97	.90	.85	.78	.70	.60	.58	.49
1.0	50	1.00	1.00	1.00	.98	.98	.95	.81	.78	.68
1.0	100	1.00	1.00	1.00	1.00	1.00	.99	.96	.94	.81
2.0	20	.94	.90	.89	.79	.76	.72	.56	.42	.35
2.0	30	1.00	1.00	1.00	.98	.98	.98	.86	.80	.58
2.0	50	1.00	1.00	1.00	1.00	1.00	1.00	.96	.93	.81
2.0	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.95

References

- Cox, D. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B* **24**, 406-424.
- Hinde, J. (1992) Choosing between non-nested models: A simulation approach. In: *Advances in GLIM and Statistical Modelling, Lecture note in Statistics 78*. 119-124, New York: Springer.
- Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. (2005) The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics* **14**, 675-699.
- Watnik, M., Johnson, J., Bedrick, E. J. (2001) Nonnested linear model selection revisited. *Communications in Statistics - Theory and Methods* **30**, 1-20.

Compositional modelling of sediment formation at the surface of Mars

J.A. Martín-Fernández¹, C. Barceló-Vidal¹, C. Kolb², and H. Lammer²

¹ Dept. Informática y Matemática Aplicada, Univ. de Girona, Girona, Spain.

² Space Research Institute, Austrian Academy of Sciences, Graz, Austria.

1 Introduction

Compositional modelling is performed on chemical compositions of Martian surface materials in order to understand weathering scenarios and the role of meteoritic accumulation. Compositional techniques are applied to elucidate the influence of remote weathering by combined analysis of several soil forming branches. The results foster evidence for the past activity of volcanic exhalation products along with that of liquid water.

One of the most exciting results of the Mars Exploration Rover (MER-A and MER-B, NASA) and Mars Express (ESA) mission is the clear indication for the presence of liquid water as an agent of chemical weathering on ancient Mars. A comprehensive interpretation of the available data is introduced in Kolb et al. (2006). In our compositional modelling we employ chemical data that were obtained in the course of the Mars Pathfinder and the MER missions to identify the processes involved in the degradation of primary planetary material and in sediment formation.

Aitchison (1986) defined a composition as a collection of D non-negative measurements, which sum to unity per weight, volume, or abundance. Such constraints are obeyed by the Simplex space geometry, which represents a D -dimensional analogue of a triangle, in contrast to the D -dimensional orthogonal Euclidean space geometry. Centered log ratio *clr* transformation translates compositional data from the constrained Simplex space to the Euclidean real space. The *clr* transformation of a compositional vector \mathbf{C} with components c_i is defined as $clr(\mathbf{C}) = [\ln \frac{c_1}{g(\mathbf{C})}, \dots, \ln \frac{c_D}{g(\mathbf{C})}]$, where $g(\mathbf{C})$ is the geometric mean of the vector \mathbf{C} . In this work we focus on visualization of data variability by means of biplot techniques (Aitchison and Greenacre, 2002). Based on the variability patterns, rock alteration and soil formation processes such as remote weathering are modelled.

2 Compositional modelling

If in the course of an alteration process a subset of elements is mobile and is added or lost from a given volume of rock, the concentrations of all components, the mobile ones but also the immobile ones, will change merely due to the closure condition. Aitchison (1986) proposed the so-called perturbation mechanism in order to model an exchange

of chemical compositions by means of compositional vectors under consideration of the Simplex geometry. The application of a compositional vector \mathbf{C} on the chemical composition \mathbf{C}^* to yield the chemical composition \mathbf{C}^{**} by means of the perturbation operation \oplus , is defined as $\mathbf{C}^{**} = \mathbf{C} \oplus \mathbf{C}^* = \left[\frac{c_1 c_1^*}{\sum c_k c_k^*}, \dots, \frac{c_D c_D^*}{\sum c_k c_k^*} \right]$. The components of the alteration vector \mathbf{C} must be a measure of change for the same parts of individual observations linked by this vector. An active change of compositions takes place if values of the corresponding vector entries are not equal to the geometric mean. Vector entries which are above this level are actively increasing, values which are below are actively decreasing. However, a passive change of compositional values could occur, although the corresponding values of vector entries equal the geometric mean.

Chemical compositions of Martian surface materials in terms of element wt% of 13 elements (*Na, Mg, Al, Si, P, S, Cl, K, Ca, Ti, Cr, Mn, Fe*) were taken from the literature (Kolb et al., 2006). Figure 1 represents a *clr*-biplot on Pathfinder and Mars Exploration Rover data, which describes 74% of the variance. In the biplot the presumed source rocks have been projected. The lines denote the distribution of *clr* transformed variables across the chemical variability of Martian surface materials. Five classes of sample suites are discernable: Domain of soil, MER-B evaporites, MER-B basalts, MER-A basalts and Mars Pathfinder basalt-andesites. The soils plot close to the origin. The presumed source rocks are located in different more peripheral positions, illustrating the chemical uniformity among soils as opposed to the rather heterogeneous nature of rock compositions. The trends allow discrimination between coated and abraded rocks which are located proximal and distal to the region of soils, respectively. The magmatic rocks plot on the side of *clr*(Si) and the evaporites are located in the area spanned by *clr*(S), *clr*(Cl), and *clr*(Mg). MER-A basalts plot in the vicinity of *clr*(Mg) and *clr*(Cr); basalt-andesites are shifted toward majority of *clr*(K) and *clr*(Si); MER-B basalts are of intermediary character. Overall, these observations are consistent with petrological models of Martian rocks. The shift between brushed MER-A basalts and abraded MER-A basalts indicates the formation of chemical weathering crusts different from fresh rocks (Fig. 1).

The alteration vector calculation is based on *clr* Principal Component Analysis (*clr*-PCA) of fresh basalts (abraded) and their crusts (brushed rocks). We transform back the first *clr*-PCA eigenvector to obtain the alteration vector $\mathbf{C} = (Na, Mg, Al, Si, P, S, Cl, K, Ca, Ti, Cr, Mn, Fe) = [7.2, 5.3, 6.4, 6.2, 6.8, 12.6, 12, 12.7, 6.2, 6.5, 6.2, 6, 5.9]$, which has geometric mean equal to $g(\mathbf{C}) = 7.3$. This vector explains 82% of variability. The relation of vector entries to the geometric mean is a measure for their degree of change. The element Na remains actively unchanged in the course of alteration, the others change to different degree. S and Cl are incorporated in crust, while Mg and Fe are removed from crust in relation to the fresh rock. This is consistent with the formation pathway of Mg-, Fe-bearing sulfates and ferric oxides upon dissolution of Mg-, Fe-bearing primary silicates. The element K in crusts stems most probably from impurities, derived from remote weathering of basalt-andesitic or other K-bearing rocks. This procedure based on *clr*-PCA analysis is applied to the others branches (Fig. 1). Applied to the entire MER-A basalt and soil data complex to derive soil formation vectors, are capable to explain high levels of variability (94%). Application of *clr*-PCA to the Mars Pathfinder (MPF) basalt-andesitic rock and soil branch provided

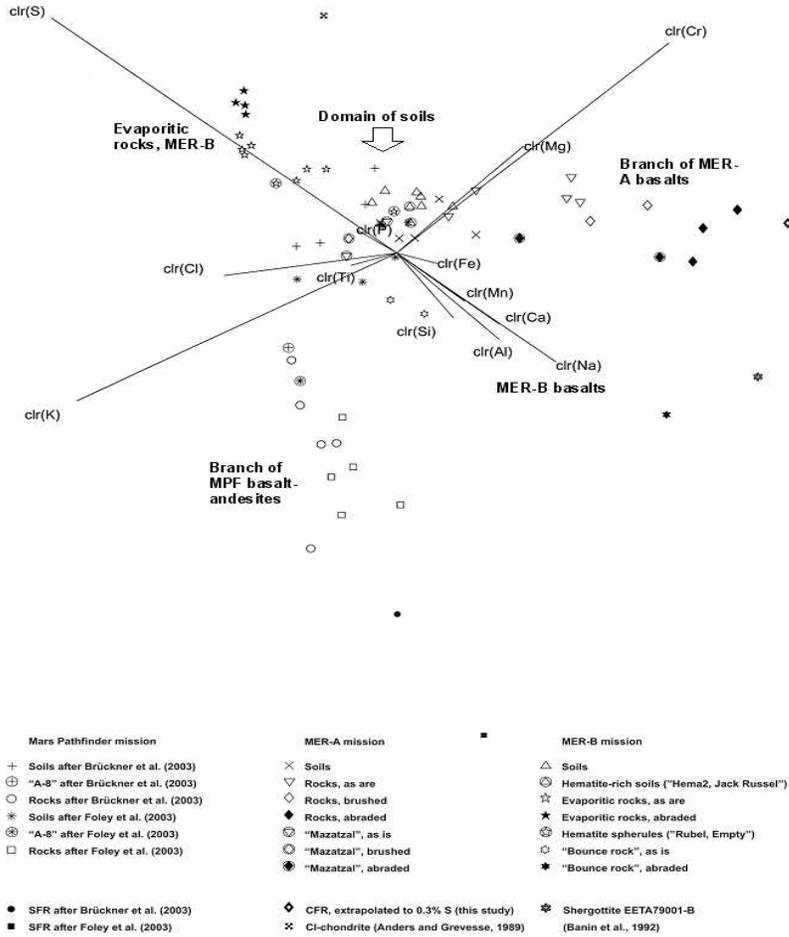


FIGURE 1. *clr*-biplot of Martian samples. The abbreviation SFR and CFR stands for Soil Free Rock and Crust Free Rock, respectively.

a vector with 92% explanation of variability. The vector along the MER-B basalt branch by means of *clr*-PCA covers 99% of variability. In Kolb et al. (2006) detailed interpretations are given.

Compositional data analysis by means of *clr*-biplot visualization techniques provides a clear separation of Martian surface samples and allows to assign individual elements to specific principal component characteristics: basalt-andesites are related to K, Si; MER-A basalts are related to Mg, Cr; MER-B basalts are of intermediary nature and MER-B evaporites are related to S, Cl, Mg. Overall, these findings are consistent with existing petrological models of Martian rocks. Furthermore, chemical weathering rinds are observed to be significant differently composed in relation to soil and fresh rock,

but appear as intermediate stage of soil formation.

Acknowledgments: Work partially financed by the A.I. of the Spanish Ministry for Science and Technology (HU2005-0006).

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall, 416 pp. Reprinted in 2003 by Blackburn Press.
- Aitchison, J., and Greenacre, M. (2002). Biplots of compositional data. *Applied Statistics* **51**, 375-392.
- Anders, E., and Grevesse, N. (1989). Abundances of the elements: Meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197-214.
- Banin, A., Han, F.X., Kan, I., and Cicelsky, A. (1997). Acidic volatiles and the Mars Soil. *J. Geophys. Res.* **102**, 13, 341-356.
- Brückner, J., Dreibus, G., Rieder, R., and Wänke, H. (2003). Refined data of Alpha Proton X-ray Spectrometer analyses of soils and rocks at the Mars Pathfinder site: Implications for surface chemistry. *J. Geophys. Res.* **108**, (E12) ROV 35-1.
- Foley, C.N., Economou, T., and Clayton, R.N. (2003). Final chemical results from the Mars Pathfinder alpha proton X-ray spectrometer. *J. Geophys. Res.* **108**, (E12) ROV 37-1.
- Kolb, C., Martín-Fernández, J.A. Abart, R. and Lammer, H. (2006). The chemical variability at the surface of Mars. *Icarus* **183**, 10-29.

Balances versus amalgamations in compositional data with an application in welfare research

Glòria Mateu-Figueras¹ and Josep Daunis-i-Estadella¹

¹ Dept. Informàtica i Matemàtica Aplicada. Univ. de Girona, Catalonia

Abstract: The amalgamation operation is a non-linear operation in the simplex with the Aitchison geometry but it is frequently used to reduce the number of parts of compositional data. The concept of balances between groups could be an alternative to the amalgamation. In this work we discuss and compare the two approaches using a real data set corresponding to behavioural measures of pregnant sows

Keywords: Aitchison Geometry; Amalgamation; Balances.

1 Algebraic geometric structure of the simplex

Compositional data are parts of some whole which give only relative information. Typical examples are parts per unit, percentages, ppm, and the like. Their sample space is the simplex, $\mathcal{S}^D = \{\mathbf{x} = (x_1, x_2, \dots, x_D) : x_1 > 0, x_2 > 0, \dots, x_D > 0; \sum_{i=1}^D x_i = \kappa\}$, where κ is a constant, generally 1 for proportions or 100 for percentages (Aitchison, 1986).

The simplex \mathcal{S}^D has a $(D - 1)$ -dimensional Euclidean space structure (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001) with the following operations. Let $\mathcal{C}(\cdot)$ denote the closure operation which normalizes any vector \mathbf{x} to a constant sum (Aitchison, 1986), and let be $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, and $\alpha \in \mathcal{R}$. Then, the inner sum, called *perturbation*, is defined as $\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(x_1x_1^*, x_2x_2^*, \dots, x_Dx_D^*)'$; the outer product, called *powering*, is defined as $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)'$; the inner product is defined as $\langle \mathbf{x}, \mathbf{x}^* \rangle_a = (1/D) \sum_{i < j} \ln(x_i/x_j) \ln(x_i^*/x_j^*)$ and the associated squared distance is $d_a^2(\mathbf{x}, \mathbf{x}^*) = (1/D) \sum_{i < j} (\ln(x_i/x_j) - \ln(x_i^*/x_j^*))^2$. The geometry here defined is known as *Aitchison geometry*, and therefore the subindex a is used.

The Aitchison inner product and its associated norm ensure the existence of an orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$, which leads to a unique expression of a composition \mathbf{x} as a linear combination, $\mathbf{x} = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a \odot \mathbf{e}_1) \oplus (\langle \mathbf{x}, \mathbf{e}_2 \rangle_a \odot \mathbf{e}_2) \oplus \dots \oplus (\langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \odot \mathbf{e}_{D-1})$. In what follows, the vector of coordinates is denoted as $\mathbf{h}(\mathbf{x})$. Note that $\mathbf{h}(\mathbf{x}^* \oplus (\alpha \odot \mathbf{x})) = \mathbf{h}(\mathbf{x}^*) + \alpha \cdot \mathbf{h}(\mathbf{x})$ and that standard real analysis can be applied to the coordinates. It could be shown that any vector of coordinates $\mathbf{h}(\mathbf{x})$ is formed by logratios. If we work with coordinates we preserve distances and our results will be coherent from a compositional point of view (Pawlowsky-Glahn and Egozcue, 2006). Like in every inner product space, the orthonormal basis is not unique but the important point is that, once an orthonormal basis has been chosen, all standard statistical methods can be applied to the coordinates and transferred to the simplex preserving their properties.

1.1 Amalgamations

If the D parts of a composition are separated into $C \leq D$ mutually exclusive and exhaustive subsets and the components of each subset are added together, the resulting C -part composition is termed an amalgamation (Aitchison, 1986). An amalgamation can be regarded as a composition in a simplex with fewer parts, and thus as a space of lower dimension. For this reason, amalgamations of parts have been extensively used to achieve reduced dimension.

Nevertheless, the amalgamation operation is a non-linear operation in the simplex with respect to the Aitchison geometry described above. It can be interpreted as a projection in a simplex of lower dimension. But the amalgamation operation does not preserve Aitchison distances under perturbation (see Egozcue and Pawlowsky-Glahn, 2005 for an illustrative example). The consequences might have important, suppose for example that we perform a cluster analysis to a compositional data set; the results might be completely different if we use the original parts or if we work with amalgamations. Moreover, when the analysis of the amalgamated parts is performed simultaneously with the analysis of the original and non-amalgamated parts, difficulties in interpretation and incompatibilities might arise.

In some cases the nature of the sampling method or the particular characteristics of our data leads to amalgamate some components. For example, Martín.Fernández et al. (1997) try to perform a classification using a 8-part compositional data set from the Darss Sill area with a large amount of zeros. As a first step an amalgamation is proposed to reduce the number of zeros. In this particular case, the zeros are concentrated in a few components. This is interpreted as a sign of overdimension concerning the number of components thus justifying the amalgamation.

1.2 Balances

The concept of balances between groups is introduced in Egozcue and Pawlowsky-Glahn (2005) as a tool to design a particular basis on the simplex in order to the corresponding coordinates are interpretable. They are based on a sequential binary partition of a D -part composition into non-overlapping groups. At each step, a group of parts is partitioned into two non-overlapping groups.

In practice, there is no need to know the exact expression of this basis, as the coordinates can be computed using a one-to-one transformation, and for values of interest the inverse transformation can be used. For example, at i -th step two groups of parts are considered, denoted here as G_{i1} and G_{i2} , then the balance is

$$b_i = \sqrt{\frac{r_i \cdot s_i}{r_i + s_i}} \ln \frac{(\prod_{x_j \in G_{i1}} x_j)^{1/r_i}}{(\prod_{x_\ell \in G_{i2}} x_\ell)^{1/s_i}}.$$

In other terms, the balance is defined as the natural logarithm of the geometric parts in each group, normalised by a coefficient to guarantee unit length of the vectors of the basis. Remember that balances are coordinates with respect to an orthonormal basis, $h(\mathbf{x}) = (b_1, b_2, \dots, b_{D-1})$, and behave like real random vectors, thus all standard methods can be applied.

Observe that using balances we could easily compare the relative behaviour between two groups of variables and using the sequential binary partition we could design the adequate groups. Thus Egozcue and Pawlowsky-Glahn (2005) propose the balances as an alternative to the amalgamations because the analysis of the whole composition and also of some lower-dimensional representations can be made. Using balances the analysis is compatible and coherent with the Aitchison geometry, in particular we have the invariance of distances under perturbation.

2 Data and methodology

The largest amount of information about the welfare of the sows is obtained from the measures of behaviour, particularly measures of activity and stereotypies. Stereotypies are related to poor welfare because they developed in situation of stress, frustration or lack of control. They reflect a past or present difficulty to cope with the environment. Therefore, the decrease in stereotypies level in group-housing systems could already be considered as a welfare improvement.

The data used in this study are obtained from the comparison among two different commercial group-housing systems and conventional stalls for pregnant sows. One hundred and eighty pregnant sows were housed in conventional stalls (Stall), in groups of 10 with trickle feeding (Trick) and in groups of 20 with an electronic sow feeder (Fitmix). Sows were observed on 11 non-consecutive days during 4 h a day. General activity and stereotypes were measured by scan-sampling observation (10-min intervals) in all the system.

Our data are a 7-part compositions observed in 177 complete individuals, corresponding to the observed frequencies of 6 oronasofacial behaviours: drinking (D), sham-chewing (S), floor manipulation (T), bar manipulation (B) -only Stall and Trick-, trough manipulation (C) -only Stall and Trick-, feeder interaction (I) -only Fitmix- and one seventh residual part (H). This data are previously studied in Chapinal (2006) and Daunis-i-Estadella et. al. (2006a).

Our data structure imposes the amalgamation of the components B, C and I; therefore, a new category of behaviour called "interaction with the equipment" (BCI) is created. In a first step, the data are homogenized and the probability of each of behaviours is estimated. The distribution of the observed behaviours may be thought of as coming from a multinomial distribution, with unknown parameters. In order to correctly estimate the probabilities, the presence of zeros have no sense, it is only related to a small time periods of recording observations, and consequently a correction have to be implemented (see Daunis-i-Estadella et al., 2006a). We have, thus, the estimated composition (bci, t, d, s, h) .

3 Balances vs amalgamation

Our objective is to compare the different activities between group-housing systems, for example applying exploratory compositional data tools (Daunis-i-Estadella et al., 2006b). But, in order to make the comparison easier, specialists suggest regrouping some behaviour. Floor manipulation is highly associated to BCI behaviour. Thus,

if t component has no difference for housing systems it may be regrouped with bci component.

To analyze if t component discriminates, two different methodologies are applied. The first one is based on balances and the second one is based on amalgamations.

Using balances we can easily compare the t component with the rest of components. At first step our composition is partitioned into two groups $\{t\}$ and $\{bci, d, s, h\}$. Thus, the first balance is

$$\frac{2}{\sqrt{5}} \ln \frac{t}{(bci \cdot d \cdot s \cdot h)^{1/4}}.$$

The other balances depend on the groups formed in the following steps but they are not used here. As balances are coordinates with respect to an orthonormal basis, the standard real methodology could be applied and the analysis of variance (ANOVA) could be used for testing the equality of the means using the housing system as the factor variable. The value of the F statistic with 2 and 174 degrees of freedom is 6.40 and the corresponding p-value is 0.002. Thus our conclusion is that there are significant differences among the 3 means, consequently there is significant evidence for a housing effect in the t component. The assumptions of homogeneity of variances and the normality of observations are checked using the Levene's test (p-value=0.914) and the Anderson-Darling test of normality (p-value=0.716). As a conclusion, we decide not to amalgamate t component with bci component and to carry on the exploratory analysis with the 5-part composition (bci, t, d, s, h) .

Another often used approach is to study the logratio between t and $1-t$. Observe that $1-t$ is obtained as the amalgamation $bci + d + s + h$, i.e. we work with composition $(t, bci + d + s + h)$. As the dimension is reduced, we could easily compare t component with the rest. Now, following the standard compositional data analysis methodology, we work with the logratio

$$\frac{1}{\sqrt{2}} \ln \frac{t}{(bci + d + s + h)},$$

that is, the coordinates of composition $(t, bci + d + s + h)$ with respect to the orthonormal basis stated in Egozcue et al. (2003). As in the previous case, the analysis of variance (ANOVA) could be applied.

The value of the F statistic with 2 and 174 degrees of freedom is 0.280 and the corresponding p-value is 0.753. Thus our conclusion is that there is no significant difference among the 3 means, that is there is no significant evidence for a housing effect in the t component. The assumptions of homogeneity of variances and the normality of observations are checked using the Levene's test (p-value=0.124) and the Anderson-Darling test of normality (p-value=0.110). At this point we decide to amalgamate t component with bci component and we follow our study with the 4-part composition $(bcit, d, s, h)$. Note that the previous amalgamation $bci + d + s + h$ is no longer considered.

4 Conclusions

Using balances or amalgamations two completely opposite conclusions are obtained. In both cases the standard compositional methodology is used, as we work with logratios or coordinates with respect to an orthonormal basis. It is important to note that in the

second case, the analysis is carried out with the amalgamated parts but a conclusion in terms of the original and not amalgamated parts is finally obtained.

If the amalgamation has a clear sense and we are only interested in studying the relative variability of the parts of the new amalgamated composition, we will have no problems. For example, in a first step, the component *bci* is created as an amalgamation. Nevertheless, when the analysis of the amalgamated parts is performed before or simultaneously with the analysis of the original and non-amalgamated parts, difficulties in interpretation and incompatibilities could arise. In this case, a balance analysis is the most appropriate technique to preserve the relationship between parts due to its compositional coherence.

Acknowledgments: Work partially financed by the Spanish Ministry for Science and Technology (MTM2006-03040)

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).
- Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association* **96(456)**, 1205-1214.
- Chapinal, N. (2006). Effect of the housing and feeding system on the welfare and productivity of pregnant sows. *Ph-D Thesis. Dept. Ciència dels Animals i dels Aliments. UAB-Barcelona* .
- Daunis-i-Estadella, J., Mateu-Figueras, G., Chapinal, N., Manteca, X. and Ruíz de la Torre, J.L (2006a). Aplicación de técnicas composicionales al estudio del comportamiento de cerdas gestantes. In: *Actas del XXIX Congreso Nacional de Estadística e Investigación Operativa (SEIO)*. J. Sicilia, C. González, M.A. González y D. Alcaide (Eds.)
- Daunis-i-Estadella, J., Barceló-Vidal, C. and Buccianti, A. (2006b). Exploratory compositional data analysis. In: *Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (eds). Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, 264, 161174.
- Egozcue, J.J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology* **37(7)**, 795-828.
- Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology* **35(3)**, 279-300.

- Martín-Fernández, J.A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1997). Different Classifications of the Darss Sill Data Set Based on Mixture Models for Compositional Data. In: *Proceedings of the Third annual conference of the IAMG*. 152-156, CIMNE, Barcelona(E).
- Pawlowsky-Glahn, V. and Egozcue, J.J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)* **15(5)**, 384-398.
- Pawlowsky-Glahn, V. and Egozcue, J.J. (2006). Compositional data and their analysis: an introduction. In: *Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (eds). Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society, London, Special Publications, 264, 110.

On split plot type experiments with subsamples

Iwona Mejza¹ and Stanisław Mejza¹

¹ Department of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28, 60-637 Poznań, Poland

Abstract: In the paper we discuss an extension of a linear model for data obtained in an experiment set up in incomplete split-plot design with respect to subsampling in subplots. The considered incompleteness concerns whole plot treatments and/or subplot treatments. The extension of the model changes its dispersion structure only what affects the statistical analysis in the subplot stratum.

Keywords: incomplete split-plot design, linear model, multistratum experiments, subsample

1 Introduction

The paper deals with a problem of model building for observations obtained in experiments set up in incomplete split-plot designs when we observe a trait on some number of units on each subplot. Usually, the mean yield from the subplot or mean yield from many plants, for example, is treated as the observation in agricultural experiments. But there are some experiments when it is important to take into account measurement on a plant as an observation. It means that multiple measurements are collected for each subplot. Such group of experimental units within subplots will be called subsample. Generally, we can assume that the sizes of subsamples within all subplots are different but the same size of subsamples has many advantages. We consider such a way of modelling observations that the linear model is strictly connected with a given experiment, i.e. with a structure of its experimental material and with a method of assigning whole plot or subplot treatments to the units.

The process of randomization plays a central role in the paper. The derivations of linear models connected with different schemes of the randomization are based on the ideas given by Neyman et. al. (1935) and Nelder (1965).

The considered schemes of the randomizations are strictly connected with the three stage nested structure of plots in each experiment. So, the three-step randomization is applied, i.e. the randomization of blocks, the randomization of whole plots within each block, and the randomization of subplots within each whole plot inside each block and randomization of units (sub-subplots) within each subplot.

In the model building the basic assumption refers to a meaning of an observed yield in the experiment. In the paper we assume (Nelder, 1965) the observed yield is a sum of three components, i.e. "*zero yield*" (conceptual yield) due to a unit, a "*pure effect*" due to a treatment combination and "*technical error*" connected with measurements (additivity assumed).

Let us consider a two-factor experiment of split-plot type in which the first factor, A , occurs on S levels A_1, A_2, \dots, A_S (whole plot treatments) and the second factor, B ,

occurs on T levels B_1, B_2, \dots, B_T (subplot treatments). Let a population of units be divided into b blocks, and let each block be additionally divided into $k_1 \leq S$ whole plots, and let each whole plot be divided into $k_2 \leq T$ subplots. Additionally each subplot is divided into some units (sub-subplots).

The approach presented in the paper is applicable to incomplete split-plot designs or complete split-plot designs (as a particular case), cf. Meta (1987, 1994).

Let D be the theoretical design of the experiment. In the paper, by a treatment we mean treatment combination $A_h B_k$, $h = 1, 2, \dots, S$, $k = 1, 2, \dots, T$, while by an effect of the treatment we mean

$$\tau_{hk} = \mu + \alpha_h + \beta_k + (\alpha\beta)_{hk}, \tag{1}$$

$$h = 1, 2, \dots, S, \quad k = 1, 2, \dots, T,$$

where μ denotes the general parameter, α_h denotes the effect of the h -th whole plot treatment, β_k denotes the effect of the k -th subplot treatment and $(\alpha\beta)_{hk}$ stand for interaction effects. All the treatment effects are considered to be fixed.

The linear model for the observed yield has then the form:

$$y_{rijt(hk)} = \mu + \tau_{hk} + \rho_r + \eta_{ri} + \varepsilon_{rij} + \theta_{rijt} + e_{rijt(hk)}, \tag{2}$$

$$r = 1, 2, \dots, b, \quad i = 1, 2, \dots, k_1, \quad j = 1, 2, \dots, k_2,$$

$$h = 1, 2, \dots, S, \quad k = 1, 2, \dots, T, \quad t = 1, 2, \dots, p$$

where $y_{rijt(hk)}$ denotes the t -th observation from the (r, i, j) unit (j -th subplot within the i -th whole plot within the r -th block), on which occurs the (h, k) -th treatment combination, τ_{hk} denotes the effect of the (h, k) -th treatment combination, ρ_r denotes the effect of r -th block, η_{ri} stands for the effect of the i -th whole plot in the r -th block (whole plot error), ε_{rij} denotes the effect (subplot error) of the j -th subplot within the i -th whole plot in the r -th block, θ_{rijt} denotes the error connected with the t -th observation on the (r, i, j) unit, e_{rijt} denotes the technical error connected with measurements on the (r, i, j, t) -th unit and p denotes the size of subsample in the (r, i, j) -th unit. All the effects (apart from the treatment effects) are random in model (2).

The statistical properties of the variables $\rho_r, \eta_{ri}, \varepsilon_{rij}, \theta_{rijt}$ result from the randomization schemes applied while the assumptions concerning the technical errors, e_{rhjt} are assumed. They all affect the expected value and the dispersion structure of the observed yield as follows:

$$E(y_{rijt(hk)}) = \mu + \tau_{hk}, \tag{3}$$

$$Cov(y_{rijt(hk)}, y_{r'i'j't'(hk)}) =$$

$$\begin{cases} \sigma_\rho^2 + \sigma_\eta^2 + \sigma_\varepsilon^2 + \sigma_{\theta_{rijt}}^2 + \sigma_e^2, & r = r', \quad i = i', \quad j = j', \quad t = t' \\ \sigma_\rho^2 + \sigma_\eta^2 + \sigma_\varepsilon^2 - (p-1)^{-1} \sigma_{\theta_{rijt'}}^2, & r = r', \quad i = i', \quad j = j', \quad t \neq t' \\ \sigma_\rho^2 + \sigma_\eta^2 - (k_2-1)^{-1} \sigma_\varepsilon^2, & r = r', \quad i = i', \quad j \neq j', \\ \sigma_\rho^2 - (k_1-1)^{-1} \sigma_\eta^2, & r = r', \quad h \neq h', \\ -(b-1)^{-1} \sigma_\rho^2, & r \neq r', \end{cases} \tag{4}$$

where the random variance components σ_ρ^2 , σ_η^2 , σ_ϵ^2 , $\sigma_{\theta rijt}^2$, σ_e^2 denote the block variance, the whole plot variance (Error I), the subplot variance (Error II), the sampling variance of the (rij) - th unit and the technical error variance, respectively.

2 Remarks on statistical analysis

In the experiments with subsamples it is necessary to take into account two kinds of errors. The first error, say experimental error, is one appropriate for treatment comparison while the second error, called observational error, measures variability within the subplot. At the beginning it is necessary to check hypothesis that subsample variances in all subplots are equal. This hypothesis can be written as follows:

$$H_0 : \sigma_{\theta 111}^2 = \sigma_{\theta 112}^2 = \dots = \sigma_{\theta bk_1k_2}^2.$$

The above hypothesis we can verify by F_{max} statistic (cf. Winer, 1971). In the case "fail to reject H_0 " we can check adequacy of the model by splitting the experimental error into two parts, i.e. within sub-subplot error and a rest.

Let us assume that the incomplete split plot design has orthogonal block structure (cf. Nelder, 1965). Applying appropriate analysis of variance for multistratum experiments with orthogonal block structure the overall analysis can be split into the stratum analyzes.

In this case, we have zero stratum (0) connected with mean estimation, inter-block stratum (1), inter- whole plot (within the block) stratum (2), inter-subplot (within the whole plot) stratum (3). The analyzes in the first and second strata do not depend on sampling considered. So, they are based on proper experimental unit means. The sampling has influence on the analysis in the third stratum only.

Let $SSE3(= SSsamp + SSrest)$ denote the sum of squares in the subplot stratum with ν_3 degrees of freedom. The sum of squares for samples, $SSsamp$, we calculate by summing the sums of squares for units within subplots. The $SSsamp$ has $\nu_s = bk_1k_2(p - 1)$ degrees of freedom. The sum of squares for the rest is equal to $SSrest = SSE3 - SSsamp$ and it has $\nu_r = \nu_3 - \nu_s$ degrees of freedom.

To check adequacy of the linear model it is necessary to calculate the F value as the ratio of mean squares, $F_c = MSrest/MSsamp$, which we compare with tabular F_T value for ν_r and ν_s degrees of freedom. When $F_c > F_T$ we use the $MSsamp$, as the denominator in the F statistic that tests hypothesis concerning treatment effects in the inter-subplot stratum. In other case instead of $MSsamp$, we use $MSE3$ in the F test.

References

- Meta, S. (1987). Experiments in incomplete split-plot designs. *Proc. of the Second. International Tampere Conf. in Statistics. Eds T. Pukkila and S. Puntanen. Univ. of Tampere, 575-584.*
- Meta, S. (1994). On modelling of experiments in natural sciences. *Biometrical Letters* **31**, 79-100.

Nelder, J.A. (1965). The analysis of randomized experiments with orthogonal block structure. *Proc. of the Royal Soc. of Lond. Ser. A* **283**, 147-178.

Neyman, J., Iwazkiewicz, K., Kolodziejczyk, S. (1935). Statistical problems in agricultural experiments (with discussion). *J. Royal Statist. Soc. Suppl.* **2**, 107-180.

Winer, B.J. (1971). *Statistical Principles in Experimental Design, 2nd Ed.* McGraw-Hill, Inc.

On a modelling environmental indexes

Stanisław Mejza¹, João T. Mexia², and Dulce Pereira³

¹ Department of Mathematical and Statistical Methods, Agricultural University, Wojska Polskiego 28, 60-637 Poznań, Poland

² Departamento de Matemática, Universidade Nova de Lisboa, Quinta da Torre, 2825 Monte da Caparica, Portugal

³ Departamento de Matemática, Universidade de Évora, CIMA, Colégio Luís António Verney, Rua Romão Ramalho 59, 7000-671 Évora, Portugal

Abstract: The paper deals with the structuring the Genotype x Environmental Interaction in an analysis of series of experiments. The analysis of regression is one of the most appropriate methods in this problem. As in regression analysis we should have two sets of variables, one characterizing genotypes while the second characterizing environments. The so-called adjusted means for genotypes constitute usually observations of dependent variable. The problem is how to model the environmental indexes being the observation of independent variable. In the paper we examine two approaches to a modelling the environmental indexes, one is based on so called adjusted means for environments while second method uses iterative (called zig-zag) algorithm for estimation of the considering indexes.

Keywords: genotype indexes, environmental indexes, adjusted means, genotype x environment interaction, zig zag algorithm

1 Introduction

Let us consider data arranged in a two-way array with b rows and J columns. The analysis of this data can be done without any reference to some applications. But in the paper we will refer the data to the series of agricultural experiments in which a set of the J genotypes were examined over the set of the b environments. The purpose of such experiments is the selection of genotypes that are consistently high yielding over the range of observed or potential environments. This is connected mainly with the analysis of Genotype x Environment Interaction (GEI) although the other sources of variation are statistically and agronomically important as well. The main problem of inference from series of experiments is connected with modelling (structuring) the GEI effects. Usually, the GEI are nonorthogonal. Hence, to its analysis it is necessary to use very advanced statistical tools (cf. e.g. Aastveit and Meta, 1992). In the paper our interest to GEI analysis is limited to two cases. First one is some modification of analysis of two-way table as we are doing in incomplete block designed experiments. The second approach is based on some application of joint regression analysis.

2 Adjusted environmental effects

In this approach the GEI will be expressed by fixed additive model with fixed effects of genotypes and environments and random error term.

Let the Y_{ij} denote the observation obtained for the i -th environment and the j -th genotype (treatment) which can be modelled as:

$$Y_{ij} = \mu + \tau_i + \beta_j + e_{ij}, \quad i = 1, 2, \dots, b, \quad j = 1, 2, \dots, J,$$

where μ denotes the general mean τ_i - the effect of the i -th environment, β_j - the effect of the j -th genotype and finally, e_{ij} - denotes the error.

In the matrix notation the above model can be written as:

$$Y = 1\mu + \Delta'\tau + D'\beta + e$$

where 1 denotes the vector of ones, Δ' and D' are design matrices for environments and genotypes corresponding to the τ and β - vectors of environment and genotype effects, e -denotes the vector of errors. In this approach the results known in the theory of block designs are used. Then so called adjusted means for environments can be calculated as:

$$\tilde{\beta} = \tilde{\mu} + G^{-1}Q,$$

where $\tilde{\mu} = n^{-1}Y'1$ - denotes the general mean, $G = k^\delta - Nr^{-\delta}N'$ - is the information matrix for estimation the environmental effects, $N = \Delta D'$ - denotes the environment x genotype incidence matrix, $k = N1$, $r = N'1$, $n = r'1$, $k^\delta = \text{diag}(k_1, k_2, \dots, k_b)$, $r^{-\delta} = \text{diag}(1/r_1, 1/r_2, \dots, 1/r_J)$, $Q = T - Nr^{-\delta}B$, $B = DY$, $T = \Delta Y$, G^{-1} - denotes the generalized inverse of the matrix G

3 Environmental index

In this approach we use joint regression to structuring (modelling) GEI (multiplicative model). The observation Y_{ij} is modelled as:

$$Y_{ij} = \alpha_j + \beta_j x_i + e_{ij}, \quad i = 1, 2, \dots, b, \quad j = 1, 2, \dots, J,$$

where the (α_j, β_j) , $j = 1, \dots, J$ are the regression coefficients and the x_i , $i = 1, \dots, b$ are the environmental indexes. The main problem in such modelling is how to estimate the parameters. One can observe that the lately proposed so called zig-zag algorithm is very efficient in finding the estimates of (α_j, β_j) and the x_i (cf. Mexia et al., 1999, Pereira and Mexia, 2002, 2003a, 2003b).

In this approach the following goal function is minimized

$$S(\alpha^J, \beta^J, x^b) = \sum_{i=1}^b \sum_{j=1}^J p_{ij} (Y_{ij} - \alpha_j - \beta_j x_i)^2.$$

Usually the weight p_{ij} is 1 [0] when cultivar j is present [absent] in the i -th environment. In the zig-zag algorithm the minimization is carried out iteratively. At the beginning it is recommended to start with some initial values for indexes. In the complete case, i.e., when all genotypes occur in each environment, the average yield for environment can be a good initial values (cf. Gusmao, 1985). In the worse case any initial values can

be taken (cf. Pereira, 2004). Then the goal function is minimized with respect to other parameters after fixing some of them in previous iteration. The process converge always but its time depends on initial values. At the end of each iteration the environmental indexes are rescaled so that the range of environmental indexes is kept unchanged. Hence, the iteration procedure is called zig-zag algorithm.

The aim of the paper is to compare the above shortly described two approaches to estimation of environmental indexes. It is impossible to compare them analytically. The comparison, to some extent, will be based on a few examples.

4 Examples

Example 1.

The starting experiment includes 20 genotypes of rye observed in 32 environments in Poland. In the paper we will compare the approaches on the basis of yield/plot observed genotypes. The data are represented in matrix of 32 environments (rows) and 20 genotypes (columns). It means that starting point experiment was complete. Then from that data set we removed $1/5$, $1/2$ and $3/5$ of observations. This made structure of the data nonorthogonal. For these three data sets we have calculated environmental indexes by both approaches. Then the coefficient of correlations were calculated which were: 0.9999, 0.9977 and 0.9986 respectively. In this examples we observe very high correlation.

Example 2.

In the second example we use the observations from series of experiments with rye in which the yield of 20 genotypes was observed in 20 environments in Poland. In each of the environment exactly 4 genotypes were observed. Together we had 80 observations. The same genotypes were observed in the same environments but in two years. The correlation coefficients between the environmental indexes calculated by two methods considered were as follows: 0.827 and 0.418.

Let us note that all correlation coefficients are significant at the significance level 0.05 but in the last case the correlation coefficient is much smaller than in other cases.

5 Discussion

In the paper the examples with rye were considered only. Rye belongs to a quiet stable variety over different environments. Hence, probably there is very good correspondence between environmental indexes obtained by both methods. In the last case the GEI was a little higher and immediately the correlation coefficient was smaller but still significant at 0.05 significance level. Our observations suggest that this method is more suitable in the case when the environments are non homogenous. Another comparison can be connected with calculations. It seems to us that receiving the environmental indexes (Section 3) is much easier and numerically more efficient. The calculation of the generalized inverse, in the case of big number of environments is numerically difficult and it is biased by numerical errors. This is a very weak point of that method.

Acknowledgments: The Research Centre for Cultivar Testing (Slupia Wielka, Poland) is thanked for use of their data in this paper. The third author of this work is member of the CIMA-UE, research center financed in the ambit of FEDER by “Programa de financiamento Plurianual”, of the Science and Technology Foundation - Portugal.

References

- Aastveit, A., and Meta, S. (1992). A selected bibliography on statistical methods for the analysis of genotype x environment interaction, *Biuletyn Oceny Odmian* **24-25**, 83-97.
- Gusmão, L. (1985). An adequate design for regression analysis of yield trials. *Theor. Appl. Genet.* **71**, 314-319.
- Mexia, J.T., Pereira, D.G., and Baeta, J. (1999). L_2 Environmental indexes. *Biometrical Letters* **36**, 137-143.
- Pereira, D.G., and Mexia, J.T. (2002). Multiple comparison in Joint Regression Analysis with special reference to variety selection. *Scientific papers of the Agricultural University of Poznan, Agriculture*, **3**, 67-74.
- Pereira, D.G., and Mexia, J.T. (2003a). Reproducibility of Joint Regression Analysis. *Colloquium Biometryczne* **33**, 279-299.
- Pereira, D.G., and Mexia, J.T. (2003b). The use of Joint Regression Analysis in selecting recommended cultivars. *Biuletyn Oceny Odmian (Cultivar Testing Bulletin)*, **31**, 19-25.
- Pereira, D.G. (2004). *Análise Conjunta Pesada de Regressões em Redes de Ensaio*. Ph'd Thesis. Universidade de Evora.

Identifiability of causal models with ignorable assignments and non-ignorable treatments

Andrea Mercatanti¹

¹ Viale Montegrappa 81, 59100 Prato, Italy. mercatan@libero.it

Abstract: this paper examines the identification problem that arise when evaluating causal treatment effects in observational studies with ignorable assignments to a binary treatment and where the non-ignorability of the treatment is due to an imperfect compliance to the assignments.

Keywords: causal inference; identifiability; ignorability.

1 The model

The concept of ignorability adopted in this paper is that characterizing the Rubin Causal Model (Holland, 1986), that is an independence condition between the assignment to a binary treatment and the couple of potential outcomes. We do not restrict the assignment to have only treatment-mediated effects on the outcome as usually done when evaluating causal non-ignorable treatment effects by the instrumental variables method. In order to characterize the likelihood function, other than supposing the assignment to be ignorable, we assume: the Stable Unit Treatment Value Assumption, a non-zero effect of the assignment on the outcome, and the absence of units doing always the opposite of their assignments (Angrist et al., 1996). Under this set of assumptions the likelihood is characterized by the presence of two mixtures of distributions; we consider the identifiability when the outcome distributions of various compliance statuses are in the same class. In this case the distribution function is in the parametric class, (Mercatanti, 2006):

$$\mathcal{F}' : \left\{ f(y_i, d_i, z_i; \theta) : I_{\zeta(1,0)} \cdot (1 - \pi) \cdot \omega_a \cdot g_{a0}^i + I_{\zeta(0,1)} \cdot \pi \cdot \omega_n \cdot g_{n1}^i + I_{\zeta(1,1)} \cdot \pi \cdot (\omega_a \cdot g_{a1}^i + \omega_c \cdot g_{c1}^i) + I_{\zeta(0,0)} \cdot (1 - \pi) \cdot (\omega_n \cdot g_{n0}^i + \omega_c \cdot g_{c0}^i) \mid \theta \in \Theta \right\}$$

where

$$\Theta : \left\{ \theta : (\pi, \omega, \eta) \mid \sum_{t: a, n, c} \omega_t = 1; \omega_t > 0, \forall t; 0 > \pi > 1 \right\}$$

and where: $I_{(\cdot)}$ is an indicator function; $\zeta(d, z)$ is the group of the units assuming treatment d and assigned to the treatment z ; π is the probability $P(z_i = 1)$; $\omega : (\omega_a, \omega_n, \omega_c)$, where ω_t is the probability of an individual being in the t group, $t : a$ (*always-takers*), n (*never-takers*), c (*compliers*); the function $g_{tz}^i : g_{tz}(y_i; \eta_{tz})$ is the outcome distribution for a unit in the t group and assigned to the treatment z ; $\eta : (\eta_{a1}, \eta_{a0}, \eta_{n1}, \eta_{n0}, \eta_{c1}, \eta_{c0})$.

2 Identifiability

Mixture models can present particular difficulties with identifiability, consequently the study of identifiability for the parametric class \mathcal{F}' , that involves two mixtures, is not straightforward. In order to explain the reasons of these difficulties, let's consider the general class of distribution functions from which the two mixtures are to be formed:

$$\mathcal{G} : \{g(y_i; \eta) | \eta \in \Upsilon, y_i \in R\} \quad (1)$$

and the general class of distribution functions of two-components mixtures of (1):

$$\mathcal{F}'' : \left\{ f(y_i, \theta) : \sum_{h=1}^2 \omega_h \cdot g(y_i; \eta_h) | g(\cdot; \eta_h) \in \mathcal{G}, \forall h; y_i \in R; \theta \in \Theta \right\} \quad (2)$$

where

$$\Theta : \{\theta : (\omega_1, \omega_2, \eta_1, \eta_2) | (\omega_1 + \omega_2) \leq 1, \omega_1 > 0, \omega_2 > 0; \eta \in \Upsilon\}.$$

In general a parametric family of densities $\mathcal{E} : \{e(y; \lambda) : \lambda \in \Lambda, y \in R\}$ is identifiable if distinct members of the parameter space Λ always determine distinct members of the family: $e(y; \lambda') \equiv e(y; \lambda'') \Leftrightarrow \lambda' = \lambda''$. It is well known (Titterington et al., 1985; McLachlan and Peel, 2000) that (2) is not identifiable, since $f(y; \theta)$ is invariant under the two permutations of the component labels h in θ . Indeed, the presence of two densities in the same class, $g(y; \eta_1)$ and $g(y; \eta_2)$, implies that $f(y; \theta) \equiv f(y; \theta^*)$ if the component labels 1 and 2 are interchanged in θ^* compared to θ . Titterington et al. (1985) propose a weak definition of identifiability for finite mixtures of distribution in the same parametric class by which a class of mixtures is identifiable if distinct members of the parameter vector Θ always determine distinct members of the family up to permutations of the label components. Under their definition, (2) is identifiable if and only if \mathcal{G} is a linearly independent set over the field of real number R .

However, and contrarily to an analysis of the mixture model $f(y_i, \theta) \in \mathcal{F}''$ at cluster purposes, we show the components labelling matters for $f(y_i, d_i, z_i; \theta) \in \mathcal{F}'$ at causal inference purposes. In order to study the identifiability of \mathcal{F}' , the more general class will be introduced:

$$\begin{aligned} \mathcal{M} : \{ & m(y, \mathbf{x}; \theta) : I_{(\mathbf{x} \in A_1)} m_1(y; \theta) + I_{(\mathbf{x} \in A_2)} m_2(y; \theta) + \dots + I_{(\mathbf{x} \in A_j)} m_j(y; \theta) \\ & + \dots + I_{(\mathbf{x} \in A_k)} m_k(y; \theta) | y \in R, \mathbf{x} \in A \subseteq R^d, A = \cup_j A_j, \cap_j A_j = \emptyset \} \end{aligned} \quad (3)$$

where the k distributions $m_j(y; \theta)$ are not necessarily in the same parametric class. A first useful result is proposed in the following proposition.

Proposition 1: a necessary and sufficient condition for parametric class (3) to be identifiable is that set $\Xi = \cap_j \Xi_j = \emptyset$; where Ξ_j is the set of pairs (θ', θ'') , $\theta' \neq \theta'' \in \Theta$ such that $m_j(y; \theta') \equiv m_j(y; \theta'')$.

Proof (Necessity): suppose that $\Xi = \cap_j \Xi_j \neq \emptyset$, then $m_j(y; \theta') \equiv m_j(y; \theta'')$, $\forall j$ and $\forall (\theta', \theta'') \in \Xi$. Consequently $m(y, \mathbf{x}; \theta') = \sum_j I_{(\mathbf{x} \in A_j)} m_j(y; \theta') \equiv \sum_j I_{(\mathbf{x} \in A_j)} m_j(y; \theta'') = m(y, \mathbf{x}; \theta'')$, $\forall (\theta', \theta'') \in \Xi$, which implies that (3) is not identifiable.

Proof (Sufficiency): if $\Xi = \cap_j \Xi_j = \emptyset$, then do not exist pairs (θ', θ'') , $\theta' \neq \theta'' \in \Theta$ such that $m_j(y; \theta') \equiv m_j(y; \theta'')$, $\forall j$. Consequently $\exists y$ such that $m(y, \mathbf{x}; \theta') = \sum_j I_{(\mathbf{x} \in A_j)} m_j(y; \theta') \neq \sum_j I_{(\mathbf{x} \in A_j)} m_j(y; \theta'') = m(y, \mathbf{x}; \theta'')$ which implies that (3) is identifiable•

Parametric class \mathcal{F}' is a particular case of (3), with $k = 4$. Proposition 2 identifies the set Ξ for \mathcal{F}' under the assumption that parametric class of the outcome distributions is a linearly independent set over the field of real number R (we omit the simple but tedious proof).

Proposition 2: if, in \mathcal{F}' , the parametric class of outcome distributions \mathcal{G} is a linearly independent set over the field of real number R , then one of the following conditions holds for any pair $(\theta', \theta'') \in \Xi \neq \emptyset$, $\theta' \neq \theta'' \in \Theta$:

$$\omega'_a = \omega'_c = \omega''_a = \omega''_c,$$

or

$$\omega'_n = \omega'_c = \omega''_n = \omega''_c,$$

or

$$\omega'_a = \omega'_c = \omega'_n = \omega''_a = \omega''_c = \omega''_n \bullet$$

3 Conclusions

Given Propositions 1 and 2, a distribution function $f(y_i, d_i, z_i; \theta)$ in \mathcal{F}' is identifiable unless: $\omega_a = \omega_c$, or $\omega_n = \omega_c$, or $\omega_a = \omega_n = \omega_c$. This is a set of less restrictive conditions compared to a simple mixture model analysis where identifiability is assured only up to permutations of the label components. Furthermore this set of equality conditions for the mixing probabilities are easily testable given the assumptions usually adopted for identifying causal effects by the instrumental variables method.

It is worth to note the restriction on the parametric class of the outcome distributions \mathcal{G} , imposed in Proposition 2, rules out the case of a binary outcome. The parametric class of binomials $Bi(N, \theta)$, $0 < \theta < 1$, is indeed a linearly independent set on R if and only if $N \geq 2T - 1$, where N is the number of independent trials for each observation (Titterington et al., 1985). Given $T = 2$ for the two mixtures in \mathcal{F}' , the condition on N is not satisfied for a binary outcome, where $N = 1 < 2T - 1 = 3$. This implies that for a binary outcome Ξ could be greater than under $N \geq 2T - 1$. This is confirmed by an application to data from a randomized community trial of the impact of vitamin A supplements on children's survival (Imbens and Rubin, 1997). The authors made a likelihood analysis of this randomized experiment with non-compliance, a binary outcome, in absence of always-takers and removing the exclusion restriction. There was no a unique solution, rather the resulting likelihood function had a set-valued maximizer.

Acknowledgments: the author thanks Andrea Galassi for useful comments and suggestions.

References

- Angrist J.D., G.W. Imbens, D.B. Rubin (1996). Identification of causal effect using instrumental variables; *J.A.S.A.* **434**, 444-455.
- Holland P.W. (1986). Statistics and Causal Inference; *J.A.S.A.* **81**, 945-970.
- Imbens G.W., D.R. Rubin (1997). Bayesian inference for causal effects in randomized experiments with noncompliance; *The Annals of Statistics* **25**, 305-327.
- McLachlan G.J., D. Peel (2000). *Finite mixture models*. John Wiley and Sons, Inc.
- Mercatanti A. (2006). A constrained maximum likelihood estimation for relaxing the exclusion restriction in causal inference; in *Proc. of the 21st I.W.S.M.* J.Hinde et al. eds.
- Titterton D.M., A.F.M. Smith, U.E. Makov (1985); *Statistical analysis of finite mixture distributions*. J. Wiley & Sons, Inc.

A comparative study of nonparametric derivative estimators

John Newell¹ and Jochen Einbeck²

¹ Department of Mathematics, National University of Ireland, Galway, Ireland

² Department of Math. Sciences, Durham University, Durham DH1 3LE, UK

Abstract: Nonparametric derivative estimation has never attracted much attention as one gets the derivative estimates as “by-products” from a local polynomial or spline fit. However, these estimates often suffer from boundary effects and are very sensitive to outliers. Apart from this, the local polynomial estimators suffer from a systematic downward bias, as we will demonstrate. This article is intended to re-establish research interest in derivative estimation, and to guide the user who needs to work with one of the available packages.

Keywords: Derivatives; Splines; Kernels

1 Motivation

Nonparametric estimation of derivatives is important in a variety of disciplines. Specifically, when considering a regression problem of type $y_i = m(x_i) + e_i$, one is often not interested in $m(\cdot)$ itself, but rather in the relative change dm/dx of m when increasing or decreasing x by a small value dx . An important special case is when x represents time, in which the 1st derivative of m has the interpretation of a speed, and the 2nd derivative of an acceleration, which is of interest in the analysis of growth curves. However, the importance of estimating derivatives goes far beyond the end in itself. Often one relies on asymptotic approximations in order to obtain bias and variance estimates, confidence intervals, optimal bandwidths, etc., and these expressions usually involve derivatives of $m(\cdot)$, which are normally unknown and have to be estimated. A further field of application for derivative estimators are change point problems. For instance, when analyzing blood lactate data of elite athletes, one is interested in the workload at which the lactate level suddenly rises, which can be detected by finding the maximum of the 2nd derivative (Newell et al., 2005).

2 On nonparametric derivative estimation

There are two main approaches to nonparametric derivative estimation. Consider firstly local polynomials of degree p . The estimator of the j^{th} derivative $m^{(j)}(x)$ ($0 < j \leq p$) at point x is given by $\hat{m}^{(j)}(x) = j! \hat{\beta}_j(x)$ according to Taylor’s theorem, where $\hat{\beta}_j(x)$ is obtained by minimizing

$$\sum_{i=1}^n K \left(\frac{x_i - x}{h} \right) \left(y_i - \sum_{j=0}^p \beta_j(x) (x_i - x)^j \right)^2$$

in terms of the vector $(\beta_0(x), \dots, \beta_p(x))$. Thereby K is a kernel function and h the bandwidth controlling the degree of smoothing. Secondly, in spline smoothing, the usual way of estimating derivatives is to take the derivatives of the spline estimate. In other words, if $\hat{m}(x)$ is an estimate of $m(x)$, one considers $\frac{d^j}{dx^j} \hat{m}(x)$ as an estimator of $m^{(j)}(x)$. Several authors have pursued this idea, using splines with (Heckman & Ramsay, 2000) or without penalization.

As these ideas are quite simple, several papers published in the mid-nineties, particularly originating from the local polynomial smoothing community, gave the impression that the entire issue of nonparametric derivative estimation is solved, and as a result the research activity about this topic stalled to some extent. This is unfortunate, as most problems are treated rather cursorily in the literature and many open questions remain. For instance, Ramsay (1998) noted that ‘typically one sees derivatives go wild at the extremes, and the higher the derivative, the wilder the behavior’, and that further problems arise when it comes to smoothing parameter (bandwidth) selection, where CV and GCV can be ‘poor guides’. In the sequel, we discuss some of these issues in the framework of a comparative study.

3 Comparison of available routines

For illustration, we consider a data set generated by contaminating the function $m(x) = x + 2 \exp(-16x^2)$, $x \in [-2, 2]$, with very small Gaussian noise ($\sigma = 0.1$). A moderate outlier at the left boundary with coordinates $(-1.97, -1.75)$ and a further outlier at $(0.95, 0)$ were added by hand, giving a total sample size $n = 60$.

3.1 Local polynomial methods

We start with considering the functions `locfit` (contained in the homonymous package) and `locpoly` in package **KernSmooth**. We use the usual default setting $p = j + 1$ as theoretically motivated by Fan & Gijbels, 1996, p. 77ff. The bandwidths are chosen such that the curves pass roughly equally well through the central part of the curve (we used the result of `locfit`’s `gcvplot` for the 2nd derivative, but undersmoothed for the first). Both functions produce a considerable bias there, which cannot be cured by modifying the bandwidth as otherwise the outlier and boundary effects get even worse. In fact, there is a systematic problem with this kind of estimators: Note that the asymptotic bias of the derivative estimate based on a quadratic fit with bandwidth h is given by

$$\text{Bias}(\hat{m}'(x)|x_1, \dots, x_n) = c \cdot m'''(x)h^2 + o_P(h^2)$$

($c > 0$ being a constant depending on kernel moments), which can be deduced from Fan & Gijbels (1996), Th^m 3.1. This implies that, where $m'(\cdot)$ is concave, the bias is negative, and where $m'(\cdot)$ is convex, the bias is positive. Hence, concave parts of the derivative will be pulled down and convex parts will be pulled up. As the concave part will usually (but not necessarily in a mathematical sense) correspond to positive and the convex part to negative derivatives, we can speak of a *downward smoothing bias* similar as observed by Stoker (1993) for density derivative estimation. This bias, clearly visible in the left panel of Figure 1, tends to increase with the derivative order j ; one reason is that the necessary bandwidth h (appearing in the bias generally as a

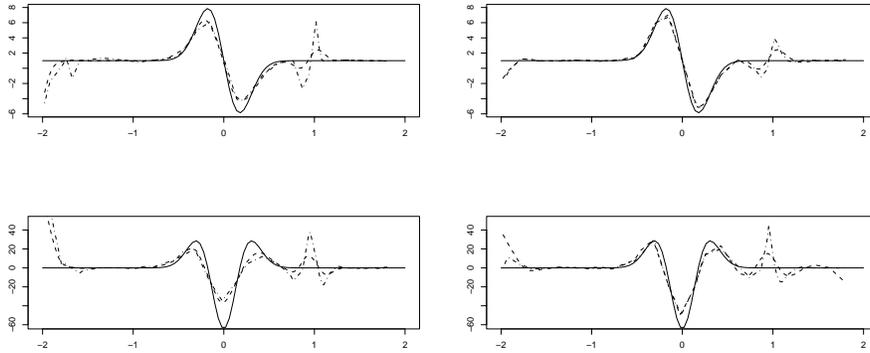


FIGURE 1. Tutorial on behavior of derivative estimators (top: 1st deriv., bottom: 2nd deriv.), left: `locpoly` (dashed), `locfit` (dashed-dotted); right: `smooth.Pspline` (dashed), `D1D2` (dashed-dotted).

factor h^{p+1-j}) increases with j . The smoothing bias diminishes when setting $p = j + 2$ as suggested by Ruppert (1997), at the expense of increased outlier and boundary effects (not shown).

3.2 Spline based methods

We consider here for comparison the functions `smooth.Pspline` (R package `pspline`) and `D1D2` (`sfsmisc`), both using penalized smoothing splines. The latter is restricted to cubic splines, whereas we use for the former a quintic and septic spline for the 1st and 2nd derivative, respectively (Ramsay, 1998). The smoothing parameter is selected for `smooth.Pspline` using the built-in GCV routine, and for `D1D2` such that the fits pass equally well through the central part. Both fits are much less biased than the local polynomial estimators, and more stable at the left boundary.

4 Conclusion

In the poster, we extend this study to include comparisons of the R packages `lokern`, `lpridge` (local), and `SemiPar` (splines). Given the overall stability, functionality, and performance (assessed through a small simulation study), our favorites are rather among the spline based methods; in particular `pspline` and `SemiPar` work generally quite well and offer several interesting options (notably, `SemiPar` is, apart from `locfit`, the only package featuring confidence bands). However, there is a general lack of *robust* derivative estimators. Further, smoothing parameter selection tools are in all packages based on optimizing the estimate of the *regression function* and not of the derivative, which can lead to serious undersmoothing (Jarrow et al., 2004). Function `D1D2` at least addresses this problem by adding a "fudge" offset to the GCV-selected smoothing parameter. As a brief guide, the capacities of the packages investigated are summarized below:

Package version	function	j_{max}	Smooth. Par.
locfit 1.5-3	locfit	2	GCV
KernSmooth 2.22-19	locpoly	no limit	—
lokern 1.0-4	glkerns	4*	plug-in**
lpridge 1.0-3	lpridge	9	—
pspline 1.0-10	smooth.Pspline	4***	CV/GCV
sfsmisc 0.95-9	D1D2	2	GCV
SemiPar 1.0-2	spm	7***	RE(ML)

*if bandwidth selected automatically, then $j_{max} = 2$. **a variant **lokerns** featuring a variable bandwidth is also implemented. ***no formal requirement, but from our experience it breaks down computationally for higher orders.

References

- Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.
- Heckman, N.E., and Ramsay, J. (2000). Penalized regression with model-based penalties. *The Canadian Journal of Statistics* **28**, 241-258.
- Jarrow, R., Ruppert, D., and Yu, Y. (2004). Estimating the term structure of corporate debt with a semiparametric penalized spline model, *JASA* **99**, 57-66.
- Newell, J., Einbeck, J., Madden, N., and McMillan, K. (2005). Model free endurance markers based on the second derivative of blood lactate curves. In: Francis et al. (Eds), *Proceedings of the 20th IWSM*, 357-364, Sydney.
- Ramsay, J. (1998). Derivative estimation. StatLib— S-News Thu, 12 March 1998, www.math.yorku.ca/Who/Faculty/Monette/S-news/0556.html
- Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *JASA*, **92**, 1049–1062.
- Stoker, T.M. (1993). Smoothing bias in density derivative estimation. *JASA* **88**, 855–863.

Censored partial regression models and the study of the determinants of survival of Russian commercial banks

Jesus Orbe¹ and Vicente Núñez-Antón²

¹ Departamento de Econometría y Estadística. Universidad del País Vasco, Lehendakari Aguirre, 83, E-48015 Bilboes, Spain. jesus.orbe@ehu.es

² Departamento de Econometría y Estadística. Universidad del País Vasco, Lehendakari Aguirre, 83, E-48015 Bilboes, Spain. vicente.nunezanton@ehu.es

Abstract: This paper studies the relevant factors in the survival of Russian commercial banks during the transition period. We propose a semiparametric AFT model that does not require to assume a distribution for the survival time and represents a relevant alternative to the PH model. Our proposal extends Stute's (1993) method by considering a partial censored regression model. This methodology represents a very general, flexible and alternative approach for this type of analysis.

Keywords: Banking; Censoring; Kaplan-Meier; Survival.

1 Introduction

The fast pace development of the Russian commercial banking industry has its origin in 1988, right after the banking reform. This rapid evolution of the market was the result of the null entry barriers that caused high rates of entry and that were consequently followed by a period of high rates of exit. As a consequence, many banks had to exit the market without refunding their deposits. Therefore, both for the banks and for the banks' depositors, it is of interest to analyze which are the relevant factors that motivate the exit or the closing of the bank. The survival of new firms has been previously studied, specially in the manufacturing industry, for firms in the United States by Audretsch and Mahmood (1995) and Dunne et al. (1989), Wagner (1994) for the German industry, Mata (1994) for Portugal, Arrighetti (1994) for Italy, Audretsch and Mahmood (2000) for the Netherlands, Segarra and Callejón (2002) and Esteve et al. (2004) for Spain. In the context of the banking industry, Shumway (2001), Wheelock and Wilson (2000), Whalen (1991) and Lane et al. (1986) used hazard models to analyze the bank failure process in the United States, and Carree (2003) used the same methodology for the Russian market. This paper proposes an alternative and flexible approach to study the direct effects of relevant factors on the survival time instead of on the hazard rate. In addition, and taking as a reference the analysis in Carree (2003), we propose the use of a different methodology that allows us to estimate the unknown functional form of a covariate's effect and that does not require to make any assumption on the probability distribution of the survival time.

2 Methodology

We use a regression model to evaluate the effect of k covariates, $\mathbf{z}_i = (z_{i1}, \dots, z_{ik})'$, on the survival time of the i -th bank, t_i :

$$\ln t_i = \mathbf{z}'_i \beta + u_i, \quad i = 1, \dots, N \quad (1)$$

One of the main reasons to consider this alternative model is the clear and easy interpretation of the results, because here we measure the direct effect of the covariates on the survival time instead of on the conditional probability of exit of the market at time t :

$$h(t | \mathbf{z}_i) = h_0(t) \exp(\mathbf{z}'_i \beta), \quad i = 1, \dots, N \quad (2)$$

Our proposed estimation method is based on the weighted least squares (WLS) methodology. Given that at the end of the study some banks do not leave the market, their survival times are censored and, as is well known, in the presence of censorship, the ordinary least squares methodology produces biased and inconsistent estimators. In order to solve this problem, we use the WLS approach where the weights take into account the effect of censorship:

$$\hat{\beta} = \left[\sum_{i=1}^N \mathbf{z}_{[i]} \mathbf{z}'_{[i]} w_{[i]} \right]^{-1} \left[\sum_{i=1}^N \mathbf{z}_{[i]} y_{(i)} w_{[i]} \right] \quad (3)$$

In this setting, and because of censorship, we have the observed duration variable Y . If the observation is not censored, we observe the duration of the bank in the market, $Y = T$. On the other hand, if the observation is censored, C , we only know that the duration is larger than this value, i.e. $T \geq C$. After ordering the observed duration, $y_{(i)}$ represents the i -th ordered observed duration; $\mathbf{z}_{[i]}$ is the $(k \times 1)$ vector that represents the values of the k covariates corresponding to $y_{(i)}$; and $w_{[i]}$ is the weight assigned to that bank that is calculated as the contribution of $y_{(i)}$ on the Kaplan-Meier estimator of the distribution function (Kaplan and Meier, 1958). An easy way to compute the weights $w_{[i]}$ can be described as:

1. Order the observed durations, $y_{(1)} < y_{(2)} < \dots < y_{(N)}$, and put the same weight, $\frac{1}{N}$, on all observations.
2. Start with the smallest value of Y , $y_{(1)}$. If it is not censored, keep its weight; otherwise, if it is censored, put zero weight on it and redistribute its weight between the observed durations that are larger than this observation.
3. Repeat step 2 until the largest value of Y , $y_{(N)}$.

This estimation method has been studied in Stute (1993) for linear regression models with censorship. In our case, we need a method that considers both linear and nonlinear effects. The motivation for this lies on the specification of the effect of the covariate representing the relative interest rate (*intrel*). Carree (2000) points out the possible quadratic relationship for this covariate on the hazard, indicating that the higher the

deposit interest rate offered by the bank, the higher its risk. In addition, banks which are offering low interest rates may not be interested in attracting customers and, thus, are trying to exit the market. Carree (2003) specified a quadratic relationship for this covariate but the estimation of the effect was not statistically significant. This paper allows for a general specification of the functional form of the effect of this covariate, without assuming any parametric form for it. Besides, it does not need to assume any distribution for the duration variable. Carree (2003) assumed a Gompertz basic hazard function $h_0(t)$. Our proposal will avoid any incorrect specification of the model and its consequences (see, e.g., Hollander and Schumacher, 2006). We use the technique of penalized weighted least squares (PWLS) and, thus, the estimators are obtained by minimizing

$$\sum_{i=1}^N \{y_{(i)} - \mathbf{z}'_{1[i]}\beta - f(z_{2[i]})\}^2 w_{[i]} + \alpha \int f''(z_2)^2 dz_2 \tag{4}$$

with respect to β and $f(\cdot)$. In equation (4), $\mathbf{z}_{1[i]}$ is a $(k - 1) \times 1$ vector of $(k - 1)$ linear covariates for the i -th bank and $z_{2[i]}$ takes the value of the covariate introduced in a nonparametric form (i.e., the nonlinear covariate). That is, there is no functional assumption on $f(\cdot)$. Equation (4) contains a penalized term that considers the “smoothness” of $f(\cdot)$, by using the integrated squared second derivatives. The parameter α controls for the relevance of the goodness-of-fit and for the smoothness of $f(\cdot)$. The purpose is to have a reasonable goodness-of-fit together with a smooth $f(\cdot)$. The effect of the censorship is handled with the $w_{[i]}$ weights. Orbs et al. (2003) studied this estimation process and used a simulation study to check on the properties of the proposed methodology.

3 Results and Conclusions

As can be seen in Table 3, the results are similar to Carree (2003); that is, smaller banks exit sooner. In addition banks that enter the market in the last part of the study have shorter durations. As expected, the more experienced banks have a longer duration in the market.

TABLE 1. Estimates of the linear covariates effects

Covariate	Coefficient estimate	95% Confidence Interval
<i>invtam</i>	-0.0003	(-0.00046, -0.00017)
<i>expe</i>	0.0707	(0.05312, 0.11671)
<i>tiempo</i>	-0.1053	(-0.15881, -0.07399)

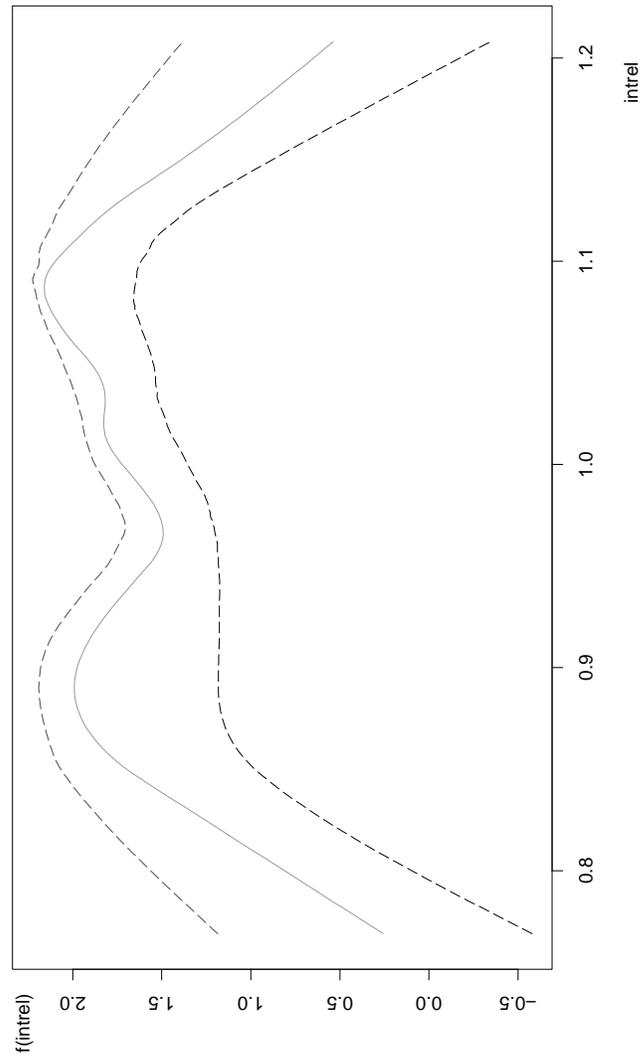


FIGURE 1. Estimate of the effect of the nonlinear covariate: *intrel*

The risk increases and the survival time decreases with the interest rate. This risk is measured by using the ratio between the interest rate offered by the bank and the mean interest rate offered at the market. This effect is captured in the nonparametric term and we obtain a U inverted functional form effect on the duration (see Figure 1). This agrees with the U quadratic effect (on the hazard function) obtained in Carree (2003) but, in his case, it was not significant. Therefore, this functional form shows that the mean survival time is lower for banks which offer extreme interest rates. It

seems that banks which offer the highest interest rates have more risk and exit sooner, and banks which offer the lowest rates are reflecting that they are not interested in attracting customers and are trying to exit the market.

As a conclusion we have to add that the results obtained strengthen the ideas put forward in Carree (2003) and add new insights to them. However, what is really important is that these results have been derived by using an alternative approach and applying a very general model. On the one hand, instead of assuming a quadratic form for the functional form of the effect of interest ratio, we introduce it in a nonparametric form. On the other hand, we do not assume any probability distribution for the survival time. Therefore, this methodology is robust under any incorrect specification of the functional form effects and also under the incorrect specification of the distribution probability and its consequences.

Acknowledgments: Research supported by BEC2003-02028, MTM 2004-00341 and MTM2006-06550 grants of the Ministerio de Ciencia y Tecnología, by Universidad del País Vasco 9/UPV 00038.321-13631/2001 and by the Econometrics Research Group UPV/EHU GIU06/77.

References

- Audretsch, D.B. and Mahmood, T. (1995). New firm survival: new results using a hazard function. *Review of Economics and Statistics* **77**, 97-103.
- Audretsch, D.B. and Mahmood, T. (2000). Firm survival in the Netherlands. *Review of Industrial Organization* **16**, 1-11.
- Carree, M.A. (2000). Interest and hazard rates of Russian saving banks. *ERIM Report Series* ERS-2000-26-STR.
- Carree, M.A. (2003). A hazard rate analysis of Russian commercial banks in the period 1994-1997. *Economic Systems* **27**, 255-269.
- Dunne, T., Roberts, M.J. and Samuelson, L. (1989). The growth and failure of US manufacturing plants. *Quarterly Journal of Economics* **104**, 671-698.
- Evans, D.S. (1987). Test of alternative theories of firm growth. *Journal of Political Economy* **104**, 671-698.
- Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica* **50**, 649-670.
- Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457-481.
- Orbs, J. (2000). Un modelo de regresión parcial censurado para análisis de supervivencia. Tesis Doctoral, Departamento de Econometría y Estadística. Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Bilboes.

- Orbs, J., Ferreira, E. and Núñez-Antón, V. (2003). Censored partial regression. *Biostatistics* **4**, 109-122.
- Segarra, A. and Callejón, M. (2002). New Firms' survival and market turbulence: new evidence from Spain. *Review of Industrial Organization* **20**, 1-14.
- Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of multivariate Analysis* **45**, 89-103.
- Stute, W. (1999). Nonlinear censored regression *Statistica Sinica* **9**, 1089-1102.
- Whalen, G. (1991). A proportional hazards model of bank failure: an examination of its usefulness as an early warning tool. *Economic Review* (Federal Reserve Bank of Cleveland) **27**, 21-31.
- Wheelock, D.C. and Wilson, P.W. (2000). Why do banks disappear? The determinants of US bank failures and acquisitions. *Review of Economics and Statistics* **82**, 127-138.

Copulas and their extremal transformations

M.I. Ortego¹ and J.J. Egozcue¹

¹ Departament de Matemàtica Aplicada III. Universitat Politècnica de Catalunya, Jordi Girona Salgado 1-3, E-08034 Barcelona (Spain).
ma.isabel.ortego@upc.edu

Abstract: In hazard assessment problems, events are assumed to occur as a Poisson process, and these events are measured by some kind of random magnitude in which the interest is focused. A problem of general concern is to describe dependence between two of these magnitudes. This dependence is modelled using copula functions. The main goal is to find out the transformation of such a copula when attention is restricted to excesses of the magnitudes over some determined thresholds, one for each magnitude; and when maxima of each variable are extracted in a fixed time.

Keywords: threshold; extremes; POT method

1 Introduction

Understanding relationships between multivariate events is a basic problem in Statistics. Dependence of random variables has been often described using correlation, although this parameter does not describe the full dependence of the variables and is often misused in contexts where the normal distribution does not play the main role. Copulas arise in this context (Nelsen 1999), characterizing joint (multivariate) distributions independently of marginal distributions.

Hazard modelling often deals with extremal events which probability distributions are skewed and non-normal and, consequently, their dependence is not fully represented by correlation. Therefore, copula functions seem to be a proper representation of dependence in this context (e.g. Coles and Tawn 1994; de Haan and de Ronde 1998).

Statistical methods dealing with hazardous phenomena are often affected by lack of data. As the size of the events increases, the occurrence rate of events decreases. When dealing with large size events, very few data are normally available to estimate the dependence between two or more characteristics of the event and its estimations are uncertain. A strategy to face this problem may be to study the dependence of such characteristic magnitudes for small size events and, then, extrapolate those features to larger sizes of events, provided they correspond to the same phenomenon. The key point is to model how dependence of the considered magnitudes is transformed for larger sizes.

The present aim is to study the change of the copula describing the dependence of excesses over a threshold when the threshold is increased. Also the copula describing dependence of maxima is transformed into a new copula when the time of extraction of maxima changes. Both transformations have been called extremal. A standard model

used in these situations is a Poisson point process. Each event is characterized by two random variables, or magnitudes. Then, the process of events and their magnitudes constitute a doubly marked Poisson processes. The above mentioned extremal transformations are studied in this context.

1.1 A model of extremes in marked Poisson processes

In the present approach, events are modelled as points in time. Interest is normally centered in occurrence probabilities of a given number of events within a time t . Particularly, attention is paid to certain classes of events whose magnitudes satisfy some condition. For instance, rainfall events whose precipitation in 5 minutes was more than 5mm per m^2 or precipitation in 30 minutes was more than 20mm per m^2 . Magnitudes X_1, X_2 are assumed to be identically distributed from event to event and their joint cumulative distribution function (cdf) is $F_{X_1 X_2}(x_1, x_2)$.

Moreover, these sizes are also assumed to be independent from event to event and from time occurrence.

This model determines the occurrence probabilities of such a class of events in a standard way. If A is an event in the (X_1, X_2) space, and $N(A)$ the number of events whose magnitudes X_1, X_2 are in A , the probability function is given by

$$P[N(A) = n \mid \lambda(A), t] = \frac{[\lambda(A)t]^n \exp[-\lambda(A)t]}{n!}, \quad n = 0, 1, 2, \dots \quad (1)$$

where $\lambda(A) = \lambda_0 P[A]$, t is the observation time, and λ_0 is the Poisson rate of events in the underlying process.

As frequently done when dealing with hazard problems, special interest is centered on the extremal behavior of magnitudes. Point over threshold (POT) methods (e.g. Embrechts et al. 1997) define the excess of a variable over a given threshold. Similarly, define bivariate excesses of the variables (X_1, X_2) over the bivariate threshold (h_1, h_2) as $(Y_1 = X_1 - h_1, Y_2 = X_2 - h_2)$, provided that $X_1 > h_1$ and $X_2 > h_2$. The change distribution of X_i under change of threshold is well-known in the univariate case, and the present approach uses an extension to the bivariate case.

For extraction of maxima from observations in a fixed time t , assume that N is the random number of events in this time t , and define Z_i , $i = 1, 2$, the maximum of the X_i corresponding to those events. Again univariate expressions relating F_{X_i} and F_{Z_i} are well known; the present approach uses the bivariate extension of such relationship.

1.2 Copulas

Copulas are bivariate (multivariate) distributions with uniform marginals. Copula functions are able to fully represent dependence between two or more variables, independently of their marginal distributions. The use of copula functions allows to treat separately marginal models and dependence of variables. The joint cdf of X_1, X_2 is modelled by a copula $C_{X_1 X_2}[\cdot, \cdot]$ (Nelsen, 1999), i.e.

$$F_{X_1 X_2}(x_1, x_2) = C_{\mathbf{X}}[F_{X_1}(x_1), F_{X_2}(x_2)], \quad (2)$$

where the subscript \mathbf{X} denotes (X_1, X_2) . This completes the model.

2 Transformation of copulas

A first goal is to study the relationship between the copula in eq. (2) and copulas of excesses over a given threshold. The copula corresponding to the joint cdf of the excesses (Y_1, Y_2) is denoted $C_{\mathbf{Y}(\mathbf{h})}$, i.e.

$$F_{Y_1 Y_2}(y_1, y_2 | X_1 > h_1, X_2 > h_2) = C_{\mathbf{Y}(\mathbf{h})}[F_{Y_1}(y_1), F_{Y_2}(y_2)] , \tag{3}$$

where the subscript $\mathbf{Y}(\mathbf{h})$ points out that excesses, (Y_1, Y_2) , are taken over the bivariate threshold $\mathbf{h} = (h_1, h_2)$. The expression which relates dependence of excesses and dependence of magnitudes is proved to be

$$C_{\mathbf{Y}(\mathbf{h})}[v_1, v_2] = \frac{C_{\mathbf{X}}[\eta_1, \eta_2] + C_{\mathbf{X}}[u_1, u_2] - C_{\mathbf{X}}[u_1, \eta_2] - C_{\mathbf{X}}[\eta_1, u_2]}{1 + C_{\mathbf{X}}[u_1, u_2] - u_1 - u_2} , \tag{4}$$

where η_i is an implicit function of u_i and v_i depend also on $C_{\mathbf{X}}$.

The second goal is to obtain the relationship between the copula of magnitudes and the copula of their maxima of observed events within a time t . Denote Z_1, Z_2 , the maxima of the magnitudes X_1, X_2 respectively for events recorded in a fixed time t . The joint cdf of Z_1, Z_2 is modelled by a copula $C_{\mathbf{Z}}$, i.e.

$$F_{Z_1 Z_2}(z_1, z_2) = C_{\mathbf{Z}}[F_{Z_1}(z_1), F_{Z_2}(z_2)] ,$$

where the subscript \mathbf{Z} is used for (Z_1, Z_2) . Transformation of $C_{\mathbf{X}}[\cdot, \cdot]$ into $C_{\mathbf{Z}}[\cdot, \cdot]$ is given by two equivalent expressions:

For $w_1 \geq w_{01}$ and $w_2 \geq w_{02}$,

$$C_{\mathbf{Z}}[w_1, w_2] = \exp \left[-\lambda t \left(1 - C_{\mathbf{X}} \left[1 + \frac{\ln w_1}{\lambda t}, 1 + \frac{\ln w_2}{\lambda t} \right] \right) \right] , \tag{5}$$

and,

$$C_{\mathbf{Z}}[\exp(-\lambda t(1 - u_1)), \exp(-\lambda t(1 - u_2))] = \exp[-\lambda t(1 - C_{\mathbf{X}}[u_1, u_2])], \tag{6}$$

using alternatively u_i or w_i as arguments, where $w_i = \exp(-\lambda t(1 - u_i))$ or, conversely, $u_i = 1 + \ln w_i / (\lambda t)$, for $i = 1, 2$. For $w_1 < w_{01}$ and $w_2 < w_{02}$, the copula can be chosen arbitrarily just preserving monotonicity and uniform marginals.

3 Conclusions

Events in a Poisson process are sized by two random magnitudes, assumed identically distributed and independent from event to event. Interest is set to bivariate excesses over a threshold and maxima within an interval. The transformation of copulas describing the change of dependence between the two magnitudes for their excesses and maxima under change of threshold or time of extraction respectively have been established.

Acknowledgments: This research has been financially supported by the Spanish Ministry of Education and Science through the project MTM2006-03040 .

References

- Coles, S.G. and Tawn, J.A. (1994). Statistical methods for multivariate extremes: an application to structural design. *Appl. Statist.* **43**, 1-48.
- de Haan, L. and de Ronde, J. (1998). Sea and wind: multivariate extremes at work. *Extremes* **1**, 7-45.
- Embrechts, P. and Klüppelberg, C. and Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer-Verlag.
- Nelsen, R.B. (1999). *An introduction to copulas*. New York, NY, USA: Springer-Verlag.

A convenient device for replacing rounded zeros in compositional data: *aln* model

J. Palarea-Albaladejo¹, J. Daunis-i-Estadella² and J.A. Martín-Fernández²

¹ Dept. Informática de Sistemas, Univ. Católica San Antonio, Murcia, Spain.

² Dept. Informática y Matemática Aplicada, Univ. de Girona, Girona, Spain.

1 Introduction

Formally, a composition is a vector $\mathbf{x} = [x_1, \dots, x_D]$ such that $x_j > 0$, $j = 1, \dots, D$, subject to the constraint $x_1 + \dots + x_D = 1$. The sample space of compositions is the unit simplex \mathcal{S}^D . Its peculiarities prevent us from applying the standard multivariate statistical techniques designed for real spaces. Log-ratio methodology (Aitchison, 1986) provides the one to one correspondences between the simplex and the real space, opening up the whole of unconstrained real space multivariate data analysis. The results can then be translated back into the compositions of the simplex.

Sometimes in practice some parts take *rounded zero* values or *trace zeros*, making it impossible to use the log-ratio methodology. From a non-parametric point of view, the *multiplicative replacement* (MR) method (Martín-Fernández et al., 2003) replaces the zeros by a small number provided by the analyst. In this work, a computationally feasible parametric method based on a modification of the EM-algorithm is proposed. Its performance is analyzed by Monte Carlo simulation.

2 *aln*: multivariate log-ratio normal model

Aitchison (1986) introduces the additive log-ratio transformation $\text{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \dots, \frac{x_{D-1}}{x_D} \right] \in R^{D-1}$. Since the alr transformation is asymmetric in the components, one must verify that the applied statistical technique is invariant under permutations of the components. In addition, the alr transformation is not an isometry. To avoid the above difficulties, an isometric log-ratio transformation (ilr) is introduced (Egozcue et al., 2003)

$$\text{ilr}(\mathbf{x}) = \mathbf{y} = [y_1, \dots, y_{D-1}] \in R^{D-1}, \text{ where } y_i = \frac{1}{\sqrt{i(i+1)}} \ln \left(\frac{\prod_{j=1}^i x_j}{(x_{i+1})^i} \right),$$

which allows to apply any multivariate technique to the coordinates from an orthonormal basis. In our strategy, the original zeros in the compositional data set \mathbf{X} are transformed in missing data in $\mathbf{Y} = \text{alr}(\mathbf{X})$. The main idea is to impute the missing part of \mathbf{Y} and transform back from R^{D-1} to \mathcal{S}^D . We select the alr transformation rather

than ilr transformation because with the alr transformation the information about the detection limit can be easily incorporated to the alr-transformed data model. Furthermore, the consistency of results is guaranteed (Aitchison, 1986) when inference is based on the likelihood of the additive logistic normal model. Recall that a random composition vector $\mathbf{x} \in S^D$ is distributed according to an additive logistic-normal (aln) model (Aitchison, 1986) when $\mathbf{y} = alr(\mathbf{x})$ is distributed according to a $(D - 1)$ -dimensional normal model with mean vector μ and covariance matrix Σ .

3 Modified EM-algorithm in combination with aln model

A rounded zero occurs when $x_{ij} < \gamma_j$, where γ_j denotes the detection limit for the component x_j . When this relationship is alr-transformed into the real space, a missing data in \mathbf{Y} is obtained when $y_{ij} < \psi_{ij}$. Note that here $\psi_{ij} = \ln(\gamma_j/x_{iD})$, being x_D a part without zero values. On the t th iteration of the *modified* EM-algorithm (*mEM*) a missing value in the position (i, j) of \mathbf{Y} is imputed (Palarea-Albaladejo et al., 2007a, 2007b) using the equation

$$E[y_j | \mathbf{y}_{-j}, y_j < \psi_j, \theta^{(t)}] = \mathbf{y}_{-j}^T \beta - \sigma_j \frac{\phi\left(\frac{\psi_j - \mathbf{y}_{-j}^T \beta}{\sigma_j}\right)}{\Phi\left(\frac{\psi_j - \mathbf{y}_{-j}^T \beta}{\sigma_j}\right)},$$

where $\theta^{(t)}$ denotes the t th estimated parameters vector $\theta = (\mu, \Sigma)$ of the aln model; ϕ and Φ the density and the distribution function, respectively, of the standard normal distribution; σ_j^2 denotes the variance of y_j , and β is the vector of coefficients of the linear regression of y_j on \mathbf{y}_{-j} . Note that imputing by this way the method takes into account the information contained in the observed variables as much as the information about the detection limit. The EM algorithm generates a sequence $\{\theta^{(t)}\}$ which converges iteratively (Dempster et al., 1977) to the maximum-likelihood estimate of θ .

4 Simulation-based numerical results

Initially, 1000 data sets of size 300×5 are generated from a 5-part random composition. The compositional geometric mean of the random composition \mathbf{c} is given by $g(\mathbf{c}) = [0.027, 0.045, 0.201, 0.605, 0.122]$, and its total variability, $\text{totvar}(\mathbf{c})$, is equal to 3.996. The value of the geometric mean ensures that the fourth part takes the highest values, and parts 1 and 2 take the smallest values. In addition, the slightly high variability introduced ensures that the simulated data sets not are too similar. These data sets are free of zeros. Following that, small values in the compositions are changed by zero. In this way, a range of 10 realistic detection limits is considered: from 0.25% to 2.5% with increments of 0.25%. Thus, in total, 10 000 data sets of compositional data with rounded zeros have been generated. Next, the data sets are sorted in ascending order according to the proportion of zeros and the MR and *mEM* strategies to replace them are applied. For multiplicative replacement, the zeros are replaced by the 65% of the corresponding detection limit. MSD and the STRESS

$$\text{MSD} = \frac{\sum_{i=1}^{300} d_a^2(\mathbf{c}_i, \mathbf{r}_i)}{300} \quad \text{and} \quad \text{STRESS} = \frac{\sum_{i < j} (d_a(\mathbf{c}_i, \mathbf{c}_j) - d_a(\mathbf{r}_i, \mathbf{r}_j))^2}{\sum_{i < j} d_a^2(\mathbf{c}_i, \mathbf{c}_j)},$$

evaluate the distortion between the data set \mathbf{C} and the *completed* data set \mathbf{R} . By d_a we denote the Aitchison distance between two compositions \mathbf{x} and \mathbf{x}^* defined as the Euclidean distance between the vectors $ilr(\mathbf{x})$ and $ilr(\mathbf{x}^*)$.

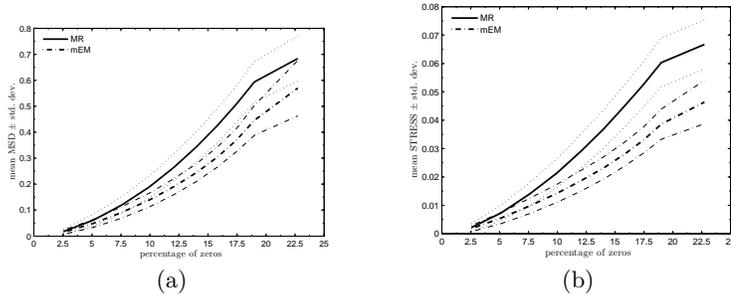


FIGURE 1. Replacement methods distortion: (a) MSD. (b) STRESS.

Figure 1 shows the patterns followed by the MR and mEM methods in relation to the proportion of zeros in the samples by means of the average of the MSD and STRESS measures (*continuous lines*), \pm their respective standard deviations (*dotted lines*), for different intervals of percentages of zeros. When the number of zeros grown the performance of mEM overcome that of MR. Since the MR method replaces all zeros by the same value, it tends to underestimate the variability in the data sets (figure 2). For all samples, the differences between the log-ratio total variabilities for both, the *completed* data set \mathbf{R} and the data set \mathbf{C} , is plotted. The mEM method also tends to underestimate the variability since it replaces zeros with an expected value, but this effect is appreciably smaller. With compositions of higher dimensions the expected result is that the mEM algorithm works better, since the information available to replace zeros by suitable values will increase. The same result will happen if the sample size is enlarged. Therefore, the yield of the mEM algorithm is bound by the size of the data matrices, as is common in all parametric strategies.

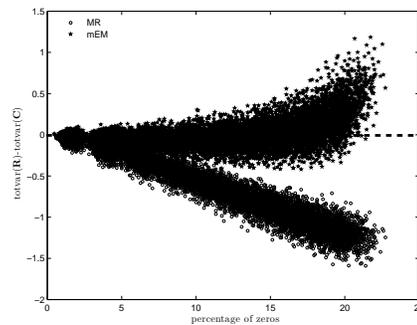


FIGURE 2. Sample variability subestimation.

Acknowledgments: Work partially financed by the D.G.I. of the Spanish Ministry for Science and Technology (MTM2006-03040).

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall, 416 pp. Reprinted in 2003 by Blackburn Press.
- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* **39**, 1-38.
- Egozcue, J.J., Pawłowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C. (2003). Isometric log-ratio transformations for compositional data analysis. *Math. Geol.* **35**, 3, 279-300.
- Martín-Fernández, J. A., Barceló-Vidal, C., Pawłowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets. *Math. Geol.* **35**, 3, 253-278.
- Palarea-Albaladejo, J., Martín-Fernández, J. A., Gómez-García, J. (2007a). A parametric approach for dealing with compositional rounded zeros. *Math. Geol.* (to appear).
- Palarea-Albaladejo, J., Martín-Fernández, J. A., Gómez-García, J. (2007b). A modified EM algorithm for replacing rounded zeros in compositional data sets. *Computer and Geosciences* (to appear).

Modelling Survival Data using Generalized Additive Models with Flexible Link

Ana L. Papoila¹ and Cristina S. Rocha²

¹ Faculdade de Ciências Médicas, Dep. de Bioestatística e Informática, Universidade Nova de Lisboa, Campo Mártires da Pátria 130, 1169-056 Lisboa, Portugal, CEAUL (apapoila@hotmail.com)

² Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Edifício C6, Piso 4, 1749-016 Lisboa, Portugal, CEAUL (cmrocha@fc.ul.pt)

Abstract: When using Generalized Linear Models (GLMs), misspecification of the link is very likely to occur due to the fact that the information, necessary to correctly choose this distribution function, is usually unavailable. To overcome this problem, new developments emerged which, simultaneously, gave rise to more flexible models. As a result, survival analysis also derived benefit from this new line of research. In fact, the gamma-logit model may be viewed as a GLM with binary response and unknown link function belonging to the one-parameter family of transformations, introduced by Aranda-Ordaz(1981). We suggest the use of flexible parametric link families in Generalized Additive Models (GAMs) with binary response and propose a generalization of the gamma-logit model, which we will denote by additive gamma-logit model. Based on the local scoring algorithm, the estimation procedure minimizes the deviance through the use of a deviance profile plot. A simulation study was carried out and the proposed methodology was applied to a real current status data set.

Keywords: Generalized additive model; unknown link function; survival analysis; gamma-logit model; current status data.

1 Introduction

With the evolution of Statistics, there has been an emphasis on the development of models with greater flexibility. That is what happened with the GLMs, in particular with the logistic model. In fact, several generalizations of this model were developed to ensure a minimization of the errors resulting from a bad choice of the link. Power transformation families were used to control symmetric and asymmetric departures from the logistic model and many parametric link classes were proposed (e.g. Pregibon (1980) and Aranda-Ordaz (1981)). As a consequence, survival analysis also benefited from these developments, due to the correspondence that can be established between models in binary regression analysis and in survival analysis (Doksum and Gasko, 1990). For instance, we may refer the gamma-logit model that, from the inferential point of view, is equivalent to a binary response

GLM, with unknown link function belonging to the Aranda-Ordaz (1981) transformations family. However, considering a GAM instead of a GLM is the natural extension of the gamma-logit model, in the sense that smooth functions may be used to establish the relation between the covariates and the response variable, often in a more realistic way. Some work has already been done to extend GAMs to a broader class of models with unknown non-parametric link function (Hastie and Tibshirani (1984) and Roca-Pardiña *et al.* (2004)). In this paper we propose the introduction of parametric link families in GAMs and, although the developed procedures may be applied to any model with a response variable whose distribution belongs to the exponential family, our paper will obviously focus the binary response case. Our proposal lies somewhere between an additive model with a fixed link and an additive model with a fully non-parametric link.

When using GAMs with flexible link, it is necessary to calculate an odds ratio curve because, unlike the GLMs, the effect of a continuous covariate on the response depends not only on the shape of the partial function but also on the functional form of the link. In our case, we have derived an estimator of the odds ratio curve and also constructed pointwise confidence intervals for the odds ratios, following Figueiras and Cadarso-Suárez (2001) and Cadarso-Suárez *et al.* (2005).

A simulation study was conducted and the new methodology was applied to a real current status data set.

2 GLMs with flexible parametric link and the gamma-logit model

The idea of using GLMs with flexible parametric link emerged as a natural consequence of the development of goodness of link tests for GLMs. In this context, Pregibon (1980) suggested a procedure to examine the adequacy of a particular hypothesized link function of a GLM, by embedding this function and the correct, but unknown, link in a family of link functions.

Let Y be a response variable with a distribution belonging to the exponential family and (X_1, \dots, X_p) a vector of p covariates. A GLM with flexible parametric link is defined by $E(Y|X_1, \dots, X_p) = h(\beta_0 + \sum_{j=1}^p \beta_j X_j, \psi)$, where h , known as the link function, is a strictly monotone differentiable function that belongs to the family $\mathcal{H} = \{h(\cdot, \psi) : \psi \in \Psi\}$, ψ represents the link parameter vector and $\beta_0, \beta_1, \dots, \beta_p$ are the regression coefficients, that must be estimated from the available data. This defines a broad class of models but, at the present, we will only focus the particular case of a model with binary response and parametric link belonging to the family proposed by Aranda-Ordaz (1981), in order to obtain the existing gamma-logit model. In a survival analysis context, this family is defined by

$$\gamma\text{-logit}(u) = \begin{cases} \log \left\{ \frac{(1-u)^{-\gamma} - 1}{\gamma} \right\} & \text{if } \gamma > 0 \\ \log[-\log(1-u)] & \text{if } \gamma = 0. \end{cases} \quad (1)$$

and h is the inverse of the function defined in (1).

3 GAMs with flexible parametric link and the additive gamma-logit model

In this paper, we propose the introduction of GAMs with a flexible parametric link, in order to obtain an extension of the gamma-logit model which we will denote by additive gamma-logit model. Let Y be a response variable with a distribution belonging to the exponential family and (X_1, \dots, X_p) a vector of p covariates. A GAM with flexible parametric link is defined by $\mu = E(Y|X_1, \dots, X_p) = h(\beta_0 + \sum_{j=1}^p f_j(X_j), \psi)$, where h , the link function, is a strictly monotone differentiable function that belongs to the family $\mathcal{H} = \{h(\cdot, \psi) : \psi \in \Psi\}$, where ψ represents the link parameter vector. The partial functions $f_j(X_j)$, $j = 1, \dots, p$, are arbitrary univariate functions that must be estimated from the data and represent the effect of the covariates on the response. As previously referred, we will only focus the particular case of a model with a binary response and parametric link belonging to the family proposed by Aranda-Ordaz (1981), in order to obtain the additive gamma-logit model defined by $F(t|\mathbf{x}) = h\left\{\gamma\text{-logit}[F_0(t)] + \sum_{j=1}^p f_j(x_j)\right\}$, where $F_0(t)$ represents the baseline distribution function.

In what concerns estimation, we added, to the Fortran program developed by Hastie and Tibshirani (1990), new routines that allowed the estimation of β_0 and of the partial functions f_1, \dots, f_p through the use of the iterative modified backfitting (Buja *et al.*, 1989) and local scoring algorithms (Hastie and Tibshirani, 1990). Cubic smoothing splines were used to model individual predictors. The amount of smoothing was defined, before fitting the model, by the specification of the degrees of freedom. In order to estimate the parameter vector ψ , we used a deviance profile plot.

To estimate the odds ratio curve we followed Cadarso-Suárez *et al.* (2005), that proposed a generalization of the odds ratio curve suggested by Figueiras and Cadarso-Suárez (2001) for the logistic GAMs. In fact, Cadarso-Suárez *et al.* (2005) defined the generalized odds ratio curve for a continuous covariate X_j at point x , and taking x_0 as the reference value, by

$$OR_j^{x_0}(x) = E_{(X_1, \dots, X_p)} \left[\frac{p(X_1, \dots, x, \dots, X_p)/(1 - p(X_1, \dots, x, \dots, X_p))}{p(X_1, \dots, x_0, \dots, X_p)/(1 - p(X_1, \dots, x_0, \dots, X_p))} \right],$$

where $p(X_1, \dots, X_p) = P(Y = 1|X_1, \dots, X_p)$ and $E_{(X_1, \dots, X_p)}$ represents the mean operator over the covariates $\{X_k\}_{k \neq j}$. Thus, if we consider a GAM with a link belonging to the Aranda-Ordaz (1981) transformations family,

we obtain the following estimator of the odds ratio

$$\widehat{OR}_j^{x_0}(x) = \frac{1}{n} \sum_{i=1}^n \frac{(1 + \hat{\psi} \times e^{\hat{\beta}_0 + \hat{f}_1(X_{i1}) + \dots + \hat{f}_j(x) + \dots + \hat{f}_p(X_{ip})})^{1/\hat{\psi}} - 1}{(1 + \hat{\psi} \times e^{\hat{\beta}_0 + \hat{f}_1(X_{i1}) + \dots + \hat{f}_j(x_0) + \dots + \hat{f}_p(X_{ip})})^{1/\hat{\psi}} - 1},$$

where $\hat{\psi}$, $\hat{\beta}_0$ and \hat{f}_j are estimates obtained from fitting our GAM. In what concerns the construction of pointwise confidence intervals for the odds ratio curve, we used bootstrap techniques (Cadarso-Suárez *et al.*, 2005).

A simulation study was carried out, not only to evaluate the quality of the link parameter estimates, but also to compare the performance of the proposed GAM with that of the GLM with the same parametric link. We concluded that the estimation process was satisfactory and that a substantial gain, in what concerns the deviance, may be achieved with our model.

4 A real case study

To apply the proposed methodology, we have studied the elapsed time from first injecting drug use until HIV infection, using a data set of 361 drug users who started using intravenous drugs between 1974 and 1997 and were admitted to the detoxification unit of the Hospital Universitari Germans Trias i Pujol in Badalona, Spain, between 1987 and 2000. For these individuals the moment of HIV infection is unknown. In fact, for 15% of the cases, the only available information about this instant is limited to the interval [instant of last negative HIV test, instant of first positive HIV test]. For the rest of the individuals, we only know their status (infected or not infected) at the date of the last HIV test (monitoring instant). This means that the data is mainly case I interval censored and so we decided to treat all the observations as current status data.

From the available data we used the variables *age* of first intravenous drug use, *gender*, the elapsed *time* (T), in months, from the instant of first intravenous drug use until the date of the last HIV test (monitoring time) and the indicator variable Y that gives us information about the result of the last HIV test (0 if the individual is seronegative or 1 if the individual is infected). We considered the model $\mu = h\{\beta_0 + f(T)\} + f_1(\textit{age}) + \beta_1 \times \textit{gender}$. The estimate of the link parameter was obtained through the minimization of the deviance, calculated for a grid of values of ψ and we considered that $\hat{\psi} = 5$ was the best estimate, for a deviance of approximately 377.2. The resulting fitted model is given by $\hat{\mu} = 1 - \left(1 + 5 \times e^{[5.18 + \hat{f}(T)] + \hat{f}_1(\textit{age}) + 2.61 \times \textit{gender}}\right)^{-1/5}$. For the variable *age*, we refer to Figure 1 for a graphical representation of the odds ratio curve, estimated for both genders and considering the mean age (19 years) as the reference value. As we can see from these two figures, the graphics are

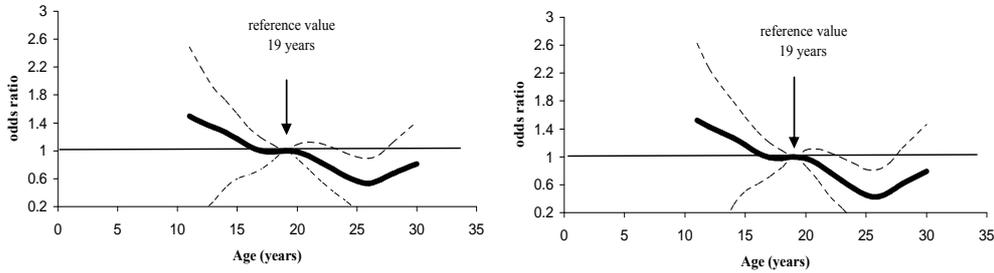


FIGURE 1. $OR_{(age)}$ estimates and corresponding 95% confidence intervals, female and male genders.

very similar. It seems to exist a lower risk of infection for the individuals who initiated their injecting drug addiction with an age of approximately 26 years old. Survival curves for both female and male were obtained and

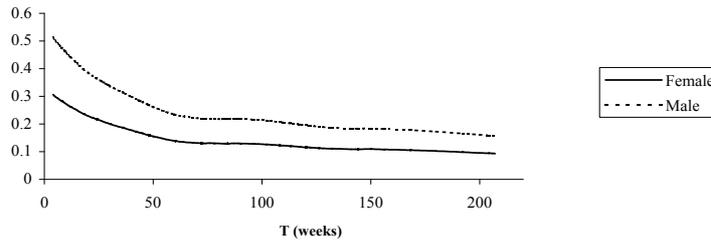


FIGURE 2. Estimates of the survival functions of time until HIV infection for females and males who initiated their drug addiction with a mean age of 19 years.

from Figure 2 we can see that time until HIV infection is longer for men. It also seems that the curves level off and the resulting plateau may indicate the existence of immune individuals in the population. In fact, it is admissible that some of the injecting drug users take the adequate precautions and consequently an HIV infection is unlikely to occur.

Finally, to evaluate the goodness-of-fit of the proposed model, the deviance residuals were examined and no serious trends, characteristic of a bad fit,

were detected. To overcome the lack of global goodness-of-fit tests for these kind of models, we used bootstrap techniques and concluded that the model was reasonably adequate. The 95% bootstrap confidence interval for the deviance (352.86, 425.69) was obtained. However, we are aware of the existence of unobserved heterogeneity among the individuals. So, we believe that the introduction of a frailty term would certainly improve the fit of the model.

Acknowledgements: This research was supported by FCT/POCI 2010. The authors would like to thank Drs. Klaus Langohr, Guadalupe Gómez and Robert Muga for making the data available.

References

- Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary regression data. *Biometrika* **68**, 357-363.
- Buja, A., Hastie, T.J. and Tibshirani, R.J. (1989). Linear smoothers and additive models (with discussion). *Annals of Statistics* **17**, 453-555 .
- Cadarso-Suárez, C., Roca-Pardiñas, J.R., Figueiras, A. and Manteiga, W. (2005). Non-parametric estimation of the odds ratios for continuous exposures using generalized additive models with an unknown link function. *Statistics in Medicine* **24**, 1169-1184.
- Doksum, K.A. and Gasko, M. (1990). On a correspondence between models in binary regression and in survival analysis. *International Statistical Review* **58**, 243-252.
- Figueiras, A. and Cadarso-Suárez, C. (2001). Application of nonparametric models for calculating odds ratios and their confidence intervals for continuous exposures. *American Journal of Epidemiology* **154**, **3**, 264-275.
- Hastie, T. and Tibshirani, R. (1984). Generalized additive models. Tech. Rep. 98, Dept. of Statistics, Stanford University.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, New York.
- Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society, series C* **29**, 15-24.
- Roca-Pardiñas, J., Manteiga, W., Bande, M., Sánchez, J., Cadarso-Suárez, C. (2004). Predicting binary time series of SO_2 using generalized additive models with unknown link function. *Environmetrics* **15** , 1-14.

Local Influence under Parameter Constraints

Gilberto A. Paula¹ and Francisco José A. Cysneiros²

¹ Universidade de São Paulo - Brazil, giapaula@ime.usp.br

² Universidade Federal de Pernambuco - Brazil, cysneiros@de.ufpe.br

Abstract: Calculations of local influence curvatures have been well developed when the parameters are unrestricted. In this paper we discuss the assessment of local influence under linear equality parameter constraints with extensions to inequality constraints. Using a penalized quadratic function we express the normal curvature of local influence for arbitrary perturbation schemes in interpretable forms, which depend on restricted and unrestricted components. The results are quite general and can be applied in various statistical models. In particular, we derive normal curvatures for generalized linear models. An application is given.

Keywords: Diagnostic methods; Generalized linear models; Inequality constraints; Leverage; Likelihood displacement; Restricted estimation.

1 Introduction

The aim of this paper is to derive local influence curvatures (Cook, 1986) under linear equality parameter constraints with extensions to inequality constraints. The theory of local influence has been largely applied under unrestricted parameters. However, few has been developed when the parameters are subject to restrictions (see, for instance, Paula, 1993; Kwan and Fung, 1998 and Paula, 1999). In this work we assume a quadratic penalty function as suggested by Nyquist (1991) for the parameter estimation under linear equality parameter constraints. Following the same procedures given in Cook (1986) we express the normal curvature in an interpretable form for arbitrary perturbation schemes. Extensions to inequality constraints are made by using the Kuhn-Tucker conditions. The methodology is applied in generalized linear models and an illustrative example is given.

2 Restricted normal curvature

For a given dataset let $L(\boldsymbol{\theta})$ be the log-likelihood function, where $\boldsymbol{\theta}$ is a $r \times 1$ parameter vector. The main interest is to maximize the log-likelihood function $L(\boldsymbol{\theta})$ subject to the linear constraints $\mathbf{C}\boldsymbol{\theta} - \mathbf{d} = \mathbf{0}$, where $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_k)^T$ and $\mathbf{d} = (d_1, \dots, d_k)^T$, with \mathbf{C}_ℓ being an $r \times 1$ vector of constants and d_ℓ are scalars, $\ell = 1, \dots, k$. Similarly to Nyquist (1991) that investigated this problem in generalized linear models, we will apply the methodology of penalty functions by considering the quadratic penalized function $P(\boldsymbol{\theta}, \boldsymbol{\tau}) = L(\boldsymbol{\theta}) - \frac{1}{2} \sum_{\ell=1}^k \tau_\ell (d_\ell - \mathbf{C}_\ell^T \boldsymbol{\theta})^2$. The procedure consists in finding the solution of $\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}, \boldsymbol{\tau})$ for positive and fixed values of τ_ℓ , $\ell = 1, \dots, k$. The derivative

of $P(\boldsymbol{\theta}, \boldsymbol{\tau})$ with respect to $\boldsymbol{\theta}$ is calculated for fixed $\boldsymbol{\tau}$ and the solution by setting the derivative to zero will be denoted by $\boldsymbol{\theta}(\boldsymbol{\tau})$. The unrestricted maximum likelihood estimate is given by $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{0})$ and the equality restricted estimate is obtained as $\tilde{\boldsymbol{\theta}} = \lim_{\tau_1 \rightarrow \infty, \dots, \tau_k \rightarrow \infty} \boldsymbol{\theta}(\boldsymbol{\tau})$. We define the perturbed penalty function by $P(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{\omega}) = L(\boldsymbol{\theta}|\boldsymbol{\omega}) - \frac{1}{2} \sum_{\ell=1}^k \tau_{\ell} \omega_{\ell} (\mathbf{d}_{\ell} - \mathbf{C}_{\ell}^T \boldsymbol{\theta})^2$, where $\boldsymbol{\omega}$ is the $s \times 1$ perturbation vector. Following the same steps given in Cook's paper with $L(\boldsymbol{\theta})$ replaced by $P(\boldsymbol{\theta}, \boldsymbol{\tau})$ we first obtain the normal curvature for fixed $\boldsymbol{\tau}$ and then by making $\tau_1 \rightarrow \infty, \dots, \tau_k \rightarrow \infty$ we find that the normal curvature for $\boldsymbol{\theta}$ at the unitary direction $\boldsymbol{\ell}$ is expressed as

$$C_{\ell}(\boldsymbol{\theta}) = 2|\boldsymbol{\ell}^T \boldsymbol{\Delta}^T (\mathbf{B}^U + \mathbf{B}^R) \boldsymbol{\Delta} \boldsymbol{\ell}|, \tag{1}$$

where $\mathbf{B}^U = \ddot{\mathbf{L}}_{\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}}^{-1}$ is the component corresponding to the unrestricted parameters whereas $\mathbf{B}^R = -\ddot{\mathbf{L}}_{\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}}^{-1} \mathbf{C}^T (\mathbf{C} \ddot{\mathbf{L}}_{\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}}^{-1} \mathbf{C}^T)^{-1} \mathbf{C} \ddot{\mathbf{L}}_{\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}}^{-1}$ is the additional component in the Cook's curvature due to the linear equality constraints, $\boldsymbol{\Delta}$ is an $r \times s$ matrix with elements $\Delta_{ji} = \partial^2 P(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{\omega}) / \partial \theta_j \partial \omega_i$, $i = 1, \dots, s$ and $j = 1, \dots, r$, evaluated at $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\omega}_0$ (no perturbation vector).

If the Kuhn-Tucker conditions are satisfied (see, for instance, Fahrmeir and Klingler, 1994) extensions to $\mathbf{C}\boldsymbol{\theta} \geq \mathbf{d}$ are straightforward. Thus, the maximization of $L(\boldsymbol{\theta})$ subject to $\mathbf{C}\boldsymbol{\theta} \geq \mathbf{d}$ can be reduced to maximize $L(\boldsymbol{\theta})$ subject to $\mathbf{C}_I \boldsymbol{\theta} = \mathbf{d}_I$, where \mathbf{C}_I is a submatrix of \mathbf{C} and \mathbf{d}_I is the corresponding subvector.

3 Generalized linear models

Generalized linear models (McCullagh and Nelder, 1989) is a well known class of nonlinear regression models which consists in assuming $g\{E(y_i)\} = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$, $g(\cdot)$ is the link function and $y_i, i = 1, \dots, n$, are independent random variables following some distribution in the exponential family indexed by the parameters $\mu_i = E(y_i)$ and $\phi^{-1} > 0$ and whose density function assumes the form $f(y_i; \mu_i, \phi) = \exp[\phi\{y_i \nu_i - b(\nu_i)\} + c(y_i, \phi)]$. We will assume the constraints $\mathbf{R}\boldsymbol{\beta} = \mathbf{d}$. Here one has $\mathbf{C} = (\mathbf{R} \ \mathbf{0})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$, $r = p + 1$ and $s = n$. In order to simplify the curvature expressions suppose $-\ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ replaced by its expected value $\mathbf{K}_{\boldsymbol{\theta}\boldsymbol{\theta}} = \text{blkdiag}\{\mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}}, \mathbf{K}_{\phi\phi}\}$, where $\mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \phi(\mathbf{X}^T \mathbf{W} \mathbf{X})$ and $\mathbf{K}_{\phi\phi} = -E\{\ddot{c}(y_i, \phi)\}$ are, respectively, the Fisher information matrices for $\boldsymbol{\beta}$ and ϕ , $\mathbf{W} = \text{diag}\{W_1, \dots, W_n\}$, $W_i = (d\mu_i/d\eta_i)^2/V_i$ and V_i is the variance function. Thus, the normal curvatures for $\boldsymbol{\beta}$ and ϕ at the unitary direction $\boldsymbol{\ell}$ reduce to $C_{\ell}(\boldsymbol{\beta}) = 2|\boldsymbol{\ell}^T \boldsymbol{\Delta}_1^T \text{Var}(\tilde{\boldsymbol{\beta}}) \boldsymbol{\Delta}_1 \boldsymbol{\ell}|$ and $C_{\ell}(\phi) = 2|\boldsymbol{\ell}^T \mathbf{K}_{\phi\phi}^{-1} \boldsymbol{\Delta}_2 \boldsymbol{\Delta}_2^T \boldsymbol{\ell}|$, where $\text{Var}(\tilde{\boldsymbol{\beta}}) = \phi^{-1}(\mathbf{X}^T \tilde{\mathbf{W}} \mathbf{X})^{-1} [\mathbf{I}_p - \mathbf{R}^T \{\mathbf{R}(\mathbf{X}^T \tilde{\mathbf{W}} \mathbf{X})^{-1} \mathbf{R}^T\}^{-1} \mathbf{R}(\mathbf{X}^T \tilde{\mathbf{W}} \mathbf{X})^{-1}]$, $\boldsymbol{\Delta}_1$ is a $p \times n$ matrix and $\boldsymbol{\Delta}_2$ is an $n \times 1$ vector of constants.

3.1 Case-weight perturbation

Suppose the log-likelihood function expressed as $L(\boldsymbol{\theta}) = \sum_{i=1}^n L_i(\boldsymbol{\theta})$, where $L_i(\boldsymbol{\theta})$ denotes the contribution of the i th observation. Under the case-weight perturbation scheme the perturbed log-likelihood function takes the form $L(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^n \omega_i L_i(\boldsymbol{\theta})$, $0 \leq \omega_i \leq 1$. Hence, the normal curvature for $\boldsymbol{\beta}$ at the unitary direction $\boldsymbol{\ell}$ can be expressed approximately as

$$C_{\ell}(\boldsymbol{\beta}) = 2|\boldsymbol{\ell}^T \mathbf{D}(\tilde{\mathbf{r}}_P)(\tilde{\mathbf{H}} - \tilde{\mathbf{G}}) \mathbf{D}(\tilde{\mathbf{r}}_P) \boldsymbol{\ell}|, \tag{2}$$

where $\mathbf{D}(\tilde{\mathbf{r}}_P) = \text{diag}\{\tilde{r}_{P_1}, \dots, \tilde{r}_{P_n}\}$ with $\tilde{r}_{P_i} = \sqrt{\tilde{\phi}(y_i - \tilde{y}_i)}/\sqrt{\tilde{V}_i}$ being the i th Pearson residual, $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}$ and $\mathbf{G} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, where $\mathbf{Z} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{R}^T$. For the particular case in which $\boldsymbol{\ell}$ is an $n \times 1$ vector of zeros with one at the i th position, the total local influence for $\boldsymbol{\beta}$ takes the form $C_i(\boldsymbol{\beta}) = 2\tilde{r}_{P_i}^2(\tilde{h}_{ii} - \tilde{g}_{ii})$, where h_{ii} and g_{ii} are the diagonal elements of the matrices \mathbf{H} and \mathbf{G} , respectively. Since $\tilde{\mathbf{M}} = \tilde{\mathbf{H}} - \tilde{\mathbf{G}}$ is a leverage matrix (see Paula, 1999), then $0 \leq \tilde{m}_{ii} \leq 1$ so that $\tilde{h}_{ii} \geq \tilde{g}_{ii}$. Thus, $C_i(\boldsymbol{\beta})$ is large if $\tilde{r}_{P_i}^2$ and (or) $(\tilde{h}_{ii} - \tilde{g}_{ii})$ are (is) large.

4 Application

Applications of generalized linear models with linear inequality parameter constraints are described by McDonald and Diamond (1990). This is for example the case of the smelter workers study in which the Poisson regression model $\log E(y_i) = \log E_i + \mathbf{x}_i^T \boldsymbol{\beta}$, $i = 1, \dots, 40$, is proposed to analyze respiratory cancer deaths among a cohort of smelter workers exposed to airborne arsenic trioxide (Breslow et. al, 1983), where $y_i \sim P(\lambda_i E_i)$, $\log E_i$ denotes an offset and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_9)$ represents the birthplace (U.S. or foreign), 5 levels of moderate arsenic and 4 levels of heavy arsenic. McDonald and Diamond (1990) argue that the death rates for each exposure might form a non-decreasing sequence. This implies imposing the inequality constraints $0 \leq \beta_3 \leq \beta_4 \leq \beta_5 \leq \beta_6$ (moderate arsenic effects) and $0 \leq \beta_7 \leq \beta_8 \leq \beta_9$ (heavy arsenic effects). It may be showed that the inequality constrained maximum likelihood estimates can be obtained for this data by fitting the Poisson model above subject to $\beta_3 = 0$, $\beta_4 = \beta_5$ and $\beta_7 = \beta_8$.

From the index plot of $C_i(\boldsymbol{\beta})$ (Figure 1) four groups appear with large influence, groups #1, #10 and #4 (U.S.-born) with large SMR (standardized mortality ratio) and group #21 (foreign-born) with an unexpected large SMR. Some restricted estimates present large changes after dropping group #21. Inferences also change at the level of 5% for the parameters β_2 , β_7 and β_8 indicating lack of robustness of the maximum likelihood estimates against the perturbation scheme.

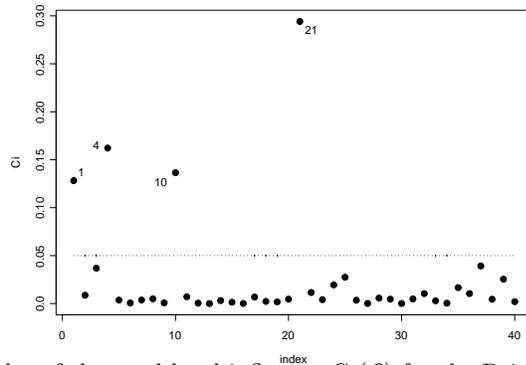


FIGURE 1. Index plot of the total local influence $C_i(\boldsymbol{\beta})$ for the Poisson model fitted to the smelter works' data. Dotted line corresponds to $2\bar{C}$.

Acknowledgments: This work was supported by CNPq and FAPESP, Brazil.

References

- Breslow, N. E.; Lubin, J. H.; Marek, P. and Langholz, B. (1983). Multiplicative models and cohort analysis. *J. Amer. Statist. Assoc.* **78**, 1-12.
- Cook, R. D. (1986). Assessment of local influence (with discussion). *J. Roy. Statist. Soc. Ser. B* **48**, 133-169.
- Fahrmeir, L. and Klinger, J. (1994). Estimating and testing generalized linear models under inequality restrictions. *Statist. Papers* **35**, 211-229.
- Kwan, C. W. and Fung, W. K. (1998). Assessing local influence for specific restricted likelihood application to factor analysis. *Psychometrika* **63**, 35-46.
- McDonald, J.M. and Diamond, I. (1990). On the fitting of generalized linear models with non-negativity parameter constraints. *Biometrics* **46**, 201-206.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall: London.
- Nyquist, H. (1991). Restricted estimation of restricted generalized linear models. *Appl. Statist.* **40**, 133-141.
- Paula, G. A. (1993). Assessing local influence in restricted regression models. *Computat. Statist. Data Anal.* **16**, 63-79.
- Paula, G. A. (1999). Leverage in inequality constrained regression models. *The Statistician* **48**, 529-538.

On the analysis of censored reliability data

Defen Peng¹ and Gilbert MacKenzie²

¹ Department of Statistics , Zhongnan University of Economics & Law, Wuhan, P.R. China

² Centre of Biostatistics, University of Limerick, Ireland

Abstract: We develop an analysis of the reliability of reliability data following a negative exponential distribution and subject to Type I censoring. Our main focus is to determine the role of the particular analytical approach in *interim analysis*. Accordingly, in this preliminary study, we aim to investigate, by means of a detailed simulation study, the effects of early stopping on decision-making in a comparative two group reliability study (eg, a two group randomised controlled clinical trial).

Keywords: Reliability, Censoring Types I & II, Interim Analysis, Designed Experiment, Simulation

1 Introduction

A particular analysis of the reliability of Type I censored reliability data was given by Finselbach & Watkins (2006). The form of the analysis suggested that the overall approach may have some role in interim analysis. Thus, in a reliability experiment with a fixed duration of time, c , we may wish to know whether it is possible to terminate the experiment early, say at time, $h < c$, without loss of information. This of course requires an evaluation of the costs of so doing - effectively we need information on the effects of early stopping on the final decision-making process.

In order to develop our methods and compare findings with those in the original paper we have mirrored the simple, expository, assumptions made there. However, we also deal with a Negative Exponential survival time random variable with parameter θ , the inverse parametrization.

Moreover, our perspective differs somewhat from the original authors. Their focus lay in investigating the asymptotic properties of their scheme, that is, in comparing the outcome at $t = c$ with $t = \infty$. On the other hand, our interest centres on comparing the outcome at $t = h < c$ with $t = c$ for a sensible range of h values.

2 The Basic Model

The model for our non-negative Exponential random variable, T , takes the form:

$$f(t; \theta) = \theta \exp(-\theta t) \quad t \geq 0 \quad (1)$$

whence $E(T) = 1/\theta$ and $V(T) = 1/\theta^2$, whereas Finselbach & Watkins (2006), hereafter FW, considered the model with $E(T) = \theta^* = 1/\theta$.

For our regression model we adopted

$$\theta_i = \exp(x_i' \beta) \quad (2)$$

where $x_i' = (x_{0i}, x_{1i})$, with $x_{0i} = 1 \forall i = 1, \dots, n$ and x_{1i} is a binary treatment indicator = 1 for treatment and = 0 for control, leading to $n/2$ ones in a balanced trial design. Here $\beta' = (\beta_0, \beta_1)$, whence β_1 is the treatment effect.

In some ways their choice of parametrization (1) is more natural, but the alternative form is often used in simulation studies and one question is simply, does the choice of parametrization matter?

2.1 Type I Censoring

Suppose at any given time, say c , the data are subject to Type I censoring and that we only have the exact lifetimes t_1, \dots, t_M of the M ($0 \leq M \leq n$) items that have failed before c , with the remaining $n - M$ items having a censored operational life of c . Thus, with $M \geq 0$, we have

$$\hat{\theta}_c = \frac{M}{S_M + (n - M)c} \quad (3)$$

where $S_M = \sum_{i=1}^M t_i$. Let the true lifetime data be $t_1, \dots, t_M, t_{M+1}, \dots, t_n$, then $S_n = S_M + \sum_{i=M+1}^n t_i$. Importantly, the link between $\hat{\theta}$ and $\hat{\theta}_c$ is

$$n\theta^{-1} = M\hat{\theta}_c^{-1} + \sum_{i=M+1}^n (t_i - c) \quad (4)$$

Where t_{M+1}, \dots, t_n denote the lifetimes of items still operational at c , and M follows a Binomial distribution with parameters n and $q_c = 1 - \exp(-\theta c)$ which is the probability that any item fails in $(0, c)$. We may proceed to show, after some algebra, that

$$\text{Corr}(\hat{\theta}, \hat{\theta}_c) \simeq \sqrt{q_c}. \quad (5)$$

The arguments leading to equation (4) follow the steps given in FW and depend on exploiting the usual asymptotic relationships which hold with Fisher Information (Watkins & John, 2004). FW also reached equation (4) but with $q_c = 1 - \exp(-\frac{c}{\theta})$. We note that the correlation does not depend on n .

3 Simulation Studies

3.1 Inverse Parametrization

Overall, the results (not shown) were similar in both models and our final conclusion is that the choice of parametrization is immaterial in practice and may be made on the grounds of convenience.

TABLE 1. Anova for the % concordance between tests of H_0 for β_{1h} and β_{1c}

Effect	df	ss	ms	F-ratio	p
β_1	2	1597.07	798.54	239.14	0.000
n	2	231.06	115.53	17.30	0.000
p	2	225.51	112.76	16.88	0.000
h	2	366.42	183.21	27.43	0.000
$\beta_1 \times n$	4	1760.62	440.15	65.91	0.000
$\beta_1 \times p$	4	160.87	40.22	6.02	0.001
$\beta_1 \times h$	4	128.66	32.16	4.82	0.004
$\beta_1 \times n \times p$	8	655.19	81.90	12.26	0.000
Residual	52	347.28	6.68		
Total	80	5472.68			

3.2 Effect of Early Stopping

In order to investigate the effects of early stopping we designed a simulation study mimicking the effect of a two group randomised controlled trial using the exponential regression model defined at (2). A full factorial experiment was organised to evaluate effects of varying: h as $h_1 = 1c/2, h_2 = 2c/3$ and $h_3 = 3c/4$; the sample size $n = 100, 500$ and 1000 ; the percentage censored (at c) $p = 0.2, 0.5$ and 0.7 and the parameter $\beta_1 = 0.2, 0, 5$ and 1 . The intercept parameter β_0 was set to zero throughout and x_i was as described above. Thus, in this preliminary investigation we have a 3^4 factorial design covering 81 scenarios.

4 Discussion

We analysed the % concordance, c , between the tests of the null hypotheses that $H_0 : \beta_{c_j} = 0$ and $H_0 : \beta_{h_j} = 0$ for $j = 1, \dots, m = 1000$. Table 1 shows the results for the same factorial design described above this time using the percentage concordance as the outcome summary measure, while Figure 1 shows the histogram of the % concordance obtained. Overall, the results are encouraging. From Table 1 we see that the % concordance varies with several factors simultaneously and the pattern of dependence is complex. Exploring the quadratic surface model space led to: $\hat{c} = 78.767(\pm 4.685) - 7.864(\pm 3.072)p - 0.009(\pm 0.003)n + 0.963(\pm 3.367)\beta + 20.402(\pm 6.055)h + 0.012(\pm 0.003)[\beta \times n]$ which explained 55.8% of the variation.

The concordance analysis was encouraging, suggesting that it may be possible to isolate scenarios in which early stopping was a viable option. However, further work is required before a firm conclusion can be reached, including the extension to the case where the data are subject to Type II Censoring and follow the Weibull distribution.

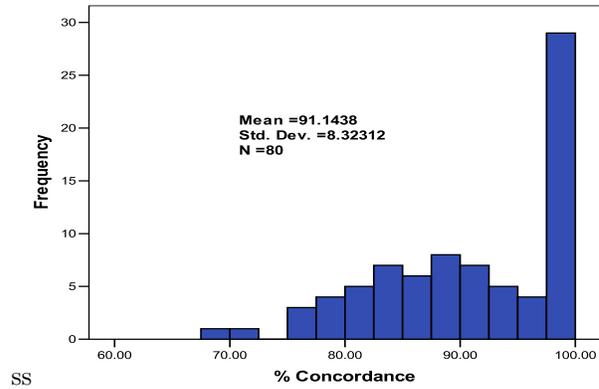


FIGURE 1. % Concordance: tests of H_0 : for β_n & β_c

Acknowledgement

The work was carried out in part in the Centre of Biostatistics in the University of Limerick, Ireland, where Professor Peng was a visiting research scholar supported by the Chinese Government.

References

- Finselbach HK & Watkins AJ (2006). Proceedings of the 21st IWSM, Galway, pp 182-189.
- Lawless JF (1982). *Statistical Models & Methods for Lifetime Data*, Wiley, New York.
- Watkins AJ & John AM (2004). On the expected Fisher Information for the Weibull distribution with Type I censored data. *International Journal of Pure & Applied Mathematics* **15**, 401-412.

Modelling IVF Data using an Extended Continuation Ratio Random Effects Model.

Ruth Penman¹, Gillian Heller² and John Tyler³

¹ Statistics Department, Macquarie University, NSW 2109, Australia,
rpenman@efs.mq.edu.au

² Statistics Department, Macquarie University, NSW 2109, gheller@efs.mq.edu.au

³ Next Generation Fertility, Parramatta, NSW 2150, jtyler@nextgenfertility.com.au

Abstract: In Vitro Fertilization (IVF) offers infertile couples the opportunity to have a baby. Many of these couples have multiple attempts to try to achieve this goal. To model the risk factors that affect the process at any of its stages, across the multiple attempts, an ordinal response and an extended continuation ratio random effects model have been used. This enabled the modelling of differential effects of covariates on different stages of the process. A significant random effect suggests that there is a couple specific effect, which may be interpreted as a fertility index.

Keywords: IVF; extended continuation ratio model; random effects model.

1 Introduction

Many couples experience fertility problems making it difficult for them to conceive a child. IVF (in vitro fertilization) offers these couples the chance to have a baby. Our aim is to develop a single model that looks at the risk factors for success for the process as a whole. In previous work we have developed a model for a single attempt using an extended continuation ratio model. In this paper we are extending that model to include multiple attempts.

The IVF process consists of a number of sequential stages. The couple start the process with the collection of the oocytes (or eggs) and the sperm (stage 1). The eggs are then fertilized to produce embryos (stage 2), some of which are transferred to the uterus (stage 3). This may result in pregnancy (stage 4), which may in turn result in the birth of a baby (stage 5).

The couple must be successful at every stage to achieve the ultimate successful outcome, the birth of a live baby. Therefore it is important to determine the risk factors that affect each stage of the process. We can consider these stages as an ordinal response, where the response is the maximum stage successfully achieved in a given attempt, with values 0 (start the process) to 5 (birth of a baby). The collection of sperm is not considered as a separate stage. If the collection of sperm is unsuccessful then fertilization will not take place and, provided the collection of oocytes has been successful, the couple will be classified as having reached stage 1.

Many couples have multiple attempts at IVF providing valuable additional information. There are additional complications for second and subsequent attempts due to

the use of frozen embryos. Embryos not used in one attempt are frozen for use in future attempts. When a frozen embryo is used the couple effectively start the process at the transfer stage. This needs to be taken into account in the model.

There is also a condition, known as OHSS, where it is dangerous for a subject to have oocytes collected and an embryo transferred in a single attempt. For these subjects the two attempts required to effectively complete an IVF cycle are combined and an indicator variable is set up to indicate the presence of this condition.

The data used included a maximum of four attempts for each subject.

2 The Extended Continuation Ratio Mixed Effects Model

2.1 The continuation ratio model

The continuation ratio model is best suited to data where the interest is in the individual categories of the response, rather than a cumulative response. It is also recommended when the outcome is irreversible, that is the subject can only go forward through the levels or stages. For these reasons this model is the most appropriate of the ordinal models for the IVF data.

For the forward continuation ratio model, the continuation ratio $\delta_k(x)$ is the conditional probability of being in category k , given that at least k is reached:

$$\begin{aligned}\delta_k(x) &= P(Y = k | Y \geq k, X = x) \\ &= \pi_k(x) / [\pi_k(x) + \pi_{k+1}(x) + \dots + \pi_K(x)], k = 0, 1, \dots, K - 1\end{aligned}$$

where K is the highest category and $\pi_k(x)$ is the probability of having response k , given covariate vector x . We model the logit of the continuation ratio giving:

$$\begin{aligned}\text{logit}(\delta_k(x_i)) &= \log \frac{\delta_k(x_i)}{1 - \delta_k(x_i)} = \alpha_k + x_i' \beta \\ \implies \delta_k(x_i) &= \frac{\exp(\alpha_k + x_i' \beta)}{1 + \exp(\alpha_k + x_i' \beta)}\end{aligned}$$

where x_i is the vector of covariates for subject i .

The advantage of the continuation ratio model is that by restructuring the data simple binary logistic regression can be used to estimate these models. The data is restructured by creating a subset of the data for each 'cut-point', where the cut-point corresponds to categories $k = 0, 1, \dots, K - 1$. The subset consists of all observations with a response greater than or equal to that cut-point. Two new variables are created in each subset, one to specify the cut-point and a binary variable that is 1 if the ordinal response is equal to that cut-point and 0 otherwise. The subsets are then concatenated into one data set and the new binary response is used as the response in a logistic model. The cut-point is included as a factor in the model to give category specific intercepts. The observations are conditionally independent given the cut-point.

2.2 The extended continuation ratio model

Since the stages of the IVF process are different biological processes, stage specific coefficients are needed. The continuation ratio model can be extended to allow different coefficients for each stage by simply including interactions between the new cut-point variable and the risk factors.

The model for the extended continuation ratio model is:

$$\text{logit}(\delta_k(x_i)) = \alpha_k + x'_i \beta_k .$$

2.3 The extended continuation ratio random effects model

Since we are analyzing multiple attempts per subject (couple) we cannot regard these observations as independent. A random effect may be added to the model to account for this correlation (or between-subject heterogeneity). This gives the following random effects model for multiple attempts:

$$\text{logit}(\delta_k(x_{ij})) = \alpha_k + x'_{ij} \beta_k + b_i$$

where $b_i \sim N(0, \sigma^2)$ and x_{ij} is the vector of covariates for subject i at their j th attempt.

Let y_{ij} be the ordinal response for subject i at attempt j , and z_{ijk} be the binary response for subject i at attempt j and cut-point k . The likelihood that subject i has response k ($k = 0, 1, \dots, 5$ for the IVF data), at attempt j is:

$$f_{ij}(y_{ij} = k | \beta, b_i) = \begin{cases} \prod_{\ell=0}^k \frac{\exp(\alpha_\ell + x'_{ij} \beta_\ell + b_i) z_{ij\ell}}{1 + \exp(\alpha_\ell + x'_{ij} \beta_\ell + b_i)} & k = 0, \dots, 4 \\ \prod_{\ell=0}^4 \frac{1}{1 + \exp(\alpha_\ell + x'_{ij} \beta_\ell + b_i)} & k = 5 \end{cases}$$

The marginal likelihood for subject i over all attempts is:

$$f_i(y_i | \beta, \sigma^2) = \int \prod_{j=1}^{J_i} f_{ij}(y_{ij} | \beta, b_i) g(b_i | \sigma^2) db_i$$

where J_i is the number of attempts for subject i ($J_i = 1, 2, 3, 4$). The total likelihood is maximized using Dual Quasi-Newton optimization and non-adaptive Gaussian Quadrature for the integration. SAS NLMIXED was used to implement this model.

Note: Since those subjects using frozen embryos start at the transfer stage, they are not included in the data for the first two cut-points.

3 Results

The standard deviation of the random effect was found to be highly significant ($\hat{\sigma} = 0.72, SE = 0.08$), suggesting a significant subject-specific effect which may be interpreted as a fertility measure for the couple. This result must be treated with some care due to possible boundary problems in the hypothesis. Empirical Bayes may be used to estimate the random effects posthoc.

The risk factors that were tested included BMI (Body Mass Index), female age, type of treatment, OHSS and attempt number. All factors were significant at one or more stages. The model was also adjusted for the year of treatment. This factor may have an effect on the outcome due to changing technology but is not relevant as a risk factor.

TABLE 1. Results - Odds Ratios for Significant Risk Factors

Risk Factor	Cut-point	Stage	OR	CI
Attempt 4	1	Egg Collection	0.45	0.20 - 0.98
Attempt 4	3	Embryo Transfer	0.73	0.55 - 0.96
BMI < 20	0	No success	2.01	1.22 - 3.31
BMI > 30	4	Pregnancy	2.32	1.52 - 3.55
Age > 35	0	No success	2.45	1.68 - 3.58
Age > 35	1	Egg Collection	1.66	1.19 - 2.30
Age > 35	2	Fertilization	1.37	1.01 - 1.85
Age > 35	3	Embryo Transfer	1.94	1.60 - 2.35
Age > 35	4	Pregnancy	2.09	1.48 - 2.96
Treatment ICSI	0	No Success	0.68	0.47 - 0.99
Treatment ICSI	1	Egg Collection	0.63	0.45 - 0.88
Treatment ICSI	2	Fertilization	0.54	0.38 - 0.76
Treatment FER	3	Embryo Transfer	1.85	1.52 - 2.25
OHSS	3	Embryo Transfer	2.09	1.38 - 3.17

BMI was categorized into 4 groups, underweight (< 20), normal (20 - 25), overweight (25 - 30) and obese (> 30). Subjects who are underweight are less likely to have oocytes successfully collected than subjects with normal BMI. Subjects who are obese are more likely to miscarry than subjects with normal BMI.

Age was categorized into 3 groups, < 30, 30 - 35 and > 35 with the 30 - 35 group being used as the referent category. Older women (> 35) were less likely to be successful at all stages of the process.

There were three treatments considered, the traditional IVF process, ICSI (Intra Cytoplasmic Sperm Injection), where the sperm is placed directly into the egg, and the use of frozen embryos. Those subjects using ICSI were more likely to be successful in the early stages of the process. Those using frozen embryos were less likely to fall pregnant, that is were more likely to stop at the transfer stage, than those using fresh embryos. However changes in technology since this data was collected have improved the likelihood of falling pregnant when using frozen embryos.

Those subjects with OHSS were also less likely to fall pregnant. These subjects use frozen embryos so this is consistent with the results when using frozen embryos.

There was no significant difference between attempt 1 and attempts 2 and 3 at any stage. At attempt 4 subjects are more likely to have eggs successfully fertilized, and more likely to fall pregnant than at attempt 1. This may be a result of the cohort remaining at this attempt.

4 Discussion

It is important to understand the risk factors that affect success at each stage of the IVF process. This will improve the chances of success, provide information for further research and inform subjects of their likelihood of success.

Using the stages of the process as an ordinal response in an extended continuation ratio random effects model, enables us to model the whole process over multiple attempts and determine which factors affect each stage. Further risk factors can be tested using this model.

The posthoc estimates of the b_i 's provide valuable information in understanding the patterns of success in IVF.

Not all couples had 4 attempts. Couples may decide not to continue with IVF for many reasons, such as having successfully had a baby, the financial or emotional costs of the process or having fallen pregnant without the assistance of IVF. For this reason we felt that the 'dropouts' could be considered to be at random and consequently ignored. Further consideration will be given to this issue.

There are a number of other issues still to be investigated. These include looking at other distributions for the random effect, in particular a non-parametric distribution, consideration of a random error term to model overdispersion (variance component model) and further investigating a test for the significance of the variance of the random effect.

References

- Bender, R., and Benner, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal* **42**, 677-699.
- Diggle, P.J., Heagerty, P., Liang, K. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.
- Dos Santos, D.M. and Berridge, D.M. (2000). A continuation ratio random effects model for repeated ordinal responses. *Statistics in Medicine* **19**, 3377-3388.
- Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Science.

Overview of Joint Regression Analysis

Dulce Pereira¹ and João T. Mexia²

¹ Department of Mathematics, CIMA-UE, University of Évora, Colégio Luís António Verney, Rua Romão Ramalho 59, 7000-671 Évora, Portugal, dgsp@uevora.pt

² Department of Mathematics, Faculty of Sciences and Technology - The New University of Lisbon, Quinta da Torre, 2825 Monte da Caparica, Portugal, jtm@fct.unl.pt

Abstract: Joint Regression Analysis (JRA) has been widely used to compare cultivars. In this technique a linear regression is adjusted per cultivar. The slope of each regression measures the ability of the corresponding cultivar to answer to variations in productivity. Presently we are mainly interested in cultivars with better responses to high productivity. To extend the application range of JRA to connected series of designs in incomplete blocks, thus going beyond the classic case of series of randomized blocks, we introduced the L_2 environmental indexes. Nowadays, comparison trials for cultivars are mainly α -designs, which have incomplete blocks. Moreover, the introduction of these indexes: enables the integration of JRA into the statistical inference for normal models; allows a better approach to the study of specific interactions. These interactions occur when a cultivar behaves abnormally well or abnormally badly, for a (location, year) pair. We will also, use JRA to obtain and update of lists of recommended cultivars. Appropriate algorithms have been developed for the adjustments: the zig zag algorithm and the double minimization algorithm.

Keywords: Joint Regression Analysis, linear regressions, L_2 environmental indexes, double minimization, zig zag algorithm.

1 Introduction

Joint Regression Analysis (JRA), is a widely used technique for evaluation of cultivars, integrating in a variable (the environmental index) the productive capacity for each (location, year) pair.

A linear regression of the yields on the environmental index was adjusted per cultivar. In the complete case, in which all cultivars are present in every block, the environmental indexes for blocks were measured by their average yields (cf. Gusmão, 1985, 1986). When incomplete blocks are used, such as is the case with α -designs, an iterative zig zag algorithm may be used to estimate simultaneously the regression coefficients and the environmental indexes (cf. Mexia et al., 1999).

The upper contour defined by the adjusted regression is a convex polygonal. The cultivars that partake in the upper contour, the dominant ones, have highest yields each for a certain range of the environmental indexes. Non dominant cultivars must be compared with dominant ones.

This technique was systematically studied by Pereira (2004) considering how to use it in cultivar selection. Later Pinto (2005) showed that it is a useful guide for managing plant breeding program.

2 Zig zag algorithm

The goal function to be minimized will be

$$S(\alpha^J, \beta^J, x^b) = \sum_{i=1}^b \sum_{j=1}^J p_{ij} (Y_{ij} - \alpha_j - \beta_j x_i)^2$$

where the $(\alpha_j, \beta_j), j = 1, \dots, J$ are the regression coefficients, the $x_i, i = 1, \dots, b$, are the block environmental indexes and the Y_{ij} denote the yield of cultivar j in block i if present. Usually the weight p_{ij} is 1 [0] when cultivar j is present [absent] from block i . When the cultivar occurs we take $p_{ij} = p_i$. These weights may differ from block to block to express differences in representativity of the blocks.

In the zig zag algorithm the minimization is carried first for the coefficients and then for the environmental indexes. At the end of each iteration the environmental indexes are rescaled so that the range of environmental indexes is kept unchanged.

When we have α -designs the initial environmental indexes for a block will be the average yield for the corresponding super-block.

3 Double minimization

The zig zag algorithm for these adjustments performs well, but it has not been established that it converges to the absolute minimum of the goal function (cf. e.g. Mexia et al., 2001, Mexia and Pereira, 2001) . We presented an alternative algorithm for the adjustment of Joint Regression Analysis and showed that, in the complete case, it converges to the absolute minimum (cf. Pereira and Mexia, 2007). Both algorithms were applied to an example in which they display an excellent agreement. We also analyzed the reason behind such agreement between both algorithms. When we apply the double minimization algorithm we obtain the conditional estimators (given the environmental indexes x_1, \dots, x_b) of the regression coefficients and the conditional minimum for the sum of squares of residuals $\bar{S}(x^b)$. To choose the environmental indexes we minimize $\bar{S}(x^b)$. We point out that the goal function keep unchanged (we get an equivalent regression) if we make a linear transformation on the controlled variable. We can restrict ourselves to vectors x^b , such that

$$\begin{cases} \sum_{i=1}^b p_i x_i &= 0 \\ \sum_{i=1}^b p_i x_i^2 &= 1. \end{cases} \tag{1}$$

As we can see in Pereira and Mexia (2007) the vector x^b which, satisfies constraints (1), and minimizes $\bar{S}^1(x^b) = \sum_{i=1}^b \sum_{j=1}^J p_i \left(Y_{ij} - \frac{\sum_{i=1}^b p_i Y_{ij}}{\sum_{i=1}^b p_i} \right)^2 - x^{bT} Z x^b$ or equivalently maximizes $x^{bT} Z x^b$ is the first eigenvector, γ_1^2 , of matrix $Z = \sum_{j=1}^J Y_j^{+b} Y_j^{+bT}$, with Y_j^{+b} the vector with components $p_i \left(Y_{ij} - \frac{\sum_{i=1}^b p_i Y_{ij}}{\sum_{i=1}^b p_i} \right), i = 1, \dots, b, j = 1, \dots, J$.

4 Cultivar comparison

As stated above the dominant cultivars are associated to ranges for whose environmental indexes they have the highest yields. When comparing a dominant cultivar with a non dominant one we use the highest [lowest] index in the associated range of the second cultivar has larger [lesser] adjusted slope than the dominant one. This comparison may be carried out using one sided t-tests or multiple comparison methods. Of these the Bonferroni method was the one that performed better (cf. Pereira and Mexia, 2002).

5 Method robustness and reproducibility

Recent unpublished work showed that the method is extremely robust. Thus using the data from 17 α -designs with 20 cultivars missing observation were randomly "placed". Starting with the complete data and then with incidences rates from 5% to 75%, for missing observations in triplicate the technique was applied. Surprisingly the selection results varied very little throughout all the study. Moreover, applying the technique to two sets of yield trials in which the same cultivars of winter rye were used we arrived at highly compatible results for both sets of field trials (cf. Pereira and Mexia, 2003a).

6 Recommended cultivars

The technique may be applied to manage lists of recommended cultivars (cf. Pereira and Mexia, 2003b). Thus it may be used to promote new selected cultivars to recommended status and down-grade previously recommended cultivars.

Acknowledgments: The Research Centre for Cultivar Testing (Slupia Wielka, Poland) is thanked for use of their data in this paper. The first author of this work is member of the CIMA-UE, research center financed in the ambit of FEDER by "Programa de financiamento Plurianual", of the Science and Technology Foundation - Portugal.

References

- Gusmão, L. (1985). *An adequate design for regression analysis of yield trials*. *Theor. Appl. Genet.* **71**, 314-319.
- Gusmão, L. (1986). *Inadequacy of blocking in cultivar yield trials*. *Theor. Appl. Genet.* **72**, 98-104.
- Mexia, J.T., Pereira, D.G., and Baeta, J. (1999). *L₂ Environmental indexes*. *Biometrical Letters* **36**, 137-143.
- Mexia, J.T., Pereira, D.G., and Baeta, J. (2001). *Weighted linear joint regression analysis*. *Biometrical Letters* **38**, 33-40.

- Mexia, J.T. and Pereira, D.G. (2001). *Joint regression analysis for winter rye cultivars using L_2 indexes*. *Colloquium Biometryczne* **31**, 207-212.
- Pereira, D.G., and Mexia, J.T. (2002). *Multiple comparison in Joint Regression Analysis with special reference to variety selection*. *Scientific papers of the Agricultural University of Poznan, Agriculture* **3**, 67-74.
- Pereira, D.G., and Mexia, J.T. (2003a). *Reproducibility of Joint Regression Analysis*. *Colloquium Biometryczne* **33**, 279-299.
- Pereira, D.G., and Mexia, J.T. (2003b). *The use of Joint Regression Analysis in selecting recommended cultivars*. *Biuletyn Oceny Odmian (Cultivar Testing Bulletin)* **31**, 19-25.
- Pereira, D.G. (2004). *Análise Conjunta Pesada de Regressões em Redes de Ensaio*. *Ph'd Thesis. Universidade de Evora*.
- Pinto, I. (2005). *Análise Conjunta de Regressões e Planos de Melhoramento*. *Ph'd Thesis. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia*.
- Pereira, D.G., and Mexia, J.T. (2007). *Double minimization for complete series of experiments in Joint Regression Analysis*. Submitted to American Journal of Mathematical and Management Sciences.

Study of the 1st, 2nd and 3rd Guided Interruption Periods in an HIV Clinical Trial

Núria Pérez-Álvarez^{1,3}, Guadalupe Gómez³, Lidia Ruiz² and Bonaventura Clotet^{1,2}

¹ Fundació de Lluita contra la SIDA, Badalona, Spain.

² Fundació Irsicaixa and HIV Clinical Unit, Badalona, Spain.

³ Universitat Politècnica de Catalunya, Barcelona, Spain.

Abstract: In an HIV clinical trial conducted to learn about the safety of an intermittent antiretroviral therapy, the description of the longitudinally measured CD4 counts and the duration time without therapy are of interest. Survival analysis methods and linear mixed models are used to study the relationship between them and the baseline and retrospective covariates. We conclude that the effect of previous immunological and virological markers on the CD4 dynamics and on the times treatment-free diminishes over time. CD4 and VL markers at the end of each stage, duration of previous stages and pharmacological history are relevant for the subsequent times treatment-free.

Keywords: Cox Model; Linear Mixed Model; Guided Treatment interruption.

1 Introduction

The TIBET study is an open-label, comparative, randomized, prospective and multicenter trial conducted to learn about the safety of interrupting the Highly Active AntiRetroviral Therapy (HAART). Patients are randomly allocated to either the control group on which the HAART therapy is maintained or to the experimental group on which HAART is interrupted until one of the following situations occurs: i) an AIDS-defining illness diagnosis; ii) $CD4 \leq 350/mm^3$ or iii) $HIV-1 RNA \geq 100,000$ copies/mL. When any of these events occurs, patients resumed HAART until $CD4 > 500$ cells/ mm^3 and $HIV RNA < 50$ copies/mL. Thereafter, antiretroviral therapy is interrupted again and subsequent cycles of therapy interruption (OFF stage) and reinitiation (ON stage) are followed according to the previously described parameters.

Our aim is to study the duration of the first, second and third period without treatment and to describe the dynamics of the CD4 cell counts in these stages. Survival analysis methods and linear mixed models are used to study the relationship between these responses and the baseline and retrospective covariates, taking as well into account the CD4 and Viral Load measurements in previous stages.

2 Methods

A Cox model is a well-recognized statistical technique that can be used to explore the relationship between the survival of a patient and several explanatory variables.

In this study we have used the Cox proportional-hazards model to assess the effect of baseline, pretherapy, preHAART and retrospective variables on the duration of the 3 first periods from interruption to treatment resumption. The advantages of this semiparametric model are that the regression coefficients are estimated independently of the baseline hazard along as its flexibility to handle time varying coefficients or staggered entry times. Schoenfeld residual plots are used to validate the multivariate models for models with covariates which hold the hazards proportionality assumption. A mixed effect model is used for the repeated measurements of CD4+ cells during the elapsed time on which the patient is without therapy. This methodology, as opposed to classical multivariate regression techniques, is applicable when the number of measurements per patient is unbalanced and fairly small and has the additional feature of allowing some of the parameters describing the CD4 dynamics to vary between patients, and combining the information from all subjects in the estimation of common parameters. To elucidate about the significance of the fixed and random effects parameters, Wald test and likelihood ratio test are used. To validate the model the value of the likelihood together with the analyzes of the residuals are used.

3 Results

The analyzes are conducted for the 100 patients in the interruption group. An interim administrative censoring in July 2004 has insured 96 weeks of follow-up.

3.1 Analyzes of the times from interruption until treatment resumption

Three independent multivariate Cox regression models are conducted for the first three periods without therapy and the most significant covariates are chosen for each case. A multivariate Cox regression model for the **1st interruption period** is fitted for all the predictive factors in the univariate analysis with $p\text{-value} < 0.20$. Using a 5% significance threshold, the final model is described via the CD4+ nadir (Hazard Rate = 0.997; 95%CI = (0.996, 0.999)), the HIV-1 RNA levels before the initiation of any antiretroviral therapy (Hazard Rate = 2.271; 95%CI = (1.514, 3.407)), and the fact that the patient had received suboptimal regimen prior to HAART (Hazard Rate = 0.517; 95%CI = (0.312, 0.856)). We conclude here that subjects with lower CD4+ nadir, higher HIV-1 RNA levels and who had received suboptimal regimen prior to HAART were more likely to resume therapy.

The same strategy is used for the **2nd interruption period** and the following covariates are found statistically significant at a 5% level in the multivariate model: subjects with higher HIV-1 RNA levels before the initiation of any antiretroviral therapy (Hazard Rate = 2.57; 95%CI = (1.48, 4.47)), with higher HIV-1 RNA levels at the starting of the previous therapy (Hazard Rate = 2.83; 95%CI = (1.19, 6.72)) and who had to restart therapy due to a decrease on CD4 counts (Hazard rate = 2.51; 95%CI = (1.15, 5.51)) were related to shorter times without therapy.

Finally, for the **3^d interruption period**, female patients (Hazard Rate = 12.447; 95%CI = 2.974, 52.094)), those subjects who had received suboptimal regimen prior to HAART (Hazard Rate = 0.517; 95%CI = (0.312, 0.856)) and have restarted therapy due

to a decrease on CD4 counts (Hazard Rate= 3.259; 95%CI =(1.0668,9.55)) were more likely to resume therapy.

3.2 Study of the CD4+ cells dynamics

The repeated measurements of the CD4+ cells during the **1st, 2nd and 3^d periods without treatment** is modelled via a linear mixed model such as

$$\log_{10} CD4_{ij} = (\alpha + b_{i0}) + (\beta + b_{i1}) * Week_j + \gamma * CD4w_{0i} + \varepsilon_{ij}$$

where ij indicates the measurement done for the i^{th} individual at week j ; b_{i0} and b_{i1} are the random effects in the intercept and week slope associated with the i^{th} individual and distributed as normal mean 0 random variables and ε_{ij} , the within-subject error term, as a normal mean 0 random variable.

For the **1st period without treatment** the model is summarized as

$$\log_{10} CD4_{ij} = (2.42 + b_{i0}) + (-0.0024 + b_{i1})Week_j + 0.00041CD4w_{0i} + \varepsilon_{ij}$$

where b_{i0} and b_{i1} are $N(0, 0.0942)$ and $N(0, 0.0022)$ respectively; and ε_{ij} is $N(0, 0.0732)$.

A multivariate model is fitted extending the previous one with the retrospective and baseline factors. The CD4 nadir, the \log_{10} preHAART and the CD4 pretherapy interacting with time effect are found significant at 5% level. The effect of having 100 CD4 cells/mm³ more at nadir or baseline imply have 1 cell/mm³ more in the studied CD4 evolution; on the contrary, the increase of 1 \log_{10} preHAART VL entail a reduction of 1 CD4 cell/mm³.

Using analogous steps for the **2nd stage without therapy** the equation is that

$$\log_{10} CD4_{ij} = (2.332 + b_{i0}) - 0.006Week_j + 0.0006CD4w_{0i} + \varepsilon_{ij}$$

where $b_{i0} \sim N(0, 0.10482)$ and $\varepsilon_{ij} \sim N(0, 0.08122)$.

CD4 at starting the 1st period on treatment and the CD4 preHAART are significantly associated with the CD4 evolution in the multivariate model. The CD4 evolution expected is 1 cell/mm³ larger per every 100 CD4 cells at starting the 1st period on, the increase is of 1 CD4 cell per month when CD4 preHAART is 100 cells/mm³ larger. Finally, for the **3^d period without treatment**, the model is summarized as

$$\log_{10} CD4_{ij} = (2.571 + b_{i0}) - 0.0062Week_j + 0.00023CD4w_{0i} + \varepsilon_{ij}$$

where $b_{i0} \sim N(0, 0.1452)$ and $\varepsilon_{ij} \sim N(0, 0.1092)$.

Once adjusting the multivariate model the following covariates are found statistically significant at 5% level: if the CD4 at the starting the 1st period on treatment is 100 cells/mm³ larger the CD4 cells expected is 1 cell/mm³ larger. One month without therapy and being female carries to 1.06 and 2 CD4 cells/mm³ less in the CD4 evolution, respectively.

4 Conclusion

Longitudinal data methods represent one of the principal research strategies employed in medical and social research. The defining feature of such studies is that subjects are measured repeatedly through time. In this paper Cox method to study the time to an event and a mixed model to analyze correlated data have been used and lead to the following clinical conclusion: the effect of previous immunological and virological markers on the CD4+ dynamics and on the times treatment-free, diminishes over time. CD4+ and VL markers at the end of each stage, duration of previous stages and pharmacological history are relevant for the subsequent times without treatment. We determine that treatment interruption was longer and with higher levels of CD4 cells in HIV infected patients with persistent high CD4 and low viral load, for younger males, and for those having a limited exposure to antiretrovirals. These subject and clinical characteristics may confer additive likeliness of safety stay during this treatment-free periods.

References

- Thernau, T.M. and Grambsch, P.M. (2000). *Modelling survival data. Extending the Cox model*. Springer-Verlag.
- Thiebaut, R., Pellegrin, I., Chene, G., Viillard, J.F., Fleury, H., Moreau, J.F., Pellegrin, J.L., Blanco, P. (2005). Immunological markers after long-term treatment interruption in chronically HIV-1 infected patients with CD4 cell count above 400×10^6 cells/l. *AIDS* **19**, 53-61.
- Verbeke, G. and Molenberghs, G. (2000). *Linear mixed model for longitudinal data*. Springer, New York.

A Discretised-Copula-Based Transition Model for Binary Longitudinal Data

Luis Carlos Pérez-Ruíz¹ and Gabriel Escarela¹

¹ Departamento de Matemáticas, Universidad Autónoma Metropolitana, Unideal Cataplasia, AT-321, Av. San Rafael Atlixco No. 186, Col. Vicentina, C.P. 09340, México D.F., Mexico, ge@xanum.uam.mx

Keywords: Maximum likelihood; Markov regression models; Missing observations; Mixture transition distribution model; Serial Correlation.

1 Introduction

The analysis of binary data collected at successive time points to examine the relationship between the probability of success and time dependent covariates can conveniently be modelled using Markov chains. Under this approach, transition probabilities given the past can be determined by a set of covariates. In this paper, we present a general parametric class of transitional models of order p for binary longitudinal data. Specifically, we extend the work of Raftery (1985), giving rise to an autoregressive mixture model that allows for covariates.

In these models the conditional distribution of the current observation given the present and past history is a mixture of conditional distributions, each of them corresponding to the current observation given each one of the p -lagged observations. This likelihood-based methodology is attractive to analyse longitudinal binary data since it includes optimality of estimators under correct model specification, the availability of inferential procedures such as likelihood ratio tests, the flexibility to allow for unequally spaced observations, and the robustness of certain missing data structures.

2 Transition models for discrete data

To establish context, consider first-order discrete time stationary Markov processes, also known as AR(1). If the state space of the process is a finite or countable set, the transition distribution function is given by

$$F_{2|1}(y_2 | y_1) = \sum_{z \leq y_2} \frac{f(y_1, z)}{f_1(y_1)},$$

where $f(y_1, y_2)$ is a joint mass function with both marginal probability functions equal to $f_1(\cdot)$. Although, there are many bivariate discrete distributions in the literature, few of these share the property of having the same marginal distributions.

A conditional probability function can be obtained to define the transition distribution of $Y_t | Y_{t-1}$ when both a copula function C and the distribution of Y_t are given. If $F(y_t)$ represents the marginal cdf of Y_t , the family of transition distributions of $\{Y_t\}$ can be characterized by the discrete transition density function $f_{2|1}(y_t | y_{t-1}) = \Pr\{Y_t = y_t | Y_{t-1} = y_{t-1}\}$ as follows (Joe, 1996):

$$f_{2|1}(y_t | y_{t-1}) = \{C[F(y_t), F(y_{t-1})] - C[F(y_t), F(y_{t-1} - 1)] - C[F(y_t - 1), F(y_{t-1})] + C[F(y_t - 1), F(y_{t-1} - 1)]\} / f(y_{t-1}).$$

In principle, it is possible to extend the transition copula models described above to higher-order representations. In practice, however, this can be quite cumbersome and computationally expensive. A more parsimonious approach than a fully parametric Markov chain can be the *mixture transition distribution* (MTD) model of order p introduced by Raftery (1985) which is characterized by the following conditional density

$$f(y_t | y^{[t,p]}) = \sum_{k=1}^p \omega_k f_k(y_t | y_{t-k}), \tag{1}$$

where $y^{[t,p]} = (y_{t-1}, \dots, y_{t-p})$, $\sum_{k=1}^p \omega_k = 1$, $\omega_k \geq 0$, $k = 1, \dots, p$, and $f_k(y_t | y_{t-k})$ denotes the one-step ahead transition density corresponding to lag y_{t-k} . Raftery showed that the lagged bivariate distributions satisfy a system of matrix equations similar to the Yule-Walker equations, and that the past values Y_{t-1}, \dots, Y_{t-p} do not interact among them in their effect on the conditional distribution of Y_t given the past.

To construct each of the one-step conditional densities in equation (1), we employ the copula transition models described above by choosing both a family of copulas and a family of distributions for the underlying marginals. In this study, we use the bivariate distribution function belonging to the Gaussian copula, which has the form

$$C_\rho(v_1, v_2) = \Phi_2[\Phi^{-1}(v_1), \Phi^{-1}(v_2)], \quad (v_1, v_2)^T \in (0, 1)^2,$$

where $\Phi_2(\cdot, \cdot)$ is the cdf of a bivariate Gaussian distribution with mean $(0, 0)^T$ and covariance matrix \mathbf{R} equal to an 2×2 non-singular matrix with the off-diagonal elements equal to ρ , with $\rho \in (-1, 1)$, and the diagonal elements equal to one, and $\Phi^{-1}(\cdot)$ is the inverse function of the standard Gaussian cumulative distribution. The use of the bivariate Gaussian copula is appealing since it encodes dependence in the same way that the bivariate normal distribution does using the dependence parameter ρ , with the difference that it does so for random variables with any arbitrary marginals.

A MTD model for a two-state Markov chain can straightforwardly be obtained by setting the marginals F_k equal to Bernoulli distributions with probability of success p_k , and the copula functions $C^{[k]}(u, v)$ equal to Gaussian copulas $C_{\rho_k}(u, v)$ with dependence parameter ρ_k . In the regression setting, it is possible to model the conditional probability functions of Y_t given the k -th lag as functions of covariates \mathbf{x}_t . Thus, a more general MTD model assumes that the k -th marginal has the following conditional cumulative distribution function of Y_t with covariates \mathbf{x}_t

$$F_k(y_t; \mathbf{x}_t) = q_k(\mathbf{x}_t) I_{[0,1)}(y_t) + I_{[1,\infty)}(y_t), \quad y_t \in (-\infty, \infty),$$

where $q_k(\mathbf{x}_t) = 1 - p_k(\mathbf{x}_t)$, $p_k(\mathbf{x}_t)$ is the probability of success given in terms of the covariates at time t , and $I_A(y)$ is the indicator function of A , which equals 1 if $y \in A$ and equals 0 otherwise.

Using the properties of the copula, the resulting one-step transition probability functions, given by

$$p_{l|m}^{(k)} = f(l | m) = \Pr\{Y_t = l \mid Y_{t-k} = m\}, \quad l, m \in \{0, 1\},$$

can be represented as

$$\begin{aligned} p_{0|0}^{(k)} &= C_{\rho_k}[q_k(\mathbf{x}_t), q_k(\mathbf{x}_{t-k})]/q_k(\mathbf{x}_{t-k}), \\ p_{0|1}^{(k)} &= \{q_k(\mathbf{x}_t) - C_{\rho_k}[q_k(\mathbf{x}_t), q_k(\mathbf{x}_{t-k})]\}/p_k(\mathbf{x}_{t-k}), \\ p_{1|0}^{(k)} &= \{q_k(\mathbf{x}_{t-k}) - C_{\rho_k}[q_k(\mathbf{x}_t), q_k(\mathbf{x}_{t-k})]\}/q_k(\mathbf{x}_{t-k}), \\ p_{1|1}^{(k)} &= \{1 - q_k(\mathbf{x}_{t-k}) - q_k(\mathbf{x}_t) + C_{\rho_k}[q_k(\mathbf{x}_t), q_k(\mathbf{x}_{t-k})]\}/p_k(\mathbf{x}_{t-k}). \end{aligned}$$

A convenient way to account for concomitant information in this model is to use the probit link in the marginal probability model. This link implies that $p_k(\mathbf{x}_t) = \Phi(\boldsymbol{\beta}_k^T \mathbf{x}_t)$, where Φ is the standard normal cumulative distribution function and $\boldsymbol{\beta}_k$ is the vector of coefficients. Using the symmetry property of the standard normal distribution, we find that $q_k(\mathbf{x}_t) = \Phi(-\boldsymbol{\beta}_k^T \mathbf{x}_t)$, and thus, under the normal copula, we obtain $C_{\rho_k}[q_k(\mathbf{x}_t), q_k(\mathbf{x}_{t-k})] = \Phi_2(-\boldsymbol{\beta}_k^T \mathbf{x}_t, -\boldsymbol{\beta}_k^T \mathbf{x}_{t-k})$; here, Φ_2 has dependence parameter equal to ρ_k .

3 Inference and diagnostics

Consider m equally spaced responses for the i -th individual $y_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, m$. The contribution to the likelihood for the i -th subject can be written as

$$L_i = f(y_{i1}; \mathbf{x}_{i1}) \prod_{k=2}^m f_k(y_{ik} \mid H_i^{[k,j]}), \quad (2)$$

where $j = \min(k-1, p)$, and $H_i^{[k,j]} = \{y_{i,k-1}, \dots, y_{i,k-j}; \mathbf{x}_{i,k}, \mathbf{x}_{i,k-1}, \dots, \mathbf{x}_{i,k-j}\}$. The mixture transition model is not linear, so numerical techniques need to be used in order to find the maximum likelihood estimators. We used the function `nlm` of the R language to minimize $-2 \times \prod_{i=1}^n L_i$, the deviance. For convenience we estimated the parameter $\text{arctanh}(\rho_k)$, which takes values in $(-\infty, \infty)$ and guarantees that $-1 < \rho_k < 1$.

Our approach to applying the transition models is to determine the order of the Markov chain prior to inference about the regression and dependence parameters. In this study we consider the Bayesian information criterion (BIC) as our main model choice criterion. Once the order of the model is fixed, the process of finding a parsimonious model can follow standard ideas based on the likelihood ratio; for example, that used in the theory of generalised linear models.

The assumptions made about a particular transitional model may be checked by calculating the conditional randomized quantile residuals proposed by Dunn and Smyth (1996). Under the assumed model, such residuals are exactly normal and are found by inverting the fitted conditional distribution function for each response values and finding the equivalent standard normal quantile. Thus, some simple plots for checking that the randomized quantile residuals are observed values of independent and normally standard distributed random variates should indicate the quality of the fit.

4 A real-data illustration

We analyse data from a clinical trial designed to assess the reduction of sun exposure in children. This data set is from the research of Lori A. Crane, Ph.D., University of Colorado Health Sciences Center (NIH CA 74592). Infants were enrolled prior to 6 months of age, and were randomized into an intervention or a control group. Parents of the treatment group received sun protection kits and advice at well-child visits between 2 and 36 months of age. Annual parent surveys at 1, 2 and 3 years of age were used to follow the intervention up. In this study, we categorize the response “avoids mid-day sun exposure” as never or seldom (coded 0) and frequently or always (coded 1). The two explanatory variables were the factor **Treatment** and the **time** of the observation. There were 680 children in the follow-up study from which, in addition to complete record cases, missing observations were obtained, so a child could have 1, 2 or 3 observations.

Using the Gaussian copula and a probit model in the Bernoulli distribution to construct each of the k one-step transition functions described above, we proceed to fit the discrete transition regression model to the sun exposure data. To handle cases where observations are missing, we used the assumption that the data are missing at random since the reason the data are missing at certain time points is independent of the outcomes at those time points (Little and Rubin, 1987).

To cover all possible situations of missingness, the likelihood in Eq. (2) has to be modified by taking into account only the indexes corresponding to observed data. For instance, when only the second observation is missing, the likelihood replaces the one-step probabilities by two-step probabilities $\Pr\{Y_{i3} = y_{i3} \mid Y_{i1} = y_{i1}\}$, which are computed by multiplying the transition matrices corresponding to each step.

Employing the BIC criterion, we found that, in presence of both covariates and the corresponding interaction, the best-fitting model is an AR(1). Since the change in deviance was negligible when the interaction term was removed from the AR(1) model, we concluded that such an interaction has no significant effects. The two remaining effects, on the other hand, showed significant effects. Thus, we concluded that the most parsimonious model includes both covariates.

The parameter estimates and their standard errors corresponding to the most parsimonious model are shown in Table 1. The results indicate that the effects corresponding to **Treatment** are significant (p -value=0.024), suggesting that the transitions between avoiding midday sun and not avoiding the midday sun differs between the intervention group and the control group. The negative coefficient corresponding to **time** and its statistical significance (p -value < 0.0001) indicates that, after accounting for **Treatment**,

TABLE 1. Maximum likelihood estimates using the AR(1) model for the sun exposure data.

Parameter	Estimate	S.E.
Intercept	1.376	0.101
time	-0.334	0.040
Treatment	0.184	0.081
$\text{arctanh}(\rho)$	0.642	0.066

the exposure tends to decrease as the number of parent interviews increases. The one year lag serial dependence ρ is estimated as 0.564 with a 95% confidence band of (0.469, 0.646), which represents an association between adjacent responses significantly different from zero.

After carrying out a residual analysis, which consists of kernel density, normal Q-Q plot, autocorrelation and partial autocorrelation functions of the randomized quantile residuals, we found that the proposed model appears to fit the data reasonably well.

5 Discussion

This paper introduces a flexible class of models for serial binary response data that permits feasible likelihood-based Markov regression analysis. Our method is a natural extension of the AR(p) model for Gaussian mixtures given by Le *et. al.* (1996). The discretised-copula model introduced here constructs separate models for the marginal response and for the dependence among longitudinal observations.

Families of copulas different from the Gaussian can also be used. Further, since there can be considerable differences in the shapes of conditional distributions generated by the discretised copula model, even when the fitted marginal distribution and estimated correlations are similar, it is important to choose correctly a family of copulas and a marginal distribution for each conditional distribution in the mixture model. How best to do this remains an open question. Hence it is, as always, very important to check the validity of a proposed model before any inferences are drawn.

Although the methods presented here are focused on equally spaced data, the discretised Gaussian copula model lends itself to formulating growth curve analysis in the form of a continuous-time autoregressive structure (eg Jones and Boadi-Boateng, 1991) to study longitudinal data when each subject is observed at different unequally spaced time points. If the Markov dependence assumption remains plausible, the methodology can be modified to permit general time spacing. This aspect warrants further research.

Acknowledgments: The second co-author wishes to thank CONACYT, Mexico, for their financial support.

References

- Dunn, P.K. and Smyth, G.K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* **5**, 236-244.
- Joe, H. (1997). *Multivariate Models and Dependence Concepts*. New York: Chapman & Hall.
- Jones, R.H. and Boadi-Boateng, F. (1991). Unequally Spaced Longitudinal Data with AR(1) Serial Correlation. *Biometrics* **47**, 161-175.
- Le, N.D., Martin, R.D. and Raftery, A.E. (1996). Modeling flat stretches, burst and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association* **91**, 1504-1515.
- Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- Raftery, A.E. (1985). A model for High-Order Markov Chains. *Journal of the Royal Statistical Society, Series B* **47**, 528-539.

Damage detection of structures by analyzing embedded time series of vibration signals

Christian Pfeifer¹, Michael Oberguggenberger² and Alexander Ostermann¹

¹ Institut für Mathematik, Technikerstraße 13, A-6020 Innsbruck

² Institut für Grundlagen der Bauingenieurwissenschaften, Technikerstraße 13, A-6020 Innsbruck

Abstract: In this study we are using time series techniques applied to vibration signals in order to identify damages of building structures. In contrast to other papers calculations are done with real data as well as with simulated data. Results based on simulated data are compared with those based on real data.

Keywords: Damage detection, time series.

1 Introduction

Detecting damages of structures in civil engineering (e.g. bridges) caused by environmental influences has become of great importance in the last years. This can be done by analyzing the vibration signals of structures. Using the vibration signals, there are roughly two different proposals in order to recognize structural changes of buildings: One way is to observe the modal frequencies of the vibration signals applying Fast Fourier Transform. Changes of modal frequencies can be seen as an indicator for structural damages (changes of stiffness). Our proposal is to analyse embedded time series of vibration signals. Changes of estimated time series parameters indicate structural changes. In [1] simulated data of a steel frame structure were analyzed using the time series approach. The authors report that damages of the structure have an influence on the parameters. We confirm these observations using real data processed in a slightly different way, supply theoretical background and compare with results of simulations.

2 Continuous model

Engineering structures are usually modelled by finite elements. In the dynamic case, this leads to the following linear differential equation:

$$\dot{X}_t = AX_t + f_t, \quad X_0 = c \quad (1)$$

where t , $t \geq 0$, denotes time and X_t is a vector comprising the displacements and velocities of the nodes. A denotes a matrix dependent on mass, stiffness and damping

of the structure. The vector f_t characterizes the external force input. In our case the force vector is assumed to be of ambient type (such as wind, traffic, seismic activity) and can be seen as a random input using white noise:

$$f_t = B\dot{W}_t \quad (2)$$

where \dot{W}_t is the derivative of a Wiener process. Thus equation (1) turns into a stochastic linear differential equation and the solution can be written as a stochastic integral:

$$X_t = e^{At}c + \int_0^t e^{A(t-s)}B \, dW_s \quad (3)$$

3 Embedded discrete stochastic process

In general it is not possible to observe the continuous stochastic process X_t , $t \geq 0$. Usually we make observations on equally spaced points of time $j\tau$, $j = 0, 1, 2, 3, \dots$, τ chosen freely. This leads to the embedded discrete process Y_j :

$$Y_j = X_{j\tau}, \quad j = 0, 1, 2, 3, \dots \quad (4)$$

X_t is an asymptotically stationary Gaussian process if the eigenvalues of A have negative real parts [2]. As a consequence of this the discrete process Y_j and its components are stationary, too. This encourages us to take ARMA(p, q) time series models [3] into consideration such as ($p = 2, q = 1$):

$$Y_j = \varphi_1 Y_{j-1} + \varphi_2 Y_{j-2} + Z_j + \theta_1 Z_{j-1} \quad (5)$$

where φ_1 , φ_2 and θ_1 denote the autoregressive and moving average parameters. Moreover, Z_j denotes a sequence of iid Gaussian random variables with zero mean. In case of a damped spring-mass oscillator the matrix A is of the form:

$$A = \begin{pmatrix} 0 & 1 \\ -r/m & -k/m \end{pmatrix} \quad (6)$$

Here m denotes the mass, k the stiffness and r the damping factor of the spring. For this simple case we can show that the embedded discrete process is equal to an ARMA(2,1) process according to (5) up to a small Gaussian error term with variance of order τ^2 . The parameters φ_1 , φ_2 and θ_1 turn out to be:

$$\varphi_1 = 2e^{-r\tau/2m} \cos \omega\tau, \quad \varphi_2 = e^{-r\tau/m}, \quad \theta_1 = -1 \quad (7)$$

where $\omega = \sqrt{\frac{k}{m} - \frac{r^2}{4m^2}}$. As we can see the autoregressive parameters φ_1 and φ_2 are dependent on the stiffness parameter k . This parameter may be used as an indicator of damage of the system. However, things become more complex if we consider arbitrary finite element models. A transformation of variables leads to simple subsystems as described in (6).

4 Real data example

In [1] time series analysis was done with simulated data generated by a `Matlab`-routine. In contrast to this our calculations are based on data of a real small steel truss bridge (span of the bridge 8 m). The vibration signals are the result of ambient excitations (wind, traffic, seismic activity). Getting signals from high precision sensors we proceed as follows:

- Choose time span τ and time points $0, \tau, 2\tau, \dots$ where samples $Y_j = X_{j\tau}$ are taken from the vibration signals.
- Split up raw data into N subsamples of length L (in our case $L = 2500$).
- Normalise each subsample according to $\frac{x-\bar{x}}{\sigma_x}$.
- Identify model orders p and q of an ARMA(p, q) process fitted to the normalized data of the undamaged structure using the Akaike Information Criterion (AIC).
- Estimate autoregressive parameters for each subsample of the undamaged and the damaged structure.
- Compare parameters of the undamaged case with parameters of the damaged case.

In contrast to [1] we use the normalized data without any filtering (see theoretical considerations above). It turned out that the choice of the time span τ has an influence on the number of ARMA parameters. If we increase the time span τ , it seems to be that the number of parameters is reduced. This observation is supported in the spring-mass oscillator case, where the coefficient φ_2 tends to zero much faster than φ_1 as τ increases, thus asymptotically reducing the ARMA(2, 1)-model to an ARMA(1, 1)-model.

Figure 1 shows results based on a vibration signal recorded in the middle of the bridge. Using time span $\tau = \frac{1}{100}$ sec we choose (AIC) an ARMA(2,2) model in this case. In Figure 1 (left hand side), boxplots of the first autoregressive coefficient φ_1 dependent on the damage level (from 0, undamaged, to 7, maximum damage level) are given. The damage levels are supposed to be cumulative. In Figure 1 (right hand side), results based on a finite element simulation model of a simple bridge are presented for comparison. Choosing an ARMA(4,2) model for these data boxplots of the first autoregressive coefficient φ_1 dependent on damage levels (0-4), which are cumulative, are given. We notice that differences of φ_1 between damage levels are much smaller in the real data case than in the simulation case. Indeed, in the real data case, there are observable differences only in the highest damage levels. The differences in the simulation case are highly significant with respect to all levels. For both cases φ_1 turns out to be a monotone function of the damage level.

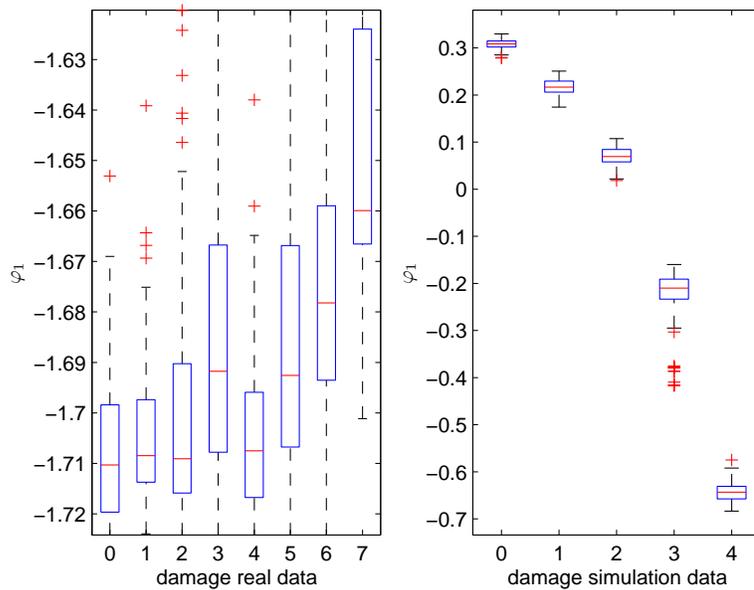


Figure 1: Boxplots of first autoregressive coefficient φ_1 dependent on damage level for real data case (left) and simulation data case (right)

5 Conclusion

We have presented a method of fitting ARMA-models to discrete processes imbedded in continuous stochastic systems. Changes of the coefficients in the continuous system are seen to produce changes in the ARMA parameters. Thus damage of the structure can be detected by observing the ARMA parameters. However, it turned out that the influence of damage in the real data example was less noticeable than in the simulation case.

References

- Arnold, L. (1974). *Stochastic Differential Equations: Theory and Applications*. Wiley, New York.
- Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting*. Springer, New York.
- Nair, W.K.K., Kiremidjian, A.S. and Law, K.H. (2006). Time series based damage detection and localization algorithm with application to the ASCE benchmark structure. *Journal of Sound and Vibration* **291**, 349–368.

A product-multinomial framework for categorical data analysis with missing responses

Frederico Z. Poletto¹, Julio M. Singer¹, and Carlos Daniel Paulino²

¹ Departamento de Estatística, Instituto de Matemática e Estatística, Universidade de São Paulo, Caixa Postal 66281, São Paulo, SP, 05311-970, Brazil

² Departamento de Matemática and CEMAT, Instituto Superior Técnico, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1049-001, Lisboa, Portugal

Abstract: We extend the multinomial modelling scenario for the analysis of categorical data with missing responses described by Paulino (1991, *Brazilian Journal of Probability and Statistics*, **5**, 1-42) to the product-multinomial setup so that the inclusion of explanatory variables is allowed. Assuming an ignorable missing data mechanism, linear and log-linear models may be fitted via maximum likelihood. Weighted least squares methodology may also be used to fit more general functional linear models, if a missing completely at random mechanism is assumed. We also consider a hybrid approach, where any missingness process is fitted by maximum likelihood in a first step, and the estimated marginal probabilities of categorization and their covariance matrix are used in a second stage to fit the model via weighted least squares, in the spirit of functional asymptotic regression methodology. All the methods were computationally implemented via subroutines written in R.

Keywords: Incomplete categorical data; MAR, MCAR, and MNAR; Linear, log-linear, and functional linear models.

1 Problem description and notation

For simplicity, we admit that the random vector \mathbf{Y} of response variables can assume R values \mathbf{y} corresponding to combinations of the levels of its components, Y_1, Y_2, \dots, Y_k . For instance, when $\mathbf{Y} = (Y_1, Y_2, Y_3)'$, and Y_1, Y_2 , and Y_3 may assume, respectively, 2, 3, and 5 different values, $R = 2 \times 3 \times 5 = 30$. Likewise, we assume that the vector \mathbf{X} of explanatory variables can take S values \mathbf{x} , corresponding to the combinations of the levels of its components, X_1, X_2, \dots, X_q . The R response categories are indexed by r and the S subpopulations, by s .

We assume that each one of the n_{s++} sampling units randomly selected from the s -th subpopulation can be independently classified into the r -th response category with the same probability $\theta_{r(s)}$, $r = 1, \dots, R$, $s = 1, \dots, S$.

For several reasons, it may not be possible to completely observe the responses of all variables from \mathbf{Y} . In these cases, only part of the n_{s++} sampling units is classified into one of the R originally defined response categories, while the remaining units are associated to some type of missingness. For subpopulation s , $s = 1, \dots, S$, we define T_s missingness patterns in the following way. The set of units with no missing data (*i.e.*, complete classification) is represented by $t = 1$ and the sets that have some degree

of missingness, by $t = 2, \dots, T_s$. We admit that the units corresponding to the t -th missingness pattern, $t = 2, \dots, T_s$, are recorded in response classes \mathcal{C}_{stc} , $c = 1, \dots, R_{st}$, constituted by at least two of the R response categories, with $\mathcal{C}_{stc} \cap \mathcal{C}_{std} = \emptyset$, $c \neq d$ and $\cup_{c=1}^{R_{st}} \mathcal{C}_{stc} = \{1, \dots, R\}$. Thus, each one of the $t = 2, \dots, T_s$ missingness patterns form partitions $\mathcal{P}_{st} = \{\mathcal{C}_{stc}, c = 1, \dots, R_{st}\}$ of the complete classification pattern $\mathcal{P}_{s1} = \mathcal{P}_1 = \{\{r\}, r = 1, \dots, R\}$ and R_{st} denotes the number of response classes with the t -th missingness pattern for the s -th subpopulation.

We assume that a sampling unit selected from the s -th subpopulation with the r -th response category is classified into the t -th missingness pattern with probability $\lambda_{t(rs)}$, $r = 1, \dots, R$, $s = 1, \dots, S$, $t = 1, \dots, T_s$. The $\{\lambda_{t(rs)}\}$ are the conditional probabilities of missingness and the $\{\theta_{r(s)}\}$ are the marginal probabilities of categorization. We assume that there are no missing values in \mathbf{X} .

2 Probabilistic model and missingness mechanisms

We assume that the observable frequencies $\{n_{stc}\}$ have a product-multinomial distribution expressed by the probability mass function

$$\prod_{s=1}^S \frac{n_{s++}!}{\prod_{t=1}^{T_s} \prod_{c=1}^{R_{st}} n_{stc}!} \prod_{r=1}^R (\theta_{r(s)} \lambda_{1(rs)})^{n_{s1r}} \prod_{t=2}^{T_s} \prod_{c=1}^{R_{st}} \left(\sum_{r \in \mathcal{C}} \theta_{r(s)} \lambda_{t(rs)} \right)^{n_{stc}}, \quad (1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}'_s, s = 1, \dots, S)'$, $\boldsymbol{\theta}_s = (\theta_{r(s)}, r = 1, \dots, R)'$, $\sum_{r=1}^R \theta_{r(s)} = 1$, $s = 1, \dots, S$, and $\sum_{t=1}^{T_s} \lambda_{t(rs)} = 1$, $r = 1, \dots, R$, $s = 1, \dots, S$. This factorization into a marginal model for the measurements, $\{\theta_{r(s)}\}$, and a conditional model for the missingness process given the measurements, $\{\lambda_{t(rs)}\}$, corresponds to the so-called selection model framework (Little and Rubin, 2002).

If it were possible to identify the response category of every observation in each of the missingness patterns, y_{str} would be the hypothetical number of sampling units from the s th subpopulation with the t th missingness pattern classified into the r th response category, $s = 1, \dots, S$, $t = 1, \dots, T_s$, $r = 1, \dots, R$. Hence, $\{y_{str}\}$ denote the augmented frequencies, which are observed only under the missingness pattern $t = 1$ (no missing data), where $n_{s1r} = y_{s1r}$, $s = 1, \dots, S$, $r = 1, \dots, R$. Under the other patterns such frequencies are non-observable and we know solely the frequencies associated to the response classes \mathcal{C}_{stc} , namely $n_{stc} = \sum_{r \in \mathcal{C}_{stc}} y_{str}$, $s = 1, \dots, S$, $t = 2, \dots, T_s$, $c = 1, \dots, R_{st}$. Therefore, the $R \sum_{s=1}^S T_s - S$ linearly independent parameters $\{\theta_{r(s)}, \lambda_{t(rs)}\}$ (associated to the augmented frequencies $\{y_{str}\}$) when faced with the $S(R - 1) + \sum_{s=1}^S l_s$ linearly independent observable frequencies $\{n_{stc}\}$ (associated to the parameters $\{\sum_{r \in \mathcal{C}_{stc}} \theta_{r(s)} \lambda_{t(rs)}\}$) highlight an overparameterization of (1) with $\sum_{s=1}^S [R(T_s - 1) - l_s]$ non-identifiable parameters, where $l_s = \sum_{t=2}^{T_s} R_{st}$.

As the interest usually lies in $\{\theta_{r(s)}\}$, reduced structures are considered for $\{\lambda_{t(rs)}\}$ to render the model identifiable. The most common way to overcome this problem is by assuming a non-informative missingness mechanism or, according to Little and Rubin (2002), a missing at random (MAR) mechanism, expressed by

$$\text{MAR} : \lambda_{t(rs)} = \alpha_{t(cs)}, \quad s = 1, \dots, S, t = 1, \dots, T_s, c = 1, \dots, R_{st}, r \in \mathcal{C}_{stc},$$

indicating that the conditional probabilities of missingness depend only on the observed response classes and, conditionally on these, they do not depend on the unobserved response categories. The missing completely at random (MCAR) mechanism, a special case of the MAR mechanism, is defined by

$$\text{MCAR} : \lambda_{t(rs)} = \alpha_{t(s)}, \quad s = 1, \dots, S, \quad t = 1, \dots, T_s, \quad r = 1, \dots, R,$$

and implies that the conditional probabilities of missingness do not depend on the response categories, being or not partially observed. The statistical model under the MAR mechanism is saturated and, under the MCAR mechanism has $S + \sum_{s=1}^S (l_s - T_s)$ degrees of freedom. Both mechanisms lead to factorizations of the likelihood involving one term depending on $\{\theta_{r(s)}\}$ but not on $\{\lambda_{t(rs)}\}$ and another term depending on $\{\lambda_{t(rs)}\}$ but not on $\{\theta_{r(s)}\}$. So, when $\{\theta_{r(s)}\}$ and $\{\lambda_{t(rs)}\}$ are functionally independent, we can ignore any of these mechanisms for likelihood inferences about $\{\theta_{r(s)}\}$. In addition, under the MCAR mechanism, the inferences about $\{\theta_{r(s)}\}$ can be based only on the distribution of $\{n_{stc}\}$ conditionally on $\{n_{st+}\}$ and, hence, this missingness mechanism can also be ignored for frequentist inferences. However, the MAR mechanism is not ignorable for frequentist inferences on $\{\theta_{r(s)}\}$. See Paulino (1991) for details in the multinomial setting.

Missing not at random (MNAR) or informative missingness mechanisms can be formulated by assuming that at least two conditional probabilities of missingness of response categories pertaining to the same class are not equal, *i.e.*, $\{a, b\} \in \mathcal{C}_{stc}$ and $\lambda_{t(as)} \neq \lambda_{t(bs)}$. Nevertheless, it is necessary to specify at least $\sum_{s=1}^S [R(T_s - 1) - l_s]$ parametric constraints to obtain an identifiable structure.

3 Inferences on structural models

We consider linear $[\mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}]$, log-linear $[\mathbf{A} \ln(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}]$, and functional linear $[\mathbf{F}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}]$ models written in the form of freedom equations or, alternatively, expressed in the equivalent constraint formulations $[\mathbf{U}\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$, $\mathbf{U} \mathbf{A} \ln(\boldsymbol{\theta}) = \mathbf{0}$, and $\mathbf{U}\mathbf{F}(\boldsymbol{\theta}) = \mathbf{0}]$, where the elements of \mathbf{A} and \mathbf{U} are usually equal to 1, 0, or -1 and must satisfy certain restrictions; \mathbf{X} is the model specification matrix; and $\mathbf{F}(\boldsymbol{\theta})$ is generally expressed as composition of linear, logarithmic, and exponential functions or addition of constants. See Koch *et al.* (1985) for further details in the context of complete data. We obtain the score vector, the hessian and the Fisher information matrices under the MAR and the MCAR mechanisms for the linear, log-linear, and saturated models allowing them to be fitted by maximum likelihood (ML). Weighted least squares (WLS) methodology can be used to fit saturated and functional linear models under the MCAR mechanism. We also use a hybrid methodology, where any missingness mechanism is fitted by ML in a first step using a saturated model for $\boldsymbol{\theta}$, and the estimates of $\boldsymbol{\theta}$ and of its covariance matrix are used in a second stage to fit a functional linear model via WLS, in the spirit of functional asymptotic regression methodology described by Imrey *et al.* (1981, 1982) for complete data. Computational subroutines written in R were developed. The required computations are automatically conducted by the designed functions when MAR or MCAR mechanisms are considered. For MNAR mechanisms, the first step must be programmed, by means of one of the built-in optimization functions in R, in order to obtain the ML estimates.

Acknowledgments: This research received financial support from Brazil (CNPq and FAPESP) and Portugal (Fundação Calouste Gulbenkian).

References

- Imrey, P.B., Koch, G.G., Stokes, M.E. *et al.* (1981, 1982). Categorical data analysis: some reflections on the log linear model and logistic regression. *International Statistical Review*, Part I: historical and methodological overview, **49**, 265-283; Part II: data analysis, **50**, 35-63.
- Koch, G.G., Imrey, P.B., Singer, J.M., Atkinson, S.S., and Stokes, M.E. (1985). *Analysis of Categorical Data*. Montréal: Les Presses de L'Université de Montréal.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: John Wiley & Sons.
- Paulino, C.D. (1991). Analysis of incomplete categorical data: a survey of the conditional maximum likelihood and weighted least squares approaches. *Brazilian Journal of Probability and Statistics* **5**, 1-42.

Regression Modelling of Competing Risks in a Bladder Cancer Study

Núria Porta¹, M.Luz Calle², Guadalupe Gómez¹ and Núria Malats³

¹ Universitat Politècnica de Catalunya, Barcelona, Spain

² Universitat de Via, Barcelona, Spain

³ Institut Municipal d'Investigació Mèdica, Barcelona, Spain

Abstract: Competing risks data usually arises in studies in which the failure of an individual may be classified into one of k ($k > 1$) mutually exclusive causes of failure. Regression modelling for competing risks is undertaken by specifying models for the cause-specific hazard functions or for the cumulative incidence functions. We review three of these methods, and we focus on their distinct methodological features, on how to interpret parameters and how to validate models. The methodologies are illustrated with data from a bladder cancer study.

Keywords: Competing risks; cause-specific hazards; cumulative incidence function.

Student Oral Presentation - Núria Porta E-mail: nuria.porta-bleda@upc.edu.

1 Introduction

The "Spanish Bladder Cancer Study" is a multicenter study with 1356 newly diagnosed bladder cancer cases. Recurrences of the tumour remain common among bladder cancer patients, and efforts to reduce them are of paramount clinical importance. The aim of the study is to characterize different courses of the disease. After the start of first-line therapy, the evolution of the disease may lead to different scenarios of interest: (i) recurrence, if the tumour reappears and is classified as superficial; (ii) progression, if the new tumour is classified as invasive and (iii) death, if the subject dies due to bladder cancer. Let T be the time from diagnose to the first failure, which may be due to recurrence, progression or death due to bladder cancer. Let C be the cause of failure. In order to identify distinct patterns of the disease it is necessary to explore the joint distribution of (T, C) .

This type of survival data belongs to what is known as competing risks data, which usually arises in studies in which the event of interest of an individual may be classified into one of k ($k > 1$) mutually exclusive causes of failure. For each individual, a time to failure, T , and a cause of failure, C , are observed. The joint distribution of (T, C) is completely specified through either the cause-specific hazards, $\lambda_j(t)$, or through the cumulative incidence functions, $F_j(t)$. Modelling these two functions leads to different types of regression models when covariates are present.

In this work we explore distinct methodologies for the regression modelling of competing risks. Three methodologies are considered: Cox's proportional hazards model (Prentice *et al.*, 1978), Aalen's additive hazards model (Aalen, 1993) and Fine and Gray's model for the subdistribution function (Fine and Gray, 1999), which are briefly

described in the next section. The three methodologies are applied to the Spanish Bladder Cancer Study and compared in terms of interpretation of the involved parameters and diagnostic tools for model assumptions.

2 Regression modelling in competing risks

Define, for each individual, the pair (T, C) , T being the failure time, and C the failure cause. T is assumed to be a continuous and positive random variable, and C takes values in the finite set $\{1, \dots, k\}$. We remark here that the individual fails from one and only one cause. The cause-specific hazard function for the j^{th} cause

$$\lambda_j(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr(T < t + \Delta t, C = j | T \geq t)}{\Delta t} \quad j = 1, \dots, k$$

represents the rate of occurrence of the j^{th} failure in the presence of all the concurrent causes. On the other hand, the subdistribution function from type j failure

$$F_j(t) = Pr(T \leq t, C = j) \quad j = 1, \dots, k$$

define the probability of the subject failing from cause j in the presence of all the competing risks. The F_j 's functions, which are typically of interest, are referred to as cumulative incidence functions (CIF's).

The following equations show how to compute the total hazard $\lambda(t)$, and the survival function $S(t)$, and the relationship between the λ_j 's and the F_j 's:

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{Pr(T < t + \Delta t | T \geq t)}{\Delta t} = \sum_{j=1}^k \lambda_j(t), \\ S(t) &= Pr(T > t) = e^{-\int_0^t \sum_{j=1}^k \lambda_j(u) du} \\ F_j(t) &= \int_0^t \lambda_j(u) S(u) du \quad j = 1, \dots, k. \end{aligned}$$

Regression modelling for competing risks is undertaken by specifying models for the cause-specific hazard functions $\lambda_j(t|\mathbf{x})$ or for the cumulative incidence functions $F_j(t|\mathbf{x})$, given a vector of covariates \mathbf{x} . In what follows, we present two methods based on the cause-specific hazards and one method based on the cumulative incidence function.

The classical analysis establishes a Cox proportional hazards model (Prentice *et al.* 1978) for each of the cause-specific hazard:

$$\lambda_j(t|\mathbf{x}) = \lambda_{0j} e^{\mathbf{x}^t \beta_j} \quad j = 1, \dots, k,$$

where β_j is a $p \times 1$ vector of regression coefficients for each cause. We could analyse each cause of failure separately, treating individuals failing from other causes as censored observations. The effect of the covariates is assumed to act multiplicatively on an unknown baseline hazard function. Estimation of the regression parameters β_j is based

on partial likelihood. The Cox methodology has been widely discussed, becoming the standard analysis to perform in regression modelling (see Lawless, 2003, for example). In long follow up studies, models with constant parameters along time, such as the Cox model, may be inappropriate, since time may influence the effect of a covariate in the hazard. An alternative is found in the methodology proposed by Aalen (1993, 2001), in which an additive hazards model for each cause-specific hazard is specified:

$$\lambda_j(t|\mathbf{Z}(t)) = \beta_{j0}(t) + \mathbf{Z}(t)^\dagger \beta_j(t) \quad j = 1, \dots, k.$$

Unlike in the Cox setting, this model assumes that the covariates act in an additive manner on the unknown baseline hazard function. This effect is assessed through the time-dependent functions $\beta_j(t)$, so its evolution along time can be explored. Least-squares techniques are employed here to derive estimates of $\mathbf{B}_j(t) = \int_0^t \beta_j(u) du$. Additive models are a convenient option when proportional hazards do not hold. Though these models are very flexible and easy to implement, they are less used than Cox model since inference regarding its nonparametric terms is not fully developed.

The modelling of the cause-specific hazards applies when the goal is to assess if a factor is associated with the risk of a specific cause of failure. However, when the goal is to compare the observed incidence of events from a given cause between groups, the cumulative incidence functions should be used. Estimates of these functions can be obtained via $\hat{F}_j(t|\mathbf{x}) = \sum_{t_i \leq t} \hat{\lambda}_j(t_i|\mathbf{x}) \hat{S}(t|\mathbf{x})$, where $\hat{\lambda}_j(t|\mathbf{x})$ are the estimated hazards resulting from Cox's or Aalen's analyses. The overall survivor function is $\hat{S}(t|\mathbf{x}) = \exp\left\{-\sum_{j=1}^k \sum_{t_i \leq t} \hat{\lambda}_j(t_i|\mathbf{x})\right\}$, and t_i denotes the distinct failure times. Although the effect of the covariates on the cause-specific hazard $\lambda_j(t|\mathbf{x})$ is directly given by β_j , the effect on the cumulative incidence function $F_j(t)$ combines the effect β_j together with the overall effect on $\hat{S}(t|\mathbf{x})$. Hence, no direct estimate for the effect of a covariate in the cumulative incidence function $F_j(t)$ is given. Some work has been done to obtain such estimates.

Fine and Gray (1999) proposed a proportional hazards model to fit the hazard $\gamma_j(t)$ derived from the cumulative incidence or subdistribution function:

$$\begin{aligned} \gamma_j(t|\mathbf{x}) &= \lim_{\Delta t \rightarrow 0} \frac{Pr(T < t + \Delta t, C = j | \mathbf{x}, \{T \geq t \text{ or } (T < t \text{ and } C \neq j)\})}{\Delta t} \\ &= \frac{f_j(t|\mathbf{x})}{1 - F_j(t|\mathbf{x})} \quad j = 1, \dots, k. \end{aligned}$$

The conditional expression reflects two different scenarios: i) the event has not occurred at time t , ii) the event has occurred from a different cause before t . Contrary to the previous analyses, a patient failing from other causes would not be removed from the risk set at his/her time of failure. The subdistribution functions are given then by $F_j(t|\mathbf{x}) = 1 - \exp(-\int_0^t \gamma_j(t|\mathbf{x}))$. A Cox model is assumed for $\gamma_j(t|\mathbf{x}) = \gamma_{0j}(t)e^{\beta_j^\dagger \mathbf{x}}$, $j = 1, \dots, k$, where the covariates are linear on a complementary log-log transformed cumulative incidence function. When censoring is absent or is always observable, Fine and Gray (1999) showed that the partial likelihood approach is valid for estimation. In the case of right-censoring, they developed a weighted score function based on the non-censored case.

3 Application to the Spanish Bladder Cancer Study

We present the results of the three methodologies applied to the bladder cancer data. The 1356 newly-diagnosed bladder cancer cases were recruited between 1997 and 2001 in 18 Spanish hospitals, and followed up to the end of 2005. In this paper, only 994 superficial bladder cancer cases, where tumour was confined to the lining of the bladder, are considered. One important question to consider is why some patients experience a progression as a first event after diagnosis instead of a recurrence. About 84% of failures are due to recurrence, 13% to progression and only 3% to death due to bladder cancer. The median time to develop a recurrence or a progression as first event is similar, 9.3 and 9.7 months, respectively suggesting that there exist distinct courses/aggressiveness of the tumour development. Which prognostic factors characterize and differentiate patients who progress from those experiencing a recurrence? In this case, where the variable of interest is time until the first event, a model for competing risks is needed.

Four factors are considered which may affect the risk of reappearance of the tumour: stage+grade, tumour multiplicity, Spanish region and treatment. The analysis was implemented in the freeware statistical package R. Modelling of each cause-specific hazard involves censoring individuals failing from other causes. Proportional hazards assumption was tested by graphical exploration of Schoenfeld residuals and diagnostic tests based on them. Validation of the additive model was assessed via graphical exploration of the cumulative martingale residuals, as stated in Aalen (1993). The parameter estimates within each model are given in Table 1. In both models, the four factors considered resulted statistically significant for recurrence, whereas only stage+grade and tumour multiplicity were so for progression. Stage+grade for death due to bladder cancer was significant in the proportional hazards model, while its significance was not clear in the additive model. Ta_T1GII tumours are associated to the risk of recurrence, while progression and death are more frequent in TaGIII and T1GIII tumours.

In Cox's model, the hazard ratio $e^{\hat{\beta}_j}$ for covariate z is interpreted as the increase of hazard relative to the reference level of the z . In Aalen's model, $\hat{\beta}_j(t)$ for z is the increase on absolute risk relative to the baseline hazard at instant t . Since $\mathbf{B}_j(t) = \int_0^t \beta_j(u) du$, a graphical exploration of these parameter estimates can be obtained by observing the slopes of $\hat{\mathbf{B}}_j(t)$ versus time, and how they can vary. As an example, Figure 1 shows the cumulative regression functions $\hat{B}_j(t)$ for covariates multiplicity and Ta_T1GII when studying recurrence. The slope for tumour multiplicity remains constant for the first 60 months approximately, indicating that its effect on recurrence is constant over time. The slope for Ta_T1GII tumours varies over time, indicating a non-constant effect.

Fine and Gray's (1999) approach, to model the subdistribution hazard for cause j , does not censor individuals failing from other causes. The same factors as in the above models resulted significant, but now the subdistribution hazard ratio $e^{\hat{\beta}_j}$ for covariate z has a direct interpretation in terms of the cumulative incidence function. The effect of T1GIII tumours over the incidence of recurrence is -0.343 while the effect of this covariate on the rate of recurrence, censoring by other causes, is -0.1764 . Though similar estimates are found in our data, this situation may be dramatically different.

TABLE 1. Parameter estimates for the three models of competing risks.

Recurrence									
Factor	Cox PH for $\lambda_j(t Z)$			Additive for $\lambda_j(t Z)$			Cox PH for $\gamma_j(t Z)$		
	β	SE	p-value	β	SE	p-value	β	SE	p-value
Stage+Grade	Ref.level= TaGI								
Ta_T1GII	0.3987	0.135	0.003	0.165	0.064	0.004	0.408	0.133	0.002
TaGIII	0.1944	0.207	0.350	0.061	0.100	0.354	0.194	0.215	0.370
T1GIII	-0.1764	0.247	0.480	-0.074	0.071	0.522	-0.343	0.247	0.170
Tumour multiplicity	Ref.level= 1 tumour								
>1 tumour	0.5358	0.128	<0.001	0.237	0.067	0.000	0.500	0.129	<0.001
Spanish Region	Ref.level= Barcelona								
Valles	0.5439	0.237	0.022	0.202	0.086	0.026	0.576	0.242	0.017
Alicante	0.1070	0.312	0.730	0.013	0.092	0.976	0.155	0.311	0.620
Tenerife	0.0497	0.260	0.850	0.000	0.079	0.834	0.097	0.263	0.710
Asturias	0.4009	0.221	0.007	0.157	0.071	0.061	0.436	0.228	0.055
Treatment	Ref.level= TUR								
TUR+BCG	-0.5836	0.171	<0.001	-0.248	0.072	0.001	-0.584	0.176	0.001
TUR+Chemo	-0.3125	0.157	0.046	-0.148	0.077	0.044	-0.292	0.154	0.058
TUR+BCG+Chemo	0.4139	0.316	0.190	0.157	0.152	0.332	0.421	0.319	0.190
Other	-1.1193	0.592	0.059	-0.338	0.101	0.001	-1.080	0.570	0.058
Progression									
Stage+Grade	Ref.level= TaGI								
Ta_T1GII	-0.128	0.540	0.810	0.002	0.015	0.768	-0.231	0.537	0.670
TaGIII	1.079	0.524	0.040	0.115	0.073	0.129	2812	0.524	0.048
T1GIII	1.999	0.423	<0.001	0.197	0.055	<0.001	2.001	0.428	<0.001
Tumour multiplicity	Ref.level= 1 tumour								
>1 tumour	0.764	0.321	0.017	0.044	0.028	0.038	0.689	0.324	0.034
Death due to bladder cancer									
Stage+Grade	Ref.level= TaGI								
TaGI	Ref			Ref			Ref		
Ta_T1GII	-0.872	-0.755	0.450	-0.008	0.008	0.417	-0.987	1.153	0.390
TaGIII	0.355	0.308	0.760	0.013	0.026	0.768	0.213	1.150	0.850
T1GIII	1.784	2.442	0.015	0.050	0.029	0.068	1.678	0.724	0.020

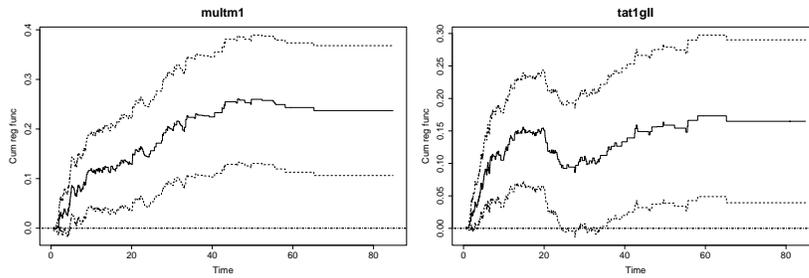


FIGURE 1. Cumulative regression functions for recurrence.

Acknowledgments: This research was partially supported by Grant 050831 from La Marató de TV3 Foundation and by grant MTM2005-0886 from Ministerio de Ciencia y Tecnología. Núria Porta is a recipient of a research fellowship from DURSI.

References

- Aalen, O.O. (1993). Further results on the non-parametric linear regression model in survival analysis *Statistics in Medicine* **12**, 1569-1588.
- Aalen, O.O. and Borgan, Ø. and Fekjær, H. (2001). Covariate Adjustment of event histories estimated from Markov chains: The additive approach *Biometrics* **57**, 993-1001.
- Fine, J.P. and Gray, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk *Journal of the American Statistical Association* **94**, 496-509.
- Lawless, J.D. (2003). *Statistical Models and Methods for Lifetime Data*. New-York: John Wiley & Sons, Inc.
- Prentice, R.L. and Kalbfleisch, J.D. *et al.* (1978). The analysis of failure times in the presence of competing risks *Biometrics* **34**, 541-554.

Robust estimation of linear models with grouped data and arbitrary errors with unknown scale parameter

Carlos Rivero¹ and Teofilo Valdes²

¹ **Communicating author.** Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Económicas y Empresariales, Universidad Complutense de Madrid, Campus de Somosaguas, 28223, Madrid, Spain, crivero@estad.ucm.es

² Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Ciudad Universitaria, 28040, Madrid, Spain, tevaldes@mat.ucm.es

Abstract: In this paper we present an iterative estimation procedure, valid to fit linear models when the dependent data may be grouped in intervals and the distribution of the errors is known but for an unknown scale parameter. The proposed procedure can be seen as an alternative to the EM algorithm, which avoids the awkward computations of the M- and E- step. The convergence of the algorithm and the stochastic asymptotic properties of the corresponding estimators are analyzed.

Keywords: Grouped data; Iterative estimation; Linear models; Unknown errors; Scale parameter.

1 Introduction

When we fit a linear model to a data set in which the dependent variable may be both non-grouped or grouped the Ordinary Least Squares (OLS) estimate can not be used, due to the grouped data. It is known that the direct application of the OLS estimate to the non-grouped data may yield an undesirable bias and inefficiency on the parameter estimate, as a result of the information loss, which increases as the grouping intervals increase in length and as the percentage of grouped data grows.

Let us assume the linear model

$$y_i = \beta^t x_i + v_i, \quad i = 1, 2, \dots, n$$

where β is an unknown m -dimensional parameter that must be estimated, x_i is a m -dimensional vector of regressors and v_i are independent and identically distributed mean-zero perturbations. We will assume that y_i may be observed or grouped (interval-censored).

When the distribution of v_i is known, methods based on Maximum Likelihood, as the EM-type algorithms, may be used to estimate β . Rivero and Valdes (2004) introduced a method, based on OLS, to estimate β , which notably reduced the awkward computations involved in the EM algorithm and which can be seen as an alternative to it.

We propose a procedure to estimate β when v_i depends on an unknown scale parameter. The aim of this paper is to show the good stochastic convergence properties of the proposed algorithm.

It will be organized as follows. Section 2 presents notation, the statistical model, and describes the problem. Section 3 is devoted to the antecedents of the proposed algorithm, when the distribution of the errors is known. Section 4 presents the proposed algorithm and, finally, the performance of the cited algorithm will be analyzed by simulations in section 5.

2 Statistical model

Let us consider a usual regression model

$$y_i = \beta^t x_i + \sigma \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

where β is an unknown m -dimensional parameter which must be estimated, σ is an unknown scale parameter, x_i is a m -dimensional vector of regressors and the errors ε_i are independent and identically distributed strongly unimodal perturbations with mean zero and variance one (not necessarily normally distributed).

It will be assumed that the dependent variable y_i may be grouped with positive probability. Then the index set $I = \{1, 2, \dots, n\}$ may be partitioned into I^g and I^{ng} containing those i 's whose values y_i have been grouped and non-grouped, respectively. The grouping classes of the grouped data are given by the following partition of the real line

$$(-\infty, c_1], (c_1, c_2], \dots, (c_{r-1}, \infty).$$

If $i \in I^{ng}$ the value y_i is observed. Otherwise y_i is missing and only the classification interval which overlaps y_i is known, $y_i \in I_i = (c_{s_i}, c_{s_i+1}]$.

The existence of grouped data makes the OLS estimation and inference on β and σ unapplicable (even assuming that the errors ε_i are normally distributed). Imputation methods, as the EM algorithm or extensions, can be certainly used to estimate β and σ (see McLachlan and Krishnan, 1997 or Watanabe and Yamaguchi, 2003). Our algorithm is a computational procedure which may be conceived as an alternative to the EM algorithm. Our proposal mainly avoids the complex computations of the M-step of the EM algorithm, by which the unknown parameters β and σ are updated (see Tanner, 1996 for details). Furthermore, the E-step of the EM algorithm is notably simplified by our algorithm.

3 Antecedents of the algorithm

Assuming that σ is known, Rivero and Valdes (2004) proposed to estimate the true vector parameter β in model (1) by means of an iterative algorithm, based on single imputations of the grouped values and OLS estimations. The estimation of the parameter β was obtained iterating the following algorithm:

INITIALIZATION

a.1 Fix an arbitrary starting vector β^0 .

ITERATIONS

a.2 Assuming the current point β^p to be known, the next point is defined by:

$$\beta^{p+1} = (X'X)^{-1} X'Y(\beta^p, \sigma).$$

where $X' = (x_1, \dots, x_n)$, $Y(\beta^p, \sigma) = (y_1(\beta^p, \sigma), \dots, y_n(\beta^p, \sigma))'$ and, for all $i \in 1, \dots, n$,

$$y_i(\beta^p, \sigma) = \begin{cases} y_i, & \text{if } i \in I^{ng} \\ x_i^t \beta^p + E(\sigma \varepsilon_i | x_i^t \beta^p + \sigma \varepsilon_i \in I_i), & \text{if } i \in I^g, y_i \in I_i. \end{cases}$$

a.3 Change β^p to β^{p+1} and return to step a.2. Iterate until convergence.

Rivero and Valdes (2004) proved that for any starting vector β^0 , the sequence β^p generated by the algorithm converges, as $p \rightarrow \infty$, to a vector $\hat{\beta}$, which does not depend on the starting vector. Consequently, the point $\hat{\beta}$ is taken as the estimate of β for the fixed sample. Rivero and Valdes (2004) states that $\hat{\beta}$ is a consistent and asymptotically normal estimation of β , and proposed a consistent estimation of the covariance matrix of the limit normal distribution.

4 Proposed algorithm

In this section, we propose an algorithm to estimate β in model (1) when σ is unknown. The estimation of the parameters β and σ is obtained iterating the following algorithm:

INITIALIZATION

b.1 Fix arbitrary values β^0 and σ^0 .

ITERATIONS

Assuming β^r and σ^r are known:

b.2 Update β^r by running steps a.1, a.2 and a.3 in the former section, assuming that σ^r is the known variance, and taking β^{r+1} as the limit point of the iterative procedure.

b.3 Update σ^r by

$$\sigma^{r+1} = \frac{1}{n} \left[\sum_{i=1}^n (\hat{y}_i - x_i^t \beta^{r+1})^2 + \sum_{i \in I^g} \hat{v}_i \right],$$

where $\hat{y}_i = y_i(\beta^{r+1}, \sigma^r)$ was defined in the former section and \hat{v}_i is the within variance of the grouped observation

$$\hat{v}_i = V(\sigma^r \varepsilon_i | x_i^t \beta^{r+1} + \sigma^r \varepsilon_i \in I_i), \quad \text{for every } i \in I^g.$$

b.4 Change σ^r by σ^{r+1} and β^r by β^{r+1} , and return to step b.2. Iterate until convergence.

5 Simulations and empirical convergence properties

We have simulated several linear models (where the error distributions, the percentage of grouped data, the magnitude of σ and the sample size vary) and we have detected the following to be the main properties:

- (1) For a fixed sample, the proposed algorithm is asymptotically globally stable. This means that for any initial points β^0 and σ^0 , the algorithm converges (as $r \rightarrow \infty$) to limit points, $\hat{\beta}$ and $\hat{\sigma}$, which do not depend on the initial points. Note that, a similar property was proved in Rivero and Valdes (2004) for the convergence of β and σ known.
- (2) The limit points $\hat{\beta}$ and $\hat{\sigma}$ are consistent estimations (as $n \rightarrow \infty$) of the unknown parameters β and σ .
- (3) The estimation $\hat{\beta}$ is asymptotically normal. Although this property is beyond our scope, given the limited length of this work, its plausibility is quite certain on seeing the results of Rivero and Valdes (2004) and the consistency of $\hat{\sigma}$ as an estimator of σ .
- (4) The covariance matrix of the limit normal distribution can be consistently estimated, thus inference on β can be carried out.
- (5) Finally, we highlight the simplicity of our algorithm, regardless of the error distribution. The simulations show that the computer time required by our algorithm reduces notably those of the EM-algorithm.

References

- Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.
- McLachlan, G. J. and Krishnan T. (1997). *The EM Algorithm and Extensions*. Wiley.
- Rivero, C. and Valdes, T. (2004). Mean-based iterative procedures in linear models with general errors and grouped data. *Scandinavian Journal of Statistics* **31**, 469-486.
- Tanner, M. A. (1996). *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer.
- Watanabe, M and Yamaguchi, K. (2003). *The EM Algorithm and Related Statistical Models*. Marcel Dekker.

Joint modelling of time-to-event and longitudinal binary data with excess zeros

Dimitris Rizopoulos¹, Geert Verbeke¹ and Emmanuel Lesaffre¹

¹ Biostatistical Centre, Catholic University of Leuven, Belgium

Abstract: Many longitudinal studies generate both the time to some event of interest and repeated measures data. This paper is motivated by a study on patients with a renal allograft, in which interest lies in the association between longitudinal proteinuria (a dichotomous variable) measurements and the time to renal graft failure. An interesting feature of the sample at hand is that nearly half of the patients were never tested positive for proteinuria (≥ 1 gr/day) during follow-up, which introduces a degenerate part in the random-effects density for the longitudinal process. In this paper we propose a two-part shared parameter model framework that effectively takes this feature into account, and we investigate sensitivity to the various dependence structures (induced by a copula function) used to describe the association between the longitudinal measurements of proteinuria and the time to renal graft failure.

Keywords: Joint Modelling; Random-Effects; Copulas; Sensitivity Analysis.

1 Introduction

Our research has been motivated by a study on 432 patients that underwent a primary renal transplantation with a graft from a cadaver or living donor, and for whom the new graft has survived for at least one year. The clinical interest lies in the long term performance of the new graft, and especially in the graft survival for a ten year period. Out of the 432 patients considered, 91 (21.1%) experienced a graft failure. The corresponding Kaplan-Meier estimate for the time to graft failure is depicted in the top-left panel of Figure 1. During the ten year follow-up period, the patients were periodically tested for the performance of the graft. One of the outcomes recorded is the presence of proteinuria. Proteinuria is the condition in which the urine contains an abnormal amount of protein (≥ 1 gr in a 24 hours urine collection), which is an indication of renal graft malfunctioning. An interesting characteristic of the sample at hand is that for the 210 patients, proteinuria of more than 1 gr/day has never been observed. Moreover, for the remaining patients with at least one positive diagnosis of proteinuria during follow-up, the sample smooth average profiles, presented in the top-right panel of Figure 1, show a steep increase for graft failure. This feature suggests that exploration of the longitudinal evolution of the probability of proteinuria could be insightful for the time to graft failure. Thus, our aim here is to investigate the association structure between these two processes.

The setting described above connects to the framework of joint modelling of longitudinal and time to event data (see Tsiatis and Davidian, 2004 for a review). Joint models are constructed under the conditional independence assumption and posit that the event process and the longitudinal responses are independent conditionally on a latent process expressed by a set of, typically normally distributed, random-effects. However, note that a normal or another smooth random-effects density might be unrealistic for

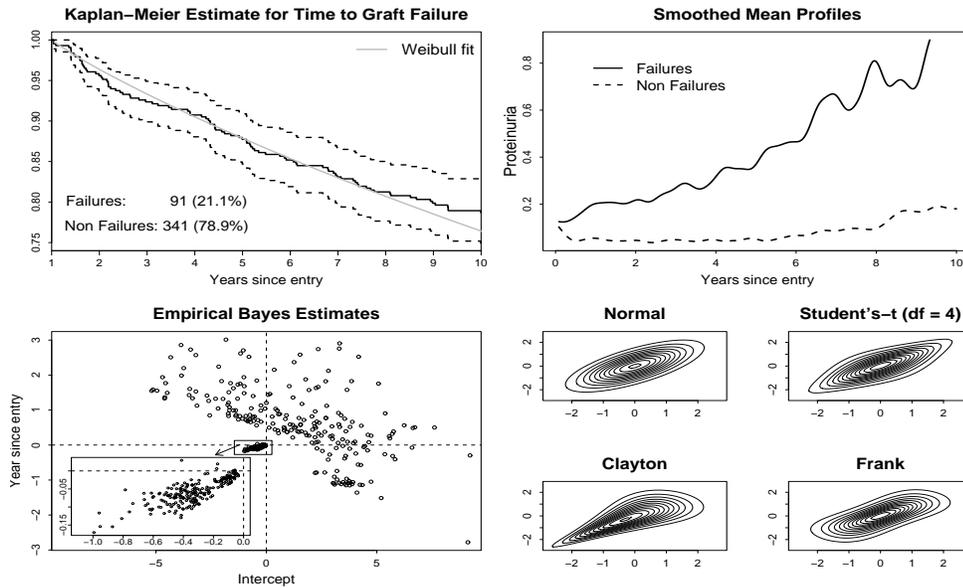


FIGURE 1. Top left panel: Kaplan-Meier estimate (with associated 95% CI) for time to graft failure, with superimposed Weibull fit. Top right panel: sample smooth average profiles (obtained using a Nadaraya-Watson kernel regression estimate) for the probability of proteinuria versus years since entry, for patients with at least one finding of proteinuria during follow-up. Bottom left panel: empirical Bayes estimates under an ignorable random slopes logistic regression for proteinuria, including all patients. The rectangle around zero contains the patients with no proteinuria history and it is magnified in the third quadrant. Bottom right panel: contour plots of the Normal, Student's- t ($df = 4$), Clayton, and Frank copula for standard normal marginals and association between the two marginals equal to 0.5 in terms of Kendall's τ .

our data, since nearly half of the patients never showed proteinuria during follow-up. This feature, induces a bimodality in the random-effects density, which is also evident in the plot of the Empirical Bayes estimates, obtained by the ignorable (i.e., ignoring the survival process) mixed-effects logistic regression, presented in the bottom-left panel of Figure 1. To overcome this problem, we propose a two-part extension of the common joint models, which assumes that the distribution of the longitudinal process is a two-component mixture with a degenerate component for patients with no proteinuria history and a mixed-effects logistic regression component for the remaining patients. This formulation allows to investigate separately the effect of first, the longitudinal evolution of proteinuria and second, the history of proteinuria, to the time to graft failure.

2 Two-Part Joint Model

Joint models typically consist of three submodels, namely the longitudinal, the survival, and the random-effects models. In our formulation however, we introduce a degenerate part that accounts for the patients with no proteinuria history. In particular, let T_i be the observed failure time for the i th patient ($i = 1, \dots, n$), which is the minimum of the true failure time T_i^* and the censoring time K_i . Set δ_i be the censoring indicator, i.e., $\delta_i = I(T_i^* \leq K_i)$, where $I(\cdot)$ is the indicator function. Let y_i denote the $n_i \times 1$ vector of binary indicators for proteinuria, and let $d_i = I(\{y_{ij} = 1; \text{ for some } j \in (1, \dots, n_i)\})$ be the indicator of proteinuria history (i.e., ever having proteinuria). The two-part shared parameter model, omitting covariates in the notation, is defined as

$$\begin{aligned} p(y_i, T_i; \theta) &= \sum_{d_i} p(d_i; \theta) p(y_i, T_i | d_i; \theta) \\ &= \sum_{d_i} p(d_i; \theta_d) \int \int \check{p}(T_i | b_{ti}, d_i; \theta_t) p(y_i | b_{yi}, d_i; \theta_y) p(b_{yi}, b_{ti} | d_i; \theta_b) db_{yi} db_{ti}, \end{aligned}$$

where $\theta^\top = (\theta_d^\top, \theta_t^\top, \theta_y^\top, \theta_b^\top)$ is the vector of the parameters in each one of the submodels with A^\top denoting the transpose of A . Further, let $p(\cdot)$ denote the appropriate probability density functions (pdf) for the longitudinal and random-effects parts, whereas for the event process we set $\check{p}(T_i | b_{ti}, d_i; \theta_t) = p(T_i | b_{ti}, d_i; \theta_t)^{\delta_i} S(T_i | b_{ti}, d_i; \theta_t)^{1-\delta_i}$, i.e., equal to either the density for the event times or the survival function for censored observations. Finally, b_{ti} and b_{yi} denote subject-specific random-effects for the two processes.

For the survival process we assume a mixed-effects accelerated failure time model defined as

$$\log T_i = w_i^\top \gamma + d_i \gamma_d + b_{ti} + \sigma_t \varepsilon_i, \quad \varepsilon_i \sim \mathcal{P},$$

where $\theta_t^\top = (\gamma^\top, \gamma_d, \sigma_t)$, and w_i is a vector of baseline covariates. The errors ε_i are assumed to follow the distribution function \mathcal{P} , with corresponding survival function S and density function p , and σ_t denotes a scale parameter (Kalbfleisch and Prentice, 2002, chpt. 3). In this work we consider parametric models for \mathcal{P} ; non-parametric alternatives in the joint modelling framework have been proposed by Tseng et al. (2005). Parameter γ_d measures the effect of proteinuria history in the logarithm of time to graft failure, which is expected to be highly significant. The random-effect b_{ti} represents a frailty term that captures unobserved heterogeneity induced, e.g., by omitted covariates (Keiding et al., 1997). The model for the longitudinal process conditionally on d_i contains a degenerate part in order to account for the fact that $y_{ij} = 0, \forall j$ when $d_i = 0$. For the patients with proteinuria history, we model the longitudinal evolution of proteinuria findings using a mixed-effects logistic regression. In particular, we assume that

$$\begin{cases} Pr(y_{ij} = 0, \forall j) = 1, & \text{if } d_i = 0 \\ Pr(y_{ij} = 1 | b_{yi}) = \exp(x_{ij}^\top \beta + z_{ij}^\top b_{yi}) / \{1 + \exp(x_{ij}^\top \beta + z_{ij}^\top b_{yi})\}, & \text{if } d_i = 1, \end{cases}$$

where $\theta_y = \beta$ is the vector of regression coefficients, y_{ij} equals one if the i th patient had a proteinuria finding at the j th time, and zero otherwise, b_{yi} are subject-specific

random-effects dictating patient's longitudinal trajectories, and X_i and Z_i are design matrices for the fixed- and random-effects, respectively. Finally, for the random-effects the common parameterization used in joint models postulates that $b_{ti} = \alpha b_{yi}$, where α denotes an association parameter. That is, the longitudinal and survival processes share the same random-effect b_{yi} , with α^2 being a rescaling factor for the variance of b_{yi} . However, this parameterization assumes perfect correlation between the underlying random-effects, which may be unrealistic in many applications. Thus, for the random-effects model we propose here a copula (Nelsen, 1999) representation for the joint distribution of b_{yi} and b_{ti} , which has the following form

$$p(b_{yi}, b_{ti} \mid d_i; \theta_b) = \begin{cases} p(b_{ti}; \omega_t), & \text{if } d_i = 0 \\ c(H_y(b_{yi}; \omega_y), H_t(b_{ti}; \omega_t); \alpha) p(b_{yi}; \omega_y) p(b_{ti}; \omega_t), & \text{if } d_i = 1, \end{cases}$$

where $c(\cdot)$ is the density of the copula $C(\cdot)$, $H_y(\cdot)$ and $p(b_{yi})$ are the marginal cumulative distribution function and the pdf for b_{yi} , respectively, and $H_t(\cdot)$ and $p(b_{ti})$ are defined analogously for b_{ti} . The parameter vector for the random-effects density is $\theta_b^\top = (\alpha, \omega_y^\top, \omega_t^\top)$, where α is the parameter of the copula, and ω_y and ω_t are the parameter vectors for the two marginals. The advantage of the copula parameterization is that it allows for separate modelling of the association structure and the marginals, thus facilitating exploration of dependence. This is illustrated in the bottom-right panel of Figure 1, which depicts the contours of four copulas assuming standard normal marginals. In particular, we observe that the copula function can significantly alter the shape of the association, even though all the other components of the bivariate densities remain the same.

3 Results

The maximum likelihood estimates for the model parameters θ are obtained using an EM algorithm, in which b_{yi} and b_{ti} are treated as missing data. For the E-step, the expectations of the form $E\{A(b_{yi}, b_{ti}) \mid y_i, T_i; \theta\}$, i.e., the expected value of any function $A(\cdot)$ of b_{yi} and b_{ti} with respect to the posterior distribution $p(b_{yi}, b_{ti} \mid y_i, T_i, d_i; \theta)$, are approximated using a Gauss-Hermite quadrature rule. For the M-step, unfortunately the complete data log-likelihood for the two-part shared parameter model does not have closed form solutions with respect to θ . Thus, the expected value of the complete data log-likelihood is numerically maximized using a quasi-Newton algorithm.

The specification of the components of the two-part shared parameter model, presented above, is as follows. First, for the history of proteinuria a logistic regression is used. Second, for the survival process a Weibull model is assumed, which seems to provide a relatively appropriate fit, according to the top-left panel of Figure 1. Third, for the longitudinal processes and based on the ignorable analysis (i.e., ignoring the event process), a mixed-effects logistic regression is adopted, with random-intercepts and -slopes. The covariate effects that are considered in all the above submodels are gender, weight, tobacco habits (no-smoker, smoker, ex-smoker), age (older than 55), and long dialysis (if dialysis before transplant). Finally, for the random-effects model and in order to investigate the influence of parametric assumptions on the size of the association between the two processes, we performed a sensitivity analysis under the Gaussian, Student's- t (df = 4), Frank, and Clayton copula functions assuming normal

marginals. All computations have been performed in R (R Development Core Team, 2006).

The results showed that the choice of the copula function has a direct impact on certain parameter estimates. For instance, the association between the survival and longitudinal processes varies from -0.179 (std. error: 0.035) for the Clayton copula to -0.535 (std. error: 0.074) for the Frank copula. Similar effects regarding the choice of

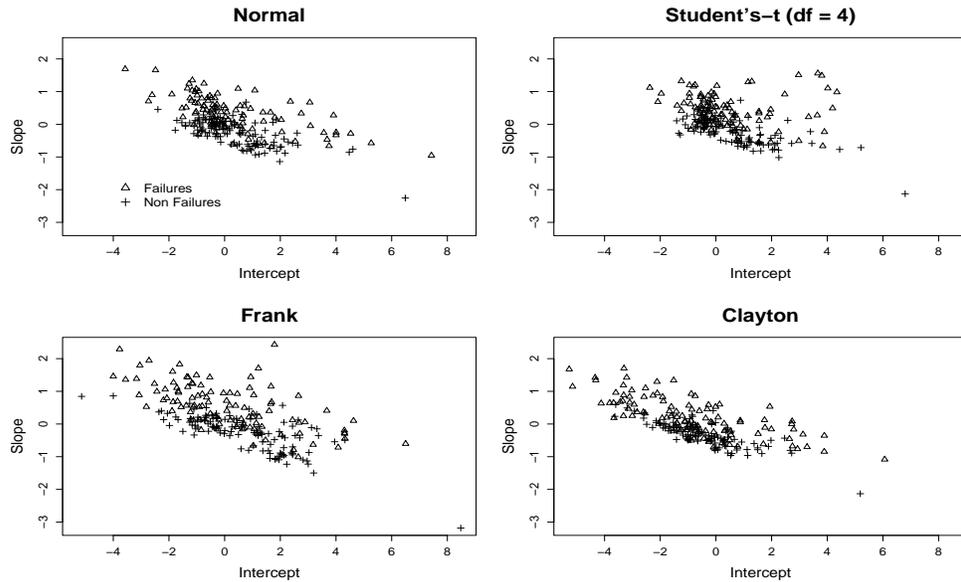


FIGURE 2. Empirical Bayes estimates for the random-effects in the longitudinal processes under the Gaussian, Student's- t ($df = 4$), Frank, and Clayton copulas, for the patients with proteinuria history.

the copula function were also apparent in the plots of the Empirical Bayes estimates for the random-effects of the longitudinal process presented in Figure 2, and the marginal survival function for the event process (figure not shown). In conclusion, the variability we observe in the overall results under the different copulas could be regarded as variability due to modelling assumptions, which is a clear indication that the common normality assumption for the distribution of random-effects may prove difficult to verify.

References

- Kalbfleisch, J., and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2nd Ed.* New York: Wiley.
- Keiding, N., Andersen, P., and Klein, J. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* **16**, 215-224.

Nelsen, R. (1999). *An Introduction to Copulas*. New York: Springer-Verlag.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**, 587-603.

Tsiatis, A., and Davidian, M. (2004). An overview of joint modelling of longitudinal and time-to-event data. *Statistica Sinica* **14**, 793-818.

Principal Component Analysis of Electoral Data

Paulo Canas Rodrigues¹ and João A. Branco²

¹ Department of Mathematics and CMA, New University of Lisbon (FCT), Quinta da Torre 2829-516, Caparica, Portugal. paulocanas@fct.unl.pt

² Department of Mathematics and CEMAT, Technical University of Lisbon (IST), Lisboa, Portugal. joao.branco@math.ist.utl.pt

Abstract: A set of electoral data is studied using Principal Component Analysis (PCA). From the results of three approaches that were attempted we concluded that Logcontrast PCA is a good choice for this type of Compositional Data (CD). It gave a useful insight to the voting structure.

Keywords: Compositional data; Principal component analysis; Logcontrast PCA; Crude PCA.

1 Introduction

The motivation for doing this work grew up during the analysis of a set of electoral data. Multiparty electoral data is essentially compositional (Katz, 1999), that is, each vote proportion x_{ij} , of a party (“political group”) i in a district (country) j , $i = 1, \dots, D$; $j = 1, \dots, n$ is such that $0 \leq x_{ij} \leq 1$ and the sum of all the vote proportions in that district (country) is 1, $\sum_i x_{ij} = 1$. These restrictions give rise to a particular structure, the CD.

CD are not rare, they appear not only in political science but also in geo-chemistry, genetics, ecology, food science and many other areas. The conventional statistical methods require that variables are independent, a condition that is not satisfied by CD since each composition is formed by mutually dependent components. Moreover, most basic concepts such as covariance and correlation of raw proportions do not have simple interpretation as they do when applied to data that is not compositional. This has strong implications to most multivariate procedures that are based on the covariance/correlation matrix, in particular to PCA which is the working method of this paper.

In the following section a very brief summary of the PCA solutions provided in Aitchison (1983, 1986) is included. The Section 3 contains the analysis of a real electoral data set. The final discussion is presented in the last section.

2 Crude PCA and Logcontrast PCA

Direct PCA of CD faces the general difficulties mentioned above together with the fact that CD often shows considerable curvature, which is not compatible with the linearity hypothesis assumed by PCA, and leads to unsatisfactory results.

Aitchison (1983, 1986) presents two methods for PCA of CD: (i) the Crude PCA, which is a PCA of the CD and (ii) the Logcontrast PCA. The Logcontrast PCA is based on the loglinear transformation

$$y_{ij} = \log \left[\frac{x_{ij}}{g(\mathbf{x}_i)} \right], \quad \text{where } g(\mathbf{x}_i) = \sqrt[x_{i1} \times \dots \times x_{iD}]{x_{i1} \times \dots \times x_{iD}}. \quad (1)$$

The role of the transformation is to remove the constraints attached to the CD. The constrained sample space associated with the CD gives way to a real multivariate unconstrained space where multivariate methods can be applied in the usual manner. The logcontrast principal components have the form $\mathbf{a}' \log(\mathbf{x})$, where \mathbf{a} is the solution of the standard equation

$$(\Gamma - \lambda \mathbf{I})\mathbf{a} = \mathbf{0}, \quad (2)$$

where Γ is the logratio covariance matrix and λ is the eigenvalue associated with the eigenvector \mathbf{a} .

3 Application

The data are the results (number of votes) of the 2004 European Parliament election in the 25 member states collected for 8 “political groups”: European People’s Party - European Democrats (EPP-ED), Group of the Party of European Socialists (PES), Alliance of Liberals and Democrats for Europe (ALDE), Union for Europe of the Nations (UEN), European Greens - European Free Alliance (Greens-EFA), European United Left - Nordic Green Left (GUE-NGL), Independence and Democracy (IND/DEM) and Others.

To perform Logcontrast PCA the package “compositions” of the software R was used. The results of the various analyses and their interpretations are now presented.

3.1 PCA

The biplot on the left side of Figure 1 relates to the PCA of the original raw data, the number of votes. Although the first two principal components explain about 81% of variance of the original data, Figure 1 does not help much to the understanding of the structure of the data because the loadings have almost the same direction and the countries overlap.

3.2 Crude PCA

The right side of Figure 1 presents the biplot for the first two crude principal components.

Although the proportions of explained variance by the first crude principal components are smaller than the proportions of explained variances by PCA (about 48.4% for the first two and 64.6% for the first three principal components), the interpretation of the right plot of Figure 1 is more illuminating. Actually we can see that the right wing

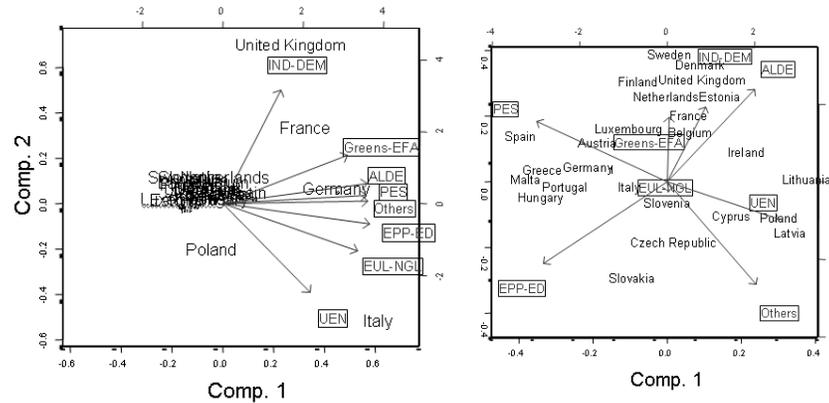


FIGURE 1. Biplots for the first two Principal Components (on the left) and the first two Crude Principal Components (on the right).

parties are on the right side, the center parties on the left and the left wing parties on the center.

In the right bottom of Figure 1 we have, mainly, the ex-URSS countries influenced by the loadings of the parties UEN and Others. On the left we have the countries that distributed his votes by the parties EPP-ED and PES. If we analyze the studied CD we can conclude that the group of countries that are in the north of Europe and in the top of Figure 1 had proportions of votes distributed by the parties EPP-ED, PES, ALDE and some influence of the left wing parties.

Although we think that this is a good interpretation of the partisan preferences of the different countries, we had a big influence of the party Others. We now present the solution of Logcontrast PCA where the importance of this variable decreases.

3.3 Logcontrast PCA

Figure 2 shows the biplot for the first two Logcontrast principal components. The first two principal components explain 59.4% of the total variance and the first three 76.8%. Watching Figure 2 we can see, on the right side, two right wing parties (ALDE and UEN) surrounded by ex-URSS countries, mainly. On the left side are the more left wing parties (Greens-EFA and EUL-NGL) together with countries that are in the north of Europe or belong to the more developed countries category. In the center of Figure 2 lies the socialist and democratic parties (PES and EPP-ED) and Others. In the bottom of the Figure 2 is the IND-DEM that incorporates EU-critics, eurosceptics and eurorealists, people who do not follow any traditional partisan trend.

4 Discussion

The nature of Electoral data is essentially compositional since the proportion of the election voting is of primary importance, rather than the number of votes.

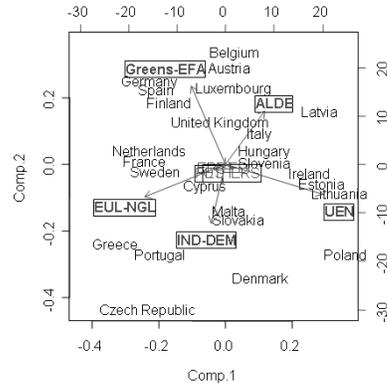


FIGURE 2. Biplot for the first two Logcontrast principal components.

Compositional data analysis techniques were used to study the 2004 European Parliament election results. Logcontrast PCA gives a clearer and more useful interpretation of the data as compared with the Crude PCA.

Not surprisingly the results of the conventional PCA of the raw data are very poor. In fact PCA is not adequate for CD.

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.
- Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* **70**, 57-61.
- Katz, J.N. and King, G. (1999). A Statistical Model for Multiparty Electoral Data. *American Political Science Review* **93**, 15-32.

Comparing different approaches to regression analysis of Receiver Operating Characteristic curves. An application to Endocrinology data

M. X. Rodríguez-Álvarez¹, I. López-de-Ullibarri² and C. Cadarso-Suárez³

¹ Dept. of Statistics and OR, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain, cotepirilampo@gmail.com

² Dept. of Mathematics, University of A Coruña

³ Dept. of Statistics and OR, University of Santiago de Compostela

Abstract: Continuous biomarkers are often used to discriminate between diseased and healthy populations. The Receiver Operating Characteristic (ROC) curve is a widely used tool for characterizing the marker accuracy. To account for covariates that might influence the test accuracy, various ROC regression methodologies have been proposed in the statistical literature, namely the induced and the direct methodologies. In this work, we compare the performance of the two methodologies, by applying the different methods to simulated data. We also apply the proposed methods to a real data set from the endocrinology field.

Keywords: Receiver operating characteristic curve; Regression Model; Accuracy measures

1 Introduction

The ROC curve is a fundamental technique in the characterization of the accuracy of continuous diagnostic tests. The performance of diagnostic tests is potentially influenced by the effect of covariates. Thus, the ROC curve (or summary indices like the area under the curve, AUC) may be of little value if important covariates are neglected. Regression modelling of either the test results or the ROC curve itself is becoming the usual method for the assessment of covariate effects.

The *induced ROC methodology* (Pepe, 1998; Faraggi, 2003) assumes that the test result Y can be expressed as a regression model on covariates X :

$$Y = \mu(D, X) + \sigma(D)\varepsilon, \quad (1)$$

where D is an indicator variable denoting the true disease status ($D = 0$, healthy; $D = 1$ diseased), $\mu(D, X)$ is the mean function, depending on D and X , $\sigma^2(D)$ is the variance, depending on D , and ε is a random variable with zero mean, unit variance and survival function S . From expression (1) the covariate-specific ROC curve can be obtained:

$$ROC_X(t) = S\left(\frac{\mu(0, X) - \mu(1, X)}{\sigma(1)} + \frac{\sigma(0)}{\sigma(1)}S^{-1}(t)\right).$$

Here, we are concerned with the method of Faraggi (2003), which assumes that Y (possibly after a transformation) is normally distributed. We refer to this method as the ‘Normal Method’.

On the other hand, the *direct ROC methodology* assumes the following regression model for the ROC curve

$$ROC_X(t) = g(h(t) + \eta(X)), \quad (2)$$

where $t \in (0, 1)$, g denotes a known link function, η is a function of the covariates and h is a monotonic function on $(0, 1)$. Different proposals for h have been suggested: in Alonzo and Pepe (2002) a parametric form is specified (‘Parametric ROC-GLM Method’); in Cai (2004) it remains unspecified (‘Semiparametric ROC-GLM Method’).

2 Simulation studies

We have conducted a simulation study to compare the performance of the induced and direct ROC methodologies. Different simulation schemes with respect to (i) the regression model specified for either the test result or the ROC curve itself, and (ii) sample size have been considered.

The simulated data, including both a continuous covariate X and a categorical covariate Z , were drawn from the following linear models

$$Y_{\bar{D}} = XZ + \varepsilon_{\bar{D}},$$

$$Y_D = .5 + Z + 4X - 4XZ + 2(X - .6)^2(X - 3)Z + \varepsilon_D,$$

where $X \sim U(0, 1)$, $Z \sim \text{Bern}(.5)$, $\varepsilon_{\bar{D}} \sim N(0, 1)$ and $\varepsilon_D \sim N(0, 1.5)$. The corresponding induced ROC curve is

$$ROC_{(X,Z)}(t) = \Phi \left(\frac{.5 + Z + 4X - 5XZ + 2(X - .6)^2(X - 3)Z + \Phi^{-1}(t)}{1.5} \right),$$

where $\Phi(t)$ is the standard normal distribution function.

We simulated 500 datasets with sample sizes $n_D = n_{\bar{D}} = 50$ and 200. As to the regression models in (1) and (2), three schemes (schemes I, II and III) were considered. In scheme I, $\eta(X, Z) = \mu(D, X, Z) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 XZ$. In order to allow for greater flexibility, in scheme II, $\mu(D, X, Z)$ and $\eta(X, Z)$ were modelled parametrically by means of natural cubic regression splines with 1 knot, at uniform quantiles of X . In scheme III, the same procedure was used, but with 4 knots. The interaction between continuous and categorical covariates was included in all cases. For direct methodology, we took $g(t) = \Phi(t)$, and $h(t) = \alpha_0 \Phi^{-1}(t)$ for the Parametric ROC-GLM Method.

For the covariate-specific ROC curve, Cramer-Von-Mises criterion was used to assess the goodness-of-fit of the estimates. In the case of AUC, we computed the mean squared error. We found that the three methods had a similar behaviour. Scheme II performed best with the first level of the categorical covariate and scheme I with the second. This is likely due to the fact that, in both schemes, the proposed regression models for (1) and (2) are closer to the true relationship between test response and covariates. Our study suggests that, in general, the Semiparametric ROC-GLM Method provides estimates with smaller variance, for both ROC curve and AUC.

3 Application to Endocrinology data

The ROC methodologies presented here have been applied to an endocrinological study aimed at the assessment of the accuracy of the Body Mass Index (BMI), adjusted by age and gender, for detecting patients with high risk of cardiovascular disease. BMI was calculated as weight (in kilograms) divided by height (in meters) squared. The study was carried out with a random sample of Galician adult population (2945 subjects, 46.2% men; age range 18-85 years). Subjects having two or more cardiovascular disease risk factors (raised triglycerides, blood pressure and plasma glucose, and reduced HDL-cholesterol) were considered as diseased.

As a first step, an exploratory analysis of the data, using non-parametric regression techniques, was done. Due to the non-linear relationship of BMI and age, and to allow enough flexibility for models (1) and (2), four knots were selected for the natural cubic splines basis. The age and gender-specific ROC and AUC estimates are shown in figure 1 and figure 2, respectively.

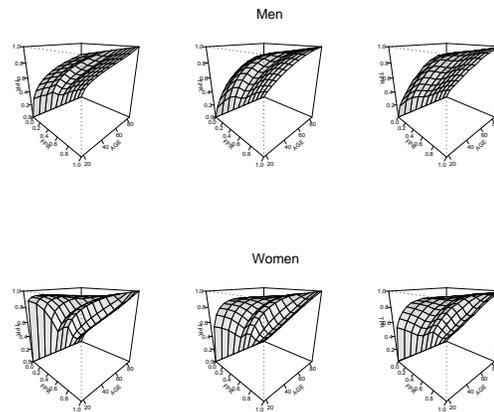


FIGURE 1. ROC estimates for BMI, by age and gender. From left to right: Normal Method, Parametric ROC-GLM Method and Semiparametric ROC-GLM Method.

From Figure 2, it is evident the good accuracy of the BMI for youngest women, with values greater than 0.7. It decreases progressively, until losing significance for the oldest women. For men, the differences between young and old people are not as important as in women. In this case, the BMI has a similar behaviour along age for detecting patients with high risk of cardiovascular disease.

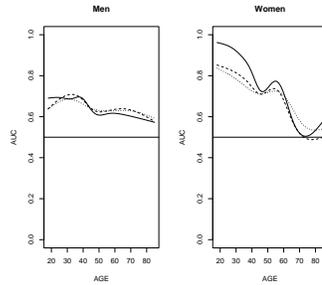


FIGURE 2. AUC estimates for BMI, by age and gender. Continuous line: Normal Method. Dashed line: Parametric ROC-GLM Method. Dotted line: Semiparametric ROC-GLM Method.

References

- Alonzo, T.A. and Pepe, M.S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**, 421-432.
- Cai, T. (2004). Semi-parametric ROC regression analysis with placement values. *Biostatistics* **5**, 45-60.
- Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* **52**, 179-192.
- Pepe, M.S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124-135.

A Local Maximum Likelihood Estimator for Logistic Regression

José António Santos¹ and M. Manuela Neves²

¹ ISEGI - New University of Lisbon; Campus de Campolide; 1070-312 Lisbon; Portugal, jsantos@isegi.unl.pt

² Department of Mathematics; ISA - Technical University of Lisbon; Tapada da Ajuda; 1349-017 Lisbon; Portugal, manela@isa.utl.pt

Abstract: A local maximum likelihood estimator based on logistic regression is presented. This semiparametric estimator is intended to be an alternative to the parametric logistic regression estimator that does not depend on regularity conditions and model specification accuracy. The use of the local likelihood procedure is illustrated on one example from the literature. This procedure is found to perform well.

Keywords: binary data; HIV data; local maximum likelihood; logistic regression; semiparametric regression

1 Introduction

The logistic regression model is the basic framework for binary data regression analysis. Its statistical properties like asymptotic efficiency, consistency and normality depend on regularity conditions and on the accuracy of the model specification. This is a shortcoming of the logistic regression and more broadly of parametric regression models.

In this work we present a semiparametric estimator for dichotomous data regression analysis, namely a local maximum likelihood estimator based on logistic regression. Its bias, variance and asymptotic distribution are not presented here but the authors would provide them for those interested in. This model has the advantage that its statistical properties do not depend on regularity conditions and model specification accuracy.

2 A Local Maximum Likelihood Model For The Logistic Regression

Tibshirani and Hastie (1987) suggested the local likelihood concept. Staniswalis (1989) and Fan, Heckman and Wand (1995) extended this concept to the kernel smoothing and local polynomial kernel regression framework.

Instead of considering the global, parametric specification of the logistic regression model one fits this model *locally* estimating a polynomial in a neighborhood of each x value of interest within the support of the covariate, and the model parameters are

estimated through weighted maximum local likelihood, whose weights are given by a particular kernel function and bandwidth. Note that if the bandwidth is suitably large the local likelihood estimator would be close to the parametric model estimator. This would be a good option if the specified parametric model is almost the true model. Otherwise, if the bandwidth is properly small the local likelihood estimator would not hang to the parametric model estimator. This would be a good option if the specified parametric model is far from the true model. See Eguchi, Kim and Park (2003) for a discussion on this topic.

For simplification purpose consider only one continuous covariate, X . The local neighborhood is selected by the bandwidth h and the kernel function K . Consider Y as a binary random variable with support $\{0, 1\}$. In the context of the logistic regression, the conditional mean of Y is:

$$E[Y|X = x_i] = p_i = \frac{\exp \{m(x_i)\}}{1 + \exp \{m(x_i)\}}, \tag{1}$$

where $m(x_i)$ is an unknown function of interest to be estimated through local polynomial smoothing.

Considering a Taylor development of degree one as an approximation to $m(x_i)$, where x_i is in a neighborhood of x , we have:

$$\lambda(x_i) \approx \exp [\beta_0 + \beta_1(x_i - x)]. \tag{2}$$

In the logistic regression framework the logarithm of the local likelihood function is:

$$\mathcal{L}_1(\beta_0, \beta_1|\mathbf{x}, \mathbf{y}, x, h) = \tag{3}$$

$$h^{-1} \sum_{i=1}^n \{(y_i \log p_i + (1 - y_i) \log(1 - p_i))K\{(x - x_i)/h\}\},$$

where the subscript 1 in \mathcal{L}_1 shows that we have chosen the degree one of the polynomial used for the local smoothing.

It is easy to see that the local maximum likelihood logistic regression estimator is obtained from solving the first order conditions for the log-likelihood function maximum:

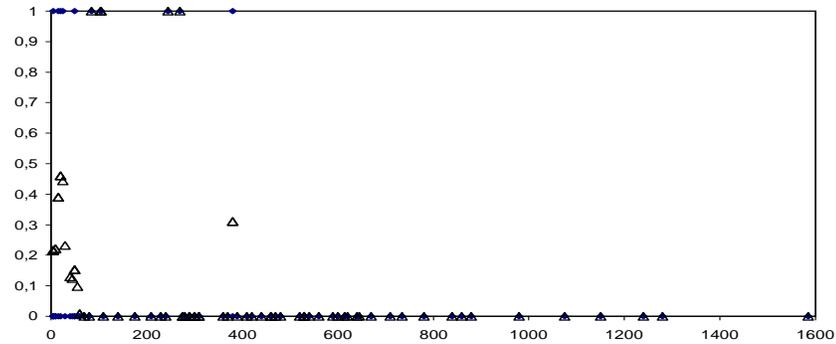
$$\sum_{i=1}^n \left\{ \left(y_i - \frac{\exp [\beta_0 + \beta_1(x_i - x)]}{1 + \exp [\beta_0 + \beta_1(x_i - x)]} \right) K\{(x - x_i)/h\} \right\} \begin{bmatrix} 1 \\ x_i - x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

There is the problem of bandwidth selection. It could be done by sample splitting or by cross-validation (Gyorfi, Kohler, Krzyzak and Walk, 2002), which means the use of part of a sample to achieve information about the other.

The bias, variance and asymptotic distribution of this estimator were obtained (see Santos, 2005). Estimation and inference procedures can be seen in Fan, Farnen and Gijbels (1998).

3 A Real Data Example

The logistic local maximum likelihood estimator was applied to model one data set from the literature. This procedure was found to perform well.

FIGURE 1. HIV data training sample (black lozenge) and curve estimate (Δ).

The data set concerned HIV data from van den Broek (1995). The binary variable relates to the occurrence of urinary tract infection concerning 98 (HIV)-infected men and the explanatory variable is the CD4+ cell counts. There is a high frequency of zero episodes (82.7%). Only 17 out of 98 men had an infection. The CD4+ cell counts ranged from 5 to 1585.

The local likelihood smoother was based on the Epanechnikov kernel function $K(z) = \frac{3}{4} (1 - z^2) 1_{\{z \in [-1, 1]\}}$ and the bandwidth was selected by sample splitting according to the criteria:

$$\hat{h}_{\text{opt}} = \arg \min_h \sum_{i=1}^m (y_i - \hat{y}_i)^2, \quad (4)$$

where $\hat{y}_i = \hat{\lambda}_i = \exp(\hat{\beta}_0) / (1 + \exp(\hat{\beta}_0))$ and m is the dimension of the testing sample. This sample was obtained from a random choice of about 20%, $m = 19$, observations of the whole sample. The remaining 79 observations comprise the training sample. The bandwidth \hat{h}_{opt} was used in the training sample to smooth the local likelihood logistic regression to the data.

The HIV data training sample curve estimates appear in Figure 1. The optimal bandwidth was $\hat{h}_{\text{opt}} = 16$ and the mean squared error was 0.069.

One can get a better fit with the whole sample. This happens because there is a loss of valuable information when some observations are put off. The HIV data whole sample curve estimates appear in Figure 2.

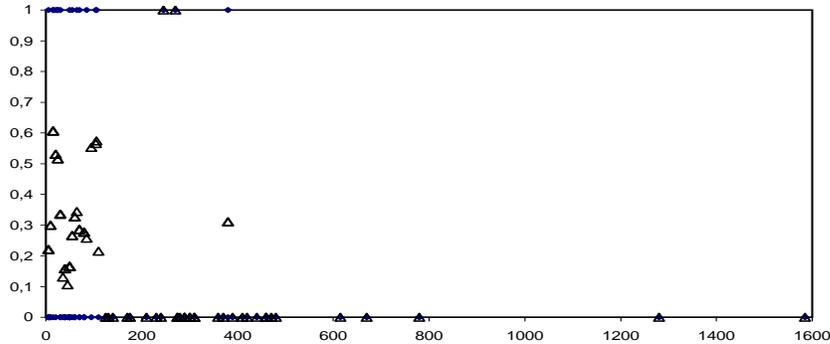


FIGURE 2. HIV data whole sample (black lozenge) and curve estimate (Δ).

Acknowledgments: This research was supported by Calouste Gulbenkian Foundation and PRODEP III.

References

- van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics* **51**, 738-743.
- Eguchi, S., Kim, T., and Park, B. (2003). Local likelihood method: a bridge over parametric and nonparametric regression. *Journal of Nonparametric Statistics* **15**, 665-683.
- Fan, J., Farnen, M., and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society B* **60**, 591-608.
- Fan, J., Heckman, N., and Wand, M. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141-150.
- Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag.
- Santos, J. (2005). *Estimação Não Paramétrica em Modelos de Regressão de Dados de Contagem com Excesso de Zeros*. PhD Thesis, ISA/Technical University of Lisbon.
- Staniswalis, J. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association* **84**, 276-283.
- Tibshirani, R., and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559-567.

Joint Modelling of a Longitudinal Variable and a Time to Event Data: Methodological and Computational Issues

Carles Serrat¹, Jaime-Abel Huertas² and Guadalupe Gómez²

¹ Dept. Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Avda Dr. Marañón 44–50, 08028–Barcelona, carles.serrat@upc.edu

² Dept. Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Edifici C5, Campus Nord, c/ Jordi Girona 1–3, 08034–Barcelona, jaime.abel@upc.edu, lupe.gomez@upc.edu

Abstract: On one hand, we analyze the existing references on joint models for longitudinal and survival data and we discuss their advantages and drawbacks, in particular the computational aspects referring to perform the estimation. On the other hand, we propose a shared parameters selection model and we develop all the procedures to apply it to an ongoing clinical trial, called TIBET, in which an intermittent therapeutic strategy has been assigned to each patient. Of special clinical interest is the lifetime that a patient needs before restarting treatment given the progression of biological markers recorded during the followup period. We conclude with a comparative analysis and the interpretation of the results.

Keywords: Joint Modelling; Longitudinal Data; Survival Analysis.

1 Joint Models in the Literature

Likelihood and Bayesian approaches rely on the specification of an appropriate likelihood for the joint model parameters; for both, much of the early literature focuses on models without autocorrelation structure for longitudinal model. Good reviews can be found in Hogan and Laird (1997), and in Tsiatis and Davidian (2004).

Analysis of the likelihood approach is recently made in Hsieh et al. (2006). Among others, Henderson et al. (2000) and Tseng et al. (2005) proposed joint modelling for survival times and longitudinal data. In Bayesian framework, Chi and Ibrahim (2006) give a model for multivariate longitudinal and multivariate survival data by using MCMC techniques.

In the computational field, Guo and Carlin (2004) give procedures for the estimating of joint modelling using standard computer packages, in both frameworks, Bayesian and Frequentist.

2 Notation

The idealized data for each subject $i = 1, \dots, n$ followed over an interval $[0, \tau]$ are $\{X_i, T_i, R_i(u), 0 \leq u \leq \tau\}$, where X_i is a vector of baseline covariates, T_i is event time, and $\{R_i(u), 0 \leq u \leq \tau\}$ is the longitudinal response trajectory.

For some set of times $t_{ij}, j = 1, \dots, n_i$, instead of the true values $R_i(t_{ij})$ we observe $Z_i(t_{ij})$. The event time may be right censored by C_i , which is independent of T_i . For subject i , observed data on the event time process consist of (Y_i, δ_i) , where $Y_i = \min\{T_i, C_i\}$ and $\delta_i = I(T_i \leq C_i)$ indicates whether Y_i is an uncensored value of T_i . The observed data $O_i = \{X_i, Y_i, \delta_i, Z_i, \tilde{t}_i\}$, where $\tilde{t}_i = (t_{i1}, \dots, t_{in_i})^T$ and $Z_i = (Z_i(t_{i1}), \dots, Z_i(t_{in_i}))^T$, are taken to be independent across i .

3 Random Effects Selection Models

Joint models can be broadly classified as either *selection models* or *mixture models*. In a selection model, the joint density function $f_{Z,T}$ is modelled as $f_{T|Z}f_Z$, and in a mixture model as $f_{Z|T}f_T$. Both types of models are applicable in either longitudinal or survival studies, although mixture models appear to be used primarily for longitudinal studies with informative dropout.

If we model Z_i only with random effects, under certain assumptions, a proposal to estimate the set of $\Psi = (\Psi_{T|Z}, \Psi_Z)$ parameters is maximizing the likelihood function with censored T_i :

$$L(\psi_{T|Z}, \psi_Z) = \prod_{i=1}^n \int_{b_i} f_Y(y_i, \delta_i | b_i; \psi_{T|Z}) f_Z(z_i | b_i; \psi_Z) f_b(b_i; \Gamma) db_i, \tag{1}$$

where,

$$f_Y(y_i, \delta_i | b_i; \psi_{T|Z}) = [f_{T|b}(y_i | b_i; \psi_{T|Z})]^{\delta_i} [1 - F_{T|b}(y_i | b_i; \psi_{T|Z})]^{1-\delta_i}. \tag{2}$$

Due to the fact that b_i appears in both models, the model for Z_i and T_i also is known as *shared parameter selection model*.

3.1 A Particular Shared Parameters Selection Model

For the longitudinal response process, a standard approach is to characterize $R_i(u)$, $u \geq 0$, only in terms of random effects b_{0i} and b_{1i} like

$$R_i(u) = b_{0i} + b_{1i}u. \tag{3}$$

However, alternatively to (3), a model for $R_i(u)$ that allows the trend to vary with time and induces a within-subject autocorrelation structure is:

$$R_i(u) = b_{0i} + b_{1i}u + Q_i(u). \tag{4}$$

where $Q_i(u)$ is a zero-mean stochastic process, usually taken to be independent of $b_i = (b_{0i}, b_{1i})^T$. The process $Q_i(u)$ describes local deviations whereas the term $b_{1i}u$ represents a constant trend.

Associations among the longitudinal and time to event processes and covariates, is characterized by the following semi-parametric model:

$$\begin{aligned} \lambda_i(u) &= \lim_{du \rightarrow 0} \Pr(u \leq T_i < u + du | T_i \geq u, R_i^H(u), X_i) / du \\ &= \lambda_0(u) \exp(\eta^T X_i + \beta R_i(u)), \end{aligned} \tag{5}$$

where $R_i^H(u) = \{R_i(t), 0 \leq t < u\}$ is the history of the longitudinal process up to time u , and the parameters are represented in β and the η vector.

If the model in (5) was expressed as $\lambda_0(u) \exp(\eta^T X_i + \beta b_{1i})$, then it reflects the belief that the hazard, conditional on the longitudinal history and covariates, is mainly associated with the assumed constant rate of change of the underlying smooth trend. If model takes $\beta R_i(u)$ as $\beta_1 b_{0i} + \beta_2 b_{1i} + \beta_3 (b_{0i} + b_{1i} u)$, the parameters β_1, β_2 and β_3 measure the association induced through the intercept, slope and current R value, respectively.

4 Estimating a Joint Model

A package with a direct procedure to estimate joint models does not exist, however it is possible to approximate the estimations with standard software, although the procedures do not cover the majority of the studies and have convergence problems. The procedures included in public or commercial software can only be adapted to the estimation of fully joint parametric models. While *SAS* has a procedure that directly approximates the joint estimations, the other packages may be used adapting their general optimization procedures. In addition to the proposal by Guo and Carlin (2004), the maximization of the likelihood in (1) is also possible with maximization procedures included in packages like *R*, *SPLUS* or *MATLAB* and numerical methods programming to approximate multiple integrals. On the Bayesian side, the *WinBugs* is a good alternative that can be applied to the joint modelling. A growth model for longitudinal and a Cox model for survival are estimated jointly by means of an *EM* algorithm proposed by Wulfsohn and Tsiatis (1997), which must be done by means of some programming tool. We have implemented the respective procedures for the estimation using standard packages for fully parametric models as well as the *EM* algorithm for the semiparametric approach. All the programming codes will be available at <http://www-eio06.upc.es/grass>.

5 Application

We apply the described techniques to the *TIBET* clinical trial. The trial contemplates the incorporation of interruption periods in the administration of an intensive therapy *HAART* (Highly Active Antiretroviral Therapy). A cohort of 100 patients enters the study with suspension of the treatment (state *OFF*). Basal and retrospective information is gathered, and every 4 weeks there is registered information of the *CD4* cell count and the viral load. If the patient's conditions deteriorate, the therapy is restarted (state *ON*), and so on. The longitudinal variable is the evolution of the *CD4* in 96 weeks. The survival time is the reinitiation time of therapy due to the *Viral Load* increase. The baseline covariate is "viral load pre-therapy" (*VL*).

We have estimated three joint models with the same growth model for the longitudinal part and different specifications for the hazards: a fully Weibull parametric model (FP)

and two semiparametric models (SP1 and SP2) according to the following expressions.

$$\begin{aligned}
 Z_i(t_{ij}) &= b_{0i} + b_{1i}t_{ij} + e_i(t_{ij}) \\
 \text{FP: } \lambda_i(t) &= \frac{1}{\sigma_s} t^{\frac{1}{\sigma_s}-1} \exp(\eta_0 + \eta_1 VL_i + \beta_1 b_{0i} + \beta_2 b_{1i}), \\
 \text{SP1: } \lambda_i(t) &= \lambda_0(t) \exp\{\beta(b_{0i} + b_{1i}t)\} \\
 \text{SP2: } \lambda_i(t) &= \lambda_0(t) \exp(\eta_1 VL_i + \beta_1 b_{0i} + \beta_2 b_{1i}),
 \end{aligned} \tag{6}$$

Tables including the results for these models and the respective comparisons and interpretations will be included.

References

- Chi, Y.-Y., and Ibrahim, J.G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432–445.
- Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modeling of longitudinal measurements and event time data. *Biostatistics* **4**, 465–480.
- Hogan, J.W., and Laird, N.M. (1997). Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistical Methods in Medical Research* **16**, 259–272.
- Guo, X., and Carlin, B.P. (2004). Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. *The American Statistician* **58**, 16–24.
- Hsieh, F., Tseng, Y.-K., and Wang, J.-L. (2006). Joint Modeling of survival and longitudinal data: likelihood approach revisited. *Biometrics* **62**, 1037–1043.
- Tseng, Y.-K., Hsieh, F., and Wang, J.-L. (2005). Joint modelling of accelerated failure time and longitudinal data. *Biometrika* **92**, 587–603.
- Tsiatis, A.A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 793–818.
- Wulfsohn, M.S., and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330–339.

Additive Survival Models with Shared Frailty

Giovani L. Silva¹ and M. Antónia Amaral-Turkman²

¹ Departamento de Matemática - IST, Universidade Técnica de Lisboa, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal, gsilvamath.ist.utl.pt

² Departamento de Estatística e Investigação Operacional, Universidade de Lisboa, Edifício C6, Campo Grande, 1749-016 Lisboa, Portugal, antonia.turkmanfc.ul.pt

Abstract: Frailty models have been proposed in order to investigate other sources of variation when the observed covariates do not fully explain the dissimilarities of the individuals in study. The frailty term can be partitioned into two or more terms in order to assess various types of frailty within the same individual. For instance, the frailty associated with a person may be divided into two random effects describing separately genetic and environmental factors, which are actually shared with other people such as mother, father, etc. The aim is to present a Bayesian analysis of additive survival models with shared or correlated frailty terms. An analysis of the adoption data described by Sørensen et al. (1988) motivates and illustrates the frailty models developed, using Markov Chain Monte Carlo methods for estimating quantities of interest.

Keywords: Shared Frailty Model; Additive Hazards Model; Survival Analysis. Bayesian Analysis; MCMC methods.

1 Introduction

In regression analysis for survival data, as the observed covariates are not fully explain the variation from individual to individual, a random effect (frailty) is included into the hazard function to take account that unobserved heterogeneity, e.g., genetic predisposition within families. In addition, the frailty can be partitioned into two or more terms in order to assess various types of frailty within the same individual. For example, the frailty of a person may be divided into two random effects describing separately genetic and environmental factors, which are shared with other people such as mother, father, etc.

Shared or correlated frailty models are herein analyzed for additive survival models (Aalen, 1980) from a Bayesian perspective (Silva and Amaral-Turkman, 2004). The additive hazards models have been presented both as a diagnostic tool and as a useful alternative to multiplicative hazards models, especially when the hazard functions are not proportional.

This work is organized as follows. Section 2 describes Aalen's additive model based on counting processes, as well as an additive frailty model with shared and correlated frailty terms. Section 3 deals with the Bayesian analysis of the additive frailty model by using Markov chain Monte Carlo (MCMC) methods for estimating quantities of interest. In section 4, we illustrate the methodology introduced here through the analysis of the adoption data described by Sørensen et al. (1988).

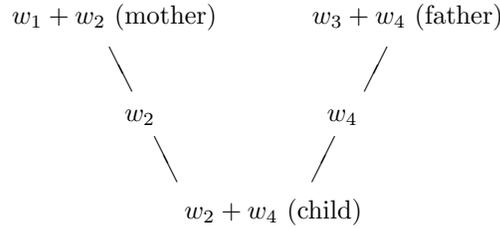


FIGURE 1. Graph of a frailty model for genetic data.

2 A shared frailty additive model

Aalen (1980) introduced an additive survival model defining the intensity of a counting process $N(t)$ - number of occurrences of a particular event up to time t - as

$$I(t|\mathbf{z}) = Y(t) \left(\alpha_0(t) + \sum_{q=1}^p \alpha_q(t) z_q \right), \tag{1}$$

where $Y(t)$ indicates whether the individual is in risk at time t , $\alpha_0(t)$ is the baseline intensity for individuals, and $\alpha_q(t)$ is the regression function that may reveal changes in the influence of the covariate z_q over time, $q = 1, \dots, p$.

In order to account for the unobserved heterogeneity, a random effect (w) is introduced into the intensity (1) additively (Rocha, 1996). Silva and Amaral-Turkman (2004) proposed a Bayesian approach for that new model that is therein so-called *additive frailty model*. Note that $\alpha_0(t)$ in that new intensity is interpreted as the baseline intensity for individuals with “null” frailty ($w = 0$).

The frailty term w for each individual may be partitioned into two or more terms, e.g., $w = w_1 + \dots + w_k$, where w_j are (correlated) frailty terms shared with other individuals, $j = 1, \dots, k$. For genetic setting, the frailty of a child may be associated with genes shared with mother and father (see Figure 1).

Assuming a multivariate counting process $\mathbf{N}(t) = (N_1(t), \dots, N_n(t))$ for n right-censored individuals (under a history \mathcal{F}_{t-}), shared frailty additive models are here defined by intensity function of $N_i(t)$, i.e.,

$$I_i(t|\mathbf{z}_i, \mathbf{w}) = Y_i(t) \left(\alpha_0(t) + \sum_{q=1}^p \alpha_q(t) z_{iq} + \mathbf{a}'_i \mathbf{w} \right), \tag{2}$$

where $\mathbf{w} = (w_1, \dots, w_k)'$ is the frailty vector and $\mathbf{a}_i = (a_{i1}, \dots, a_{ik})'$ is the vector of frailty indicator functions for the i -th individual, $i = 1, \dots, n$ (Silva and Amaral-Turkman, 2004). Petersen (1998) also showed a version of the model (2) for multiplicative frailty intensities.

3 An Bayesian approach of the current model

Partitioning the time axis into m disjoint intervals $B_j = [t_{j-1}, t_j)$, $j = 1, \dots, m$, independent gamma prior processes are assumed for the increments of the cumulative functions $\Omega_q(t) = \int_0^\infty \alpha_q(u)du$, i.e., the increment $\Omega_{qj} \equiv d\Omega_q(t)$ in B_j has gamma distribution with shape and scale parameters $c_q\Omega_{qj}^*$ and c_q , $j = 1, \dots, m$, $q = 0, \dots, p$. Notice that Ω_{qj}^* is interpreted as a prior guess of Ω_{qj} with degree of precision c_q . Let $\mathcal{D} = \{(N_i(t), Y_i(t), \mathbf{z}_i)\}$ be the survival data with n right-censored individuals. Assigning independent gamma priors for $\Omega_q(t)$, the posterior of the frailty model (2), denoted by $\pi(\Omega, \mathbf{w}, \delta|\mathcal{D})$, is proportional to

$$\prod_{j=1}^m \left[\prod_{i=1}^n \left(I_{ij}^{N_{ij}} e^{-I_{ij}} \right) \prod_{q=0}^p \left(\Omega_{qj}^{c_q\Omega_{qj}^* - 1} e^{-c_q\Omega_{qj}} \right) \right] \tau(\mathbf{w}|\delta) \tau(\delta), \tag{3}$$

where $I_{ij} \equiv \int_{t_{j-1}}^{t_j} I_i(t) dt = Y_{ij}(t)(\mathbf{z}_i' \Omega_j + \mathbf{a}_i' \mathbf{w} dt_j)$, $N_{ij} \equiv dN_i(t_j)$, $\Omega = (\Omega_{11}, \dots, \Omega_{pm})'$, $\Omega_{qj}^* = r_q dt_j$, r_q is a proposed value for $\alpha_q(t)$, $dt_j = t_j - t_{j-1}$, $\tau(\mathbf{w}|\delta)$ is the frailty distribution and $\tau(\delta)$ is a prior for hyperparameter δ .

The frailty distribution is traditionally gamma with hyperparameter δ , which measures the degree of unobserved heterogeneity through, e.g., via its standard deviation (σ_W). The posterior (3) is awkward to work with, since the marginal posterior distributions of Ω and δ are not easy to obtain explicitly. Nevertheless, these posteriors can be evaluated using Markov chain Monte Carlo (MCMC) methods.

4 Illustration

Using the model (3) for the adoption data (Sørensen et al., 1988) with 125 families 1924-1987, the intensities of death by infection (e.g., pneumonia) for biological mother, son and adoptive mother are, respectively,

$$\begin{aligned} I_{i1}(t|\mathbf{w}) &= Y_{i1}(t)[\alpha_{01}(t) + w_{i1} + w_{i2}] \\ I_{i2}(t|\mathbf{w}) &= Y_{i2}(t)[\alpha_{02}(t) + w_{i1} + w_{i3} + w_{i4}] \\ I_{i3}(t|\mathbf{w}) &= Y_{i3}(t)[\alpha_{03}(t) + w_{i3} + w_{i5}]. \end{aligned} \tag{4}$$

For simplicity, the posterior (3) is here associated with gamma frailties $(1, \delta_l)$, non-informative priors for δ_l , 65 intervals B_j 's, $c_q = 0.001$ and $r_q = 0.1$, $q = 1, 2, 3$, $l = 1, \dots, 5$. After 6000 iterations simulated, including 1000 for burn-in period, some quantities of interest were estimated for the shared additive frailty model (4).

The estimates in Table 1 indicate little unobserved heterogeneity both shared genetic ($\hat{\sigma}_{G_s}$) and environment ($\hat{\sigma}_{E_s}$) factors, shared environment factors explain 20.3% more of the variability than the shared genes, while non-shared effects have 10.5% less importance than genes.

TABLE 1. Estimates of variance components for frailties.

parameter	mean	s.d.	CI(2.5%)	CI (97.5%)
σ_{G_s}	0.022	0.0021	0.0178	0.0261
σ_{E_s}	0.023	0.0023	0.0191	0.0281
$\sigma_{E_s}^2/\sigma_{G_s}^2$	1.203	0.3337	0.6772	1.9830
$\sigma_{EG_{ns}}^2/\sigma_{G_s}^2$	0.895	0.2633	0.4796	1.4970
$\sigma_{EG_{ns}}^2/(\sigma_{G_s}^2 + \sigma_{E_s}^2)$	0.407	0.1058	0.2397	0.6564

Acknowledgments: Special Thanks to Thorkild Sørensen for supplying the adoption data. This paper was partially supported by Fundação para a Ciência e Tecnologia.

References

- Aalen, O.O. (1980). A Model for Nonparametric Regression Analysis of Counting Processes. *Lectures Notes in Statistics* **2**, 1-12.
- Petersen, J.H. (1998). An additive frailty model for correlated life times. *Biometrics* **54**, 646-661.
- Rocha, C.S. (1996). Survival Models for Heterogeneity Using the Non-Central Chi-Squared Distribution with Zero Degrees of Freedom. In: *Lifetime Data: Models in Reliability and Survival Analysis* edited by Jewell, N.P.; Kimber, A.C.; Lee, M.T., and Whitmore, G.A.. 275-279, Kluwer Academic Publishers.
- Silva, G.L., and Amaral-Turkman, M.A. (2004). Bayesian analysis of an additive survival model with frailty. *Communications in Statistics – Theory and Methods* **33**, 2517-2533.
- Sørensen, T.I.A., Nielsen, G.G., Andersen, P.K. and Teasdale, T.W. (1988). Genetic and environmental influences on premature death in adult adoptees. *New England Journal of Medicine* **312**, 727-732.

Statistical modelling of development of executive function in early childhood

I. Solis-Trapala¹, P. Diggle¹ and C. Lewis²

¹ Department of Medicine, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK

² Department of Psychology, Fylde Building, Lancaster University, Lancaster, LA1 4YF, UK

Abstract: We develop likelihood-based statistical inference to assess change in performance of young children on repeated executive function tests. To do this we build dynamic regression models that take into account the inter-relationships between tests, and propose various approaches to extrapolate the results over time.

Keywords: Dynamic regression model; Executive function; Latent variables.

1 Motivation

We are devising latent variable models for the analysis of complex multivariate repeated measurement data in which the sampling distributions associated with the individual measurements may take non-standard forms. This work has been motivated by a recent study carried out by Prof. C. Lewis and Dr. K. Shimmon (Shimmon, 2004) who investigate the development of executive control in young children. Children in the study were presented with a battery of executive function tests. Lewis and Shimmon considered four domains of executive function, namely inhibitory control, attentional flexibility, working memory and planning. At each time-point, they measured each of the above domains by repeatedly administering to children two versions of a given task, specifically designed to represent the relevant domain. The repeated measurement character of the resulting data therefore manifests itself on two different time-scales: tests were administered at three different *times* over an 18 month period; and at each of the three times, each test involves a *sequence* of trials.

The data set thus generated is large and complex and contains a mixture of discrete and continuous measures. The main aims of the study are a) to examine the effects of task modifications on the performance of children; b) to assess changes on performance within sequences of repeated trials; c) to explore interrelationships between executive tasks and d) to model performance of children over time.

2 Statistical inference

Data are in the form of two or more series of outcomes (typically, binary and continuous) gathered for each domain of executive function, at three time-points. Initially, we explore child-specific patterns of cognitive development within each domain at a single time-point. This involves joint modelling of two or more series of dependent outcomes,

not necessarily of the same type, within each domain. For each child, it will be assumed that there is an underlying development profile which is not observable. More specifically, at this stage, we assume that the series of outcomes, corresponding to different versions of each task, are conditionally independent given an individual-level latent variable. In addition, we look at the dynamic development of the series of trials by conditioning on covariates that are functions of previous observations as defined by Aalen *et al.*, 2004. The results of this stage will allow comparison of performance of children on the different versions of tasks and prediction of child's underlying development profiles for each domain. Once a multivariate representation of each domain has been attained, we investigate the extrapolation of results over time.

3 Example

In the Lewis and Shimmon study, working memory was measured by three different tasks. The *boxes* tasks (scrambled and stationary) were administered to 115 children at three time periods, and the *digit span* task at two time periods. The latter task yielded scores that can be assumed to follow a Gaussian distribution. We elaborate on the boxes tasks to exemplify modelling of executive function data generated from a non-standard sampling distribution. For simplicity of exposition, we here restrict our attention to data collected at a single time.

For the boxes tasks, six boxes were decorated with different colours and designs. While the child was watching, a sweet was placed in each box and all lids closed. A 10 second delay was imposed (a black screen was lowered to hide the boxes from the child's sight), after which the child was allowed to select one of the boxes and retrieve its content. The lid on that box was closed and another delay imposed. The child was again allowed to open one of the boxes. This continued until the child had retrieved the six sweets. In the scrambled version of the this task, the boxes were randomly switched to different positions behind the screen during the delay. Thus participants must try to remember the design of the boxes already selected in order to be successful on subsequent trials. The stationary version of the box task was identical to the scrambled version, except that the boxes were left in the same position.

We denote by $\mathbf{z}_i = (z_{i1}, \dots, z_{in_i})$ a vector with binary entries that represent failure (0) or success (1) of child i at each of the n_i trials that the participant took in order to retrieve all sweets. Let $s_{ij} = 5 - \sum_{l=1}^j z_{il}$ be the number of sweets that remain to be retrieved at trial j^{th} . The probability of finding a sweet at the first trial is one, hence we model the probability of success given that there are five or less sweets to be retrieved.

Let $P_{ik} \equiv \Pr(z_{ij} = 1 \mid s_{ij} = k)$, for $k = 1, \dots, 5$, the probability of success given that k sweets remain to be found. Conditional on the degree of difficulty to find a sweet as reflected by the number of sweets remaining to be retrieved, we assume P_{ik} to be of logistic form. In addition we assume that the probability of success is affected by the participant's working memory ability. We introduce this effect as a latent variable. Therefore we model the probability of success as

$$\text{logit}(P_{ik}) = \alpha_k + \mathbf{x}_i' \beta + U_i, \quad (1)$$

where α_k is a fixed effect that accounts for the increasing difficulty of finding a sweet at trial j^{th} as s_{ij} decreases; U_i is a latent variable that represents the participant's executive ability, and \mathbf{x}_i is a vector of covariates which include the type of task (scrambled or stationary).

Statistical inference for the fixed effects in (1) can be based on the likelihood function

$$L(\{\alpha_k\}, \beta; \mathbf{z}_i) = \prod_{i=1}^{115} \int \prod_{k=1}^5 \left[\prod_{A_i} (1 - P_{ik}) \right] P_{ik} f(U_i; \theta) dU_i, \quad (2)$$

where $A_i = \{z_{ij} \in \mathbf{z}_i \mid z_{ij} = 0 \ \& \ s_{ij} = k\}$ and $f(U_i; \theta)$ is the density function of the latent variable U_i .

One of the questions of scientific interest in this study involves prediction of the child's underlying working memory development. As noted earlier, performance on working memory was also measured through the digit span task. Under the assumption that performance on this task and on the two boxes tasks are conditionally independent given U_i , we can build a joint likelihood function based on the combined scores. This likelihood function will be the integral of the product of the three conditional densities and the latent variable density $f(U_i)$.

4 Relationships between executive functions

We use the approach of Wermuth and Cox (2006) to investigate causal relationships amongst the four executive functions. Each of these components are part of a core structure that can be expressed through graphical models. Consider for example the following two competing models:

- I. Inhibitory control and attentional flexibility are closely related skills that form the basis of planning and are underpinned by working memory.
- II. Planning is the superordinate executive skill requiring inhibitory control, attentional flexibility and working memory.

The four nodes in the centre of Figures 1 and 2 represent latent variables for inhibitory control, attentional flexibility, planning and working memory. The nodes outside the square represent the outcome variables from the tasks. Each model induces a particular pattern of inter-dependence amongst the individual measurements on each child. We are currently developing likelihood-based methods of inference for fitting and comparing models of this kind, in which the latent variable structure is specified graphically whilst the observed measurements conditional on the latent structure are modelled according to sampling distributions which respect the experimental protocols for the various tests.

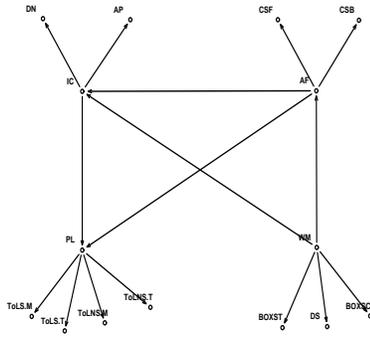


FIGURE 1. Model I.

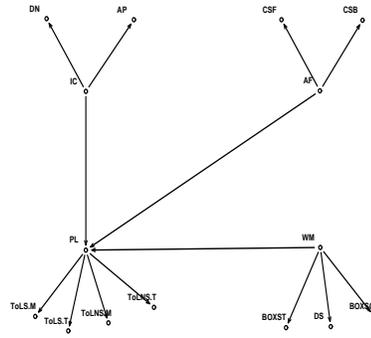


FIGURE 2. Model II.

Acknowledgments: This work is supported by a grant from the Economic and Social Research Council for the development of statistical modelling in the Social Sciences.

References

Aalen, O., Fosen J., Weedon-Fekjær, H., Borgan, Ø., and Husebye E. (2004) Dynamic analysis of multivariate failure time data. *Biometrics* **60**, 764-773.

Shimmon, K. L. (2004). *The development of executive control in young children and its relationship with mental-state understanding: a longitudinal study*. Ph.D. Thesis, Lancaster University.

Wermuth, N., and Cox, D.R. (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society B*, **66**, 687-717.

Spatial Modelling of Field Experiments: Sample Variogram and Enhanced Diagnostics

Katia Stefanova¹, Alison Smith² and Brian Cullis²

¹ Department of Agriculture and Food WA, Perth WA 6151, Australia

² Wagga Wagga Agricultural Institute, Wagga Wagga NSW 2650, Australia

Abstract: In this paper we present the analysis of a uniformity field trial using the technique proposed by Gilmour, Cullis and Verbyla (1997). In particular we clarify the role of the sample variogram and present a range of enhanced graphical diagnostics to assist with the spatial modelling process.

Keywords: Spatial Modelling; Variogram; Mixed Model; Residual Maximum Likelihood; Coverage Intervals.

1 Introduction

Spatial analysis is routinely used for the analysis of plant variety evaluation trials in Australia. The key to the efficient estimation of variety effects is the appropriate choice of the plot error variance model. Gilmour, Cullis and Verbyla (1997) presented a sequential approach to modelling and acknowledged the presence of three main types of spatial variation. The sample variogram plays a key role in this approach but may be difficult to interpret. We review the use of diagnostic tools described in Gilmour *et al.* (1997) and present an enhanced graphical representation based on the use of 95% coverage intervals.

2 Statistical Methods

2.1 Spatial Linear Mixed Model

In this section we present a model for the data from small plot field experiments which encompasses extra sources of variation other than that caused by natural variation. The model decomposes error variation into three components, namely global (spatial) variation, local (spatial) variation and extraneous variation. We assume there are yield data for $n=rc$ plots, where r and c are the numbers of rows and columns respectively. In these trials the plots are contiguous, that is, they consist of a single array. Denote the vector of plot yields $y_i(\mathbf{s}_i)$, $i = 1, 2, \dots, n$, where \mathbf{s}_i is a two cell vector of the Cartesian coordinates of the (row, column) plot centroid. The model for \mathbf{y} is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{e} = \boldsymbol{\xi} + \boldsymbol{\eta}$ is the vector of plot errors, $\boldsymbol{\tau}^{(t \times 1)}$ is the vector of fixed effects with $\mathbf{X}^{(n \times t)}$ design matrix, $\mathbf{u}^{(b \times 1)}$ is the vector of random effects with $\mathbf{Z}^{(n \times b)}$ design matrix,

$\xi^{(n \times 1)}$ is a spatially dependent random error vector and $\eta^{(n \times 1)}$ is a zero mean random vector whose elements are pairwise independent. The latter is often referred to as a measurement error. We assume that the joint distribution of (\mathbf{u}, \mathbf{e}) is Gaussian with zero mean and variance matrix

$$\sigma^2 \begin{bmatrix} \mathbf{G}(\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\phi) \end{bmatrix}$$

where γ and ϕ are vectors of variance parameters. The marginal distribution of \mathbf{y} is then

$$\mathbf{y} \sim \mathbf{N}((\mathbf{X}\tau, \sigma^2(\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})).$$

2.2 Diagnostics and Tests Used in the Modelling Process

The spatial modelling process commences by assuming that the variance model for local trend ξ is a separable process involving a first order autoregressive model for both rows and columns. The residuals from this model then provide the basis for identification of global and extraneous variation, as well as assessing the adequacy of the variance structure for local trend. Apart from the usual diagnostic tools relevant for the examination of the distributional assumptions in a standard linear model, the adequacy of the assumed variance (or correlation) model for the errors is examined. The key issues we focus on are: the presence of global variation and/or non-stationarity; the presence of extraneous variation; the adequacy of the correlation model for local trend and the need for the measurement error component. Firstly we examine a trellis plot of residuals, which is a plot of residuals against row (column) number conditional on column (row) number. Examination of this plot often reveals data anomalies and the presence of global trend. The next graphical diagnostic involves the sample variogram. We also present enhancements to the sample variogram by the implementation of 95% coverage intervals. We consider the two "faces" of the sample variogram, the slices corresponding to zero row/column displacement, augmented with approximate 95% coverage intervals. They are obtained via simulations of the current model in an analogous manner to Atkinson(1985), who constructed envelopes for half normal plots. The sample variogram is calculated for each simulation, then the mean, the 2.5% and 97.5% percentiles are obtained for each displacement. This assists with appropriate model selection reducing the risk of overfitting. Formal tests of models with nested variance structures but the same fixed effects model are provided by REML likelihood ratio tests (REMLRTs).

3 Example

The data relate to a field experiment conducted at Wongan Hills (WH-W trial), Western Australia. The field trial comprising 25 rows and 12 columns was a uniformity trial planted to one genotype with a harvested area for each plot of $3m \times 1.25m$. Management practices are usually aligned with rows and columns. For example, the trials were sown with a cone seeder which sows two plots at once and is driven in one direction until reaching the end of the row. It then returns in the opposite direction.

TABLE 1. Overview of models fitted to the WH-W trial

Model	Variation		Variance Parameters	REMLLL \ddagger Pr(D)	
	Global/Extraneous \dagger	Local			
1		AR1 \times AR1	3	317.3	
2	lin(row)	AR1 \times AR1	3	318.0	
3	lin(row) + ran(row)	AR1 \times AR1	4	330.4	< .001
4	lin(row) + cone	AR1 \times AR1	3	341.8	
5	lin(row) + cone + ran(col)	AR1 \times AR1	4	355.3	< .001

\dagger lin(row) represents the linear regression of yield on the row index, lin(row) and cone are included in \mathbf{X} , ran(col) represents a factor based on the column index and is included in \mathbf{Z} .

\ddagger REMLLL indicates REML log-likelihood and $D = -2$ times the change in REMLLL between two nested models. The REMLRTs may not be used to compare models separated by a horizontal line.

Covariates were defined for sowing/harvesting direction and cone side. The analyzes were performed using the statistical software ASREML. Table 1 presents an overview of the sequence of models for the WH-W trial. We begin by modelling the spatial variation with a separable process involving a first order autoregressive model for both rows and columns. Figure 1 presents the sample variograms for three of the models presented in Table 1. The variogram from model 1 depicted in Figure 1(a) shows clear

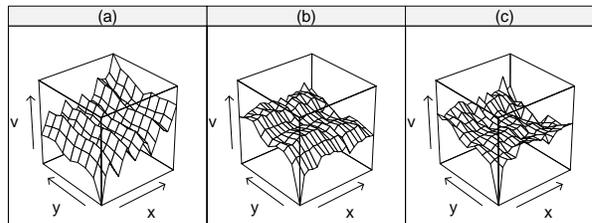


FIGURE 1. Wireframe plot of sample variograms for three models for the WH-W trial: (a) model 1, (b) model 4 and (c) model 5

departure from the theoretical variogram, which is smooth and increases exponentially in both the x and y directions to a common sill (asymptote). A key feature is the steady increase in semi-variance as the displacement in the row direction (x) increases (Figure 1(a)), which implies that a linear drift is present in the residuals (see Model 2). Moreover, the sample variogram looks very different from the mean from the 100 simulations of model 1 and falls outside the 95% coverage interval at several displace-

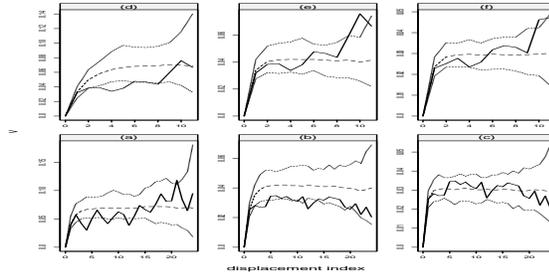


FIGURE 2. Trellis plot of faces of the sample variogram (solid line) and mean (dashed) and approximate 95% coverage interval (dotted) for three models for the WH-W trial: (a) model 1, row face (b) model 4, row face (c) model 5, row face (d) model 1, column face (e) model 4, column face (f) model 5, column face

ments, Figure 2(a). There is also a distinctive saw-tooth appearance to the variogram in Figure 1(a), therefore a cone covariate is included in Model 4, Figure 1(b). The final model (Figure 1(c)) is reasonably consistent with the theoretical variogram.

Acknowledgments: We gratefully acknowledge the financial support of the Grains Research and Development Corporation.

References

- Atkinson, A.C. (1985). *Plots, Transformations and Regressions*. Oxford: Clarendon Press.
- Gilmour, A.R., Cullis, B.R. and Verbyla, A.P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics* **2**, 269-273.

Hierarchical and empirical Bayes estimators in the analysis of insurance claims

George Streftaris¹ and Bruce J. Worton²

¹ Corresponding author: School of Mathematical and Computer Sciences, Maxwell Institute of Mathematical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK,
G.Streftaris@ma.hw.ac.uk

² School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, JCMB, King's Buildings, Edinburgh EH9 3JZ, UK

Keywords: Bayes risk; Effective sample size; Insurance claims; Markov chain Monte Carlo; Poisson regression

1 Introduction

The estimation of event counts using various Poisson-related models has attracted considerable attention, not least due to problems associated with overdispersion in the observed data. In this paper we consider an efficient and accurate approximate Markov chain Monte Carlo (MCMC) methodology for full Bayesian analysis in a Poisson/log-normal model, which can be used in the analysis of insurance claims (Streftaris and Worton, 2006). We investigate its efficiency and Bayes risk performance (e.g. Carlin and Louis, 2000) in comparison with other hierarchical and empirical Bayes approaches under various loss functions and prior assumptions, through suitably designed Monte Carlo simulation studies. Empirical Bayes methodology has traditionally been attractive in actuarial work due to its computational convenience and also because of the difficulty in specifying prior distributions in hierarchical Bayes models (e.g. Makov et al, 1996; Pai, 1997).

2 The model

In an actuarial context, we assume that given the parameters $\theta_1, \theta_2, \dots, \theta_m$, the counts Y_i represent the number of actual claims in group $i = 1, \dots, m$, and are conditionally independent Poisson variables with respective means $\theta_i E_i$, i.e.

$$Y_i | \theta_i \sim \text{Poisson}(\theta_i E_i), \quad i = 1, \dots, m,$$

where E_i , $i = 1, \dots, m$, is the total exposure of group i to a specific policy. The parameters θ_i give the rate of occurrence of claims and depend on p covariates $\mathbf{x}_i^T = (x_{1i}, x_{2i}, \dots, x_{pi})$, related to group i (e.g. age), in a regression structure modelled through a log-normal prior distribution, i.e. $\lambda_i = \log(\theta_i) \sim N(\mathbf{x}_i^T \mathbf{b}, \sigma^2)$, $i = 1, \dots, m$,

where $\mathbf{b} = (b_1, b_2, \dots, b_p)^T$ is a vector of unknown coefficients. Vague hyperprior distributions are employed to reflect the uncertainty associated with the log-normal parameters when little prior knowledge is available. Our approach may also be extended to a larger class of related models, where other prior distributions are assumed for the Poisson parameters, as for example in compound Poisson sampling models or mixed Poisson models (e.g. Grandell, 1997).

3 Approximate hierarchical Bayes methodology

We employ an efficient and accurate methodology for approximate posterior analysis under a hierarchical Bayesian framework (Streftaris and Worton, 2006). It involves a close approximation to the conditional distribution of the Poisson rates θ_i given all other model parameters and the data. The approximation is based on a log-normal/gamma mixture density which matches the first three moments of the original distribution. For the computation of the moments of the target distribution we use a method relying on entropy distance (Kullback-Liebler divergence) minimization. The resulting density is then employed in a Gibbs sampling scheme for inference. Effective sample size calculations (e.g. Brooks et al, 2003) are illustrated in Figure 1, and demonstrate that this approach mixes more efficiently than an exact algorithm, while providing the same posterior estimates.

For comparison purposes we evaluate the risk of the hierarchical Bayes (HB) estimator resulting from an exact MCMC algorithm, and two empirical Bayes (EB) estimators which are based on a linear shrinkage rule. The first EB estimator is derived in a way such that it minimizes the Bayes risk among all linear estimators of the same form, and is given by $\hat{\theta}_i^{\text{EB}} = (1 - c) y_i + c \bar{y}$, where \bar{y} denotes the sample mean of the data, and $c \in [0, 1]$ is given as $\frac{E(\theta_i)}{\text{var}(\theta_i) + E(\theta_i)}$ and is estimated in the EB context by $\min\left\{\frac{(m-1)\bar{y}}{\sum_{i=1}^m (y_i - \bar{y})^2}, 1\right\}$. The second EB estimator multiplies the coefficient c by a factor of $\frac{m-3}{m-1}$, and is a modification to the above estimator, proposed by Morris (1983) to resemble the shrinkage behaviour of a HB approach.

4 Risk performance of estimators

We investigate the performance of the considered estimators in terms of Bayes risk, given as $E_{\theta} E_{Y|\theta}\{L(\hat{\theta}, \theta)\}$, where $L(\hat{\theta}, \theta)$ is a loss function of the estimator $\hat{\theta}$. We consider loss functions of the form $L(\hat{\theta}, \theta) = \frac{1}{m} \sum_{i=1}^m \frac{(\hat{\theta}_i - \theta_i)^2}{\theta_i^k}$ for $k = 0$ (squared error loss) and $k = 1$ (normalized squared error loss). We also investigate a similar loss function, where averaging over the m components in the form above form is replaced by their maximum (maximum component squared error loss). Various prior specifications are considered. The results are shown in Table 1, and demonstrate that the risk properties of the approximate hierarchical estimator (AHB) are often better than those of the exact HB algorithm; this can be explained by the faster convergence of the approximate method. Additionally, it is shown that the AHB estimator has smaller risk than the EB methods, although additional simulations (not presented here) suggested

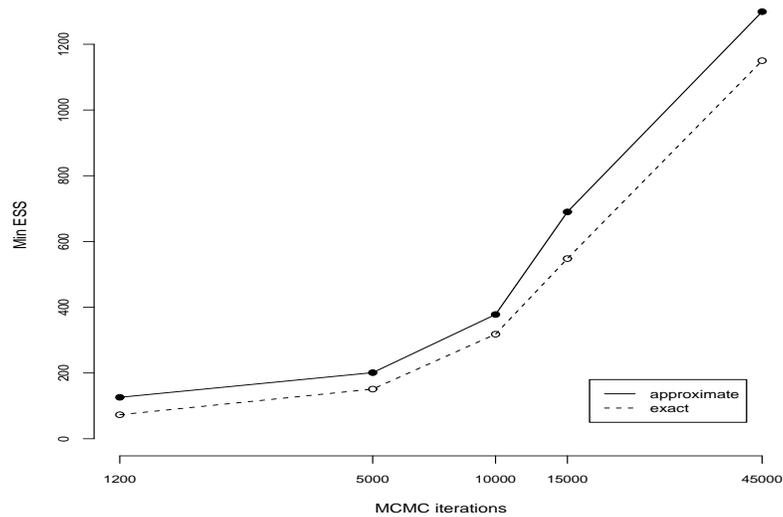


FIGURE 1. Minimum effective sample size (ESS) for the approximate Gibbs sampling approach (solid line) and an exact MCMC algorithm (dashed line). A logarithmic scale has been used on the x axis.

that for small m (e.g. $m = 10$) this result holds mainly when the true variance of $\theta_i, i = 1, \dots, m$, is large (as expected due to the vague hyperpriors used). Simulations with alternative models also showed that our method is robust under the assumption that the data come from other distributions, e.g. a Poisson/gamma model.

5 Extensions

In current work we investigate the development and performance of exact EB estimators using various possibilities regarding the specification of the parameters of prior distributions. These include algorithms involving the expectation-maximization (EM) and importance sampling (IS) techniques that provide efficient estimation when analytical results are not available.

TABLE 1. Estimated Bayes risk under squared error loss ($k = 0$), normalized squared error loss ($k = 1$) and maximum component squared error loss. Number of observations is $m = 30$ and $N = 10^4$ Monte Carlo simulations were used.

$E(\theta_i)$ $\text{var}(\theta_i)$	5.0		10.0	
	2.5	10.0	5.0	20.0
Squared error loss				
Approximate HB	1.907	3.431	3.810	6.954
Exact HB	1.948	3.436	3.864	6.934
Empirical Bayes	1.973	3.499	3.931	7.014
EB (Morris)	1.967	3.483	3.917	6.991
Normalized squared error loss				
Approximate HB	0.379	0.686	0.382	0.694
Exact HB	0.387	0.684	0.386	0.694
Empirical Bayes	0.397	0.710	0.395	0.709
EB (Morris)	0.394	0.703	0.393	0.704
Max. component squared error loss				
Approximate HB	13.095	28.397	23.857	49.812
Exact HB	13.292	28.289	24.090	49.371
Empirical Bayes	13.368	28.852	24.441	49.954
EB (Morris)	13.227	28.345	24.255	49.297

References

- Brooks, S.P., Giudici, P., Roberts, G.O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). *J. Roy. Statist. Soc., B* **65**, 3–55.
- Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Boca Raton: Chapman & Hall/CRC.
- Grandell, J. (1997). *Mixed Poisson Processes*. New York: Chapman & Hall/CRC.
- Makov, U.E., Smith, A.F.M. and Liu, Y-H. (1996). Bayesian methods in actuarial science. *The Statistician* **45**, 503–515.
- Morris, C.N. (1983). Discussion of ‘Construction of improved estimators in multiparameter estimation for discrete exponential families’, by Ghosh M., Hwang J.T., and Tsui K.W. *Ann. Statist.* **11**, 372–374.
- Pai, J.S. (1997). Bayesian analysis of compound loss distributions. *J. Econometr.* **79**, 129–146.
- Streftaris, G. and Worton, B.J. (2006). Efficient and accurate approximate Bayesian inference with an application to insurance data. Submitted.

Bayesian model selection criteria: a comparative study through simulation

Júlia Teles¹ and Maria Antónia Amaral Turkman²

¹ Departamento de Métodos Matemáticos/CIPER, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa, Estrada da Costa, 1495-688 Cruz Quebrada-Dafundo, Portugal, jteles@fmh.utl.pt

² Departamento de Estatística e Investigação Operacional/CEA, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal, antonia.turkman@fc.ul.pt

Abstract: Variable selection in regression models is a particular case of model selection. A simulation study, regarding variable selection in linear regression models, is performed to compare several Bayesian model selection criteria, namely, the estimated Bayesian information criterion, the deviance information criterion, the full sample log-score criterion and the cross validation log-score criterion.

Keywords: Bayesian model selection criteria; Linear regression; Simulation study.

1 Model selection criteria

Model selection is one of the most important statistical problems, therefore the study of performance of model selection criteria is a useful task. In this work we compare the performance of two penalized likelihood criteria, the estimated Bayesian information criterion and the deviance information criterion, and two versions of the log-score criterion, namely the full-sample version and the cross-validated version.

Consider a model m_i and let θ_i be the vector of model parameters and p_i the number of parameters in the model. We represent the likelihood by $L(\theta_i|y, m_i)$, where y is the observed data vector, with n observations.

Advocating that penalized likelihood measures are easy to obtain, Carlin and Louis (2000, p. 220) suggest the use of an estimated Bayesian information criterion. This criterion consists in selecting the model m_i that minimizes

$$\widehat{\text{BIC}}_{m_i}(y) = -2\hat{\ell}_i + p_i \ln n, \quad (1)$$

where $\hat{\ell}_i = E[\ln f(y|\theta_i, m_i)|y]$ is the posterior expectation of the log-likelihood. The first term in the second member of equation (1) is an overall measure of model fit and the second term is a penalty for model complexity.

In the case of complex hierarchical models the number of parameters is not well defined, making it difficult to use the Bayesian information criterion. Hence, Spiegelhalter et al. (2002) propose the use of deviance information criterion. Let $D_{m_i}(\theta_i) = -2 \ln f(y|\theta_i, m_i) + 2 \ln g(y|m_i)$ be the “Bayesian deviance”, where $f(y|\theta_i, m_i)$ is the

likelihood function for the observed data y and $g(y|m_i)$ is a function of the data y . In model selection context it is usual to consider $g(y|m_i) = 1$.

The deviance information criterion consists in selecting the model m_i that minimizes

$$\text{DIC}_{m_i}(y) = \overline{D_{m_i}} + p_{D,m_i}, \tag{2}$$

where $\overline{D_{m_i}}$ represents the posterior expectation of the deviance, that summarizes the fit of the model, and p_{D,m_i} is a measure of model complexity, defined as the difference between the posterior expectation of the deviance and the deviance calculated at the posterior expectation of parameter vector.

Besides these two penalized likelihood criteria, we also consider the two scoring rule based criteria (*e.g.*, Draper and Krnjajić, submitted). One of them is the full-sample log-score criteria that consists in selecting the model m_i that maximizes

$$\text{LSFS}_{m_i}(y) = \frac{1}{n} \sum_{j=1}^n \ln p^*(y_j|y, m_i), \tag{3}$$

where $p^*(\cdot|y, m_i)$ is the posterior predictive distribution for a future observation, based on the vector y . The other criterion is the cross-validated log-score criterion that consists in selecting the model m_i that maximizes

$$\text{LSCV}_{m_i}(y) = \frac{1}{n} \sum_{j=1}^n \ln p(y_j|y_{(-j)}, m_i), \tag{4}$$

where $p(\cdot|y_{(-j)}, m_i)$ is a posterior predictive distribution for a future observation, based on the vector $y_{(-j)}$, obtained removing y_j from the vector y .

2 Simulation study

We compare the performance of DIC, $\widehat{\text{BIC}}$, LSFS and LSCV criteria in the context of variable selection in Bayesian linear regression model.

2.1 Bayesian linear regression model

Consider the linear regression model, $Y|\beta, \sigma^2, X \sim N_n(X\beta, \sigma^2 I_n)$, where $Y = (Y_1, \dots, Y_n)'$, $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$, X is a $n \times (p+1)$ design matrix, $\epsilon = (\epsilon_1, \dots, \epsilon_n) \sim N_n(0, \sigma^2 I_n)$ and I_n is the $n \times n$ identity matrix.

We consider a non informative prior distribution for the parameter vector, $h(\beta, \sigma^2) = (\sigma^2)^{-1}$, for $\beta \in \mathcal{R}^{p+1}$ and $\sigma^2 > 0$. Therefore the joint posterior distribution for β and σ^2 can be factorized as $h(\beta, \sigma^2|y) = h(\beta|\sigma^2, y) h(\sigma^2|y)$. The conditional posterior distribution of β , given σ^2 , is $N_{p+1}(\hat{\beta}, \sigma^2 V_\beta)$, where $\hat{\beta} = (X'X)^{-1}X'y$ and $V_\beta = (X'X)^{-1}$. The marginal posterior distribution of σ^2 is $\chi^2\text{-Inv}(k, s^2)$, where $k = n-p-1$ and $s^2 = \frac{1}{k}(y - X\hat{\beta})'(y - X\hat{\beta})$ (*e.g.*, Gelman et al., 1995).

2.2 Details of the simulation study

In this simulation study we consider four potential predictors, x_1, \dots, x_4 , that are obtained as independent standard normal vectors. The dependent variable is generated according to the model

$$Y = \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_4 + \epsilon, \quad (5)$$

where ϵ is a vector of independent normal observations with zero mean and standard deviation equal to 2.5, and δ_j is a binary indicator variable determining if the predictor x_j is included or not included in the model, for $j = 1, 2, 3, 4$. The 16 possible models are indicated by m_1, \dots, m_{16} : m_1 for the model that includes all predictors; m_2, \dots, m_5 for models that include three predictors; m_6, \dots, m_{11} for models that include two predictors; m_{12}, \dots, m_{15} for models that include one predictor and, finally, m_{16} for the model that only includes a constant. The predictors included in each model m_i can then be identified by the vector $\delta = (\delta_1, \delta_2, \delta_3, \delta_4)$.

The dependent variable is generated according to the model m_i , for $i = 1, \dots, 16$, and all possible models are adjusted; the DIC, $\widehat{\text{BIC}}$, LSFS and LSCV criteria are used to select the best model. This procedure is replicated 100 times in the cases $n = 200$ and $n = 100$.

2.3 Results and discussion

The performance of DIC, $\widehat{\text{BIC}}$, LSFS and LSCV criteria are compared using the frequency of correct classification of the data generated from the 16 possible models. The results are presented in Table ??.

In the case $n = 200$, it seems that $\widehat{\text{BIC}}$ is, in general, the method with the best performance, while DIC has a weak performance for the data generated from the models with smaller number of predictors. On the other hand, the two log-score criteria have a weak performance for the data generated from the models with more predictors. The LSCV has in general worse performance than the LSFS.

The main difference in the results achieved in the case $n = 100$ is the small frequency of correct classification for all criteria, and the more similar performances of DIC and $\widehat{\text{BIC}}$.

Besides the results presented in Table ??, we also observed, with our simulation study, that: (i) DIC usually leads to overfitted models; (ii) $\widehat{\text{BIC}}$ is the criterion that leads to more parsimonious models; (iii) LSFS and LSCV favor models with less predictors than the predictors considered in the model that gave raise to the data. The results observed with LSCV can be explained by the instability of the harmonic mean used to estimate (??) (Carlin and Louis, 2000, p. 207).

TABLE 1. Frequency of correct classification.

model	δ	$n = 200$				$n = 100$			
		DIC	BIC	LSFS	LSCV	DIC	BIC	LSFS	LSCV
m_1	(1, 1, 1, 1)	100	98	44	8	96	74	21	0
m_2	(1, 1, 1, 0)	92	98	34	0	71	78	10	0
m_3	(1, 1, 0, 1)	87	97	33	0	83	67	5	0
m_4	(1, 0, 1, 1)	84	99	38	2	75	77	10	0
m_5	(0, 1, 1, 1)	87	98	28	1	84	74	7	1
m_6	(1, 1, 0, 0)	59	97	68	17	60	83	32	5
m_7	(1, 0, 1, 0)	73	97	76	26	63	88	56	14
m_8	(1, 0, 0, 1)	64	95	65	17	61	85	37	8
m_9	(0, 1, 1, 0)	71	94	59	11	72	84	44	2
m_{10}	(0, 1, 0, 1)	66	94	62	10	68	72	32	6
m_{11}	(0, 1, 0, 1)	74	98	69	13	62	80	53	10
m_{12}	(1, 0, 0, 0)	60	96	98	98	65	89	92	80
m_{13}	(0, 1, 0, 0)	49	95	95	97	49	84	85	73
m_{14}	(0, 0, 1, 0)	56	96	96	96	52	94	93	84
m_{15}	(0, 0, 0, 1)	64	94	87	93	55	86	91	81
m_{16}	(0, 0, 0, 0)	58	97	13	25	44	92	18	29

References

- Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis (Second Ed.)*. New York: Chapman and Hall.
- Draper, D. and Krnjajić, M. Bayesian model specification. Submitted. Available from <http://www.ams.ucsc.edu/draper/writings.html>.
- Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society B* **58**, 583-639.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.
- Spiegelhalter, D. J., Best, N., Carlin, B. P. and van der Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B* **56**, 501-514.

A Bayesian alternative to Indicator Kriging

R. Tolosana-Delgado¹, G. Mateu-Figueras², V.Pawlowsky-Glahn² and J.J. Egozcue³

¹ Universität Göttingen, Dept. Sedimentology and Environmental Geology, Goldschmidtstr. 3, D-37077 Göttingen, Germany, raimon.tolosana@geo.uni-goettingen.de

² Universitat de Girona, Dept. Computer Science and Applied Mathematics, Campus Montilivi, Ed. P-IV, E-17071, Girona, Spain, gloria.mateu@udg.es, vera.pawlowsky@udg.es

³ Universitat Politècnica de Catalunya, Dept. Applied Mathematics III, Campus Nord, Ed. C-2, E-08034 Barcelona, Spain, juan.jose.egozcue@upc.edu

Abstract: Indicator kriging results, obtained applying conventional linear techniques to indicator transformations of the observed data, are usually interpreted as approximations to the conditional probability distribution (cpdf) of a regionalized variable at an unsampled location. The method is widely used in spite of the fact that results are quite commonly impossible probability vectors. This contribution proposes a Bayesian alternative to estimate the discrete cpdf (i.e. a probability vector parameter), where the interpolations are interpreted as the log-likelihood of a multinomial variable, updating a Jeffreys Dirichlet prior. The Dirichlet distribution is characterized using the Aitchison measure, which is compatible with the geometry for probability vectors.

Keywords: Dirichlet distribution; simplex; compositional data.

1 Motivation

The delineation of the border of an oil field in an area shared by several owners is a complex issue, where available information is typically sparse, and its results may imply very significant differences in the way the income is distributed among owners. In our motivation example, several wells in the Lyons West oil field, Kansas (USA), are available: we know which wells intersect the oil column (i.e., are *productive*) and which are dry. The goal is the estimation, at every location, of the probability of being within or outside the oil field.

2 Classical indicator kriging

Indicator kriging (IK) is a geostatistical technique born to approximate the conditional probability density function (cpdf) at every node x of a grid, based on the correlation structure of indicator transformed data (Journel, 1983). The IK algorithm applied to our case study is as follows: (i) encode the available information in two complementary Bernoulli variables

$$J_s(x) = \begin{cases} 1 & \text{productive} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad J_f(x) = \begin{cases} 0 & \text{productive} \\ 1 & \text{otherwise} \end{cases} ; \quad (1)$$

with probability of success, p_s , the sought probability; (ii) estimate the semivariogram of $J_s(x)$ (or $J_f(x)$, being complementary) and fit a valid semivariogram model; here, the isotropic hole effect variogram has been used,

$$\gamma(h) = c_0 + c \cdot \left(1 - \cos\left(\frac{h}{\alpha r}\right) \cdot \exp\left(-\frac{h}{r}\right) \right), \tag{2}$$

with parameters $c_0 = 0.005$ (nugget), $c = 0.245$ (sill), $r = 1.2$ (range in miles), and $\alpha = 1$ (period) (Figure 1); (iii) apply simple kriging to estimate $J_s(x_0)$ at every node x_0 of a grid (Figure 2, left); (iv) interpret the obtained result as the probability that x_0 is a productive location. The major drawback of IK in this fashion is that it might yield impossible estimates, such as negative probabilities (as in Figure 2), and probabilities not summing up to one. Several methods exist to *reduce* or *correct* the occurrence of such non-sense estimations. To our knowledge, the only one that *avoids* them by construction is simplicial IK (Tolosana-Delgado, 2006). To use this alternative, one needs to define a probability of error b in the determination of failure or success at sampled locations, which is here estimated as twice the nugget effect. Then, the classical and simplicial IK estimators, denoted respectively with the vectors \mathbf{j}_0^c and \mathbf{j}_0^* , are related by

$$\mathbf{j}_0^* = [j_s^*, j_f^*] = \mathcal{C} \left(\exp \left(\ln \frac{1-b}{b} \cdot \mathbf{j}_0^c \right) \right), \tag{3}$$

where $\mathcal{C}(\cdot)$ is the closure operation, forcing its argument to sum up to one.

3 A bayesian interpretation

Instead of taking the simplicial IK results as estimates of the probability of being within the field, one might interpret them as the equivalent number of successes/failures (be within/outside the field) of a Bernoulli variable. Then, for every location, one might take a prior distribution for the pair of probabilities $\mathbf{p} = (p_s, p_f = 1 - p_s)$. This is classically a beta (i.e. 2-part Dirichlet) distribution, with a vector of positive parameters (α_1, α_2) . We choose the Jeffreys prior (characterized by $\alpha_1 = \alpha_2 = 1/2$), due to its invariance properties (Leonard and Hsu, 1999). Taking the likelihood from the simplicial IK results (Eq. 3), the posterior is also a beta distribution, with parameters $(\alpha_1 + j_s^*, \alpha_2 + j_f^*)$. From this posterior, we choose as characteristic value the expectation with respect to the Aitchison measure of the simplex (as explained in the next section). Results are included in Figure 2, showing that obtained probabilities are always valid and reasonable, given the data configuration (Figure 1, right). But not only so: we have available the full distribution at any location, thus we could obtain an uncertainty measure with its variance, or a risk assessment with some high-order quantiles. Note that the proposed methodology easily extends to more than two categories, by using the multinomial-Dirichlet distributions.

4 The Dirichlet distribution

Let S^D denote the D -part simplex, i.e. the set of D -component real vectors $\mathbf{x} \in R^D$ fulfilling $x_i \geq 0$ and $\sum x_i = 1$. The D -part Dirichlet distribution is classically

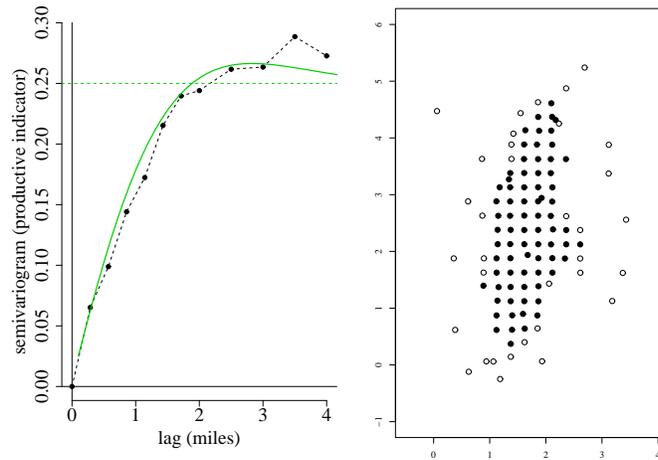


FIGURE 1. Experimental variogram, and fitted model for the indicator $J_s(x)$ (left), and data configuration (right, with filled circles corresponding to $J_s(x) = 1$ and empty ones to $J_s(x) = 0$).

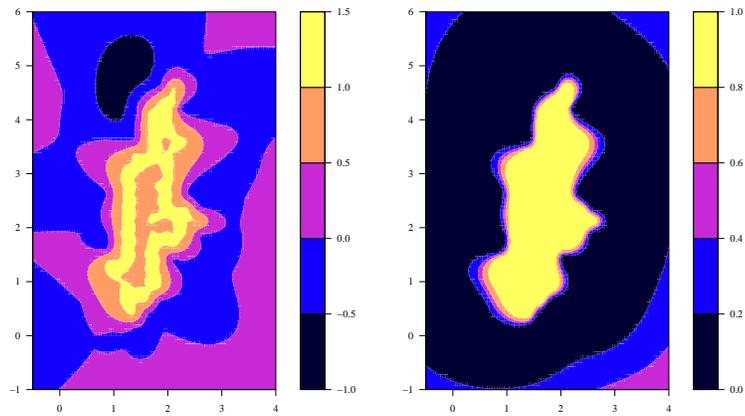


FIGURE 2. Estimates of the probability of being within the field, with classical indicator kriging (left), and with the proposed Bayesian alternative (right), using a Jeffreys prior distribution and the expectation in the simplex as representative criterion. Note the negative estimates in the left figure, which would require a correction to become valid probabilities.

defined as the closure of D independent, equally scaled Gamma distributions. If a vector $\mathbf{x} \sim \mathcal{D}i(\alpha_1, \dots, \alpha_D)$, then its density is $f(\mathbf{x}) \propto x_1^{\alpha_1-1} \dots x_D^{\alpha_D-1} I(\mathbf{x} \in S^D)$ with respect to the Lebesgue measure of R . Its mode is $\mathcal{C}(\alpha_1 - 1, \dots, \alpha_D - 1)$, and its mean $E[\mathbf{X}] = \mathcal{C}(\alpha_1, \dots, \alpha_D)$. Mateu-Figueras and Pawlowsky-Glahn (2005) give the density of the Dirichlet distribution with respect to the Aitchison measure of the simplex as $f(\mathbf{x}) \propto x_1^{\alpha_1} \dots x_D^{\alpha_D} I(\mathbf{x} \in S^D)$. Its mode is $\mathcal{C}(\alpha_1, \dots, \alpha_D)$, and its mean $E_{S^D}[\mathbf{X}] = \mathcal{C}(\exp \psi(\alpha_1), \dots, \exp \psi(\alpha_D))$, where $\psi(\cdot)$ is the digamma function. The Aitchison measure of the simplex is equivalent to the Lebesgue measure on the coordinates of the simplex, when the Aitchison geometry is considered for this sample space. The choice of which reference measure to use for the Dirichlet has in general a slight influence, *except if some of the parameters α_i are small*. This is the case of this contribution.

Acknowledgments: This research is funded by the Spanish Ministry for Science and Technology through the project MTM2006-03040.

References

- Journel, A.G. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology* **15**, 445-468.
- Leonard, T., Hsu, J.S.J. (1999). *Bayesian Methods: an analysis for statisticians and interdisciplinary researchers*. Cambridge University Press.
- Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2005). The Dirichet distribution with respect to the Aitchison measure on the simplex — a first approach. In: *Proceedings of the 2nd international workshop on compositional data analysis*. Cd-Rom.
- Tolosana-Delgado, R. (2006). *Geostatistics for constrained variables: positive data, compositions and probabilities*. PhD Thesis, Universitat de Girona.

Marginalized Semi-Parametric Shared Parameter Models for Incomplete Ordinal Responses

Roula Tsonaka¹, Dimitris Rizopoulos¹, Geert Verbeke¹ and Emmanuel Lesaffre¹

¹ Biostatistical Centre, Catholic University of Leuven, Belgium

Abstract: A new Shared Parameter Model is proposed to analyze incomplete ordinal responses subject to non-ignorable non-monotone missingness where marginal interpretation of the model parameters can be derived. Usually strong distributional assumptions are made for the shared random effects that can compromise inference when they are violated. To avoid the impact of parametric assumptions for the random effects distribution to estimates and standard errors we propose to leave this unspecified and estimate it using non-parametric maximum likelihood.

Keywords: Non monotone missingness; Vertex Exchange Method.

1 Introduction

In this work we are concerned with the analysis of longitudinal ordinal responses subject to informative non-monotone missingness. In particular, these longitudinal profiles arise from a randomized study on 895 patients suffering from Rheumatoid Arthritis who have been randomized to five treatment groups. Here we concentrate on one of the primary endpoints, which is the evaluation of the disease activity by the clinician using an ordinal scale with four categories. During a four month followup period five visits have been planned, but 48% of the patients failed to show up resulting in 19% non-monotone missingness.

Common modelling approaches for handling informative missingness are the Selection and Pattern mixture models. However, the use of such models for non-monotone missingness can prove computationally intensive due to either the high dimensional integrations involved or the large number of missingness patterns. An alternative modelling framework is the Shared parameter model (SPM), which provides a computationally convenient method for analyzing longitudinal data subject to non-monotone missingness. In SPMs a random effects component is assumed to account for the dependence between the longitudinal responses and the stochastic mechanism that describes the missingness. Such a postulation is often scientifically justifiable since e.g., in biomedical studies the patient's health status can affect both his longitudinal responses and his visiting behavior.

However, with SPMs there are mainly two issues that should be carefully considered. First, certain assumptions about the random effects distribution can possibly affect the inference. Second, when categorical responses are considered the use of a random effects model leads to a conditional interpretation of the model parameters, which is not

desirable in our study. In this work, we propose a new SPM that simultaneously deals with the above issues. In particular, to avoid the impact of parametric assumptions for the random effects distribution to parameter estimates and standard errors, we leave this distribution completely unspecified. In addition, to achieve marginal interpretation for the model parameters we extend the work of Heagerty and Zeger (2000) on marginalized multilevel models to the analysis of incomplete longitudinal ordinal responses. Thus, the proposed model produces valid likelihood-based inferences for the study population under a Missing Not At Random (MNAR) mechanism while the unspecified random effects distribution protects against model misspecification that can arise by violations of the assumptions for the random effects distribution or omitted covariates.

2 The Marginalized Shared Parameter Model

Suppose that Y_i , ($i = 1, \dots, n$) with elements y_{ij} ($j = 1, \dots, n_i$), denotes the vector of ordinal longitudinal responses for the i th individual, and R_i is the corresponding sequence of response indicators $r_{ij} \in \{0, 1\}$, with 1 denoting that y_{ij} is observed and 0 otherwise. The SPM that is proposed here assumes that a set of random effects b_i induces the association between the Y and R processes, and is formulated as

$$f(Y_i, R_i) = \int f(Y_i | b_i)f(R_i | b_i)dG(b_i), \tag{1}$$

where $b_i \sim G$ and $f(\cdot)$ denotes a probability density function. Typically, G is taken to be the normal distribution but here we make no parametric assumption for G and leave it unspecified.

To complete the definition of the marginalized SPM, the following three regression models are considered. Regarding the measurement process $f(Y_i | b_i)$, we assume that the ordinal response y_{ij} is associated with a continuous latent variable u_{ij} such that $y_{ij} = k$ if $\gamma_{k-1} < u_{ij} < \gamma_k$, where $k = 1, \dots, K$ with K the number of ordered categories and $\gamma = (\gamma_1, \dots, \gamma_{K-1})$ is a set of threshold values with $\gamma_0 = -\infty$ and $\gamma_K = \infty$. A marginal model is defined for the latent response u_{ij} by

$$u_{ij} = x_{ij}^T \beta^M + \epsilon_{ij}, \tag{2}$$

where β^M denotes the marginal regression coefficient vector, x_{ij} is the j th row of the fixed effects design matrix X_i , ϵ_{ij} are the model residuals with $\epsilon_{ij} \sim f_L$, where f_L is the standard logistic density, and the superscript T denotes the matrix transpose. This implies that the marginal probability $Pr(y_{ij} \leq k | x_{ij}) = F_L(\gamma_k - x_{ij}^T \beta^M)$ with F_L the standard logistic cdf. The second model describes the conditional on the random effects probability and is given by $Pr(y_{ij} \leq k | x_{ij}, b_i) = F_L(\Delta_{ijk} - z_{ij}^T b_i)$, where $\Delta_{ijk} = \gamma_k^C - x_{ij}^T \beta^C$ with γ_k^C and β^C denoting the conditional on the random effects thresholds and regression parameters respectively, z_{ij} the j th row of the random effects design matrix Z_i and b_i is the q -dimensional random effects vector. Finally, to link the marginal with the conditional parameters we use the deconvolution equation

$$F_L(\gamma_k - x_{ij}^T \beta^M) = \int F_L(\Delta_{ijk} - z_{ij}^T b_i)dG(b_i), \tag{3}$$

i.e., the marginal mean parameters are expressed as the expectation of the conditional mean model over the random effects. Thereby, regression parameters with a population averaged interpretation are derived.

For the missingness process R , the probability of response, $p_{ij} = Pr(r_{ij} = 1 | b_i)$, is modelled using a mixed effects logistic regression

$$\log\{p_{ij}/(1 - p_{ij})\} = w_{ij}^T \alpha + \delta z_{ij}^T b_i,$$

where w_{ij} is the j th row of the fixed effects design matrix W_i , α the regression coefficient vector and z_{ij} the j th row of Z_i . The measurement and missingness processes are linked through the random effects term and their association is quantified by the parameter vector δ .

3 Estimation

An iterative two-step procedure is proposed for obtaining the optimal set (θ, G) that maximizes the observed data log-likelihood $\ell_n(\theta, G)$ of (1) with $\theta = (\beta, \gamma, \alpha, \delta)$. In the first step, for fixed θ , $\ell_n(G)$ is maximized with respect to G , whereas in the second step for the estimated G of the first step, θ is updated.

To proceed with the estimation of G we use the theoretical results of Laird (1978) and Lindsay (1983) stating that the non-parametric maximum likelihood estimate (NPMLE) of a distribution in the class of all distributions is a discrete distribution with finite support (i.e., at most equal to the number of distinct likelihood contributions). Thus the random-effects distribution can be assumed to be a discrete distribution with unknown but finite support. However, estimating simultaneously the support points and size is susceptible to numerical difficulties due to flat likelihood surfaces. To overcome this problem we follow the proposal of Böhning (1999) who showed that an approximate NPMLE of G can be found by prespecifying a *dense* grid for its support points $\mu = (\mu_1, \mu_2, \dots)$, and $\ell_n(G)$ is then maximized in the set $\Omega_{\mathcal{M}} = \{(\pi_1, \mu_1), \dots, (\pi_C, \mu_C)\}$ where $\pi = (\pi_1, \dots, \pi_C)$ denote weights and C is large, by means of the Vertex-Exchange algorithm (VEM). The VEM is a gradient based algorithm that is built on the idea of searching at each iteration the direction, among a prespecified grid of support points, that increases the likelihood. More precisely, the VEM at each iteration exchanges weights between the support points that contribute the least and the most to the likelihood increase. These points are identified as the points that respectively minimize and maximize the gradient function that is given by

$$d(\mu, \hat{G}^*) = \frac{1}{n} \sum_{i=1}^n \frac{f(Y_i, R_i; \mu, \hat{\theta}^*)}{f(Y_i, R_i; \hat{G}^*, \hat{\theta}^*)}, \quad (4)$$

where \hat{G}^* and $\hat{\theta}^*$ denote the estimated mixing distribution and parameter vector of the previous iteration, respectively. However, in the proposed model estimating both γ_1^C and the mean of the random effects leads to identifiability difficulties. Thus, a constrained VEM step is adopted to estimate G subject to the constraint $E(G) = 0$. To achieve this we use the penalized log-likelihood, for fixed penalty λ

$$\ell^P(\theta, G) = \ell(\theta, G) - \lambda \|\phi(G) - h\|^2,$$

where $\ell(\theta, G)$ is the log-likelihood of model (1), and in our case $\phi(G) = \sum_c \pi_c \mu_c$ and $h = 0$. Accordingly the gradient function (4) takes the form

$$d(\mu, \hat{G}^*) = \frac{1}{n} \sum_{i=1}^n \frac{f(Y_i, R_i; \mu, \hat{\theta}^*)}{f(Y_i, R_i; \hat{G}^*, \hat{\theta}^*)} - \frac{2}{n} \lambda \left(\mu - \sum_c \pi_c \mu_c \right) \sum_c \pi_c \mu_c.$$

In the second step, and for the estimated G of the first step, θ is estimated using the Newton-Raphson algorithm, which requires the computation of partial derivatives of $\ell_n(\theta, G)$. These partial derivatives are easily computed for α, δ , but for γ and β^M involve an application of the chain rule, i.e.,

$$\frac{\partial \ell(\Delta_{ijk}(\theta))}{\partial \theta} = \frac{\partial \ell(\Delta_{ijk}(\theta))}{\partial \Delta_{ijk}} \cdot \frac{\partial \Delta_{ijk}}{\partial F_L(\eta_{ijk})} \cdot \frac{\partial F_L(\eta_{ijk})}{\partial \eta_{ijk}} \cdot \frac{\partial \eta_{ijk}}{\partial \theta},$$

where $\eta_{ijk} = \gamma_k - x_{ij}^T \beta^M$. In the above equation, Δ_{ijk} needs to be estimated as a function of the marginal mean parameters β^M, γ , and the random effects parameters π by solving numerically (3). These two steps are repeated until convergence. The algorithm has converged when both conditions: (1) $\max_{\mu} d(\hat{G}, \mu) < 1 + \epsilon$, which guarantees that $\ell(\hat{G} | \hat{\theta}^*) - \ell(\hat{G}^* | \hat{\theta}^*) < \epsilon$, and (2) $\ell(\hat{G}, \hat{\theta}) - \ell(\hat{G}^*, \hat{\theta}^*) < \epsilon'$ for small ϵ, ϵ' (i.e., 10^{-3} and 10^{-8} , respectively), have been satisfied.

4 Conclusion

We have extended the work of Heagerty and Zeger (2000) in the ordinal case while allowing for valid likelihood-based inferences under a MNAR mechanism. The proposed SPM effectively handles non-monotone missingness and is robust to misspecifications of the random effects component since its distribution is left unspecified. This is also corroborated by a set of simulation studies with varying distributional assumptions for the random effects. Finally, the proposed model is exemplified on the ordinal outcome of the randomized study on patients with Rheumatoid Arthritis. The scientific interest lies in estimating the marginal treatment effect while adjusting for the informative non-monotone missingness.

References

- Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*. Chapman & Hall/CRC: Monographs on Statistics and Applied Probability.
- Heagerty, P.J. and Zeger, L. S. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science* **15**, 1-26.
- Laird, N. (1978). Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association* **73**, 805-811.
- Lindsay, B. G. (1983). The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics* **11**, 86-94.

MSE of the log-risk predictor in a mixed Poisson model with spatial dependence

M.D. Ugarte¹, A.F. Militino¹ and T. Goicoa¹

¹ Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra, Campus de Arrosadia, 31006 Pamplona, Spain, lola@unavarra.es

Abstract: The use of models to borrow information from neighbouring regions smoothing the crude mortality risks is necessary in disease mapping. One of the most common models is a mixed Poisson model which incorporates spatial dependence. The main objective of this work is to estimate the MSE of the relative risk predictor obtained from this model. In particular, second-order approximations to the MSE are brought from the small area literature. The final interest is to check if second order approximations of the prediction error improve confidence intervals for the relative risks. The well-known Scottish lip cancer data is used for illustrative purposes. A simulation study is also conducted indicating that there is no much gain in efficiency when using second order corrections.

Keywords: Prediction error; disease mapping; Risks Confidence intervals

1 Introduction

Disease mapping is a small area problem and models are essential to produce reliable estimates by borrowing information from neighbouring areas. There is a huge amount of research in small area estimation, but disease mapping has been approached somewhat differently from other small area applications because sampling is not involved. More precisely, the focus is on making predictions of the relative risks, and assessing the prediction error, which is of great relevance to build confidence intervals for the relative risks. Later, one may decide whether the regions exhibit extreme risks and recover the “true” risk surface. That is, if the lower (upper) bound of the interval is greater (smaller) than 1, the region is classified as a high (low) risk region.

The estimation of the prediction errors is always a challenge in small area applications. Several authors have developed different techniques to assess the prediction error. Prasad and Rao (1990) provide an asymptotic approximation to the mean squared error (MSE) under a linear mixed model, and also obtain a second order correct estimator. Petrucci and Salvati (2006) extend the Prasad and Rao formula to applications with spatial linear models. Bootstrap estimators have been also provided (see, for example, González Manteiga *et al.*, 2007 and Hall and Maiti, 2006). Ainsworth and Dean (2006) consider a first order Empirical Bayes (EB) variance estimator and MacNab *et al.* (2004) propose a bootstrap method (bootstrap-adjusted variance estimator).

In this work, prediction error estimators, used in EB disease mapping, are compared with alternative proposals from the small area estimation literature to determine

whether second order approximations are necessary in this framework. The performance of the above cited estimators is assessed in terms of coverage probabilities of the corresponding confidence intervals for the relative risks via a simulation study. The well-known Scottish lip cancer data will be used for illustrative purposes.

2 Model and estimation technique

Let us suppose that the area under study is divided into I contiguous regions labelled $i = 1, \dots, I$. Conditional on the random region effects r_i , the number of deaths in each area, C_i , is assumed to be Poisson distributed with mean $\mu_i = e_i r_i$, where r_i represents the unknown relative risks of mortality from a rare disease, and e_i is the expected number of deaths. Then

$$C_i | r_i \sim \text{Poisson}(\mu_i = e_i r_i), \tag{1}$$

and

$$\log \mu_i = \log e_i + b_i, \quad b_i = \log r_i = \alpha + u_i. \tag{2}$$

Here, α plays the role of the logarithm of the overall risk. The vector of random effects $\mathbf{u} = (u_1, \dots, u_I)'$ is assumed to follow a multivariate normal distribution, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$, where $\mathbf{D} = \sigma^2(\lambda \mathbf{Q} + (1 - \lambda)\mathbf{I})$, the matrix \mathbf{Q} is determined by the neighbourhood structure, \mathbf{I} is the identity matrix, and the parameter λ determines the relative weight between the spatial and the unstructured variation. To estimate the model parameters, the PQL technique, popularized by Breslow and Clayton (1993), is considered.

3 Alternative measures for the prediction error

3.1 The MSE

The prediction error can be measured through the mean squared error, defined as

$$MSE[\hat{b}_i(\boldsymbol{\theta})] = E[(\hat{b}_i(\boldsymbol{\theta}) - b_i)^2] = E[(\tilde{b}_i(\boldsymbol{\theta}) - b_i)^2] + E[(\hat{b}_i(\boldsymbol{\theta}) - \tilde{b}_i(\boldsymbol{\theta}))^2], \tag{3}$$

where $\boldsymbol{\theta} = (\sigma^2, \lambda)^t$, $\tilde{b}_i(\boldsymbol{\theta})$ is the predictor of $b_i(\boldsymbol{\theta})$ assuming that the variance components are known, and $\hat{b}_i(\boldsymbol{\theta})$ is the corresponding predictor when the variance components are unknown. The first term in the right hand side of Equation (3) is easy to calculate, but the second one is usually untractable and approximations must be given. Prasad and Rao (1990) derived an approximation for this term leading to the final MSE approximation

$$MSE[\hat{b}_i(\boldsymbol{\theta})] \approx g_{1i}(\boldsymbol{\theta}) + g_{2i}(\boldsymbol{\theta}) + g_{3i}(\boldsymbol{\theta}).$$

An analytical MSE estimator Petrucci and Salvati (2006) develop a second order correct MSE estimator for spatially correlated data extending the results of Prasad and Rao (1990). They propose the following MSE estimator

$$\widehat{MSE}[\hat{b}_i] = g_{1i}(\hat{\sigma}^2, \hat{\lambda}) + g_{2i}(\hat{\sigma}^2, \hat{\lambda}) + 2g_{3i}(\hat{\sigma}^2, \hat{\lambda}) := mse^{PS}, \tag{4}$$

where the terms g_1 , g_2 and g_3 are calculated considering the spatial correlation.

A simple bootstrap MSE estimator A small area wild bootstrap approximation to the MSE (proposed by González-Manteiga *et al.*, 2007), correct up to first order, is given by

$$\widehat{MSE}^*[\hat{b}_i] = \frac{1}{J} \sum_{j=1}^J (\hat{b}_i^{*(j)} - b_i^{*(j)})^2 := mse_1, \quad (5)$$

where J is the number of bootstrap populations, and for the j th bootstrap population, $\hat{b}_i^{*(j)}$ is the predictor of $b_i^{*(j)}$, and $b_i^{*(j)}$ is the corresponding true bootstrap value.

Double bootstrap MSE estimators Hall and Maiti (2006) proposed the following double bootstrap estimators

$$mse^{H_1} = \begin{cases} mse_1 + \arctg\{I(mse_1 - mse_2)\}/I & \text{if } mse_1 \geq mse_2 \\ (mse_1)^2/[mse_1 + \arctg\{I(mse_2 - mse_1)\}/I] & \text{if } mse_1 < mse_2 \end{cases}, \quad (6)$$

$$mse^{H_2} = \begin{cases} 2mse_1 - mse_2, & \text{if } mse_1 \geq mse_2 \\ (mse_1)\exp\{-(mse_2 - mse_1)/mse_2\}, & \text{if } mse_1 < mse_2 \end{cases}. \quad (7)$$

where mse_2 is defined as

$$\widehat{MSE}^{**}[\hat{b}_i] = \frac{1}{J} \sum_{j=1}^J \frac{1}{K} \sum_{k=1}^K (\hat{b}_i^{**(jk)} - b_i^{**(jk)})^2 := mse_2. \quad (8)$$

Here, K is the number of second bootstrap populations for the j th first bootstrap population, and $\hat{b}_i^{**(jk)}$ and $b_i^{**(jk)}$ are similarly defined as $\hat{b}_i^{*(j)}$ and $b_i^{*(j)}$ but applying the bootstrap twice.

3.2 The EB variance

The EB variance is given by

$$\text{var}(b_i|\mathbf{C}) = E_{\alpha, \theta|\mathbf{C}}[\text{var}(b_i|\mathbf{C}, \alpha, \theta)] + \text{var}_{\alpha, \theta|\mathbf{C}}[E(b_i|\mathbf{C}, \alpha, \theta)], \quad (9)$$

and the PQL variance estimate $\widehat{\text{var}}(b_i|\mathbf{C}, \hat{\alpha}, \hat{\theta}) = g_{1i}(\hat{\sigma}^2, \hat{\lambda}) + g_{2i}(\hat{\sigma}^2, \hat{\lambda})$ only approximates the first term.

An analytical EB variance estimator Ainsworth and Dean (2006) consider the following first order variance estimator for the first and second terms in Equation (9)

$$\widehat{\text{var}}(b_i|\mathbf{C}) = \widehat{\text{var}}(b_i|\mathbf{C}, \hat{\alpha}, \hat{\theta}) + \mathbf{m}_i^t \left(\frac{\partial \hat{\mathbf{u}}}{\partial \hat{\theta}} \right)^t \widehat{\text{var}}(\hat{\theta}) \left(\frac{\partial \hat{\mathbf{u}}}{\partial \hat{\theta}} \right) \mathbf{m}_i := \text{var}^D, \quad (10)$$

where $\mathbf{m}_i^t = (0, \dots, \overset{i}{1}, \dots, 0)$ is a $1 \times I$ vector with 1 in the i th component and 0 elsewhere.

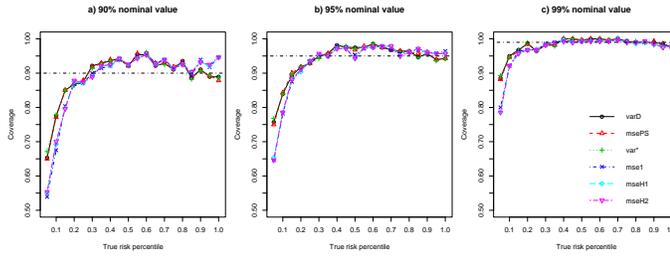


FIGURE 1. Coverage probabilities by relative risks size.

A bootstrap-adjusted EB variance estimator To estimate the EB variance, MacNab *et al.* (2004) propose a different bootstrap method. They consider the following bootstrap estimator of the EB variance (9)

$$\widehat{\text{var}}^*(\hat{b}_i) = \frac{\sum_{j=1}^J \widehat{\text{var}}(\hat{b}_i^{*(j)})}{J} + \frac{\sum_{j=1}^J (\hat{b}_i^{*(j)} - \bar{b}_i^*)^2}{J - 1}, \quad (11)$$

where, for the i th local area, $\widehat{\text{var}}(\hat{b}_i^{*(j)})$ is the PQL variance for the j th bootstrap log-risk estimate, and $\bar{b}_i^* = \sum_{j=1}^J \hat{b}_i^{*(j)} / J$. To estimate the standard error of one particular region i , a bootstrap sample set for the area consists of the observed data for the whole neighbourhood and generated samples for the rest of areas.

4 Illustration

In this section the different techniques are applied to the well known Scottish lip cancer data. Militino *et al.* (2001) showed that a full spatial autocorrelation model (Models (1) and (2) with $\lambda = 1$, so $\mathbf{D} = \sigma^2 \mathbf{Q}^-$) is appropriate. The parameters estimates (using PQL) are $\hat{\alpha} = 0.126$ (0.051) and $\hat{\sigma}^2 = 0.737$ (0.233), where the values in brackets are the standard errors. The different prediction error estimators described in Section 3 have been used to build confidence intervals for the relative risks. In general, for regions with a high estimated relative risk, the confidence intervals based on the bootstrap MSE estimators (single and double bootstrap) (5), (6) and (7) are wider than those based on the analytical expressions (4) and (10) and on the bootstrap-adjusted variance estimator given by (11). On the contrary, for regions with low estimated relative risks, the confidence intervals based on the bootstrap estimators are narrower than those based on the analytical expressions and the bootstrap-adjusted estimator.

5 Simulation study

To assess the performance of the different prediction error estimators, in terms of coverage probabilities of the corresponding confidence intervals for the relative risks, a simulation study has been conducted. $M = 100$ data sets have been generated under the spatial model given by (1) and (2) with $\lambda = 1$. For each data set we have computed the analytical prediction error estimators (4) and (10). We have also assessed the bootstrap MSE estimator (5) and the bootstrap-adjusted variance estimator (11) based on $J = 100$ bootstrap samples. To obtain the double bootstrap estimators (6) and (7), $K = 50$ bootstrap samples have been generated for each of the 100 first bootstrap replicates. The analytical and the bootstrap-adjusted estimators lead to mean coverage probabilities practically equal to the nominal value. The bootstrap (single and double) MSE estimators yield mean coverage probabilities slightly below the nominal value. However, it is interesting to display the coverage rates by true relative risk size. Figure 1 displays the 90%, 95% and 99% coverage probabilities by true relative risks size for the six estimators. It is remarkable that the coverage rates for the 25% lowest risks fall typically below the nominal rates for all the estimators. This undercoverage is more severe for the simple and double bootstrap MSE estimators mse_1 , mse^{H_1} and mse^{H_2} . On the other hand, for the 25% highest risks, the coverage rates are near the nominal values for the estimators var^D , \widehat{var}^* and mse^{PS} . However, the coverage probabilities of the confidence intervals based on the bootstrap MSE estimators mse_1 , mse^{H_1} and mse^{H_2} are above the nominal values 90% and 95%. This is problematic because a region with relative risk significantly greater than 1 can be classified as a region with relative risk equal to 1 (a false negative). From our empirical results, we highlight that, with spatially correlated data, the analytical estimators and the bootstrap-adjusted variance estimator are more appropriate than the bootstrap (simple and double) MSE estimators. Nevertheless, further research about theoretical properties of the single and double bootstrap MSE estimators for spatially correlated data seems to be needed.

Acknowledgments: We gratefully acknowledge support from the Spanish Ministry of Science and Education (project MTM2005-00511)

References

- Ainsworth L.M., Dean C.B. (2006). Approximate inference for disease mapping. *Computational Statistics & Data Analysis* **50**, 2552-2570.
- Breslow NE., Clayton DG. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.
- González-Manteiga W., Lombardía MJ., Molina I., Morales D., Santamaría L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analysis* **51**, 2720-2733.

- Hall P., Maiti T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society Series B* **68**, 221-238.
- MacNab YC., Farrel PJ., Gustafson P., Wen S. (2004). Estimation in Bayesian disease mapping. *Biometrics* **60**, 865-873.
- Militino AF., Ugarte MD., Dean CB. (2001). The use of mixture models for identifying high risks in disease mapping. *Statistics in Medicine* **20**, 2035-2049.
- Petrucci A., Salvati N. (2006). Small area estimation for spatial correlation in watershed error assessment. *Journal of Agricultural, Biological and Environmental Statistics* **11**, 169-182.
- Prasad, NGN., Rao, JNK. (1990). The Estimation of Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association* **85**, 163-171.

Fast Implementation For Mixed Effects Models with Censored Response

Florin Vaida and Lin Liu^{1 2}

¹ Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, CA 92093, USA, vaida@ucsd.edu,

² For Oral Presentation

Abstract: We propose here an EM algorithm for computing the maximum likelihood and residual maximum likelihood for linear and non-linear mixed effects models with censored response. In contrast with previous developments, this EM uses closed-form expressions for the mean and variance of a truncated multi-normal at the E-step as opposed to Monte Carlo simulation, which leads to an improvement in the speed of computation of an order of magnitude. The statistical problem is motivated by an AIDS application.

1 Introduction

Modeling censored observations using linear and non-linear mixed effects models (N/LME) occurs often in biomedical applications, in particular in pharmacokinetics (PK) and for viral load data, where left-censored observations occur below the limit of quantitation of the assay. Hughes (1999) proposed a Monte Carlo EM algorithm (MCEM) for LME with censored response (LMEC). Vaida, Fitzgerald and DeGruttola (2007) proposed a Monte Carlo EM (MCEM), and extended it to NLME with censored data (NLMEC). However, by its nature MCEM is an expensive proposition, due to the combination of Monte Carlo simulation with the iterative procedure (Ruppert, 2005). In this paper we propose an implementation of the EM algorithm for N/LMEC with greatly improved speed and precision. We show that the E-step reduces to computing the first two moments of certain truncated multinormal distributions. The needed formulas were derived by Tallis (1961) and Finney (1962). They require the multinormal CDF, for which we use the `mvtnorm` package in R (Genz, 1992). The likelihood function is easily computed as a bi-product of the E-step and is used for monitoring convergence and for model selection (AIC, likelihood ratio test). We apply the algorithm to modelling HIV viral load data occurring in a clinical study of acute and early HIV-1 infection.

2 Linear mixed effects with censored response

The EM algorithm for LMEC was proposed by Hughes (1999), with computational improvements given in Vaida *et al.* (2007). To summarize, consider the Laird-Ware model

$$y_i = X_i\beta + Z_ib_i + e_i, \quad b_i \sim N(0, \sigma^2 D), \quad e_{ij} \sim N(0, \sigma^2), \quad (1)$$

$i = 1, \dots, m$ where $e_i = (e_{i1}, \dots, e_{in_i})^\top$ and b_i, e_i are independent for all i and independent of each other. D is a positive definite matrix depending on a vector of parameters γ . Put $n = \sum_{i=1}^m n_i$, $\sigma^2 D = \Psi$ and $V_i = \text{var}(y_i) = Z_i \Psi Z_i^\top + \sigma^2 I$. The response y_{ij} is not fully observed for all i, j . Assuming left censoring, let the observed data for the i^{th} subject be (Q_i, C_i) , where Q_i represents the vector of uncensored readings or censoring levels, and C_i the vector of censoring indicators:

$$y_{ij} \leq Q_{ij} \text{ if } C_{ij} = 1; \quad y_{ij} = Q_{ij} \text{ if } C_{ij} = 0. \tag{2}$$

In the EM we update β, σ^2 with $\{y_{ij} : C_{ij} = 1\}$ as missing data, and Ψ using $\{y_{ij} : C_{ij} = 1\}$ and b_i as missing data (Vaida *et al.*, 2007). Decompose $D^{-1} = \Delta^\top \Delta$ and write: $\delta = (\beta^\top, b_1^\top, \dots, b_m^\top)^\top$, $\tilde{y} = (\tilde{y}_1^\top, \dots, \tilde{y}_m^\top)^\top$,

$$(\tilde{y}_i \quad \tilde{X}_i \quad \tilde{Z}_i) = \begin{pmatrix} y_i & X_i & Z_i \\ 0 & 0 & \Delta \end{pmatrix}, \quad \text{and} \quad M = \begin{pmatrix} \tilde{X}_1 & \tilde{Z}_1 & & \\ \vdots & & \ddots & \\ \tilde{X}_m & & & \tilde{Z}_m \end{pmatrix}. \tag{3}$$

The M-step updates are:

$$\hat{\delta} = (M^\top M)^{-1} M^\top E(\tilde{y}) \tag{4}$$

$$\hat{\sigma}^2 = \frac{1}{n} \|E(\tilde{y}) - M\hat{\delta}\|^2 + \frac{1}{n} \sum_{i=1}^m \text{tr}\{\text{var}(y_i)\} - \frac{1}{n} \sum_{i=1}^m \text{tr}\{W_i Z_i^\top \text{var}(y_i) Z_i\} \tag{5}$$

$$\hat{\Psi} = \frac{1}{m} \sum_{i=1}^m E(b_i) E(b_i)^\top + \frac{1}{m} \sum_{i=1}^m \text{var}(b_i) \tag{6}$$

where $W_i = (Z_i^\top Z_i + D^{-1})^{-1}$, $E(b_i) = W_i Z_i^\top \{E(y_i) - X_i \beta\}$, $\text{var}(b_i) = \sigma^2 W_i + W_i Z_i^\top \text{var}(y_i) Z_i W_i$, and $E(y_i)$, $\text{var}(y_i)$ are the mean and variance conditional on $\{C_i, Q_i; i = 1 \dots m\}$, taken at the current parameter value $\theta = (\beta, \sigma^2, D)$. The update for unstructured Ψ is given by (6). If Ψ is diagonal the right-hand side of (6) is replaced by the diagonal matrix with same diagonal elements as (6).

The computations use dimension reduction based on QR decomposition which take advantage of the sparse nature of the matrix M (Pinheiro and Bates, 2000). The key feature is that the number of columns of the matrices to be decomposed does not increase with the number of clusters m or the number of data points N .

From (4)-(6) it is clear that the E-step reduces to the computation of $E(y_i|Q_i, C_i, \theta)$ and $\text{var}(y_i|C_i, Q_i, \theta)$. These are determined as follows. Partition y_i into the observed and censored parts: $y_i^\top = (y_i^o{}^\top, y_i^c{}^\top)$, i.e. $C_{ij} = 0$ for all elements in y_i^o , and 1 for all elements in y_i^c ; write accordingly $Q_i = (Q_i^o{}^\top, Q_i^c{}^\top)$. Ignoring censoring for the moment, we have that marginally $y_i \sim N(X_i \beta, \Sigma = \sigma^2(I + Z_i D Z_i'))$. Then $y_i^o \sim N(X_i^o \beta, \Sigma_{oo})$, $y_i^c | y_i^o \sim N(\mu_i, S_i)$, where

$$\begin{aligned} \mu_i &= X_i^c \beta + \Sigma_{co} \Sigma_{oo}^{-1} (y_i^o - X_i^o \beta) \\ S_i &= \Sigma_{cc} - \Sigma_{co} \Sigma_{oo}^{-1} \Sigma_{oc} \end{aligned}$$

and $\Sigma = \begin{pmatrix} \Sigma_{oo} & \Sigma_{oc} \\ \Sigma_{co} & \Sigma_{cc} \end{pmatrix}$. It follows that $E(y_i|Q_i, C_i, \theta) = (y_i^o{}^\top, \mu_i^c{}^\top)^\top$, $\text{var}(y_i|Q_i, C_i, \theta) = \begin{pmatrix} 0 & 0 \\ 0 & S_i^c \end{pmatrix}$, where $\mu_i^c = E(U)$, $S_i^c = \text{var}(U)$ and $U = (y_i^c | Q_i^c, y_i^o)$

follows a multinormal distribution $N(\mu_i, S_i)$ left-truncated at Q_i^c . Let B_i be a diagonal matrix with diagonal elements the square roots of the corresponding diagonal elements in S_i . Put $X = B_i^{-1}(U - \mu_i)$. Then X has a multinormal distribution $N(0, R_i)$ left-truncated at $a_i = B_i^{-1}(Q_i^c - \mu_i)$ and $R_i = B_i^{-1}S_iB_i^{-1}$ is the correlation matrix corresponding to S_i . Then $\mu_i^c = B_iE(X) + \mu_i$, $S_i^c = B_i\text{var}(X)B_i$ and calculation of μ_i, S_i^c reduces to computing the mean and variance of X .

Formulas for $E(X)$, $\text{var}(X)$ were developed by Tallis (1961) and Finney (1962), and are available in close form, depending on the multinormal CDF. The latter is available in R through the `pmvnorm()` function from the `mvtnorm` package (Genz, 1992). (This package, in turn, uses numerical integration for the computation of the CDF.)

The variance of the MLE $\hat{\theta}$, estimated at convergence, is adjusted for the censored information using Louis' formula (Louis, 1982).

The likelihood function. Put $\Phi_n(u; A)$ and $\phi(u; A)$ be respectively the left-tail probability (component-wise) and the probability density function of the $N(0, A)$ distribution, computed at u . Let $\alpha_i = P(y_i^c < Q_i^c | y_i^o) = \Phi_{n_i^c}(a_i; R_i)$. The log-likelihood function for the observed data is given by

$$l(\theta) = \sum_{i=1}^m \{ \log \alpha_i + \log \phi_{n_i^c}(Q_i^o - X_i^o \beta; \Sigma_{oo}) \}.$$

This can be computed at each step of the EM algorithm without additional computational burden, since α_i 's are computed at the E-step, and it is used to monitor the convergence of the algorithm.

2.1 Nonlinear case

The NLME (Lindstrom and Bates, 1990) is given by $y_{ij} = f(\beta, b_i) + e_{ij}$ where $f(\beta, b_i) = f(\beta, b_i, x_{ij})$ is a non-linear function of the fixed β and random effect b_i ; x_{ij} is a vector of covariates, and b_i and e_{ij} are given by (1). The approximate MLE $(\hat{\beta}, \hat{\sigma}^2, \hat{\gamma})$ and predictors for the random effects \hat{b}_i are computed by iterative linearization (L) of the conditional mean function. The L-step yields the LME $w_i = X_i^* \beta + Z_i^* b_i + e_i$, $i = 1 \dots m$, where

$$w_i = y_i - \{ f_i^* - \left(\frac{\partial f_i^*}{\partial \beta} \right) \beta^* - \left(\frac{\partial f_i^*}{\partial b_i} \right) b_i^* \}, \tag{7}$$

$X_i^* = \frac{\partial f_i^*}{\partial \beta}$, $Z_i^* = \frac{\partial f_i^*}{\partial b_i}$, y_i is the n_i -vector dependent variable for the i^{th} subject, f_i , e_i are respectively the corresponding mean function and error n_i -vectors, and the starred terms are computed at the current parameters (β^*, b_i^*) . For censored response the linearized model is an LME with censored data, which is solved as in the previous section. More precisely, the algorithm iterates to convergence between L, E and M steps.

Comparison with MCEM. Figure 1 illustrates the convergence of the proposed algorithm in comparison with the MCEM. Notice the smooth convergence of the log-likelihood and the much improved computation time.

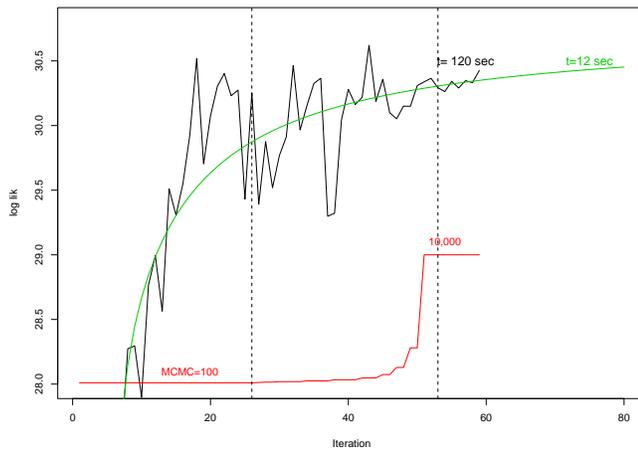


FIGURE 1. Comparison of convergence of MCEM (jagged line) and the proposed EM (smooth line). The convergence times were respectively 120 and 12 seconds in this example.

TABLE 1. Analysis of primary HIV infection. The parameters are from the random-intercept logistic model and logistic model with random intercept and random linear decrease after 50 days, respectively.

	Setpoint model		Five-parameter model	
	Estimate	SE	Estimate	SE
β_1	1.575	0.014	1.609	0.014
β_2	0.4240	0.0933	0.1441	0.0950
β_3	3.561	0.034	3.526	0.024
β_4	1.547	0.228	1.060	0.267
β_5		$-3.48 \cdot 10^{-3}$	$1.43 \cdot 10^{-3}$	
σ	0.554		0.512	
σ_{b1}	0.139		0.133	
σ_{b5}			$7.10 \cdot 10^{-3}$	
ρ_{b12}			0.17	

3 Modeling HIV-1 RNA viral load

Our application concerns 320 untreated individuals with acute HIV infection from the AIEDRP Program, a large multi-center observational study. Of the 830 recorded observations, 185 (22%) were above the limit of quantification of the assay (right-censored). In absence of treatment, following acute infection the HIV RNA decreases and then varies around a setpoint value. This setpoint value may differ between individuals, and is of central interest here. The viral setpoint characterizes the severity of infection, it may relate to the strength of the subject’s immune system and it may predict clinical

progression of the disease. Our analysis considers three models for these data, starting with a four-parameter logistic model

$$y_{ij} = \alpha_{1i} + \alpha_2[1 + \exp\{(t_{ij} - \alpha_3)/\alpha_4\}]^{-1} + e_{ij} \quad (8)$$

where y_{ij} is the \log_{10} HIV RNA for subject i at time t_{ij} . This is an inverted S-shaped curve, with the constant value for the later times representing the subject-specific setpoint. The setpoint α_{1i} was taken to be random. The results of the analysis show evidence of further viral decay after reaching the setpoint and will be presented. The parameter values are reported in Table 1. The individual profiles, together with the mean model fits are shown in Figure 2.

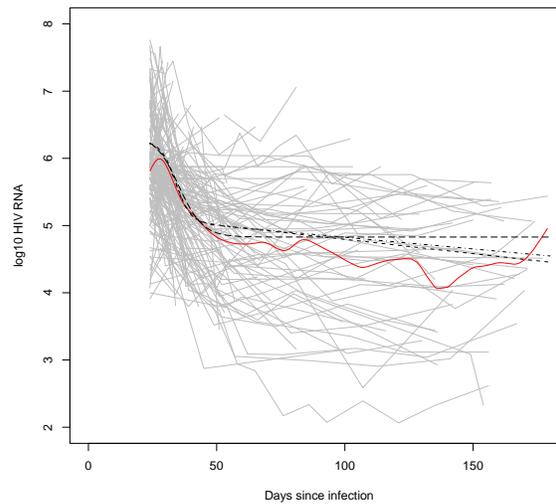


FIGURE 2. AIEDRP data and model fits from (i) random intercept logistic model (—); (ii) random intercept logistic model with linear decrease after 50 days (- · -); (iii) logistic model with random intercept and random linear decrease after 50 days (- · -). Solid line: a smooth fit of the observed data, with censored observations excluded.

References

- Finney, D. J. (1962). Cummulants of truncated multinormal distribution. *Journal of the Royal Statistical Society, Series B*, **24**, 535–536.
- Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics* **1**, 141–150.
- Hughes, J. P. (1999). Mixed effects models with censored data with applications to HIV RNA levels. *Biometrics* **55**, 625–629.

- Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–687.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *JRSS B* **44**, 226–233.
- Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New-York.
- Ruppert, D. (2005). Discussion of Maximization by parts in likelihood inference by Song, Fan, and Kalbfleisch. *JASA* **100**, 1161–1163.
- Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *JRSS B* **23**, 223–229.
- Vaida, F., Fitzgerald, A., and DeGruttola, V. (2007). Efficient hybrid EM for nonlinear mixed effects models with censored response. *Computational Statistics and Data Analysis*. To appear.

Study of ewe's milk composition using a combination of multivariate techniques and linear mixed models with random effects

O. Valero¹, A. Espinal¹, P. Jaramillo² and A. Trujillo²

¹ Servei d'Estadística, Universitat Autònoma de Barcelona

² Dpt. de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona

Abstract: Guirra ewe milk is used in the elaboration of cheese. In order to study its technological characteristics, a principal components analysis was carried out to summarize the variables of milk composition in a reduced number of factors, which will be used in a linear mixed model with random effects and repeated measures in order to compare the qualities of the milk of two breeds of ewe: Guirra and Manchega.

Keywords: principal components; mixed model; random effects; repeated measures.

1 Introduction

Guirra ewe breed is historically associated to the Comunidad Valenciana's coastal region. Its milk is usually used in the elaboration of some traditional types of cheese which have specific denominations in each region. Although information about the production and composition of the milk of the Guirra ewe breed already exists (Rodríguez et al., 2002), very little is known about the technological characteristics of this milk and the evolution of its composition and coagulation capability during lactation. Manchega ewe breed is raised for milk production in the semiarid, region of central Spain and is composed of 1.3 million adults in about 3300 herds. Its milk is mainly used in the production of "Manchego" cheese, a legally registered name and highly appreciated by consumers.

The main objective of the paper is to compare the results of the technological characteristics of Guirra ewe milk with those obtained from Manchega ewe milk.

2 Data

The study was carried out with 43 ewes, 30 Guirras and 13 Manchegas. The measured variables were: fat (F), protein (P), true protein (TP), casein (C), lactose (L), dry matter (DM), cheese yield (Y), cheese whey (W), somatic cells counts (SC), time of coagulation (TC), gel hardness (G), aggregation rate (AR) and pH. Each ewe was measured from the 9th to the 21st week of lactation, having between one and four repeated measures per ewe, a total of 113 observations. The parity effect was also taken into account, differentiating whether it was the first one or not.

3 Reduction of dimensionality

A principal components analysis was carried out to study the variables together with the objective of reducing the initial number of variables. This analysis tends to explain a set of observed variables X with a set of unobserved variables, the factors, enabling to transform the original variables into new unrelated variables and making the interpretation of the data easier.

Neither variable pH nor SC were included in the analysis due to its low correlation with the rest of the variables (lower than 0.4). The principal components were calculated with the rest of the variables (Peña, 2002). Varimax rotation was used with the objective of finding factors with more meaning when indicating a clear negative or positive association between the variables and the factors.

4 Modelization

A linear mixed model was established to compare Guirra ewe milk with that of Manchega for each of the factors defined by the factorial analysis, and for pH and SC variables. The set of covariates used were: breed, parity (first birth or subsequent) and lactation week, included as a random effect (because ewes were observed in different weeks). Repeated measures in each ewe was taken into account (Verbeke and Molenberghs, 1997). The model is:

$$Y_{ij} = X_i + Z_i b_i + \varepsilon_i \quad j = 1, \dots, 5$$

$$b_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \Sigma_i), \quad b_1, \dots, b_N, \quad \varepsilon_1, \dots, \varepsilon_N \quad \text{independent}$$

where Y_{ij} is the n_i dimensional response vector for subject i , X_i and Z_i are $(n_i \times 2)$ and $(n_i \times 1)$ dimensional matrices of known covariates (breed and parity, and lactation week), β is the 2 dimensional vector containing the fixed effects, b_i is the 1 dimensional vector containing the random effects, and ε_i is the n_i dimensional vector of residual components. Σ was chosen as an unstructured matrix of dimension n_i .

5 Results

The multivariate analysis reduced the initial variables into three components, which provides the optimal linear prediction of the initial variables' set. With these components, 80.9% of the initial variability was explained. Table 1 shows the correlations between the variables used in the factorial analysis and the three obtained factors, and manages to interpret these components (correlations lower than 0.4 were suppressed and varimax rotation was used to make the interpretation easier).

The first factor is highly correlated with fat, lactose (negatively), dry matter and cheese yield, so it can be interpreted as milk's performance. The second factor is correlated with cheese whey, time of coagulation (negatively), gel hardness and aggregation rate,

TABLE 1. Correlations between rotated components and factors.

	F	P	TP	C	L	DM	Y	W	TC	G	AR
F1	0.95				-0.62	0.87	0.90				
F2								0.88	-0.81	0.84	0.89
F3	0.85	0.85	0.66								

so it summarizes the coagulation. The third and last factor has a high correlation with both protein variables and with casein, so it represents milk's quality.

Using these three factors in the models we obtained that first factor, which refers to milk's performance, depends on parity and lactation week: the performance was smaller for the first ewe's parity, and it increased throughout time (Figure 1). For milk coagulation the result obtained is that Guirra milk is higher than Manchega's, and it decreases with the weeks (Figure 2). The third model showed that Guirra milk quality is higher than Manchega, and that it rises throughout the weeks (Figure 3). For the two response variables pH and SC, week and breed were respectively significant. Then the number of somatic cells was higher in Guirra milk than in Manchega's.

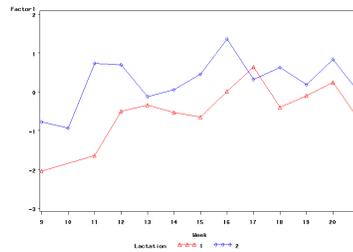


FIGURE 1. Caption text BELOW the figure.

6 Conclusions

Through the combination of multivariate analysis with linear regression with repeated measures the reduction of initial dimensionality into three factors was achieved, making the final analysis of comparison between Guirra milk and Manchega's easier, and finding statistical differences which were not found in the individual analyzes: Guirra ewe milk has better qualities than Manchega ewe milk.

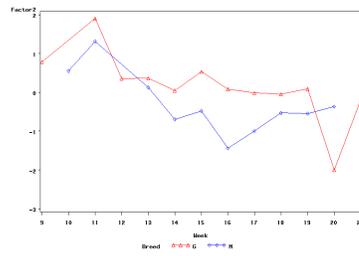


FIGURE 2. Caption text BELOW the figure.

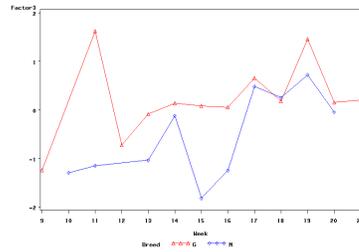


FIGURE 3. Caption text BELOW the figure.

References

Peña, D. (2002). *Análisis de datos multivariantes*. Mc Graw Hill.

Rodríguez, M., Hidalgo, M., Althaus, L., Molina, P., Peris, C. & Fernández, N. (2002). Primeros resultados de producción y composición de leche de oveja Guirra In: *Libro de actas de las XXVI jornadas científicas y VI jornadas internacionales de la sociedad española de ovinotecnia y caprinotecnia*. (904-909).

Verbeke, G., and Molenberghs, G. (1997). *Linear mixed models in practice*. Springer.

Pairwise likelihood inference in dynamic models for longitudinal ordinal outcomes

Cristiano Varin¹ and Claudia Czado²

¹ University Ca' Foscari - Venice, Italy

² University of Technology, Munich, Germany

Abstract: Longitudinal data with ordinal outcomes routinely appear in medical applications. An example is the analysis of clinical diaries where patients are asked to score the severity of their symptoms. In this contribution, we consider pseudolikelihood inference for dynamic ordinal models motivated by a study on the dependence between migraine and weather characteristics.

Keywords: longitudinal data, ordinal outcomes, pairwise likelihood.

1 Migraine data

This work is motivated by the desire to evaluate if the migraine is sensitive to weather conditions. Studies reported in the medical literature lead to contradictory conclusions on possible relationships (Prince et al., 2004). Here, we analyze a longitudinal study conducted by the psychologist T.Kostecki-Dillon on clinical records of 120 Canadian migraine sufferers who enrolled in a non-drug program on how to cope with pain. Patients were asked to keep a diary where to score four times per day - morning, noon, afternoon and bedtime - their migraine severity on a ordinal scale. This involves six categories ranging from absence of symptoms to the most invalidating headache. The data set is characterized by different and irregular observation periods for each patient, which recorded his symptoms from few days up to some months. Weather conditions have been independently collected from the closest weather station. They include sunshine, humidity, wind direction and speed, pressure, precipitation levels, and many others. A generous list of personal and clinical covariates is available as well. There is also information regarding the kind of headache, *e.g.* with or without aura or tensive migraine. It is plausible that different dynamic patterns are associated with different kind of headache, as well as with the sex of the patient. The data involves a total amount of 16,473 observations, 1,262 of which are missing, and 50 covariates.

2 Dynamic models for ordinal outcomes

Most studies on migraine severity diaries in the medical literature ignore correlation among measurements on the same patient, see Prince et alt. (2004) and references therein. This seems unmotivated. A more appropriate approach is given by generalized estimating equations (Piorecky et alt., 1996). Alternatively, one may consider

likelihood based regression time series. The benefit from a complete statistical model specification is the usage of standard model comparison techniques and forecasts in dynamic models. Unfortunately, dynamic extensions of classical cumulative probit models (*e.g.* Liu and Agresti, 2005), as the autoregressive ordinal probit model of Czado and Müller, (2005), appear impossible to handle because of the large dimensional intractable integrals involved.

Motivated by the increasing interest around composite likelihood methods (*e.g.* Molenberghs and Verbeke (2005), Varin and Vidoni, 2005), we propose instead a pairwise modelling approach, where only models for bivariate margins are specified, leaving the joint distribution of the data unspecified. This has evident advantages in terms of robustness against model assumptions, since the specification of bivariate aspects of the data may be driven by exploratory tools, differently from a difficult to assess full model construction.

Our model assumes that a pairs of ordinal outcomes is obtained by clipping a bivariate hidden Gaussian variable. The margins of the proposed model correspond to the cumulative ordinal probit model, while the correlation follow an autoregressive process of order one. The model includes a patient-specific random effect for modelling heterogeneity factors not described by the covariates, as well as the possibility to include different autoregressive coefficients depending on different groups of patients. This seems appropriate since the persistence of symptoms varies accordingly to the different types of headache and the gender.

Parameters estimates are obtained through the pairwise likelihood of order m (Varin and Vidoni, 2006). This composite likelihood is constructed from the marginal probabilities of observed pairs of outcomes far apart not more than m units. In fact, several simulations contained in Varin and Vidoni (2006, 2007) suggest that the inclusion in the pseudolikelihood of pairs formed from excessively distant observations not only increase the computational cost, but may also reduce the statistical efficiency.

Another advantage from our likelihood-type strategy is the availability of formal model selection criteria similar to the standard Akaike information criterion (Varin and Vidoni, 2005).

3 Results

Our analysis suggest that weather conditions have a very small effect on severity. Neither Seasons and the non-drug program on how to cope with pain seem having any effect. We find instead that education, job and the clinical story have an impact on the migraine. Our model suggests also the presence of different dynamics patterns for headache type and gender.

Acknowledgments: The authors would like to thank T. Kostecki-Dillon for providing the migraine data.

References

- Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: an overview and a survey of recent developments (with discussion). *Test* **14**, 1–73.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete longitudinal Data*. Springer Series in Statistics.
- Müller, G. and Czado, C. (2005). An autoregressive ordered probit model with application to high frequency financial data. *Journal of Computational and Graphical Statistics* **14**, 320–338.
- Piorecky, J., W. Becker, and M. Rose (1996). The effect of chinook winds on the probability of migraine headache occurrence. *Headache* **37**, 153–158.
- Prince, P., A. Rapoport, F. Sheftell, S. Tepper, and M. Bigal (2004). The effect of weather on headache. *Headache* **44**, 596–602.
- Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519–528.
- Varin, C. and Vidoni, P. (2006). Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics and Data Analysis*, **51**, 2365–2373.
- Varin, C. and Vidoni, P. (2007). Pairwise likelihood inference for general state space models. *Econometric Reviews*. To appear.

Auxiliary Mixture Sampling for Non-normal data

Helga Wagner¹, Regina Tüchler² and Sylvia Frühwirth-Schnatter¹

¹ Department of Applied Statistics (IFAS), Johannes-Kepler-University Linz, Altenbergerstrasse 69, A-4040 Linz, Austria, helga.wagner@jku.at (corresponding author), sylvia.fruehwirth-schnatter@jku.at

² Department of Statistics and Mathematics, University of Economics and Business Administration, Vienna, Augasse 2-6, A-1090 Vienna, Austria, regina.tuechler@wu-wien.ac.at

Abstract: We present a new MCMC method which allows Bayesian estimation of several generalized linear models, in particular models for discrete observations as Poisson counts, binary and multinomial data and for survival data. The method is based on two steps of data augmentation, where the goal of the first step is to eliminate non-linearity of the original model. In the second step the density of the error term in the resulting linear model is approximated by a mixture of normal distributions, and the component indicators of these mixtures are introduced as further latent variables. This leads to a convenient auxiliary mixture sampler requiring only draws from standard distributions like normal or exponential distributions and, in contrast to Metropolis-Hastings approaches, needs no tuning. The method is particularly useful for models with latent variables, e.g. random effects models, state space models and models involving variable selection as it renders multi-move-sampling of all effects feasible. Application of the method is illustrated by a data set from road safety.

Keywords: MCMC, generalized linear models, finite mixture approximation, discrete data, random effects model, state space model, variable selection

1 Introduction

Applied statisticians often have to deal with non-normal data, e.g. count data, binary or multinomial variables or survival data. The effect of exogenous variables on non-normal response variables is usually modelled using a generalized linear model. To account for dependency which occurs e.g. in repeated measurements, longitudinal, spatial or time series data, parameter-driven models can be used where correlation is modelled via a latent process. This latent process might be independent random effects as in generalized linear mixed models or a correlated process as in state space models. Estimation of parameter-driven models for non-normal data is a challenging problem as in contrast to the respective models for normal data, the marginal likelihood where the latent process is integrated out, often is not available in closed form.

In this paper we present an auxiliary mixture sampler for Bayesian estimation of generalized linear models for several types of non-normal data. Extensions to generalized linear mixed effects models with or without variable selection as well as to generalized linear state space models are easily included in the modelling framework.

The method is based on two steps of data augmentation: In the first step non-linearity of the original model is eliminated. The density of the error term in the resulting linear model is approximated by a mixture of normal distributions, and the component indicators of this mixture are introduced as further latent variables. Thus the original model can be represented as an equivalent partially Gaussian model where straightforward Gibbs sampling of all effects is possible.

The rest of the paper is organized as follows. We start by defining the regression models for which our sampler can be applied in section 2 and then describe the data augmentation steps and the sampling scheme for fixed effects models in sections 3-5. More complex model specifications are discussed in section 6. Finally an application of the auxiliary mixture sampler using a Poisson state space model is presented in section 7.

2 Regression Models for Non-normal Data

The auxiliary mixture sampling method presented in this paper can be applied to generalized linear regression models for Poisson, binary and exponential data, $y = (y_1, \dots, y_n)$. Data y_i follow one of these distributions with parameter μ_i . A link function links μ_i to a linear predictor λ_i . For the moment let us assume a standard regression model with covariates X_i and model parameters α so that we obtain the linear predictor

$$\lambda_i = X_i\alpha.$$

For Poisson and exponential data we use the log-link-function

$$\mu_i = \exp(X_i\alpha),$$

whereas for binary data we consider the logit regression model

$$\mu_i = \frac{\exp(X_i\alpha)}{1 + \exp(X_i\alpha)}.$$

The regression effects are assumed to have a normal prior distribution.

3 Auxiliary Mixture Sampling for Poisson Data

3.1 Data Augmentation Steps

Poisson count data y_i can be regarded as the number of jumps of an unobserved Poisson process with intensity μ_i , in the time interval $[0,1]$. In the first step of our data augmentation scheme we introduce, for each observation y_i , the inter-arrival times τ_{ij} , $j = 1, \dots, y_i + 1$ of this Poisson process as missing data. As these inter-arrival times are distributed as $Exp(\mu_i)$, the original Poisson regression model can be transformed into the linear model

$$-\log \tau_{ij} = X_i\alpha + \varepsilon_{ij},$$

where the error distribution is a type I extreme value distribution. To obtain a model that is conditionally Gaussian, this non-normal density can be approximated by a mixture of ten normal components with densities $f_N(\varepsilon; m_r, s_r^2)$ as in Frühwirth-Schnatter and Wagner (2006):

$$p_\varepsilon(\varepsilon) = \exp(-\varepsilon - e^{-\varepsilon}) \approx q_{R,\varepsilon}(\varepsilon) = \sum_{r=1}^{10} w_r f_N(\varepsilon; m_r, s_r^2).$$

Since the parameters for the means m_r , the variances s_r^2 and the weights w_r are fixed we only have to introduce the component indicators r_i for each observation in the second data augmentation step. Thus the non-normal, nonlinear Poisson regression model reduces to the linear, Gaussian model

$$-\log \tau_{ij} = X_i \alpha + m_{r_i} + \varepsilon_{r_i}, \quad \varepsilon_{r_i} \sim N(0, s_{r_i}^2).$$

3.2 Auxiliary Mixture Sampling Steps

The Gibbs type sampling scheme for Poisson regression models involves the following steps:

- (I) Linear predictor: sample α from the multivariate normal posterior $p(\alpha|\tau, R)$, where $R = (r_1, \dots, r_n)$, $\tau = (\tau_{11}, \dots, \tau_{n, y_n+1})$
- (II) Latent variables coming from first data augmentation step: sample τ conditional on the data y and α .
- (III) Indicators coming from second data augmentation step: sample R conditional on τ and α from the respective discrete distribution with 10 categories.

Note that step (I) and (III) will be identical for all types of non-normal data that can be dealt with the auxiliary mixture sampler. Step (II) is specific to the respective data distribution, however for all types of data considered here, the full conditional distribution of the auxiliary variables from the first data augmentation step will only depend on the data y and the model parameters α but not on the component indicators R .

To sample the interarrival times τ from the conditional distribution $p(\tau|y, \alpha)$ we interpret each count y_i as the number of jumps of a separate Poisson process in the unit time interval and use well-known properties of the Poisson process. Given there are y_i jumps in the unit time interval, the waiting times are distributed as the order statistics of y_i random variables with uniform $U_{[0,1]}$ -distribution. Sampling the interarrival times $\tau_{i1}, \dots, \tau_{i, y_i}$ therefore is accomplished by drawing y_i $U_{[0,1]}$ - random variables, sorting them and building the differences between two subsequent order statistics.

Given y_i jumps in the time interval $[0, 1]$, the waiting time to the $y_i + 1$ -th jump is greater than 1, which implies that conditional on $\tau_{i1}, \dots, \tau_{i, y_i}$, the last inter-arrival time τ_{i, y_i+1} is truncated at $1 - \sum_{j=1}^{y_i} \tau_{ij}$ and thus is drawn from a truncated exponential distribution with mean $1/\mu_i$.

4 Auxiliary Mixture Sampling for Logit Models

For binary data $y_i \in \{0, 1\}$ auxiliary mixture sampling is based on the interpretation of logit models in terms of utilities: Let u_i^0 be the utility of choosing category 0 and u_i be the utility of choosing category 1, which is modelled as depending on covariates X_i as

$$u_i = X_i\alpha + \varepsilon_i.$$

Then category 1 is observed, i.e. $y_i = 1$, iff $u_i > u_i^0$, otherwise $y_i = 0$. The binary logit regression model results as the marginal distribution of y_i , if u_i^0 and ε_i follow a type I extreme value distribution. Thus in our first data augmentation step we introduce the latent utilities as missing variables. Introducing the component indicators r_i for each ε_i as in the Poisson regression model leads to a normal regression model for the latent utilities with heteroscedastic errors but known variances

$$u_i = X_i\alpha + m_{r_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, s_{r_i}^2).$$

We obtain a similar Gibbs type sampling scheme as in Section 3.2, where in step (II) the latent utilities u_i have to be sampled. This however requires only draws V_i and W_i from the standard exponential distribution as conditional on $\exp(X_i\alpha)$ and y_i , the latent utility u_i is given as

$$u_i = -\log\left(\frac{V_i}{1 + \exp(X_i\alpha)} + \frac{W_i}{\exp(X_i\alpha)} I_{\{y_i=0\}}\right)$$

Extension to multinomial logit models, where y_i take a value in one of $m+1$ unordered categories is straightforward. For each observation y_i , however $m+1$ latent utilities $(u_i^0, u_i^1, \dots, u_i^m)$ have to be introduced as missing data, see Frühwirth-Schnatter and Frühwirth (2006).

5 Auxiliary Mixture Sampling for Exponential Survival Data

Obviously the auxiliary mixture sample can also be applied to analyse regression models for exponential survival data. Actually for completely observed exponential survival times, i.e. $y_i \sim \text{Exp}(\mu_i)$ only the second data augmentation step is necessary to obtain the normal regression model

$$-\log(y_i) = X_i\alpha + m_{r_i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, s_{r_i}^2).$$

This means that only steps (I) and (III) in the Gibbs sampling scheme of Section 3.2 have to be carried out.

Usually however survival times are not completely observed but are subject to right censoring. For right censored observations therefore in a first data augmentation step the unobserved complete survival times are introduced as missing data. Due to the no-memory property of the exponential distribution, unobserved survival times τ_i for censored observations y_i , can be generated as

$$\tau_i = y_i + \xi_i$$

where $\xi \sim \text{Exp}(\mu_i)$, see Wagner (2006). In this case step (II) of the Gibbs sampling scheme again only requires draws from the exponential distribution. Extensions to other types of missing information, e.g. interval-censored data are straightforward.

6 Extensions to More Complex Models

The data augmentation scheme for standard regression models is easily extended to deal with more complex models: In Frühwirth-Schnatter and Frühwirth (2006) random effects are included in logit models, where

$$\lambda_i = X_i\alpha + Z_i\beta_i \quad \beta_i \sim N(0, Q). \quad (1)$$

In Tüchler (2007) a stochastic search variable approach is introduced into a logit model with linear predictor (1). This allows us to select a subset of variables from X_i . Furthermore covariance selection yields a sparse structure for Q and makes it possible to determine fixed versus random effects.

A linear predictor with state space form for time series observations where parameters are allowed to change over time is included for Poisson data in Frühwirth-Schnatter and Wagner (2006):

$$\begin{aligned} \lambda_i &= X_i\alpha + Z_i\beta_i \\ \beta_i &= \beta_{i-1} + \omega_i, \quad \omega_i \sim N(0, Q), \quad \beta_0 \sim N(b_0, B_0). \end{aligned}$$

As – conditional on the latent variables introduced in the data augmentation steps – we are dealing with a linear Gaussian model, multi-move sampling of all effects α and $\beta_i, i = 1, \dots, n$ can be accomplished as for the corresponding *linear Gaussian* random effects or state space model. One further step has to be added in the Gibbs sampling scheme if the covariance matrix Q of the random effects β_i respectively the innovations ω_i is treated as a parameter. Using a conjugate inverted Wishart prior for Q this requires only draws from an inverted Wishart distribution.

In general, our data augmentation scheme allows to estimate models for non-Gaussian data for any form of the linear predictor where Gibbs sampling for the equivalent model with Gaussian errors is feasible.

7 Application to Road Safety Data

Auxiliary mixture sampling was applied to analyze a time series of monthly counts of killed or injured children aged 6-10 in Linz from 1987-2005. A new law intended to increase road safety came into force in Austria on October 1, 1994, since when pedestrians who want to use a pedestrian crossing have to be allowed to cross. To analyze the effect of this law on the risk μ_t of being killed or seriously injured as a child living in Linz state space modelling as in Harvey and Durbin (1986) seems quite natural but as counts are small, not exceeding 5, an analysis using normal state space models is clearly inappropriate. We fitted a basic structural model including an immediate intervention effect and found a pronounced decrease for the children's risk after the change in law. The seasonal pattern turned out to be constant over time with significantly lower risk than the annual average in the holiday months of July and August and higher in June and October.

References

- Frühwirth-Schnatter, S. and Frühwirth, R. (2006). Auxiliary Mixture Sampling with Applications to Logistic Models. To appear in *Computational Statistics and Data Analysis*.
- Frühwirth-Schnatter, S. and Wagner, H. (2006): Auxiliary Mixture Sampling for Parameter-driven Models of Time Series of Small Counts with Applications to State Space Modelling. *Biometrika* **93**, 827-841.
- Harvey, A. C. and Durbin, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling, *Journal of the Royal Statistical Society, Series A* **149**, 187-227
- Tüchler, R. (2007). Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling. *Research Report Series, WU Vienna* , <http://epub.wu-wien.ac.at>.
- Wagner, H. (2006). Auxiliary Mixture Sampling for Dynamic Survival Models. *IFAS Research Report*, <http://www.ifas.jku.at>.

Use of functional data analysis and longitudinal latent class analysis to investigate the developmental origins of disease

Robert M. West and Mark S. Gilthorpe

¹ Biostatistics Unit, University of Leeds, LS2 9LN, UK

Keywords: Functional data analysis; Longitudinal latent class analysis; Developmental origins of health and disease.

1 Introduction

Longitudinal data pose substantive challenges to linear regression analyses (De Stavola et al., 2006). A common problem is that different study participants are measured at different intervals or on a different number of occasions. Also, the repeated measurements on each study participant are often highly correlated. Traditional approaches often involve excluding participants or summarizing data across time periods, thereby reducing statistical power. Multilevel modelling accommodates the clustering of measurements within individuals, but cannot address large serial correlation of observations as covariates: collinearity.

One potential solution to these problems – one that not only recognizes the serial correlation involved but actually requires this – is ‘functional data analysis’. This approach reduces the multiple measurements made on any given individual to a single temporal ‘function’ observed over time. Analyses are then conducted on these observational ‘functions’ rather than the separate measurements involved (Ramsay and Silverman, 1997). Functional data analysis is not however common for longitudinal data (De Stavola et al., 2006), and the aim here is to demonstrate how it might address the challenges posed by such data using, as an example, a ‘lifecourse’ analysis exploring the developmental origins of health and disease (DOHaD).

Research into DOHaD evolved from two complementary strands of research. The first focused on the foetal origins of adult disease (FOAD) and drew on evidence linking growth retardation in utero with chronic diseases in later life (Barker et al., 1989). The second sought to explore the independent, additive and/or interactive effects of events throughout the lifecourse on subsequent variation in morbidity (Kuh and Ben-Shlomo, 1997). Combined, these two strands aim to establish the relative importance of different types of events (genetic, ontogenetic and pathological) at different periods of the lifecourse (prenatal, childhood, adolescence, adulthood and old age) and thereby identify ‘critical periods’ when therapeutic or prophylactic interventions are best targeted.

1.1 Example data

To demonstrate how functional data analysis and longitudinal latent class analysis might be combined, the following analyses draw on a data set deposited in the public domain for the 3rd International Congress on the developmental origins of Health and Disease. This data set was based on a sample of 1000 30-year old men and provided measurements of birth weight (kg), current weight (kg) and current height (m) – subsequently used to calculate current body mass index (BMI: kg/m^2). Additional postnatal measurements of body weight had been recorded for each participant up until the age of 10, and data were also provided on ‘2-hour glucose concentration in mmol/l ’ after ‘a standard glucose tolerance test’.

2 Methods

2.1 Functional data analysis

The package ‘fda’ in the statistical package ‘R’ (see R Development Core Team, 2005) was used to fit fourth-order B splines to the participants’ weight measurements, with knots placed at each measurement and a penalty term imposed on the second derivative to ensure that the curves produced were smooth. As the first derivative of the function was to be modelled, the smoothness penalty was applied to the second derivative.

The first derivatives of the weight curves (that is the slopes or rate of weight gain) were then calculated at intervals of 250 days from birth onwards. This provided interpolated data on the rate of weight gain at regular spacings so that the fda package could be re-applied and its full range of techniques exploited. One of the most powerful is functional principal component analysis, see Ramsay and Silverman (1997). The impact of the first, second and third functional principal components on the mean function was then summarized using graphs of weight gain against age. These graphs were used to identify key periods during childhood when each of the principal components had their strongest impact on the variance in weight gain.

2.2 Longitudinal latent class analysis

Longitudinal analysis will now proceed with data only from these key periods, the important values being interpolated childhood weights and fitted weight derivatives. An important step in the clinical analysis of this population is to identify phenotypes, especially any phenotype that is susceptible to disease later in life.

The values at key periods cannot in general be considered to be uncorrelated, but since they were derived from functional PCA stronger correlations were avoided. Thus collinearity was reduced.

2.3 Regression on latent classes

The outcome from the glucose tolerance test (GTT) is a continuous variable. Rather than regress upon the values at key periods, regression is performed simply on the latent classification weighted by probabilistic assignment. This identifies those latent classes/potential phenotypes for which elevated GTT identify insulin resistance or diabetes.

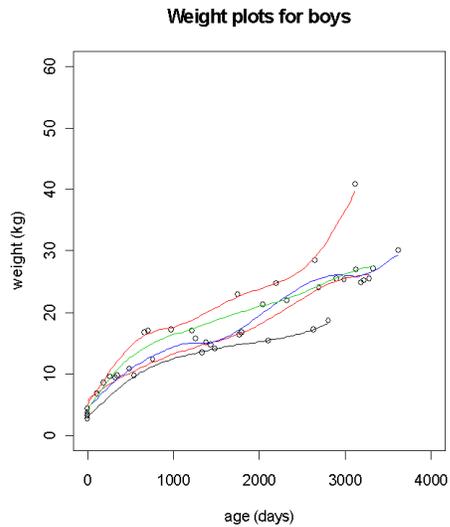


FIGURE 1. Fitting smooth functions to a sample of boys weight trajectories.

3 Results

Figure 1 shows the fitting of smooth functions to a sample of boys weight trajectories. Note the variable number of measurements. There were a number of boys for which measurements beyond 2500 days were not available. Rather than extrapolate the data or reduce the number of boys used in the analysis, fitting was only considered up to 2500 days (around age 7 years).

Some of the boys had very few measurements and again a pragmatic decision was needed as to inclusion criteria. It was decided that weight measurements were required at least at three time points which restricted the dataset from 100 boys to 861. Data inspection revealed no obvious bias in applying this criterion.

From functional PCA on the weight derivative, the key periods were identified as 0 (birth), 1250, 2250, and 2500 days. In addition the weight height and BMI data at age 30 years was also considered. These are currently being analysed with latent-class software (Latent GOLD and Mplus). Initial results reveal interesting classes: those with small and large birth weights, but more importantly those who gain weight rapidly later in childhood (2500 days). Regression on probabilistic latent class membership reveals those classes for whom the individuals have raised GTT outcomes and so are more likely to suffer insulin resistance or diabetes.

4 Conclusions

The dataset analysed is typical in lifecourse epidemiology and raises key issues. This work demonstrated the benefits of combining FDA with LLCA: variable numbers of measurements made at differing time points were accommodated, colinearity issues were reduced and important latent classes/potential phenotypes were identified.

References

- Barker D.J.P, Osmond C., Golding J., Kuh D. , and Wadsworth M.E.J. (1989) Growth in utero, blood pressure in childhood and adult life, and mortality from cardiovascular disease. *British Medical Journal* **298**, 564–7.
- De Stavola B.L., Nitsch D., Silva I.D., McCormack V., Hardy R., Mann V., Cole T.J., Morton S. and Leon D.A. (2006) Statistical issues in life course epidemiology. *American Journal of Epidemiology* **163**, 84–96.
- Kuh D. and Ben-Shlomo Y. (1997) Chapter 1: Introduction – a life course approach to the aetiology of adult chronic disease. In Kuh DA and Ben-Shlomo Y (Eds). *A life course approach to chronic disease epidemiology*. Oxford, Oxford University Press.
- R Development Core Team. (2005) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL , <http://www.R-project.org> [8 February 2007].
- Ramsay J.O. and Silverman B.W. (1997) *Functional data analysis*, Springer, New York.

A Hybrid Test for Non-Nested Models

Paul Wilson¹

¹ Department of Mathematics, National University of Ireland, Galway

Abstract: Compared to tests for nested models, little attention has been given to methods that test the log-likelihood ratios of non-nested models. We outline two such methods, Cox's and Vuong's, highlighting the advantages and disadvantages of both. We propose a hybrid test that combines the advantages both methods, without their disadvantages.

Keywords: Non-nested models; Cox's test; Vuong's test; model discrimination.

1 Introduction

Standard statistical theory provides us with a range of tools for choosing between nested models. However, in many practical data analysis problems we wish to choose between non-nested models, i.e. models where neither model is a special case of the other. The problem of choosing between non-nested models arises in many areas of scientific research. Recent examples include, in environmental science, Dobbie and Welsh (2001); in agricultural science, Allcroft and Glasby (2003); and in political science, Smith (1999).

This problem was first considered by Cox (1961,1962), who developed an analytic test; a further analytic test was later considered by Vuong (1989). Williams (1970) and Hinde (1992) have proposed simulation based alternatives to Cox's approach.

2 Cox's Method

Let M_f and M_g be non-nested models for observations Y_t , $t = 1, 2, \dots, n$, conditional on covariates X_t and Z_t , and with parameters θ and γ , respectively. Let $\hat{\theta}$ and $\hat{\gamma}$ be the maximum likelihood estimators of θ and γ , and let $\hat{\gamma}_\theta$ be the maximum likelihood estimator of γ if, in fact, M_f is the correct model. Let $LR_n(\hat{\theta}_n, \hat{\gamma}_n)$ be the log-likelihood ratio: $\sum_{t=1}^n \log \frac{f(Y_t|X_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}$.

Cox's test is based on the statistic:

$$T_f = \left\{ LR_n(\hat{\theta}_n, \hat{\gamma}_n) - E_f \left(LR_n(\hat{\theta}_n, \hat{\gamma}_{\theta_n}) \right) \right\} \quad (1)$$

i.e. the difference between the observed and the expected values of the log-likelihood ratio of the data where the null hypothesis, H_f , is that M_f is the true model, as opposed to M_g .

Cox (1962) shows that under H_f , T_f is asymptotically normally distributed with approximate mean zero and variance

$$n \left\{ V_f \left(\log \frac{f(Y_t|X_t; \hat{\theta}_n)}{g(Y_t|Z_t; \hat{\gamma}_n)} \right) - \frac{C_f^2 \left(\log \frac{f(Y_t|X_t; \hat{\theta}_n)}{g(Y_t|Z_t; \hat{\gamma}_n)}, \frac{\partial}{\partial \theta} \log f(Y_t|X_t; \hat{\theta}_n) \right)}{V_f \left(\frac{\partial}{\partial \theta} \log f(Y_t|X_t; \hat{\theta}_n) \right)} \right\} \quad (2)$$

where V_f is the expected value of the variance under H_f , C_f is the expected value of the covariance under H_f , and $\frac{\partial}{\partial \theta} \log f(Y_t|X_t; \hat{\theta}_n)$ is the score statistic under H_f . We may therefore calculate the p -value associated with a given value of T_f , “small” p -values indicating rejection of T_f .

Similarly, reversing the roles of f and g and θ and γ in the above we may calculate the p -value associated with a given value of T_g . Combining these two results we obtain the range of conclusions summarised by Table 1.

TABLE 1. Possible outcomes of Cox’s test

		$H_0 : M_f$ is the true model		
		<i>small</i>	<i>medium</i>	<i>large</i>
p -value	<i>small</i>	Neither	M_f	Neither
	$H_0 : M_g$	<i>medium</i>	M_g	Both
	<i>large</i>	Neither	M_f	–

Cox’s method may be performed both analytically or by simulation. Analytic evaluation can be complicated, while simulation requires the refitting of the model for each resample and this can be very time-consuming.

3 Vuong’s Test

Vuong’s test considers the null hypothesis:

$$H_0 : E \left[LR_n(\hat{\theta}_n, \hat{\gamma}_n) \right] = 0 \quad (3)$$

i.e. that the expected value of the log-likelihood ratio under H_0 is zero, (and hence models M_f and M_g are equivalent). The alternative hypotheses are thus that M_f is “better” than M_g and vice-versa. The variance of LR_n can be estimated by the empirical variance:

$$\omega_n^2 \equiv \frac{1}{n} \sum_{t=1}^n \left[\log \frac{f(Y_t|X_t; \hat{\theta}_n)}{g(Y_t|Z_t; \hat{\gamma}_n)} \right]^2 - \left[\frac{1}{n} \sum_{t=1}^n \log \frac{f(Y_t|X_t; \hat{\theta}_n)}{g(Y_t|Z_t; \hat{\gamma}_n)} \right]^2 \quad (4)$$

Vuong shows that, under fairly general conditions, $\frac{LR_n(\hat{\theta}_n, \hat{\gamma}_n)}{\omega_n \sqrt{n}} \xrightarrow{D} N(0, 1)$ under the null hypothesis, and to $\pm\infty$ otherwise.

Vuong's test is undoubtedly quick and simple to execute. It is however very conservative, (see, for example, Table 2), and in those cases where the null hypothesis is rejected, this only indicates that, say, M_f is preferable to M_g , not necessarily that M_f is suitable. (See Tables 3 and 4).

4 The Hybrid Test

Clearly it would be advantageous to develop a hybrid of Cox's and Vuong's test that combines the ease of use of the latter with the accuracy of the former. We do this by replacing the single null hypothesis of Vuong with the double null hypotheses of Cox, and adjusting equation (4) to be the expected value of the variance under each null hypothesis respectively. This is equivalent to applying the analytic version of Cox's test with the right-hand term of (2) omitted when calculating the variance, or applying a simulation-based Cox's test where the model parameters are *not* refitted at each resample. Given that the requirement to refit the parameters of each resample is by far the most time consuming aspect of simulation-based versions of Cox's test, the practical benefits of not having to do so are enormous. Such a "resampling without refitting" approach has been used by Allcroft and Glasby(2003). Given that the variance of the hybrid test is greater than that of Cox's test, it is necessarily more conservative than Cox's test. Clearly the larger the right-hand term of equation (2) relative to the left-hand term, the greater the conservatism of the hybrid test when compared to Cox. Note that the ratio of the right and left hand terms of (2) is the ratio of the expected value of the covariance of the log-likelihood ratio and the score statistic under H_f to the product of the expected values of their variances. We denote the value of this correlation coefficient type statistic by r_f^2 for $H_0 : M_f$ and r_g^2 for $H_0 : M_g$. In general, these two values are different, indicating that the conservatism of the hybrid test is not symmetrical. Note that for the example of Table 4 below r_g^2 ($H_0:geometric$) is more than five times greater than r_p^2 ($H_0:Poisson$), and hence the hybrid test is more conservative when rejecting a geometric distribution than it is when rejecting a Poisson distribution, in relation to Cox's test. Tables 2 to 4 each show the results obtained when the three tests were each used to classify 1,000 samples of size 50 taken from data that followed a geometric distribution with mean 0.8, a binary distribution consisting of equal numbers of zeros and ones, and a geometric distribution with mean 6, respectively. For example, Cox's test classified 35 of the 1,000 samples taken from geometric(0.8) data as Poisson, 764 as geometric, 158 as possibly both, and 43 as neither. We see that the hybrid test performs well in relation to Cox's test, and considerably better than Vuong's test.

5 Examples

We further illustrate the usefulness of the hybrid test by comparing its performance with that of the simulation-based version of Cox's test, (1,000 resamples), and Vuong's test, when applied to two "real life" data sets. Please note that the purpose of these

TABLE 2. Classification of 1,000 samples drawn from Geometric(0.8)

Test	Pois	Geom	Both	Neither
Vuong	5	223	768	–
Cox	35	764	158	43
Hybrid	24	755	181	40
mean value of $r_p^2 = 0.105$				
mean value of $r_g^2 = 0.147$				

TABLE 3. Classification of 1,000 samples drawn from binary distribution

Test	Pois	Geom	Both	Neither
Vuong	1000	0	0	–
Cox	4	0	0	996
Hybrid	4	0	0	996
mean value of $r_p^2 = 0.092$				
mean value of $r_g^2 = 0.109$				

TABLE 4. Classification of 1,000 samples drawn from Geometric(6)

Test	Pois	Geom	Both	Neither
Vuong	0	995	5	–
Cox	0	907	0	93
Hybrid	0	962	0	38
mean value of $r_p^2 = 0.061$				
mean value of $r_g^2 = 0.309$				

examples is to emphasise the merits of the hybrid test for comparing models, not to determine a suitable model.

Firstly, we consider data from Leroux and Puterman (1992) that gives the number of movements made by a fetal lamb in each of 240 consecutive 5-second intervals (Table 5). Clearly this data is overdispersed, hence two possible “candidate” models are the negative-binomial (Poisson-Gamma) and the Neyman-A (Poisson-Poisson). Due to the presence of infinite sums in its probability density function, algorithms for fitting Neyman-A models are slow, even in the absence of covariates. As shown in Table 5, the (elapsed) time taken to complete the hybrid test is over 150 times less than that of Cox’s test, both tests concluding that, at $\alpha = 0.05$, we may reject the Neyman-A model in favour of the negative binomial. Note that whilst Vuong’s test is practically instantaneous, it fails to distinguish between the models.

Secondly, we look at data from Ridout, Demétrio, and Hinde (1998) describing the number of roots produced by 270 micropropagated shoots of the apple cultivar *Trajan*. Two covariates were present. *Period*, at 2 levels, and *Hormone* at 4. Ridout et al. show

TABLE 5. *p*-values for Fetal Lamb Data

Vuong (< 1 second)		Cox (143 minutes)		Hybrid (52 seconds)	
Neyman-A	Neg-Bin	Neyman-A	Neg-Bin	Neyman-A	Neg-Bin
0.183	0.817	0.019	0.863	0.025	0.828
$r_{NA}^2 = 0.101$			$r_{NB}^2 = 0.082$		

that *hormone* has little effect. In general, the presence of covariates increases the time taken for model-fitting. This is not of major consequence if the number of covariates in the model is small and efficient algorithms for fitting the model in question exist. For example, a comparison of two zero-inflated negative-binomial models: $roots \sim period$ and $roots \sim hormone$, (where both the mean and the over-inflation parameter are fitted by the given covariate), where both models were fitted by the R-package *Zicounts*, took approximately 12 minutes to complete for Cox’s test, as opposed to approximately 11 seconds using the hybrid test. If many covariates are present, or efficient model-fitting algorithms do not exist, then Cox’s test may prove impractical. An example is that of Table 6 which illustrates the results obtained when the three tests were used to compare Neyman-A and zero-inflated Poisson models of the form $roots \sim period$, (all parameters varying over *period*). Vuong’s test failed to reject either model. Cox’s test proved impractical: less than a quarter of the 1,000 resamples had occurred after two days, and the test was abandoned, whereas the hybrid test completed in under 2 minutes, rejecting both models.

TABLE 6. *p*-values for Trajan Apple Data, ($roots \sim period$)

Vuong (< 1 second)		Cox (estimate: 9 days)		Hybrid (117 seconds)	
Neyman-A	ZIP	Neyman-A	ZIP	Neyman-A	ZIP
0.506	0.494	—	—	0.018	0.000
$r_{NA}^2 = 0.101$			$r_{ZIP}^2 = 0.082$		

6 Conclusion

The hybrid test is a suitable alternative to Cox’s test and Vuong’s test, being much quicker than the former, and more decisive than the latter. The results above are dependent upon models being non-nested, and the extension of the hybrid test to models where this is not the case is a possible area of future research.

Acknowledgement

The Author wishes to thank Prof. John Hinde for the generous support and assistance he provided with the preparation of this paper.

References

- Allcroft D.J., and Glasby, C.A. (2003) A simulation-based study for model evaluation. *Statistical Modelling* **3**, 1–13.
- Cox D.R. (1961) Tests of Separate Families *Proceedings of the fourth Berkeley Symposium* **1**, 105–123.
- Cox D.R. (1962) Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society. Series B* **24**, 406–423.
- Dobbie M.J., & Welsh A.H. (2001) Models for zero-inflated count data using the Neyman type A distribution. *Statistical Modelling* **1**, 65–80.
- Hinde John P. (1992) Choosing Between Non-nested Models: a Simulation Approach. In Fahrmeir L, Francis B, Gilchrist R, Tutz G eds. *Advances in Glim and Statistical Modelling: Proceedings of the Glim92 Conference and 7th International Workshop on Statistical Modelling*. New York: Springer.
- Leroux, B.G. & Puterman, M.L. (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545–558.
- Ridout, M.S, Demétrio, C.G.B., & Hinde, J.P. (1998) Models for count data with many zeros. *Proceedings of the XIXth international Biometrics conference, Cape Town, Invited Papers*. 179–192.
- Smith, A (1999) Testing Theories of Strategic Choice: The example of Crisis Escalation. *American Journal of Political Science* **43**, 1254–1283.
- Vuong, Q.H. (1989) Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57**, 307–333.
- Williams, D.A. (1970) Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics* **28**, 23–32.

Author Index

- Adamska, E, 360
Adamski, T, 39
Aerts, M, 43, 253
Aguilar, L, 81
Alberts, DS, 350
Alfó, M, 48
Almansa, J, 52
Alonso, A, 21
Alonso, AM, 56, 60
Alonso, J, 52
Amaral-Turkman, MA, 540
Amaral-Turkman, MA, 556
Areira, A, 65
Arostegui, I, 69
Assaf, A, 73
Assam, P, 21
Augustin, T, 245
Bailey, TC, 227
Barber, A, 77
Barber, X, 77, 117
Barceló-Vidal, C, 81, 427
Barreiros, J, 154
Barry, S, 87
Bartolucci, F, 93
Batchelor, A, 99
Bayarri, MJ, 3
Bécue-Bertaut, M, 103
Berger, J, 3
Beutels, P, 43
Blagojevik, M, 107
Blance, A, 111
Böckenholt, U, 205
Boj, E, 249
Bolfarine, H, 114
Bollaerts, K, 43
Botella, F, 117
Bowman, A, 87
Branco, JA, 524
Brewer, MJ, 121
Brijs, T, 364
Broner, S, 126
Brown, JA, 130
Buil, A, 136
Burzykowski, T, 21
Buyse, M, 21
Caballero-Águila, R, 140, 144
Cabri, J, 154
Cadarso-Suárez, C, 528
Calle, ML, 508
Camarda, CG, 148
Carita, AI, 154
Carvalho, J, 111
Casado, D, 56
Casella, G, 8
Cazzaro, M, 158
Cegielska-Taras, T, 360
Ceranka, B, 164, 168
Chua, SJ, 172
Claeskens, G, 185
Clotet, B, 490
Colombi, R, 158
Conde, S, 177
Conesa, D, 181
Consentino, F, 185
Cordeiro, C, 181
Correal, M, 195
Costa, MJ, 199
Crainiceanu, C, 300
Cribari-Neto, F, 211
Cruyff, M, 205
Cullis, B, 548
Currie, I, 370
Cysneiros, AHMA, 211
Cysneiros, FJA, 473
Czado, C, 584
Daunis-i-Estadella, J, 431, 463
Delicado, P, 126, 277
Diggle, J, 544
Dittrich, R, 215
Dorta-Guerra, R, 287
Dryden, IL, 221
Durbán, M, 396
Economou, T, 227
Egozcue, JJ, 459, 560
Eilers, PHC, 148, 233, 239, 390
Einbeck, J, 245, 449
Elliot, MR, 400
Escarela, G, 494

- Espinal, A, 580
Esteve, A, 249
Faddy, MJ, 265
Faes, C, 43, 253
Falguerolles, A de, 259
Fenlon, JS, 265
Finselbach, HK, 172
Firth, D, 99
Fontcuberta, J, 136
Fortiana, J, 249
Fox, J, 269
Francis, B, 215
Friedl, H, 333
Friendly, M, 269
Frühwirth-Schnatter, S, 587
Gadelha de Araújo Jr., CA, 211
Gampe, J, 148, 239
García-Ligero, MJ, 273
Geys, H, 253
Gilthorpe, MS, 111, 593
Giordano, S, 158
Giraldo, R, 277
Girón, FJ, 8
Goicoa, T, 568
Gómez, G, 283, 490, 508, 536
González-Dávila, E, 287
González-Sierra, MA, 287
González-Yanes, A, 287
Graczyk, M, 164, 168
Grané, A, 291
Greenacre, M, 295
Greven, S, 300
Grilli, L, 306
Guolo, A, 310
Ha, ID, 412, 314
Hatzinger, R, 215
Heiner, K, 319
Heller, G, 323, 481
Hens, N, 43
Hermoso-Carazo, A, 140, 144, 273, 329
Hinde, J, 319, 339
Hofrichter, J, 333
Holian, E, 339
Hsu, CH, 350
Huertas, J-A, 536
Jaramillo, P, 580
Jiménez-López, JD, 140, 144
Jowaheer, V, 354
Julià, O, 283
Kaczmarek, Z, 39, 360
Kapelán, Z, 227
Karlis, D, 364
Katzenbeisser, W, 215
Keiding, N, 16
Kirkby, J, 370
Kolb, C, 427
Komárek, A, 376
Kosmidis, I, 382
Küchenhoff, H, 300
Kulinskaya, E, 386
Lambert, P, 390
Lammer, H, 427
Lee, D-J, 396
Lee, Y, 314
Lesaffre, E, 376, 518, 564
Lewis, C, 544
Li, B, 221
Li, Y, 400
Linares-Pérez, J, 140, 144, 273, 329
Liu, F, 3
Liu, L, 574
Long, Q, 350
López-de-Ullibarri, I, 528
López-Pintado, S, 56
López-Quílez, A, 117, 181
Lupparelli, M, 93
Lynch, J, 404
Machado, C, 408
MacKenzie, G, 177, 404, 412, 477
MacNab, YC, 417
Mair, P, 423
Malats, N, 508
Marques, CJ, 154
Martín-Fernández, JA, 81
Martínez-Beneito, MA, 181
Martín-Fernández, JA, 427, 463
Maruotti, A, 48
Marx, BD, 239
Matawie, KM, 73
Mateu, J, 277
Mateu-Figueras, G, 431, 560
Matthews, FE, 344

- Mayoral, AM, 77, 117
Mejza, I, 437
Mejza, S, 437, 441
Mercatanti, A, 445
Mexia, JT, 65, 486
Militino, AF, 568
Molenberghs, G, 21, 253
Morales, J, 77, 117
Moreno, E, 8
Nakamori, S, 140, 144, 273
Neves, MM, 191, 532
Newell, J, 449
Nunes, F, 408
Núñez-Antón, V, 69, 453
Obergruppenberger, M, 500
Oliveira, MM, 65
Orbe, J, 453
Ortego, MI, 459
Ostermann, A, 500
Palarea-Albaladejo, J, 463
Papola, AL, 467
Paula, GA, 473
Paulino, CD, 408, 504
Paulo, R, 3
Pawlowsky-Glahn, V, 560
Peng, D, 477
Penman, R, 481
Peña, D, 60
Pereira, D, 441, 486
Pérez-Álvarez, N, 490
Pérez-Ruiz, LC, 494
Peters, A, 300
Pfeffermann, D, 28
Pfeifer, C, 500
Poblet, M, 103
Poletto, FZ, 504
Porta, N, 508
Rampichini, C, 306
Rau, R, 239
Rea, WS, 130
Reale, M, 130
Rigby, R, 323
Rivero, C, 514
Rizopoulos, D, 518, 564
Rocha, CS, 467
Rodrigues, PC, 524
Rodríguez, J, 60
Rodríguez-Álvarez, MX, 528
Romo, J, 56
Ruiz, L, 490
Sacks, J, 3
Sánchez-Zapata, JA, 117
Santos, JA, 532
Sebastián, E, 117
Sermaidis, GJ, 364
Serrat, C, 536
Shaw, JEH, 199
Shen, L, 221
Shkedy, Z, 43
Silva, GL, 540
Singer, JM, 245, 504
Smith, A, 548
Solis-Trapala, I, 544
Soria, JM, 136
Soulsby, C, 121
Souto, JC, 136
Stasinopoulos, M, 323
Staudte, RG, 386
Stefanova, K, 548
Streftaris, G, 552
Surma, M, 39
Sutradhar, B, 354
Taylor, JMG, 350, 400
Teles, J, 556
Tetzlaff, D, 121
Teuns, G, 253
Tilahun, A, 21
Tolosana-Delgado, R, 560
Trujillo, A, 580
Tsonaka, R, 564
Tüchler, R, 587
Turner, HL, 99
Tyler, J, 481
Ugarte, MD, 568
Vaida, F, 574
Valdes, T, 514
Valero, O, 580
Van Damme, P, 43
van den Hout, A, 344
van der Heijden, P, 205
Varin, C, 584
Veiga, H, 291

Verbeke, G, 518, 564
Vilagut, G, 52
Wagner, H, 587
Waldron, S, 121
Watkins, AJ, 172
West, RM, 593
Wilson, P, 597
Worton, BJ, 552
Zeileis, A, 423