

**Proceedings of the  
23d International  
Workshop  
on Statistical Modelling**

**July 7-11, 2008**

**Utrecht**

**Paul H. C. Eilers  
(editor)**

Proceedings of the 23d International Workshop on Statistical Modelling.  
Utrecht, July 7-11, 2008  
Paul H. C. Eilers, editor  
Utrecht 2008.

**Editor:**

Paul H. C. Eilers  
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

*Printed by Ipskamp Partners, Enschede*

## **Scientific Programme Committee**

- Ulf Böckenholt (McGill University, Montréal)
- Dankmar Böhning (University of Reading)
- María Durbán (Universidad Carlos III de Madrid, Leganes)
- Paul Eilers (Chair) (Utrecht University)
- John Hinde (National University of Ireland, Galway)
- Mikis Stasinopoulos (London Metropolitan University)



## Preface

Welcome to the 23d International Workshop on Statistical Modelling, welcome to Utrecht. In this volume you will find all contributions by invited speakers, contributing speakers, and poster presenters. Following a long tradition, the proceedings will be ready at the beginning of the workshop. For the first time the proceedings are also provided as a PDF file on a USB stick. This allows you to see color figures and to enlarge details in graphs.

The International Workshop on Statistical Modelling (IWSM) has been held in Europe, USA and Australia since 1986. It has always been focused on papers that are motivated by real-life data and that make novel contributions to the subject of Statistics. Statistical Modelling is an important cornerstone in many scientific disciplines, and the workshop has consistently provided a rich environment for cross-fertilization of ideas from different statistical areas. The workshop has brought together scientists from different nationalities with different backgrounds and experience, and has always promoted contributions from students early in their careers, allowing time for discussion and interchange between junior and senior scientists.

When you browse these proceedings, you will be struck by the enormously wide range of applications and models. You might not be interested in all of them, but the format of the Workshop, with no parallel sessions, will give you an opportunity to get a taste of unfamiliar areas. Hopefully this will give you ideas for your own research. Don't be afraid to ask questions, during the sessions or during the breaks; lively interaction between statisticians is a major goal of the Workshop.

Participants have attended from all corners of the globe, and workshops have travelled around Europe - to Perugia (1987), Vienna (1988), Trento (1989), Toulouse (1990), Utrecht (1991), Munich (1992), Leuven (1993), Exeter (1994), Innsbruck (1995), Orvieto (1996) and Biel/Bienne (1997); to the USA - New Orleans (1998); and back to Europe - Graz (1999), Bilbao (2000), Odense (2001), Chania (2003), Leuven (2003) and Florence (2004); to Australia - Sydney (2005); and back again to Europe - Galway (2006), and Barcelona (2007).

I wish you a fruitful and pleasant stay in Utrecht.

Paul Eilers  
Utrecht, June 2008

## Part 1. Invited papers

A. AGRESTI ET AL. Good Confidence Intervals for Discrete Statistical Models	3
J.P. FOX Bayesian Item Response Models For Complex Survey Data	19
J.H. FRIEDMAN Fast Sparse Regression and Classification	27
T. STIJNEN ET AL. Random effects meta-analysis in the framework of the general(ized) linear mixed model	59
G. TUTZ Boosting Strategies in Semiparametrically Structured Regression	73

## Part 2. Contributed papers

M. AERTS ET AL. Semi-parametric regression to identify farms with high <i>Salmonella</i> infection burden	89
T. ANZAI ET AL. Modelling the Duration of Necessary and Non-necessary Activities in Daily Life: Fitting Mixture Models to Japanese Time Use Survey Data	93
C. ARMERO ET AL. Developing an expert system for predicting Legionella outbreaks in evaporative installations by using Bayesian hierarchical models	99
G. BAIO ET AL. Bayesian hierarchical model for the prediction of football results	104
F. BARTOLUCCI ET AL. A multidimensional latent Markov IRT model	109
C. BELITZ ET AL. Complex additive penalties for generalized structured additive regression	115
F. BOLAND Hierarchical Generalised Linear Models Analysis of Bovine Tuberculosis on Milk Data	121
K. BOLLAERTS ET AL. Dose-Illness Models for Human Salmonellosis Based on Outbreak Data	125
A.R. BRETNALL ET AL. Empirical random-effects models to predict the amount individuals withdraw at cash machines	131
M.J.S. BRINKHUIS ET AL. Student monitoring using Chess ratings	137
S. CAKMAK ET AL. Do pollution time-series studies contain residual confounding by personal risk factors for acute health events?	143
C.G. CAMARDA ET AL. A Warped Failure Time Model for Human Mortality	149
P. CANAS RODRIGUES ET AL. Cd and Cu migration during electro-dialysis: biregression modeling	155
A.I. CARITA ET AL. Logistic Regression Model: continuous independent variables and linearity in the logit	159
S. CECERE ET AL. On the Bayesian 2-stage procedure for parameter estimation in copula models	163

P. ROCHA CHELLINI ET AL. Simulation-based results for bioequivalence studies using the 2x2 crossover design	169
S. J. CHUA ET AL. The Reliability of Type II Censored Reliability Analyses for Weibull Data	173
N. COFFEY ET AL. Estimating Functional Principal Components using the Linear Mixed Effects Model.	178
S. CONDE ET AL. Search Algorithms for Log-Linear Models in Contingency Tables. Comorbidity Data	184
M.J.L.F. CRUYFF ET AL. Analyzing Randomized Response Data using a Doubly Zero-Inflated Poisson Model	188
I. CURRIE Smoothing overparameterized regression models	194
D. DEJARDIN ET AL. Joint Modeling of Progression Free Survival and Death	200
J.J DE ROOI ET AL. Smoothing zeros and small counts in meta-analysis of clinical trials	204
D. DRAGHICESCU Modelling space-time quantiles of ground-level ozone	210
T. ECONOMOU ET AL. A hidden semi-Markov model for the occurrences of water pipe bursts	216
A. ESPINAL ET AL. On parsimonious higher-order binary Markov chain models	221
C. FERGUSON ET AL. Nonparametric Principal Components Analysis for Ecological Data	226
P.A. FILIPE ET AL. Modelling Individual Animal Growth in Random Environments	232
J. GAMPE ET AL. Variation in Mortality: Estimation via a Meta-Analytic Approach	238
R. GILCHRIST ET AL. An Application of GAMLSS: An Insurance Type Model for the Health Cost of Cold Housing	244
D. GOMES ET AL. Cold and Heat waves: Modelling the mortality in Évora — Portugal	250
M.J.HAWKINS ET AL. Modelling Changes in Irish Forest Carbon Stocks	252
T.A.M. HENDRICK ET AL. The Behaviour of the Self-Protective Parameter Estimation in Models for Randomized-Response Data.	255
D.J. HESSEN Constant Latent Odds-Ratios Models and Marginal Maximum Likelihood Estimation	259
J.J. HOUWING-DUISTERMAAT ET AL. Methods for aggregation and linkage analysis of human longevity in selected families.	265
M. HUBREGTSE ET AL. Explaining Self-Protective "No"-Saying in Randomized Response	271
I. HUDSON ET AL. Climate impacts on Sudden Infant Death Syndrome: a GAMLSS approach	277

M.A. JONKER ET AL. Correlated gamma frailty model for linkage analysis on twin data with application to interval censored migraine data	281
S.W. KIM ET AL. Modelling and synchronization of four Eucalyptus species via MTD and EKF	287
B. KLINGENBERG Dose-response modeling with bivariate binary data under model uncertainty	293
E.J.H. KORENDIJK ET AL. The Robustness of the Parameter and Standard Error Estimates in Trials with Partially Nested Data. A Simulation Study	299
C. KOU ET AL. Variable Selection in Joint Modelling of Mean and Covariance Structures for Longitudinal Data	305
D.-J. LEE ET AL. <i>P</i> -spline ANOVA-type interaction models for spatio-temporal smoothing	311
I. LÓPEZ-DE-ULLIBARRI ET AL. ROC.Regression: A new R software for ROC Regression Analysis	317
J. LYNCH ET AL. Piece-wise exponential PH models	321
G. MACKENZIE ET AL. Robustness of the Regression Parameter in PH and Non-PH Frailty Survival Models	327
N. MARTÍN New Estimators for Mortality Ratios through Loglinear Modelling with Linear Constraints	331
B.D. MARX ET AL. Bilinear Varying-Coefficient Models for Seasonal Time Series and Tables	335
S. MEJZA ET AL. A note on a modelling environmental indexes	341
I. MEJZA ET AL. Model building for series of block designed experiments	345
R.X. DE MENEZES ET AL. Integrated statistical analysis to identify associations between DNA copy number and gene expression in microarray data	349
I. MITRA ET AL. An Estimator of the Variance of Measurement Error	354
V.M.R.. MUGGEO Estimation of linear errors-in-variables models with error-free covariates: a backfitting approach	360
G. NEUBAUER ET AL. A Generalized Poisson Model for Underreporting	364
R.C.A. RIPPE ET AL. Models for Fluorescence Signals on SNP Arrays	370
I. SAVVALA ET AL. Quintile stratification on propensity scores and standardization for a T-test	376
R. SCHEUFELE ET AL. A simple method of estimating relative risk from a prevalence study and observations of disease duration	380
S.K. SCHNABEL ET AL. Optimal expectile smoothing	386
A.L. SIQUEIRA ET AL. Comparing crossover designs in average bioequivalence studies	392

J. SLEEP ET AL. Comparison of Self-Organizing Maps, Mixture, K-means and Hybrid Approaches to Risk Classification of Passive Railway Crossings	396
N. SOFRONIOU ET AL. League tables for literacy survey data based on random effect models	402
J. TELES ET AL. Power of normality tests under a mixture of gaussian distributions: a simulation study	406
J. TIJMSTRA ET AL. Testing Manifest Monotonicity and Weak Item Independence for the Constant Latent Odds-Ratios Model	410
H.-W. UH ET AL. Haplotype Frequency Estimation with the Penalized Composite Link Model	416
O. VALERO ET AL. Analysis of labor spells in Social Security contributors	422
A. VAN DEN HOUT ET AL. A Bayesian three-state model to estimate life expectancies	426
F.A. VAN EEUWIJK ET AL. Modeling Spatial Effects in Field Trials	432
H. WAGNER ET AL. Bayesian estimation of random effects models for multivariate responses of mixed data	438
P. WILSON Bias estimation of $p$ -values in analytic and simulated Cox Tests for non-nested models	443
P. WILSON The Dragnet Test: A New Approach to Choosing Between Models	448
J. WYSE A physiological application of Bayesian linear regression with a change-point	454
I. YAHAV ET AL. Reducing Error by Increasing Focus: Multivariate Monitoring of Biosurveillance Data	458



**Part 1**  
**Invited papers**



# Good Confidence Intervals for Discrete Statistical Models

Alan Agresti<sup>1</sup>, Euijung Ryu<sup>2</sup>

<sup>1</sup> Department of Statistics, University of Florida, Gainesville Florida 32611, USA

<sup>2</sup> Division of Biostatistics, Mayo Clinic, Rochester, Minnesota 55905, USA

**Abstract:** We survey good methods for constructing confidence intervals for parameters in discrete statistical models, with emphasis on categorical data. The method of inverting score tests for parameter values performs well, usually much better than inverting Wald tests and often better than inverting likelihood-ratio tests. Exact small-sample methods are conservative inferentially, but inverting a test using the mid- $P$  value provides a sensible compromise. For some models ordinary score inferences are impractical, such as when the likelihood function is not an explicit function of the model parameters. For such cases, we propose pseudo-score inference based on a Pearson-type chi-squared statistic that compares fitted values for a working model with fitted values of the model when the parameter of interest takes a fixed value. Finally, we briefly summarize a different pseudo-score approach that approximates score intervals for proportions and their differences by adding artificial observations before forming simple Wald confidence intervals.

**Keywords:** Categorical data; Multinomial models; Pearson chi-squared; Score test.

## 1 Introduction

Confidence intervals for a parameter can be constructed by inverting significance tests about the value of that parameter. Section 2 summarizes research that has found that for many categorical data analyses, inverting the large-sample score test performs well. For small samples, Section 3 illustrates how a slightly adjusted method based on inverting score tests using the mid  $P$ -value performs well.

Score confidence intervals are sometimes difficult to construct, such as when the likelihood function is not an explicit function of the model parameters. For interval estimation of a parameter in a multinomial model, we propose a “pseudo-score” method that inverts a modified Pearson statistic comparing the fitted values for the model to the fitted values assuming a particular value of that parameter. Section 4 introduces this method and Section 5 outlines potential generalizations for development in future research.

The Wald-test-based confidence interval is simple and is commonly taught in introductory statistics courses and used in practice, but it often performs poorly. Section 6 summarizes simple adjustments of Wald intervals

for proportions and their differences such that the intervals resemble and perform similarly to score intervals.

## 2 Score-Test Based Confidence Intervals

For a generic parameter  $\beta$ , consider a confidence interval (CI) based on inverting a two-sided significance test of  $H_0: \beta = \beta_0$ . For example, the 95% CI is the set of  $\beta_0$  values for which the  $P$ -value  $> 0.05$ . A common approach inverts one of three large-sample chi-squared tests: the likelihood-ratio test proposed by Sam Wilks in 1938, the Wald test proposed by Abraham Wald in 1943, or the score test proposed by C. R. Rao in 1948.

For a log likelihood function  $L(\beta)$  (using a single parameter here for notational simplicity), denote the maximum likelihood (ML) estimate by  $\hat{\beta}$ , the score function by  $u(\beta) = \partial L(\beta)/\partial\beta$ , and the information by  $\iota(\beta) = -E[\partial^2 L(\beta)/\partial\beta^2]$ . The Wald test uses

$$[(\hat{\beta} - \beta_0)/SE]^2 = (\hat{\beta} - \beta_0)^2 \iota(\hat{\beta}),$$

where  $\iota(\hat{\beta})$  denotes  $\iota(\beta)$  evaluated at  $\hat{\beta}$ . For example, the 95% Wald CI is  $\hat{\beta} \pm 1.96(SE)$ . The likelihood-ratio (LR) test statistic is

$$-2[L(\beta_0) - L(\hat{\beta})].$$

The score test statistic is

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial\beta_0]^2}{-E[\partial^2 L(\beta)/\partial\beta_0^2]},$$

where the partial derivatives are evaluated at  $\beta_0$ . For canonical generalized linear models, the score statistic is a standardization of the sufficient statistic for  $\beta$ . The three methods are asymptotically equivalent under  $H_0$  (Cox and Hinkley 1974).

Many popular tests in categorical data analysis are score tests. Examples are the Pearson chi-squared test of independence in two-way contingency tables, the McNemar test for binary matched pairs, the Cochran–Mantel–Haenszel test of conditional independence for stratified  $2 \times 2$  tables, and the Cochran–Armitage trend test for several ordered binomial samples. However, the only score CI known by most practitioners is Wilson’s (1927) CI for a binomial parameter  $\pi$ , based on inverting the asymptotic standard normal test,

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}.$$

Score CIs are less well known for other cases, even for basic parameters for  $2 \times 2$  contingency tables  $\{n_{ij}\}$  summarizing independent binomial samples. Mee (1984) showed that the score CI for the difference of proportions  $\pi_1 - \pi_2$

inverts the test of  $H_0: \pi_1 - \pi_2 = \beta_0$  using test statistic that is the square of

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \beta_0}{\sqrt{[\hat{\pi}_1(\beta_0)(1 - \hat{\pi}_1(\beta_0))/n_1] + [\hat{\pi}_2(\beta_0)(1 - \hat{\pi}_2(\beta_0))/n_2]}}$$

where  $\hat{\pi}_1$  and  $\hat{\pi}_2$  are the sample proportions (i.e., unrestricted ML estimates) and  $\hat{\pi}_1(\beta_0)$  and  $\hat{\pi}_2(\beta_0)$  are the ML estimates under the constraint  $\pi_1 - \pi_2 = \beta_0$ . (When  $\beta_0 = 0$ ,  $z^2$  is the Pearson chi-squared statistic.) For interval estimation of an odds ratio, for a given  $\beta_0$  let  $\{\hat{\mu}_{ij}(\beta_0)\}$  have the same row and column margins as  $\{n_{ij}\}$  and

$$\frac{\hat{\mu}_{11}(\beta_0)\hat{\mu}_{22}(\beta_0)}{\hat{\mu}_{12}(\beta_0)\hat{\mu}_{21}(\beta_0)} = \beta_0.$$

Let  $\chi_{1,a}^2$  denote the  $100(1-a)$  percentile of a chi-squared distribution with  $df = 1$ . Cornfield (1956) noted that the set of  $\beta_0$  satisfying

$$X^2 = \sum (n_{ij} - \hat{\mu}_{ij}(\beta_0))^2 / \hat{\mu}_{ij}(\beta_0) \leq \chi_{1,a}^2$$

form a  $100(1-a)\%$  conditional score CI for the odds ratio. Likewise, score CIs exist for the relative risk (Koopman 1984), logistic regression parameters, and generic measures of association (Lang 2005).

More generally now, let  $\{n_i\}$  denote cell counts for a multinomial model for a contingency table of arbitrary dimensions. Let  $\{\hat{\mu}_i\}$  be the ML fitted values for the model. For testing goodness of fit, the score test statistic is the Pearson statistic,

$$X^2 = \sum \frac{(n_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

This was observed by Cox and Hinkley (1974 p. 326) and then extended to a corresponding statistic for generalized linear models by Smyth (2003). When the model is the special case of the saturated model that provides a particular value for a parameter  $\beta$ , then inverting the Pearson chi-squared test of  $H_0: \beta = \beta_0$  yields the score CI.

There is now considerable evidence that score inference performs well, usually better than Wald inference and often better than LR inference for small sample sizes, in terms of actual error probabilities being close to nominal levels. See, for example, Koehler and Larntz (1980) for testing independence in two-way contingency tables, Newcombe (1998a) and Agresti and Coull (1998) for CIs for binomial proportions, Miettinen and Nurminen (1985), Newcombe (1998b), and Agresti and Min (2005a) for CIs for the difference of proportions and relative risk, Tango (1998) and Agresti and Min (2005b) for inference about the difference of proportions for dependent samples, Miettinen and Nurminen (1985) and Agresti and Min (2005a) for CIs for the odds ratio, Agresti and Klingenberg (2005) for multivariate comparisons of proportions, Agresti et al. (2008) for simultaneous CIs comparing several binomial proportions, and Ryu and Agresti (2008) for ordinal effect measures.

In practice, Wald CIs and likelihood-ratio-test based CIs (e.g., profile likelihood CIs) are commonly used and are easily accessible with statistical software. By contrast, score CIs are not as well known as they deserve to be, given how well they perform, and they are rarely available in software.

### 3 Small-Sample Score Confidence Intervals

Using the score (or other) test statistic, it is possible to apply small-sample distributions, rather than large-sample approximations, to obtain  $P$ -values and CIs. To illustrate, consider inference for a parameter of a logistic regression model. For subject  $i$  with binary outcome  $y_i$  and values for  $k$  explanatory variables  $x_{i0} = 1, x_{i1}, x_{i2}, \dots, x_{ik}$ , the model is

$$\text{logit}[P(y_i = 1)] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}.$$

Score statistics are based on the sufficient statistic, which is  $T_j = \sum_i y_i x_{ij}$  for  $\beta_j$ . Starting with the usual binomial likelihood, one can base a test on the conditional distribution of  $T_j$  after eliminating the other parameters by conditioning on their sufficient statistics. For example, with the tail method, bounds  $(\beta_{1L}, \beta_{1U})$  of a 95% CI for  $\beta_1$  satisfy

$$P(T_1 \geq t_{1,obs} | t_0, t_2, \dots, t_k; \beta_{1L}) = 0.025$$

$$P(T_1 \leq t_{1,obs} | t_0, t_2, \dots, t_k; \beta_{1U}) = 0.025.$$

Software is readily available, such as LogXact (Cytel Software 2005).

Because of discreteness, error probabilities for small-sample inference do not exactly equal the nominal values. Inverting a test with actual size  $\leq 0.05$  for all  $\beta_0$  guarantees that the CI has actual coverage probability  $\geq 0.95$ . In this sense, inferences are *conservative*. In practice, the actual coverage probability varies for different  $\beta$  values and is unknown. When the conservatism is problematic, there are ways of alleviating it (Agresti 2003). One way is to invert a single two-sided test instead of two equal-tail one-sided tests. Agresti and Min (2001) inverted two-sided score tests for the difference and ratio of proportions, and this method is available in the StatXact software (Cytel 2005). Another approach with few parameters (such as for  $2 \times 2$  tables) uses an unconditional approach to eliminate nuisance parameters, because the conditional approach exacerbates the discreteness. For  $H_0: \beta = \beta_0$  with nuisance parameter  $\psi$ , let  $p(\beta_0; \psi)$  be the  $P$ -value for a given value of  $\psi$ . The unconditional  $P$ -value is  $\sup_{\psi} p(\beta_0; \psi)$  and the  $100(1 - a)\%$  CI consists of  $\beta_0$  for which  $\sup_{\psi} p(\beta_0; \psi) > a$ . Agresti and Min (2002) found that this approach worked well for the odds ratio, using a two-sided score statistic as the criterion.

The discreteness issue can be avoided completely using randomized inference. With a discrete test statistic  $T$ , let  $\mathcal{U}$  be a uniform(0,1) random

variable. A CI with actual coverage probability exactly 0.95 has endpoints  $(\beta_L, \beta_U)$  satisfying

$$P_{\beta_U}(T < t_{obs}) + \mathcal{U} \times P_{\beta_U}(T = t_{obs}) = 0.025$$

$$P_{\beta_L}(T > t_{obs}) + (1 - \mathcal{U}) \times P_{\beta_L}(T = t_{obs}) = 0.025.$$

Stevens (1950) suggested this approach for the binomial parameter. It is a historical curiosity that he and other statisticians actually thought this approach would be adopted for applied work. For example, Pearson (1950) argued that statisticians may well come to accept randomization *after* performing an experiment just as they had come to accept Fisher's ideas about randomization *before* performing the experiment. Stevens (1950) argued that an advantage of eliminating the uncertainty about the actual coverage probability is that one would obtain a narrower CI than with standard small-sample methods.

These days, randomized inference of this type is not used, although a *fuzzy inference* approach portrays graphically all such possible CIs (Geyer and Meeden 2005). We find the approach useful for motivating an alternative method based on inverting tests using the *mid-P-value* (Lancaster 1961). For  $H_a : \beta > \beta_0$ , the mid- $P$ -value is

$$P_{\beta_0}(T > t_{obs}) + (1/2)P_{\beta_0}(T = t_{obs}).$$

Unlike the randomized  $P$ -value, it depends only on the data. Under  $H_0$ , the ordinary  $P$ -value is stochastically larger than uniform, but  $E(\text{mid-}P\text{-value}) = 1/2$ . The sum of right-tail and left-tail  $P$ -values is  $1 + P_{\beta_0}(T = t_{obs})$  for the ordinary  $P$ -value but 1 for the mid- $P$ -value. Using the small-sample distribution, a 95% CI based on the mid- $P$ -value is determined by

$$P_{\beta_U}(T < t_{obs}) + (1/2) \times P_{\beta_U}(T = t_{obs}) = 0.025.$$

$$P_{\beta_L}(T > t_{obs}) + (1/2) \times P_{\beta_L}(T = t_{obs}) = 0.025.$$

Although the coverage probability is not guaranteed to be  $\geq 0.95$ , in practice it is usually close to that value and a bit conservative in an average sense.

To illustrate, suppose  $T$  is a binomial random variable. Using this construction with binomial probabilities and  $(1/2)$  replaced by 1.0 yields the standard Clopper–Pearson exact (conservative) 95% CI. Considering this method and the mid- $P$ -based CI over all the possible parameter values between 0 and 1, Figure 1 from Agresti and Gottard (2007) shows the quartiles of the coverage probabilities as a function of  $n$ . The median coverage probability (represented in each case by the middle of the three curves) is much closer to the nominal level for the mid- $P$ -based CI. It would be useful if standard software could provide mid- $P$ -based small-sample CIs at least for binomial proportions, odds ratios, and logistic regression parameters.

FIGURE 1. Quartiles of coverage probabilities for Clopper–Pearson (—) and mid-P (- - -) small-sample confidence intervals for binomial parameter, from Agresti and Gottard (2007).

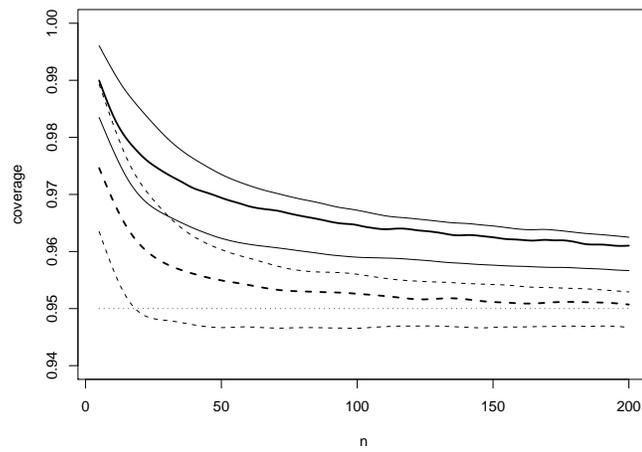


TABLE 1. Contingency table used to illustrate pseudo-score inferences

$y_1 = \text{Health Care}$	$y_2 = \text{Environment}$			Total
	1	2	3	
1	199	81	83	363
2	129	167	112	408
3	164	169	363	696
Total	492	417	558	1467

#### 4 Pseudo-Score Inference Using Pearson Chi-Squared

Although the method of inverting score tests to obtain CIs works well for basic parameters, it is often difficult or even infeasible to implement. A prime example is the set of models for which the likelihood function is not an explicit function of the model parameters. To illustrate, consider Table 1, based on responses in the 2006 General Social Survey in the U.S. to the questions “How successful is the government in (1) Providing health care for the sick? (2) Protecting the environment?” with categories 1 = successful, 2 = mixed, 3 = unsuccessful. To detect whether responses tend to be more positive for one question than the other, we could compare the marginal distributions using the cumulative logit marginal model for the responses  $(y_1, y_2)$ ,

$$\text{logit}[P(y_1 \leq j)] = \alpha_j, \quad \text{logit}[P(y_2 \leq j)] = \alpha_j + \beta, \quad j = 1, 2.$$

The multinomial log likelihood, in terms of cell probabilities  $\{\pi_{ij}\}$  and cell counts  $\{n_{ij}\}$ , is

$$L(\boldsymbol{\pi}) \propto \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \cdots \pi_{33}^{n_{33}},$$

but the model parameters refer to marginal probabilities. Other models for this table for which the score method would be difficult to implement are the random effects analog of this marginal model, the association model that specifies a common value for the four global odds ratios, and a model by which the mean for one response changes linearly across categories of the other response, for a particular choice of category scores.

For a multinomial model for cell counts  $\{n_i\}$  with ML fitted values  $\{\hat{\mu}_i\}$ , let  $\{\hat{\mu}_{i0}\}$  denote fitted values for a simpler “null” model (e.g., with a certain parameter  $\beta = \beta_0$ ). For testing the simpler model against the full model, the LR statistic is

$$G^2 = 2 \sum_i \hat{\mu}_i \log(\hat{\mu}_i / \hat{\mu}_{i0}).$$

The profile likelihood 95% CI for  $\beta$  is

$$\{\beta_0\} \text{ such that } G^2 \leq \chi_{1,0.05}^2.$$

This is available in standard software, such as in SAS with the LRCI option in PROC GENMOD and in R with the CONFINT function applied to an object such as the fit of a GLM.

We now propose a method that parallels this one, but using the Pearson statistic, with the purpose of making score-type CIs available when the score CI itself is not easily obtainable. Rao (1961) suggested the Pearson-type statistic for comparing models,

$$X^2 = \sum_i \frac{(\hat{\mu}_i - \hat{\mu}_{i0})^2}{\hat{\mu}_{i0}}.$$

This is a quadratic approximation for  $G^2$ . From a Taylor series expansion,  $X^2$  has the same limiting null distribution as  $G^2$  even under sparse asymptotics in which the number of cells in the contingency table grows with the sample size, as is the case when at least one explanatory variable is continuous (Haberman 1977). Since the score CI often out-performs the profile likelihood CI for simple contingency table problems with small  $n$ , as an alternative for general categorical data modeling we propose the CI for  $\beta$ ,

$$\{\beta_0\} \text{ such that } X^2 \leq \chi_{1,a}^2.$$

When the full model is saturated, this yields the score CI. When the model is unsaturated,  $X^2$  is not the score statistic. We refer to the test using  $X^2$  to compare models in that case as a *pseudo-score test* and the CI as a *pseudo-score confidence interval*. (In the case of a generalized linear model with canonical link function, see Lovison (2005) for a formula for the score statistic that resembles the Pearson statistic, being a quadratic form comparing fitted values for the two models.)

For Table 1, we can easily fit the marginal model for various fixed  $\beta_0$  (taking that value times the margin indicator as an offset) by using the R function `mph.fit` available from Prof. Joseph Lang at the Univ. of Iowa. The 95% pseudo-score CI for  $\beta$  is (0.2898, 0.5162). Here,  $n$  is large and results are similar to those for the profile likelihood CI of (0.2900, 0.5162).

We believe that pseudo-score methods are useful for two reasons: First, for some models, such as many models that are not generalized linear models with canonical link function, ordinary score methods are not practical but the pseudo-score methods are easily implemented. Second, when available, ordinary score inferences have been observed to perform well for discrete data. To check whether this is true for pseudo-score methods, we have conducted several simulations, comparing this method to Wald and profile likelihood CIs. To summarize, we have found that with small sample sizes, the pseudo score method performed either similarly to the profile likelihood interval or a bit better. Following is a discussion of one such simulation for which results seem to be typical.

Consider the cumulative logit marginal model we applied to Table 1, namely,  $\text{logit}[P(y_i \leq j)] = \alpha_j + \beta I(i = 2)$ . We simulated CIs for  $\beta$  for various val-

TABLE 2. Simulation result for interval estimation of  $\beta$  in marginal cumulative logit model with sample size  $n$  and underlying bivariate normal distribution with correlation 0.0, 0.4, or 0.8

$n$	Method	Correlation, with $\beta = 0.0$			Correlation, with $\beta = 0.5$		
		0.0	0.4	0.8	0.0	0.4	0.8
20	Pseudo score	0.925	0.920	0.940	0.921	0.913	0.947
	Profile like.	0.915	0.888	0.844	0.907	0.882	0.902
50	Pseudo score	0.943	0.941	0.927	0.950	0.939	0.937
	Profile like.	0.940	0.936	0.913	0.946	0.934	0.924

ues of  $\beta$ ,  $n$ , and association between the variables. For the cases  $\beta = 0$  and  $\beta = 0.5$ , Table 2 shows the estimated coverage probability (based on 100,000 simulations) for nominal 95% CIs for  $\beta$  when  $n = 20$  or 50 and the correlation = 0, 0.40, or 0.80 for an underlying bivariate normal distribution for producing joint probabilities having given margins. While there is no guarantee that similar behavior would occur for other models, the pseudo-score CI is simple to construct and shows promise of being a good, general-purpose method for inference with categorical response data.

## 5 Generalizations of Pseudo-Score Inference

This section presents generalizations of pseudo-score inference. We highlight some potential areas for future research.

### 5.1 Confidence Intervals Based on Power Divergence Statistics

For testing goodness of fit of a multinomial model with counts  $\{n_i\}$ , Cressie and Read (1984) presented a family of *power divergence statistics*,

$$P_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum n_i [(n_i/\hat{\mu}_i)^\lambda - 1], \quad \text{for } -\infty < \lambda < \infty.$$

These statistics have the same asymptotic null distribution as  $X^2$  and  $G^2$ . Mimicing the statistics for comparing a model to a simpler null model, we could use the power divergence statistic to test that a particular parameter takes value  $\beta_0$  (with corresponding fitted values  $\{\hat{\mu}_{i0}\}$ ),

$$P_\lambda = \frac{2}{\lambda(\lambda + 1)} \sum \hat{\mu}_i [(\hat{\mu}_i/\hat{\mu}_{i0})^\lambda - 1].$$

Cressie, Pardo, and Pardo (2003) used this statistic for hypothesis testing comparing pairs of nested loglinear models. We define a *power divergence*

*confidence interval* to be the set of  $\beta_0$  values having  $P_\lambda \leq \chi_{1,a}^2$ . This encompasses profile likelihood CIs ( $\lambda = 0$ ) and pseudo-score CIs ( $\lambda = 1$ ). For various models, perhaps the power divergence CI tends to work especially well for a particular  $\lambda$  value, in terms of having coverage probability near the nominal value.

## 5.2 Pseudo-Score Inference for Discrete Distributions

Suppose  $\{y_i, i = 1, \dots, n\}$  are independent observations assumed to have a specified discrete distribution. A Pearson-type statistic for comparing models has the form

$$X^2 = \sum_i \frac{(\hat{\mu}_i - \hat{\mu}_{i0})^2}{v(\hat{\mu}_{i0})} = (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0)' \hat{\mathbf{V}}_0^{-1} (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0),$$

where  $v(\hat{\mu}_{i0})$  denotes the estimated variance of  $y_i$  assuming the null distribution for  $y_i$  and  $\hat{\mathbf{V}}_0$  is the diagonal matrix containing such values (Lovison 2005). For some cases, the pseudo-score methods for multinomial responses may extend to parameters of models for discrete data using this generalized statistic. One context in which this statistic applies directly is inference about parameters of Poisson regression models.

## 5.3 Quasi-Likelihood Inference for Marginal Modeling

Pseudo-score inference using pairs of fitted values may extend to a quasi-likelihood analysis. A possible application is marginal modeling of clustered categorical responses. A popular approach for marginal modeling uses the method of generalized estimating equations (GEE). Because of the lack of a likelihood function with this method, Wald methods are commonly employed, together with a sandwich estimator of the covariance matrix of model parameter estimates. For binary data, let  $y_{it}$  denote observation  $t$  in cluster  $i$ , for  $t = 1, \dots, T_i$  and  $i = 1, \dots, n$ . Let  $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})'$  and let  $\boldsymbol{\mu}_i = E(\mathbf{y}_i) = (\mu_{i1}, \dots, \mu_{iT_i})'$ . Let  $\mathbf{V}_i$  denote the  $T_i \times T_i$  covariance matrix of  $\mathbf{y}_i$ .

For a particular marginal model, let  $\hat{\boldsymbol{\mu}}_i$  denote an estimate of  $\boldsymbol{\mu}_i$ , such as the ML estimate under the naive assumption that the  $\sum_i T_i$  observations as independent. Let  $\hat{\boldsymbol{\mu}}_{i0}$  denote the corresponding estimate under the constraint that a particular parameter  $\beta$  takes value  $\beta_0$ . Let  $\hat{\mathbf{V}}_{i0}$  denote an estimate of the covariance matrix of  $\mathbf{y}_i$  under this null model. The main diagonal elements of  $\hat{\mathbf{V}}_{i0}$  are  $\hat{\mu}_{it0}(1 - \hat{\mu}_{it0})$ ,  $t = 1, \dots, T_i$ . Separate estimation is needed for the null covariances, which are not part of the marginal model. With categorical explanatory variables, an estimate of  $\text{Cov}(y_{it}, y_{iu})$  is the sample mean value of  $(y_{at} - \hat{\mu}_{at0})(y_{au} - \hat{\mu}_{au0})$  for the set of all clusters  $a$  that have the same values of between-cluster explanatory variables as cluster  $i$ . This is also the sample estimate of the covariance for the multinomial

distribution for the  $2 \times 2$  joint distribution of  $(y_{at}, y_{au})$  for all such clusters. Now, consider

$$X^2 = \sum_i (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{i0})' \hat{\mathbf{V}}_{i0}^{-1} (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_{i0}).$$

With categorical explanatory variables,  $X^2$  applies to two sets of fitted marginal proportions for the contingency table obtained by cross classifying the multivariate binary response with the various combinations of explanatory variable values. The set of  $\beta_0$  values for which  $X^2 \leq \chi_{1,a}^2$  is a CI for  $\beta$ . Unlike the GEE approach, this method does not require using the sandwich estimator, which can be unreliable unless the number of clusters is quite large (Kauermann and Carroll 2001). Even with consistent estimation of  $\mathbf{V}_{i0}$ , however, the limiting null distribution of  $X^2$  need not be exactly chi-squared because the fitted values result from inefficient estimates. However, based on preliminary simulations, we conjecture that the chi-squared often provides a good approximation.

#### 5.4 Conditional and “Exact” Pseudo-Score Inference

When the sample size is very small or the number of parameters grows with the sample size, such as with stratified matched-pairs data, inference for canonical models for categorical responses is often applied to a conditional likelihood that eliminates nuisance parameters by conditioning on their sufficient statistics. This can also be done with pseudo-score methods. For example, when  $\{\hat{\boldsymbol{\mu}}_i\}$  and  $\{\hat{\boldsymbol{\mu}}_{i0}\}$  are obtained using the relevant conditional distributions, the set of  $\beta_0$  values for which  $X^2 \leq \chi_{1,a}^2$  is a *conditional pseudo-score confidence interval* for  $\beta$ . Likewise, for inferences about parameters that compare two unsaturated models using small-sample distributions, the pseudo-score statistic would be a sensible criterion that is an alternative to the likelihood-ratio statistic when the ordinary score statistic is difficult to obtain.

## 6 Pseudo-Score CIs that Adjust Wald CIs

Of the three types of tests that are inverted to construct CIs, the Wald is the least desirable. However, in some cases simple adjustments of Wald CIs approximate score CIs or have similar behavior. To illustrate, suppose  $y$  has a binomial distribution with parameter  $\pi$ , and let  $\hat{\pi} = y/n$ . Agresti and Coull (1998) noted that in the 95% case, finding all  $\pi_0$  such that  $|\hat{\pi} - \pi_0| / \sqrt{\pi_0(1 - \pi_0)/n} < 2$  provides the score CI of form  $M \pm 2s$  with

$$M = \left( \frac{n}{n+4} \right) \hat{\pi} + \left( \frac{4}{n+4} \right) \frac{1}{2} = \frac{y+2}{n+4}$$

and

$$s^2 = \frac{1}{n+4} \left[ \hat{\pi}(1 - \hat{\pi}) \left( \frac{n}{n+4} \right) + \frac{1}{2} \left( \frac{4}{n+4} \right) \right].$$

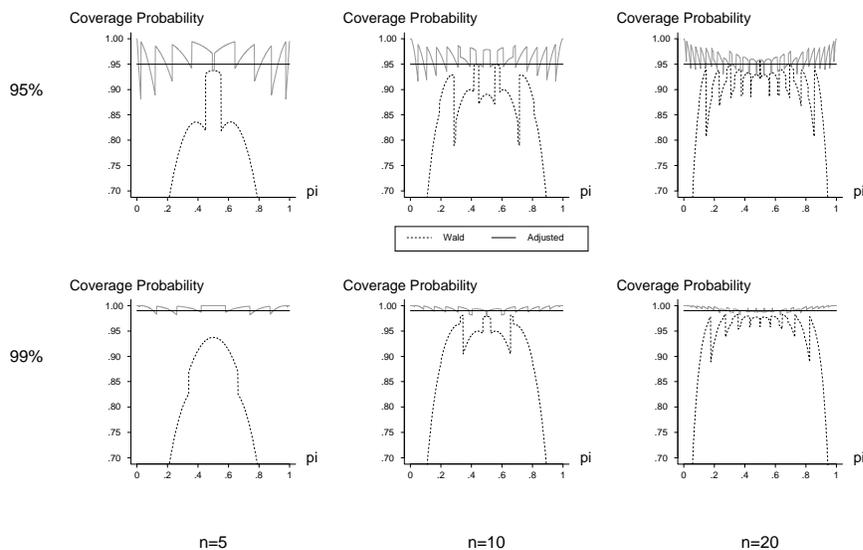


FIGURE 2. Coverage probabilities for Wald and Pseudo-Score (Adjusted Wald) confidence intervals for a binomial parameter

They used this to motivate the 95% *adjusted Wald CI*

$$\tilde{\pi} \pm 2.0 \sqrt{\tilde{\pi}(1 - \tilde{\pi})/\tilde{n}}$$

with  $\tilde{\pi} = (y + 2)/(n + 4)$  and  $\tilde{n} = n + 4$ . This has the same midpoint as the 95% score CI (when we round the normal percentile 1.96 to 2) but is slightly wider by Jensen's inequality, because the variance is found at the weighted average of  $\hat{\pi}$  and 1/2 instead of using a weighted average of variances. In fact, such simple adjustments have much improved performance. Figure 2 shows the coverage probabilities for the case of a single proportion, for various sample sizes and for 95% and 99% CIs. Agresti and Caffo (2000) showed that adding 4 outcomes yields a much better CI for the difference between two proportions for independent samples, and Agresti and Min (2005) showed a similar result for dependent samples.

Brown, Cai, and Das Gupta (2001) showed further evidence of the poor performance of the Wald method. For example, when  $\pi = 0.01$  or 0.99, the value of  $n_0$  needed in order for the coverage probability of a nominal

95% Wald CI to exceed 0.94 uniformly in  $n \geq n_0$  is about 8000, compared to 1 for the adjusted CI. The poor performance of the Wald CI is due to centering at the point estimate when the parameter space is bounded, not because the CI is too short. In fact, the Wald CI has greater length than an adjusted CI unless the parameters are relatively near the boundary of the parameter space.

The shrinkage form of the adjusted CIs also suggests that CIs resulting from the Bayesian approach can also perform well in a frequentist sense. This was shown with relatively diffuse prior distributions for a single proportion by Brown et al. (2001) and for the difference of proportions, relative risk, and odds ratio by Agresti and Min (2005).

## 7 Summary

For basic categorical data analyses, inverting the large-sample score test provides CIs having coverage probabilities near the nominal level. For small-sample distributions, inverting score tests based on the mid- $P$ -value also provides good CIs. Score CIs should be added to standard statistical software. Functions for the free software R are available for many such CIs (and Bayesian CIs) at [www.stat.ufl.edu/~aa/cda/software.html](http://www.stat.ufl.edu/~aa/cda/software.html).

Ordinary score tests and CIs are often infeasible, so we proposed a pseudo-score CI for a multinomial model parameter based on inverting a test using the Pearson statistic to compare the model to special cases in which the parameter takes fixed values. Our goal here was to present a simple unified method that performs well in a wide variety of settings and is simple for methodologists to implement with ordinary model-fitting software.

## References

- Agresti, A. (2003). Dealing with discreteness: Making ‘exact’ confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods in Medical Research*, **12**, 3-21.
- Agresti, A., Bini, M., Bertaccini, B., and Ryu, E. (2008). Simultaneous confidence intervals for comparing binomial parameters. *Biometrics*, to appear.
- Agresti, A., and Caffo, B. (2000). Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures, *The American Statistician*, **54**, 280-288.
- Agresti, A., and Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial parameters. *American Statistician*, **52**, 119-126.

- Agresti, A., and Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Applied Statistics*, **54**, 691-706.
- Agresti, A., and Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions, *Biometrics*, **57**, 963-971.
- Agresti A, and Min Y. (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics*, **3**, 379-386.
- Agresti, A., and Min, Y. (2005a). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2x2 contingency tables, *Biometrics*, **61**, 515-523.
- Agresti, A., and Min, Y. (2005b). Simple improved confidence intervals for comparing matched proportions, *Statistics in Medicine*, **24**, 729-740.
- Brown, L. D., T. T. Cai, and A. Das Gupta. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101-133.
- Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, **4**, 135-148.
- Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- Cressie, N., Pardo, L., and Pardo, M. (2003). Size and power considerations for testing loglinear models using  $\phi$ -divergence test statistics. *Statistica Sinica*, **13**, 555-570.
- Cressie, N., and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society*, **B 46**, 440-464.
- Cytel (2005). *StatXact 7 User Manual*, volumes 1 and 2, and *LogXact 7 User Manual*. Cambridge, Massachusetts: Cytel Inc.
- Geyer, C. J., and Meeden, G. D. (2005). Fuzzy and randomized confidence intervals and P-values. *Statistical Science*, **20**, 358-366. , [.1in]
- Haberman, S. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, **5**, 1148-1169.
- Kauerman, G., and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, **96**, 1387-1396.
- Koehler, K., and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, **75**, 336-344.

- Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics*, **40**, 513-517.
- Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association* **56**, 223-234.
- Lang, J. B. (2005). Profile confidence intervals for contingency table parameters. Technical report no. 351, Department of Statistics and Actuarial Science, University of Iowa.
- Lovison G. (2005). On Rao score and Pearson  $X^2$  statistics in generalized linear models. *Statistical Papers*, **46**, 555-574.
- Mee, R. W. (1984). Confidence bounds for the difference between two probabilities (letter). *Biometrics*, **40**, 1175-1176.
- Miettinen, O., and Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, **4**, 213-226.
- Newcombe, R. (1998a). Two-sided confidence intervals for the single proportion. *Statistics in Medicine*, **17**, 857-872.
- Newcombe, R. (1998b). Interval estimation for the difference between independent proportions: Comparison of eleven methods, *Statistics in Medicine*, **17**, 873-890.
- Pearson, E. S. (1950). On questions raised by the combination of tests based on discontinuous distributions. *Biometrika*, **37**, 383-398.
- Rao, C. R. (1961). A study of large sample test criteria through properties of efficient estimates. *Sankhya*, **A23**, 25-40.
- Ryu, E., and Agresti, A. (2008). Modeling and inference for an ordinal effect size measure. *Statistics in Medicine*, to appear.
- Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic. In *Science and Statistics: A Festschrift for Terry Speed*, ed. D. R. Goldstein, IMS Lecture Notes-Monograph Series, Vol. 40, pp. 115-126, Institute of Mathematical Statistics, Hayward, CA.
- Stevens, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika*, **37**, 117-129.
- Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, **17**, 891-908.
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.



# Bayesian Item Response Models For Complex Survey Data

Jean-Paul Fox<sup>1</sup>

<sup>1</sup> Twente University, Faculty of Behavioural Sciences, Enschede, The Netherlands

**Abstract:** IRT methods have become an important tool in analyzing large-scale survey data. The application of the common IRT models raises several issues like the implicit assumption of conditionally independent observations, handling collateral information, and dealing with misreporting. It is shown that the Bayesian IRT approach leads to a very flexible modeling framework for analyzing large-scale survey data. The Bayesian IRT models are extended to provide a better fit to the data and to extract richer information from the survey data. A variety of extensions will be discussed.

**Keywords:** Bayesian; Complex Surveys; IRT; Hierarchical Models

## 1 Introduction

The common item response theory (IRT) methods (Lord and Novick, 1968) are standard tools for the analysis of large-scale survey data. For example, in educational survey research, the National Assessment of Educational Progress (NAEP) is primarily focused on scaling the performances of a sample of students in a subject area (e.g., mathematics, reading, science) on a single common scale, and measuring change in educational performance over time. Further, the Organization for Economic Cooperation and Development (OECD) organizes the Program for International Student Assessment (PISA) that is focused on measuring and comparing the abilities in reading, mathematics, and science of 15-year-old pupils over 32 countries in 2000. Another example is the large international survey Trends in International Mathematics and Science Study (TIMSS) conducted by the International Association for the Evaluation of Educational Achievement (IEA) also to measure trends in students' mathematics and science performances.

IRT methods provide a set of techniques for estimating individual ability (e.g., attitude, behavior, performance) levels and item characteristics from observed discrete multivariate response data. The ability levels cannot be observed directly but are measured via a questionnaire or test. Item response theory (IRT) is, in particular, useful for large-scale survey response data where (1) the observations often have an ordinal character, (2) the sampling designs are complex with individuals responding to different sets

(booklets) of questions, (3) booklet effects are present (the performance on items depends on an underlying latent variable but also on the positioning of the items in a test), and (4) missing data occur. The essential idea of IRT is that the effects of the persons and the items on the response data are modeled by separate sets of parameters. The person parameters are usually referred to as the latent variables, and the item parameters are usually labeled item difficulties or thresholds, item discrimination parameters and guessing parameters.

The common IRT models are not directly applicable to analysing large-scale survey data for comparative research. There are several measurement issues connected to survey research that need to be addressed since ignoring them may lead to inferential errors. Further, there is often a wide variety of additional information available besides the observed response data. More accurate inferences can be made when the different sources of information can be combined.

Three topics will be considered. First, the multistage sampling design since respondents are nested in classrooms, classrooms in schools, schools within countries and so on. In a Bayesian modeling approach, a hierarchical population distribution for the respondents is easily specified that accounts for the fact that respondents are nested within clusters. Common IRT models assume a priori independence between individual abilities but homogeneity of results of individuals in the same school is to be expected since pupils in the same school share common experiences. Second, collateral information can be used when response times are observed besides the response patterns. Response times on test items are easily collected in modern computerized testing. When collecting both (binary) responses and (continuous) response times on test items, it is possible to measure the accuracy and speed of respondents. The observed response times can be informative with respect to the latent individual abilities. Third, the collection of data through surveys on personal and sensitive issues may lead to answer refusals and false responses, making inferences difficult. Respondents often have a tendency to agree rather than disagree (acquiescence) and a tendency to give social desirable answers (social desirability). A multivariate randomized response sampling design can be used to improve the quality of the survey data. It is shown that a Bayesian IRT model is easily adjusted to handle the multivariate randomized (item) response data.

## 2 Bayesian IRT Models for Binary Response Data

An IRT model for binary response data defines the probability of a correct or positive response to item  $k$  ( $k = 1, \dots, K$ ) for individual  $i$  ( $i = 1, \dots, n$ ) given the item characteristics, denoted as  $\xi_k = (a_k, b_k)^t$ , and the individual ability level,  $\theta_i$ . The well known probit version of the two-parameter IRT model is also known as the normal ogive model where the probability of

success is defined via a cumulative normal distribution,

$$P(Y_{ik} = 1 \mid \theta_i, a_k, b_k) = \Phi(a_k \theta_i - b_k) = \int_{-\infty}^{a_k \theta_i - b_k} \varphi(z) dz, \quad (1)$$

where  $\Phi(\cdot)$  and  $\varphi(\cdot)$  is the cumulative normal distribution function and the normal density function, respectively. The  $a_k$  is referred to as the discrimination parameter and the  $b_k$  as the item difficulty parameter.

The Bayesian approach towards IRT modeling starts with the specification of prior distributions. In most cases there is not much information about the values of the item parameters. Without a priori knowledge to distinguish the item parameters it is reasonable to assume a common distribution for them.

An intuitive assumption of an IRT model is that the higher a respondent's ability level the more likely it is the respondent scores well on each item. This so-called monotonicity assumption implies that  $P(Y_{ik} = 1 \mid \theta_i)$  is nondecreasing in  $\theta_i$ , for binary response data, which is satisfied when the discrimination parameter is restricted to be positive. A common dependency structure of item parameters is specified via a hierarchical structured prior. A multivariate normal prior distributed prior is assumed for the item parameters. It follows that,

$$(a_k, b_k)^t \sim \mathcal{N}(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi) I_{\mathcal{A}_k}(a_k), \quad (2)$$

where the set  $\mathcal{A}_k = \{a_k \in \mathcal{R}, a_k > 0\}$  with hyper prior parameters

$$\boldsymbol{\Sigma}_\xi \sim \mathcal{IW}(\nu, \boldsymbol{\Sigma}_0) \quad (3)$$

$$\boldsymbol{\mu}_\xi \mid \boldsymbol{\Sigma}_\xi \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\xi / K_0), \quad (4)$$

for  $k = 1, \dots, K$ . The truncated multivariate Normal distribution in Equation (2) is the exchangeable prior for the set of K item parameters  $\boldsymbol{\xi}_k$ . The joint hyper prior distribution for  $(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi)$  is a Normal inverse Wishart distribution, denoted as  $\mathcal{IW}$ , with parameters  $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 / K_0; \nu, \boldsymbol{\Sigma}_0)$  where  $K_0$  denotes the number of prior measurements, and  $\nu$  and  $\boldsymbol{\Sigma}_0$  describe the degrees of freedom and scale matrix of the inverse-Wishart distribution. These parameters are usually fixed at specified values. A proper vague prior is specified with  $\boldsymbol{\mu}_0 = \mathbf{0}$ ,  $\nu = 2$ , a diagonal scale matrix  $\boldsymbol{\Sigma}_0$  with elements 100 and  $K_0$  a small number.

In general, the respondents are assumed to be sampled independently and identical distributed from a large population. So, an independent prior distribution is specified for the ability parameter,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \quad (5)$$

for  $i = 1, \dots, n$ . A Normal inverse Gamma prior is the conjugate prior for the Normal distribution with unknown mean and variance. Therefore, a

hyper prior distribution is specified as,

$$\sigma_\theta^2 \sim \mathcal{IG}(g_1, g_2) \quad (6)$$

$$\mu_\theta \mid \sigma_\theta^2 \sim \mathcal{N}(\mu_0, \sigma_\theta^2/n_0), \quad (7)$$

where  $g_1$  and  $g_2$  are the parameters of the inverse Gamma distribution denoted as  $\mathcal{IG}$  and  $n_0$  presents the number of prior measurements.

### 3 Heterogeneity of the Respondent Population

Educational survey research is often concerned with exploring differences within and between schools. The objective is to investigate the relationship between explanatory and outcome factors. This involves choosing an outcome variable, such as student's ability, and studying differences among schools after adjusting for relevant background variables. A general acceptable statistical model in the assessment requires the deployment of multilevel analysis techniques. The student's ability is considered to be an outcome variable of the multilevel regression model. This outcome variable is not directly observable but is known to be a latent variable. The idea is to integrate the IRT model for measuring the individual abilities with a (structural) multilevel model that explains differences at different levels of abilities. The IRT measurement model defines the relationship between the ability and the corresponding observed response data. The structural multilevel model describes the nested structure of individual abilities in the population.

The respondents at level-1 are nested in clusters and indexed  $i = 1, \dots, n_j$  for  $j = 1, \dots, J$  clusters. Let level-1 respondent-specific covariates be denoted by  $\mathbf{x}_{ij}$ . The level-1 prior distribution for the ability parameter  $\theta_i$  is specified as

$$\theta_{ij} \mid \beta_j \sim \mathcal{N}(\mathbf{x}_{ij}^t \beta_j, \sigma_\theta^2), \quad (8)$$

and the level-2 covariates are denoted as  $\mathbf{w}_{qj}$  for  $q = 0, \dots, Q$ , such that the level-2 prior is specified as

$$\beta_j \sim \mathcal{N}(\mathbf{w}_j \boldsymbol{\gamma}, \mathbf{T}), \quad (9)$$

An inverse-gamma prior distribution and an inverse-Wishart prior distribution is specified for the variance components  $\sigma_\theta^2$  and  $\mathbf{T}$  respectively. The extension to more levels is easily made. The IRT measurement model with a multilevel population model for the ability parameters is called a multilevel IRT model (MLIRT). An MCMC algorithm can be used to concurrently estimate all model parameters (e.g., Fox, 2007; Fox and Glas, 2001).

Several advantages can be given of the MLIRT modeling framework. The multilevel population model parameters are estimated from the item response data without having to condition on estimated ability parameters. In

empirical multilevel studies, estimated ability parameters are often considered to be measured without an error and treated as an observed outcome variable. Ignoring the uncertainty regarding the estimated abilities may lead to biased parameter estimates and the statistical inference may be misleading. The modeling framework allows the incorporation of explanatory variables at different levels of hierarchy. The inclusion of explanatory information can be important in various situations. The use of explanatory information may lead to more accurate item parameter estimates. Another related advantage of the model is that it can handle incomplete data in a very flexible way.

#### 4 The Use of Response Times as Collateral Information

Bassili and Fletcher (1991) introduced a methodology for measuring accurately response time within the context of a telephone survey. They showed how response times can be measured precisely and reliably and that the data from such measurement offer insight into the processes underlying survey responses applied to most types of computer-assisted telephone interviewing (CATI). Nowadays, computer-based testing has become a popular mode for retrieving information and response times can be collected automatically. In educational research, when the test takers' responses as well as their response times on the items are recorded, the relationship between response times and response accuracies can be explored. This relationship is complex at different hierarchical levels and it takes the form of a tradeoff between speed and accuracy at the level of a fixed person but may become a positive correlation for a population of test takers. The response times can provide information about the items' characteristics and the individual response process. More specific, they can be used to identify bad items using, for example, the average response times as an indicator of question problems, and they may serve as indicators of uncertainty and response error. This way, the response times contain information about the item and person characteristics.

A log-normal distribution is used taking account of the natural lower-bound at zero to model the response times. Each respondent complete the items at a certain level of speed denoted as  $\zeta_i$ . The time needed to complete an item  $k$  also depends on item characteristic parameters. They are denoted as  $\phi_k$  and  $\lambda_k$ , and can be seen as a discrimination and time-intensity parameter, respectively. The log of the response time,  $\log T_{ik}$ , is normal distributed with mean  $-\phi_k\zeta_i + \lambda_k$  and variance  $\sigma_t^2$ . It follows that

$$P(t_{ik} \leq t'_{ik}) = \Phi((\log t'_{ik} - (\lambda_k - \phi_k\zeta_i)) / \sigma_t), t'_{ik} > 0. \quad (10)$$

Hence, increasing the time intensity  $\lambda_k$  leads to a positive shift of the location of the time distribution on the item. Likewise, an increase in the speed parameter  $\zeta_i$  leads to a negative shift.

Assume the normal ogive response model, Equation (1), for the response data for measuring the ability levels. At the individual level, a bivariate normal distribution is defined for the ability and speed parameter of the test taker,

$$\begin{pmatrix} \theta_i \\ \zeta_i \end{pmatrix} \sim \mathcal{N}_2 \left( \begin{pmatrix} \mu_\theta \\ \mu_\zeta \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \rho \\ \rho & \sigma_\zeta^2 \end{pmatrix} \right), \quad (11)$$

where parameter  $\rho$  denotes the covariance between the person parameters. This population distribution reflects the individual speed and ability levels in the population of test takers. This conjoint Bayesian IRT model constitutes a two-parameter normal ogive, Equation (1), for the multivariate response patterns and a normal model for the multivariate log transformed response times, Equation (10), and, at a lower level, a multivariate normal model is specified for the underlying ability and speed parameters, Equation (11). The model enables the simultaneous analysis of speed and accuracy given response times and patterns. More about MCMC methods for parameter estimation can be found in Fox, Klein Entink, van der Linden (2007), and van der Linden (2007) that also includes a detailed description of the model.

## 5 Asking Sensitive Questions

Survey researchers that are dealing with sensitive topics are often confronted with misreporting of respondents leading to biased estimates. The sensitive questions asked in the survey may lead to social desirable response behavior where respondents edit the information they report to avoid embarrassing themselves. Sensitive questions can also be seen as intrusive by the respondents or raise concerns about the possible repercussions of disclosing the information. The extent of misreporting depends on the design of the survey and whether the respondent has anything embarrassing to report. Asking sensitive questions in survey research usually affects the response rates, the item nonresponse rates, and the response accuracy. Self-administration, collecting the data in private, and confidentiality assurances are several design features that positively influences the accuracy of reports on sensitive topics. Warner (1965) introduced the randomized response technique for improving the accuracy of estimates from survey data. A popular variation on Warner's method is the unrelated-question method of Greenberg, Abu-Ela, Simmons, and Horvitz (1969) where the essential idea is that the interviewer is unaware whether the respondent is answering the sensitive question or the non-sensitive question. A randomizing device (dice, coin) is used such that with probability  $p_1$  a respondent is confronted with the sensitive question. The univariate randomized response technique enables the computation of (aggregated) estimated proportions without revealing the significance of the individual answers.

Fox (2005) introduced a multivariate randomized response technique with the two-parameter normal ogive model in Equation (1) as the response model for the randomly selected question. Let  $Y_{ik}$  and  $\tilde{Y}_{ik}$  denotes the observed randomized response and the latent response to the sensitive question, respectively, of respondent  $i$  to item  $k$ . The probability of observing a positive randomized response equals,

$$\begin{aligned} P(Y_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k) &= p_1 P(\tilde{Y}_{ik} = 1 \mid \theta_i, \boldsymbol{\xi}_k) + (1 - p_1)p_{2,i} \\ &= p_1 \Phi(a_k \theta_i - b_k) + (1 - p_1)p_{2,i}, \end{aligned} \quad (12)$$

where  $p_{2,i}$  denotes the known probability of a positive response to the non-sensitive question. In an alternative method the response to the non-sensitive question is simulated via a randomizing device that determines the respondent's answer and  $p_{2,i}$  is defined by the properties of the randomizing device.

The multivariate randomized response model makes it possible to measure the underlying sensitive behavior  $\theta_i$  of the respondents. At a lower level, the underlying sensitive behavior of the respondent can be related to other respondent or group characteristics. Therefore, assume the structural multilevel model for  $\theta_{ij}$ , Equation (8) and (9). The likelihood of interest of  $\boldsymbol{\Omega} = (\boldsymbol{\xi}, \sigma_\theta^2, \boldsymbol{\gamma}, \mathbf{T})$  given the randomized response data can be expressed as

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\Omega}) &= \prod_{j=1}^J \left[ \int \left[ \prod_{i=1|j}^{n_j} \int \prod_{k=1}^K [p_1 \Phi(a_k \theta_{ij} - b_k) + (1 - p_1)p_{2,i}]^{Y_{ijk}} \right. \right. \\ &\quad \left. \left. [p_1 (1 - \Phi(a_k \theta_{ij} - b_k)) + (1 - p_1)(1 - p_{2,i})]^{1 - Y_{ijk}} \right] \right. \\ &\quad \left. p(\theta_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}_j, \sigma_\theta^2) d\theta_{ij} \right] p(\boldsymbol{\beta}_j \mid \mathbf{w}_j, \boldsymbol{\gamma}, \mathbf{T}) d\boldsymbol{\beta}_j. \end{aligned}$$

MCMC methods makes it possible to estimate simultaneously all parameters (Fox, 2005).

## 6 Conclusions

The Bayesian IRT framework provides a set of powerful tools for the analysis of large-scale complex survey data. The Bayesian IRT model can be extended in different ways to handle measurement issues involved in large-scale survey research. It is shown that the population distribution of the respondents is easily extended to take account of a nested structure. Additional information can be incorporated via prior specifications. The framework can also be extended to a multivariate framework with different link functions to relate multivariate discrete and/or continuous observations with multiple underlying latent variables. This makes it possible to conduct

a simultaneous analysis of multiple tests each measuring a different latent variable with an underlying correlation structure. The framework can handle different complex sampling strategies to collect reliable response data which includes the randomized response sampling design.

MCMC methods can be used to estimate simultaneously the Bayesian IRT model parameters. The MCMC estimation methods make it possible to add additional complexity in a straightforward way. This includes the specification of different priors, constraints on parameters, and different distributional assumptions. The powerful estimation methods can also be used for the computation of a Deviance Information Criteria that can be used for comparing the fit of different Bayesian IRT models.

## References

- Bassili, J.N., and Fletcher, J.F. (1991). Response-time measurement in survey research. A method for CATI and a new look at nonattitudes. *The Public Opinion Quarterly*, **55**, 331-346.
- Greenberg, B.G., Abul-Ela, A., Simmons, W.R., and Horvitz, D.G. (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association*, **64**, 520-539.
- Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics*, **30**, 189-212.
- Fox, J.-P. (2007). Multilevel IRT modeling in practice. *Journal of Statistical Software*, **20**, Issue 5.
- Fox, J.-P., and Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, **66**, 269-286.
- Fox, J.-P., Klein Entink, R.H., and van der Linden, W.J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, **20**, Issue 7.
- Lord, F.M., and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- van der Linden, W.J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, **72**, 287-308.
- Warner, S.L. (1965). Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

# Fast Sparse Regression and Classification

Jerome H. Friedman<sup>1</sup>

<sup>1</sup> Department of Statistics, Stanford University, Stanford, CA 94305  
(jhf@stanford.edu)

**Abstract:** Regularized regression and classification methods fit a linear model to data, based on some loss criterion, subject to a constraint on the coefficient values. As special cases, ridge-regression, the lasso, and subset selection all use squared-error loss with different particular constraint choices. For large problems the general choice of loss-constraint combinations is usually limited by the computation required to obtain the corresponding solution estimates, especially when non convex constraints are used to induce very sparse solutions. A fast algorithm is presented that produces solutions that closely approximate those for any convex loss and a wide variety of convex and non convex constraints, permitting application to very large problems. The benefits of this generality are illustrated by examples.

**Keywords:** Regularization, variable selection, bridge-regression, lasso, elastic net,  $l_p$ -norm penalization.

## 1 Introduction

Linear structural models are among the most popular for fitting data. One is given  $N$  observations of the form

$$\{y_i, \mathbf{x}_i\}_1^N = \{y_i, x_{i1}, \dots, x_{in}\}_1^N \quad (1)$$

considered to be a random sample from some joint (population) distribution with probability density  $p(\mathbf{x}, y)$ . The random variable  $y$  is the “outcome” or “response” and  $\mathbf{x} = \{x_1, \dots, x_n\}$  are the predictor variables. These predictors may be the original measured variables and/or selected functions constructed from them. The goal is to estimate the joint values for the parameters  $\mathbf{a} = \{a_0, a_1, \dots, a_n\}$  of the linear model

$$F(\mathbf{x}; \mathbf{a}) = a_0 + \sum_{j=1}^n a_j x_j \quad (2)$$

for predicting  $y$  given  $\mathbf{x}$ , that minimize the expected loss (“risk”)

$$R(\mathbf{a}) = E_{\mathbf{x}, y} L(y, F(\mathbf{x}; \mathbf{a})) \quad (3)$$

over future predictions  $\mathbf{x}, y \sim p(\mathbf{x}, y)$ . Here  $L(y, F)$  is a loss criterion that specifies the cost of predicting the value  $F$  when the actual value is  $y$ . Popular loss criteria include squared-error

$$L(y, F) = (y - F)^2, \quad (4)$$

and Bernoulli negative log-likelihood

$$L(y, F) = \log(1 + e^{-yF}), \quad y \in \{-1, 1\} \quad (5)$$

associated with logistic regression. For a specified loss criterion the optimal parameter values are from (3)

$$\mathbf{a}^* = \arg \min_{\mathbf{a}} R(\mathbf{a}). \quad (6)$$

Since the population probability density  $p(\mathbf{x}, y)$  is unknown, a common practice is to substitute an empirical estimate of the expected value in (3) based on the available data (1) yielding

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a}) \quad (7)$$

as an estimate for  $\mathbf{a}^*$ , where

$$\hat{R}(\mathbf{a}) = \frac{1}{N} \sum_{i=1}^N L \left( y_i, a_0 + \sum_{j=1}^n a_j x_{ij} \right). \quad (8)$$

## 2 Regularization

It is well known that  $\hat{\mathbf{a}}$  (7) (8) often provides poor estimates of  $\mathbf{a}^*$ ; that is  $R(\hat{\mathbf{a}}) \gg R(\mathbf{a}^*)$ . This is especially the case when the sample size  $N$  is not large compared to the number of parameters  $(n + 1)$ . This is caused by the high variability of the estimates (7) when (8) is evaluated on different random samples drawn from the population distribution. A common remedy is to modify (7) in order to stabilize the estimates by placing a restriction on the joint solution values. That is,

$$\hat{\mathbf{a}}(t) = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a}) \quad \text{s.t.} \quad P(\mathbf{a}) \leq t. \quad (9)$$

Here  $P(\mathbf{a})$  is a non negative function of the parameters specifying the form of the constraint and  $t \geq 0$  regulates its strength. Setting  $t = 0$  produces maximum restriction by requiring the solution values to exactly satisfy  $P(\mathbf{a}) = 0$ , thereby producing the least variance. Setting  $t \geq P(\hat{\mathbf{a}})$  produces the unrestricted solution (7) with maximal variance. Intermediate values  $0 < t < P(\hat{\mathbf{a}})$  provide degrees of restriction between these two extremes,

thereby regulating the stability (variance) of the estimates  $\hat{\mathbf{a}}(t)$  with respect to different training samples (1) drawn from  $p(\mathbf{x}, y)$ .

For a given data set (1), loss criterion  $L(y, F)$  (3) (8), and constraint function  $P(\mathbf{a})$ , the solution to (9) depends on the value chosen for  $t$ . Varying its value induces a family of solutions, each member being indexed by a particular value of  $t \in [0, P(\hat{\mathbf{a}})]$ . This same family of solutions can be obtained through the equivalent (penalized) formulation of (9)

$$\hat{\mathbf{a}}(\lambda) = \arg \min_{\mathbf{a}} [\hat{R}(\mathbf{a}) + \lambda \cdot P(\mathbf{a})] \quad (10)$$

where  $P(\mathbf{a})$  is the constraining function in (9), here called the penalty, and  $\lambda > 0$  regulates its strength. Setting  $\lambda = \infty$  produces the totally constrained solution ( $t = 0$ ) whereas  $\lambda = 0$  yields the unrestricted solution ( $t \geq P(\hat{\mathbf{a}})$ ). Each value of  $0 \leq \lambda \leq \infty$  in (10) produces one of the solutions  $0 \leq t \leq P(\hat{\mathbf{a}})$  in (9) with smaller values of  $\lambda$  corresponding to larger values of  $t$ . Thus (10) produces a family of estimates in which each member of the family is indexed by a particular value for the strength parameter  $\lambda$ . This family lies on a one-dimensional path of finite length in the  $(n + 1)$ -dimensional space of all joint parameter values.

## 2.1 Model selection

The optimal parameter values  $\mathbf{a}^*$  (6) also represent a point in the parameter space. For a given penalty, the goal is to find a point on its corresponding path  $\hat{\mathbf{a}}(\lambda^*)$  that is closest to  $\mathbf{a}^*$ , where distance is characterized by the prediction risk (3)

$$D(\mathbf{a}, \mathbf{a}^*) = R(\mathbf{a}) - R(\mathbf{a}^*). \quad (11)$$

This is a classic model selection problem where one attempts to obtain an estimate  $\hat{\lambda}$  for the optimal value of the strength parameter

$$\lambda^* = \arg \min_{0 \leq \lambda \leq \infty} R(\hat{\mathbf{a}}(\lambda)) \quad (12)$$

through

$$\hat{\lambda} = \arg \min_{0 \leq \lambda \leq \infty} \tilde{R}(\hat{\mathbf{a}}(\lambda)) \quad (13)$$

where  $\tilde{R}(\mathbf{a})$  is a surrogate model selection criterion whose minimum is intended to approximate that for the actual risk (3).

There are a wide variety of model selection criteria each developed for a particular combination of loss (3) and penalty  $P(\mathbf{a})$ . Among the most general, applicable to any loss-penalty combination, is cross-validation. The data are randomly partitioned into two subsets (learning and test). The path is constructed using only the learning sample. The test sample is then used as an empirical surrogate for the population density  $p(\mathbf{x}, y)$  to compute the corresponding (estimated) risk in (3). These estimates are then used in (13) to obtain the estimate  $\hat{\lambda}$ . Sometimes the risk used in (13) is estimated by averaging over several ( $K$ ) such partitions (“ $K$ -fold” cross-validation).

## 2.2 Penalty Selection

Given a model selection procedure, the goal is to construct a path  $\hat{\mathbf{a}}(\lambda)$  in parameter space such that some of the points on that path are close to the point  $\mathbf{a}^*$  (6) representing the optimal solution. If no points on the path come close to  $\mathbf{a}^*$ , as measured by (11), then no model selection procedure can produce accurate estimates  $\hat{\mathbf{a}}(\hat{\lambda})$ . Since the path produced by (10) depends on the data, different randomly drawn data sets (1) will produce different paths for the same penalty. Thus the paths are themselves random, and one seeks a penalty  $P(\mathbf{a})$  that produces paths  $\hat{\mathbf{a}}(\lambda)$  such that

$$[E_T R(\hat{\mathbf{a}}(\lambda^*)) - R(\mathbf{a}^*)] / R(\mathbf{a}^*) = \text{small} \quad (14)$$

with  $T$  being repeated data samples (1) drawn randomly from the joint density  $p(\mathbf{x}, y)$ , and  $\lambda^*$  is given by (12). This will depend on the particular  $\mathbf{a}^*$  (6) associated with the application. Therefore, penalty choice is governed by whatever is known about the properties of  $\mathbf{a}^*$ .

## 2.3 Sparsity

One property of  $\mathbf{a}^*$  that is often suspected is sparsity. That is, only a small fraction of the input variables  $\{x_j\}_1^n$  are influencing predictions, with the identities of those influential variables being unknown. The degree of sparsity  $S(\mathbf{a})$  of a parameter vector  $\mathbf{a}$  can be defined as

$$S(\mathbf{a}) = \frac{1}{n} \sum_{k=1}^n I(|a_k| \leq \eta \cdot \max_j |a_j|) \quad (15)$$

with  $\eta \ll 1$ . If the predictor variables are all standardized to have similar scales then  $S(\mathbf{a}^*)$  represents the fraction of non influential variables characterizing the problem.

If  $\hat{\mathbf{a}}(\lambda^*) \simeq \mathbf{a}^*$  (14) then  $S(\hat{\mathbf{a}}(\lambda^*)) \simeq S(\mathbf{a}^*)$ , and in the absence of other information it is reasonable to choose a penalty that produces solutions  $\hat{\mathbf{a}}(\lambda)$  with sparsity similar to that of  $\mathbf{a}^*$  at  $\lambda = \lambda^*$ . Since the actual sparsity of  $\mathbf{a}^*$  is generally unknown, one can define a family of penalties  $P_\gamma(\mathbf{a})$ , where  $\gamma$  indexes particular penalties in the family that produce solutions of differing sparseness, and then use model selection (Section 2.1) to jointly choose good values for  $\gamma$  and  $\lambda$ . That is,

$$\hat{\mathbf{a}}_\gamma(\lambda) = \arg \min_{\mathbf{a}} [\hat{R}(\mathbf{a}) + \lambda \cdot P_\gamma(\mathbf{a})] \quad (16)$$

$$(\hat{\gamma}, \hat{\lambda}) = \arg \min_{\gamma, \lambda} \tilde{R}(\hat{\mathbf{a}}_\gamma(\lambda)). \quad (17)$$

This approach is referred to as “bridge-regression” (Frank and Friedman 1993).

**Power family** One such family of penalties is the power family defined as

$$P_\gamma(\mathbf{a}) = \sum_{j=1}^n |a_j|^\gamma; \quad \gamma \geq 0. \quad (18)$$

This is the  $l_\gamma$ -norm of the parameter vector  $\mathbf{a}$  raised to the  $\gamma$  power.

Using squared-error loss (4), special cases of (8) (10) (18) include several popular regularized regression methods, namely  $\gamma = 2$ : ridge-regression,  $\gamma = 1$ : lasso,  $\gamma = 0$ : all-subsets regression. Ridge-regression (Hörel and Kennard 1970) produces dense solutions,  $S(\hat{\mathbf{a}}(\lambda)) \simeq 0$  (15), over its entire path  $\infty \leq \lambda \leq 0$  while heavily shrinking the coefficient absolute values  $\{|\hat{a}_j(\lambda)| \ll |a_j^*|\}_1^n$  for larger values of  $\lambda$ . At the other extreme, all-subsets regression produces the sparsest solutions along its path (set of distinct points) by forcing many of the coefficient estimates to be zero and applying no shrinkage to the non zero estimates. The number of non zero coefficient estimates is regulated by the value of  $\lambda$ ; larger values of  $\lambda$  produce fewer non zero coefficients. The lasso (Tibshirani 1996) produces paths intermediate between these two extremes, setting some coefficients to zero and applying shrinkage to the absolute values of the others. As  $\lambda$  increases along the path both the degree of shrinkage and the number of zero valued coefficients increase.

For  $0 \leq \gamma \leq 2$  the power family (18) represents a continuum of penalties between all-subsets regression (sparsest solutions) and ridge-regression (dense solutions). For  $\gamma > 1$  all coefficient estimates are strictly non zero at all points along the path,  $\{|\hat{a}_j(\lambda)| > 0\}_1^n$  for  $0 \leq \lambda < \infty$ . However, their dispersion (coefficient of variation) at corresponding path points decreases with increasing  $\gamma$ . Note that for  $\gamma \geq 1$  all penalties in the power family are convex functions of their argument  $\mathbf{a}$ , so that for convex risk  $\hat{R}(\mathbf{a})$  (8) the problems represented by (16) are convex optimizations. For  $\gamma < 1$  the penalties are non convex requiring (more difficult) non convex optimization techniques.

**Generalized elastic net** The power family (18) is not the only possibility for bridging all-subsets and ridge regression. For bridging the lasso and ridge-regression Zou and Hastie 2005 proposed the elastic net family of penalties which can be expressed as

$$P_\beta(\mathbf{a}) = \sum_{j=1}^n (\beta - 1) a_j^2 / 2 + (2 - \beta) |a_j|; \quad 1 \leq \beta \leq 2. \quad (19)$$

Here the parameter  $\beta$  indexes family members with  $\beta = 2$  yielding ridge-regression and  $\beta = 1$  the lasso. For  $1 < \beta < 2$ , penalties in this family represent a mixture of the ridge and lasso penalties generating alternatives in between these two extremes.

An extension of this family to non convex members producing sparser solutions than the lasso is

$$P_\beta(\mathbf{a}) = \sum_{j=1}^n \log((1 - \beta) |a_j| + \beta); \quad 0 < \beta < 1. \quad (20)$$

As  $\beta \rightarrow 0$  this approaches the all-subsets penalty ( $\gamma = 0$  in (18)) and as  $\beta \rightarrow 1$  it yields the lasso penalty ( $\gamma = 1$  in (18)). Values of  $\beta$  between these extremes bridge all-subsets and the lasso, so that the entire family (19) (20) bridges all-subsets and ridge-regression for  $0 < \beta \leq 2$ .

For the power family (18) members indexed by a value for  $\gamma$  are “dual” to those indexed by  $2 - \gamma$  in the sense that

$$\frac{\partial P_\gamma(\mathbf{a})}{\partial a_k} = \left[ \frac{\partial P_{2-\gamma}(\mathbf{a})}{\partial a_k} \right]^{-1}.$$

The choice (20) maintains this duality between the members of the generalized elastic net (19) (20) indexed by  $\beta$  and  $2 - \beta$ .

The power and generalized elastic net families produce a similar spectrum of penalties. The principal differences occur at very small coefficient values  $|a_j| \simeq 0$ . For  $\gamma > 1$  all members of the power family have  $[\partial P_\gamma(\mathbf{a})/\partial |a_j|]_{a_j=0} = 0$ , and for  $\gamma < 1$ ,  $[\partial P_\gamma(\mathbf{a})/\partial |a_j|]_{a_j=0} = \infty$ . The former causes all coefficient estimates to be non zero at every point on the path for all convex members except the lasso, and the latter property causes the coefficient paths  $\hat{\mathbf{a}}_\gamma(\lambda)$  to have discontinuities as a function of  $\lambda$  for all non convex members. For the generalized elastic net  $[\partial P_\beta(\mathbf{a})/\partial |a_j|]_{a_j=0}$  is non zero and finite for all  $0 < \beta < 2$ . This causes the coefficients to enter (initially become non zero) sequentially with decreasing  $\lambda$  for all  $\beta < 2$ . It also produces strictly continuous paths for  $\beta \geq 1/2$ , and smaller discontinuities (jumps) for  $0 < \beta < 1/2$ . This increases the stability (reduces variance) of the coefficient estimates (Fan and Li 2001).

### 3 Direct path seeking

A principal limitation of the bridge-regression strategy (16) (17) is the computational burden of obtaining the solutions to (16) for an adequate number of different penalties and corresponding path points at which to perform (17). One approach that mitigates this burden is direct path seeking. The goal is to sequentially construct a path directly in the parameter space that closely approximates that for a given penalty  $P(\mathbf{a})$ , without having to repeatedly solve numerical optimization problems.

With direct path seeking, solution points on the path  $\hat{\mathbf{a}}(\nu)$  are indexed by path length  $\nu$ . Starting at  $\nu = 0$  with some initial point  $\hat{\mathbf{a}}(0)$  (usually  $\hat{\mathbf{a}}(0) = 0$ ) each successive point  $\hat{\mathbf{a}}(\nu + \Delta\nu)$  is obtained from the previous one  $\hat{\mathbf{a}}(\nu)$  by

$$\hat{\mathbf{a}}(\nu + \Delta\nu) = \hat{\mathbf{a}}(\nu) + \mathbf{d}(\nu) \cdot \Delta\nu; \quad \nu \leftarrow \nu + \Delta\nu. \quad (21)$$

Here  $\mathbf{d}(\nu)$  is a vector characterizing a direction in the parameter space and  $\Delta\nu > 0$  is a specified distance along that direction. These iterations continue until a point  $\nu_{\max}$  of minimum empirical risk (8) is reached

$$\nu_{\max} = \arg \min_{\nu > 0} \hat{R}(\hat{\mathbf{a}}(\nu)). \quad (22)$$

Different path seeking methods, each intended for a particular loss–penalty combination, specify different prescriptions for calculating  $\mathbf{d}(\nu)$  and  $\Delta\nu$  at each path point  $\hat{\mathbf{a}}(\nu)$ . Popular path seekers based on squared–error loss (4) include partial least squares regression (PLS, Wold *et al* 1984) which approximates the ridge–regression path (Frank and Friedman 1993), forward stepwise regression intended to approximate the all–subsets path, and least angle regression (Efron *et al* 2004) approximating the lasso path. Gradient boosting (Friedman 2001, Hastie *et al* 2007) is another direct path seeker for the lasso that can be used with any convex loss criterion.

### 3.1 Generalized path seeking

In order to perform bridge-regression, fast methods are required for inducing (approximate) paths for a wide variety of penalties, such as all those in the power (18) or generalized elastic net (19) (20) families. In addition, it would be desirable to be able to employ a variety of loss criteria inducing risk functions (8) corresponding to likelihoods for a variety of probability models.

Consider penalties of the form

$$P(\mathbf{a}) = \sum_{j=1}^n P_j(a_j) \quad (23)$$

where each  $P_j(a_j)$  is a (possibly different) function of only of  $a_j$ . Furthermore suppose

$$\frac{\partial P_j(a)}{\partial |a|} > 0 \quad (24)$$

for all values of  $a$ . These conditions define a class of penalties where each member in the class is an additive function of its arguments (23), and each additive term  $P_j(a_j)$  is a monotone increasing function of the absolute value of its single argument (24). All members of the power family (18) and generalized elastic net (19) (20) are included in this class along with many other penalties as well. The following generalized path seeking algorithm (GPS) can be used to approximate the path corresponding to any penalty in this class in conjunction with any (differentiable) convex loss.

Let  $\nu$  measure length along the path and  $\Delta\nu > 0$  be a *small* increment. Define

$$g_j(\nu) = - \left[ \frac{\partial \hat{R}(\mathbf{a})}{\partial a_j} \right]_{\mathbf{a}=\hat{\mathbf{a}}(\nu)}, \quad (25)$$

$$p_j(\nu) = \left[ \frac{\partial P_j(a_j)}{\partial |a_j|} \right]_{a_j=\hat{a}_j(\nu)} \quad (26)$$

and

$$\lambda_j(\nu) = g_j(\nu) / p_j(\nu). \quad (27)$$

Here  $g_j(\nu)$  is the  $j$ th component of the negative gradient of the empirical risk (8) evaluated at the path point  $\hat{\mathbf{a}}(\nu)$ , and  $p_j(\nu)$  is the corresponding component of the gradient of  $P(\mathbf{a})$  with respect to  $|a_j|$ . Note that by assumption (24) all  $\{p_j(\nu) > 0\}_1^n$ . The components of the vector  $\lambda(\nu)$  are the component wise ratios of these two gradients at  $\hat{\mathbf{a}}(\nu)$ . These lamdas (27) are used to drive the generalized path seeking (GPS) algorithm.

#### GPS Algorithm

```

1  Initialize:  $\nu = 0$ ;  $\{\hat{a}_j(0) = 0\}_1^n$ 
2  Loop {
3    Compute  $\{\lambda_j(\nu)\}_1^n$ 
4     $S = \{j \mid \lambda_j(\nu) \cdot \hat{a}_j(\nu) < 0\}$ 
5    if ( $S = \text{empty}$ )  $j^* = \arg \max_j |\lambda_j(\nu)|$ 
6    else  $j^* = \arg \max_{j \in S} |\lambda_j(\nu)|$ 
7     $\hat{a}_{j^*}(\nu + \Delta\nu) = \hat{a}_{j^*}(\nu) + \Delta\nu \cdot \text{sign}(\lambda_{j^*}(\nu))$ 
8     $\{\hat{a}_j(\nu + \Delta\nu) = \hat{a}_j(\nu)\}_{j \neq j^*}$ 
9     $\nu \leftarrow \nu + \Delta\nu$ 
10 } Until  $\lambda(\nu) = 0$ 

```

Line 1 initializes the path. At each step the vector  $\lambda(\nu)$  is computed via (25–27) (line 3). At line 4, those coefficients  $\hat{a}_j(\nu)$  with sign opposite to that of their corresponding  $\lambda_j(\nu)$  are identified. If there are none (usual case) the coefficient corresponding to the largest component of  $\lambda(\nu)$  in absolute value is selected (line 5). If one or more  $\lambda_j(\nu) \cdot \hat{a}_j(\nu) < 0$ , then the coefficient with corresponding largest  $|\lambda_j(\nu)|$  within this subset is instead selected (line 6). The selected coefficient  $\hat{a}_{j^*}(\nu)$  is then incremented by a small amount in the direction of the sign of its corresponding  $\lambda_{j^*}(\nu)$  (line 7) with all other coefficients remaining unchanged (line 8), producing the solution for the next path point  $\nu + \Delta\nu$  (line 9). Iterations continue until all components of  $\lambda(\nu)$  are zero (line 10). Since each step (lines 7–8) reduces the empirical risk (8),  $\hat{R}(\hat{\mathbf{a}}(\nu + \Delta\nu)) < \hat{R}(\hat{\mathbf{a}}(\nu))$ , the algorithm will reach an unregularized solution (7) where all  $\{\lambda_j(\nu) = 0\}_1^n$  (24–27).

### 3.2 Motivation

In this section motivation is provided to explain why one might expect the GPS algorithm to closely track the paths produced by (9) (10) for convex risk (8) and penalties satisfying (23) (24). Actual comparisons are presented in Section 4.

Consider the constrained formulation (9). Let  $\hat{\mathbf{a}}(t)$  be a solution to (9) at a path point indexed by a value of the constraint threshold  $t$ , and  $\hat{\mathbf{a}}(t + \Delta t)$  be the solution when the constraint is relaxed by a small amount  $\Delta t > 0$ . Then  $\Delta\hat{\mathbf{a}}(t) = \hat{\mathbf{a}}(t + \Delta t) - \hat{\mathbf{a}}(t)$  is the solution to

$$\Delta\hat{\mathbf{a}}(t) = \arg \min_{\Delta\mathbf{a}} [\hat{R}(\hat{\mathbf{a}}(t) + \Delta\mathbf{a}) - \hat{R}(\hat{\mathbf{a}}(t))] \quad \text{s.t. } P(\hat{\mathbf{a}}(t) + \Delta\mathbf{a}) - P(\hat{\mathbf{a}}(t)) \leq \Delta t. \quad (28)$$

Suppose the path  $\hat{\mathbf{a}}(t)$  is a continuous function of  $t$

$$\left\{ \left| \frac{d\hat{a}_j(t)}{dt} \right| < \infty \right\}_1^n, \quad t > 0. \quad (29)$$

Then as  $\Delta t \rightarrow 0$ , assuming (23) (24), (28) can be expressed to first order

$$\begin{aligned} \Delta\hat{\mathbf{a}}(t) = \arg \max_{\{\Delta a_j\}_1^n} & \sum_{j=1}^n g_j(t) \cdot \Delta a_j \\ \text{s.t. } & \sum_{\hat{a}_j(t)=0} p_j(t) \cdot |\Delta a_j| + \sum_{\hat{a}_j(t) \neq 0} p_j(t) \cdot \text{sign}(\hat{a}_j(t)) \cdot \Delta a_j \leq \Delta t \end{aligned} \quad (30)$$

where

$$g_j(t) = - \left[ \frac{\partial \hat{R}(\mathbf{a})}{\partial a_j} \right]_{\mathbf{a}=\hat{\mathbf{a}}(t)}$$

and

$$p_j(t) = \left[ \frac{\partial P_j(a_j)}{\partial |a_j|} \right]_{a_j=\hat{a}_j(t)}.$$

Furthermore, suppose that all coefficient paths  $\{\hat{a}_j(t)\}_1^n$  are monotonic functions of  $t$

$$\{ |\hat{a}_j(t + \Delta t)| \geq |\hat{a}_j(t)| \}_1^n \quad (31)$$

so that  $\{\text{sign}(\hat{a}_j(t)) = \text{sign}(\Delta\hat{a}_j(t))\}_{\hat{a}_j(t) \neq 0}$ . Under this (additional) constraint (30) becomes

$$\Delta\hat{\mathbf{a}}(t) = \arg \max_{\{\Delta a_j\}_1^n} \sum_{j=1}^n g_j(t) \cdot \Delta a_j \quad \text{s.t. } \sum_{j=1}^n p_j(t) \cdot |\Delta a_j| \leq \Delta t.$$

This is a linear programming problem with solution

$$\begin{aligned} j^*(t) &= \arg \max_{1 \leq j \leq n} |g_j(t)| / p_j(t) \\ \Delta\hat{a}_{j^*}(t) &= [g_{j^*}(t) / p_{j^*}(t)] \cdot \Delta t \\ \{\Delta\hat{a}_j(t) = 0\}_{j \neq j^*}. \end{aligned} \quad (32)$$

From (25–27) one sees that the GPS algorithm (lines 5 and 7–8) follows the strategy implied by (32) provided  $\text{sign}(\lambda_j(\nu)) = \text{sign}(\hat{a}_j(t))$  for all  $\hat{a}_j(t) \neq$

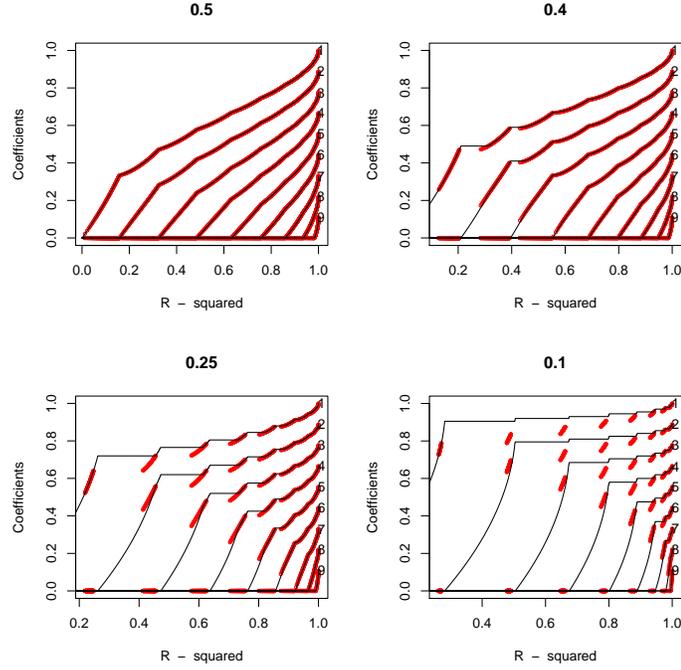


FIGURE 1. Exact (red) and GPS (black) paths for elastic net non convex penalties  $\beta \in \{0.5, 0.4, 0.25, 0.1\}$  with orthogonal predictors.

0. This will be the case at all points for which the GPS and exact paths coincide, as a consequence of the Karush-Kuhn-Tucker (KKT) optimality conditions

$$\lambda_j(t) = \lambda(t) \cdot \text{sign}(\hat{a}_j(t)), \quad \hat{a}_j(t) \neq 0, \quad (33)$$

where  $\lambda(t) > 0$  is the value of  $\lambda$  in (10) corresponding to  $t$ . At the beginning, the exact ( $t = 0$ ) and GPS ( $\nu = 0$ ) paths coincide by construction (line 1). Therefore as long as the exact path (9) remains continuous (29) and monotonic (31) for  $t \leq t_0$ , in the limit  $\Delta\nu \rightarrow 0$  ( $\Delta t \rightarrow 0$ ) the GPS and exact paths will coincide for  $t \leq t_0$ .

If the exact path (10) is continuous and monotonic along its entire path ( $\infty \leq \lambda \leq 0$ ), as is often the case, then the GPS algorithm produces the exact path ( $0 \leq \nu \leq \nu_{\max}$ ) (22) as  $\Delta\nu \rightarrow 0$ .

A sufficient (but far from necessary) condition for such total monotonicity

is orthogonality of the predictor variables over the training sample (1)

$$\sum_{i=1}^N x_{ij}x_{ik} = 0, j \neq k. \tag{34}$$

In this case the GPS algorithm produces the exact path provided the latter is continuous.

**Discontinuity** With the generalized elastic net family (19) (20), all members for which  $\beta \geq 1/2$  produce continuous paths. For  $\beta < 1/2$  the paths are not continuous. There can be jumps at those points ( $\lambda$  values (10)) where each successive variable enters (coefficient initially becomes non zero). This is caused by the variables entering with finite non zero coefficient values at those points. This is illustrated in Fig. 1 for  $\beta \in \{0.5, 0.4, 0.25, 0.1\}$  in the case of orthogonal (standardized) predictor variables (34) and squared-error loss (4). Here the exact path solutions for nine coefficients are shown as thick (red) points plotted in terms of fraction of explained risk (8)

$$r(\lambda) = [\hat{R}(\hat{\mathbf{a}}(\infty)) - \hat{R}(\hat{\mathbf{a}}(\lambda))] / \hat{R}(\hat{\mathbf{a}}(\infty)) \tag{35}$$

which is monotonically increasing with decreasing  $\lambda$  along the path  $\infty \leq \lambda \leq 0$ . For squared-error loss  $r(\lambda)$  is the fraction of explained variance  $R^2(\lambda)$  of the data values  $\{y_i\}_1^N$  (1) at each path point indexed by  $\lambda$ . In Fig. 1 the abscissa is the fraction of explainable variance  $r(\lambda)/r(0)$ .

As seen in Fig. 1 the coefficient paths for  $\beta = 0.5$  (upper left panel) are continuous. For  $\beta = 0.4$  (upper right panel) discontinuities appear in the coefficient paths for  $0 \leq r(\lambda) \lesssim 0.6$ ; there are values of  $r(\lambda)$  in this range at which no exact solution exists. For smaller values of  $\beta$  (lower panels) the discontinuities increase in magnitude and number. For  $\beta = 0$  representing all-subsets regression (not shown) there are no continuous sections of the exact path and it reduces to a set of discrete points for each coefficient as a function of  $r(\lambda)$  (35).

The thin (black) curves in Fig. 1 show the corresponding paths produced by the GPS algorithm for the same penalty. By construction these paths are continuous at all points for all coefficients. For  $\beta = 1/2$  the GPS and exact paths coincide at all points as a consequence of the continuity of the latter. For  $\beta < 1/2$  the GPS and exact paths coincide in those regions where the exact paths for *all* coefficients are continuous. In regions where this is not the case the GPS algorithm provides continuous approximations that fairly closely (but not exactly) track the exact paths where solutions for the latter exist. The sparseness properties of the two sets of paths are seen to be quite similar. In the case  $\beta = 0$  (all-subsets, not shown) the GPS and exact paths coincide at the (discrete) points representing solutions for the latter. At other path points GPS provides continuous paths that interpolate between the corresponding exact solution points.

As pointed out by Fan and Li 2001, discontinuities in the coefficient paths are undesirable because they lead to instability (increased variance) in the coefficient estimates. In this sense the continuous GPS paths might be preferred on statistical grounds even when the exact paths can be calculated as here in the orthogonal case (34).

**Non monotonicity** When all exact coefficient paths  $\hat{a}_j(t)$  (9) are continuous, the GPS paths coincide with the exact ones as long as all  $\hat{a}_j(t)$  remain monotonic (31). In this case one has from the KKT conditions (33)

$$\lambda_j(\nu) \cdot \text{sign}(\hat{a}_j(\nu)) = \max_k \lambda_k(\nu) > 0, \quad \hat{a}_j(t) \neq 0, \quad (36)$$

for all non zero coefficients. If at some point  $t_0$  ( $\nu_0$ ) one or more exact paths  $\hat{a}_j(t)$  become non monotonic ( $|\hat{a}_j(t + \Delta t)| < |\hat{a}_j(t)|$ ), then (33) remains valid for  $t > t_0$ , whereas (36) need not hold for all GPS coefficient paths. There may be no single variable increment (lines 7–8) that produces the exact solution for  $\nu > \nu_0$ . So long as all  $\{\text{sign}(\lambda_j(\nu)) = \text{sign}(\hat{a}_j(\nu))\}_1^n$  GPS will continue to monotonically update the coefficients that satisfy (36), leaving those for which  $\lambda_j(\nu) < \max_k \lambda_k(\nu)$  constant. These are the coefficients for which the exact solutions have become non monotonic. This continues until the corresponding  $\lambda_j(\nu)$  for one or more of these variables changes sign. At that point  $\lambda_j(\nu) \cdot \text{sign}(\hat{a}_j(\nu)) < 0$  and the GPS algorithm (line 6) chooses the coefficient  $\hat{a}_{j^*}(\nu)$  corresponding to the most negative  $\lambda_j(\nu) \cdot \text{sign}(\hat{a}_j(\nu))$  for updating. This update (line 7) causes  $|\hat{a}_{j^*}(\nu + \Delta\nu)| < |\hat{a}_{j^*}(\nu)|$  thereby (belatedly) introducing non monotonicity into the GPS paths of these coefficients.

As long as the set  $S$  (line 4) is not empty the coefficients  $\{\hat{a}_j(\nu)\}_{j \in S}$  will continue to decrease in absolute value for successive steps while the other coefficients  $\{\hat{a}_j(\nu)\}_{j \notin S}$  remain constant. Each update (line 7) causes  $|\lambda_{j^*}(\nu + \Delta\nu)| < |\lambda_{j^*}(\nu)|$  since to first order

$$\Delta|\lambda_{j^*}(\nu)| = -[(h_{j^*}(\nu) + \lambda_{j^*}(\nu) q_{j^*}(\nu) \text{sign}(\hat{a}_{j^*}(\nu)))/p_{j^*}(\nu)] \cdot \Delta\nu \quad (37)$$

where  $h_{j^*}(\nu)$  and  $q_{j^*}(\nu)$  are the corresponding diagonal elements of the Hessians of  $\hat{R}(\mathbf{a})$  (8) and penalty  $P(\mathbf{a})$  respectively. Since  $h_{j^*}(\nu) > 0$  by convexity,  $p_{j^*}(\nu) > 0$  by assumption (24) (26), and  $|\lambda_{j^*}(\nu)|$  is small when  $\lambda_{j^*}(\nu)$  initially becomes negative, this quantity (37) is initially small and negative and stays that way as a result of (37). Thus the largest  $|\lambda_j(\nu)|$ ,  $j \in S$ , is decreased at each step (line 7) until another (if any)  $|\lambda_l(\nu)|$ ,  $l \in S$ , becomes larger. In this way, all coefficients  $\{\hat{a}_j(\nu)\}_{j \in S}$  are repeatedly updated, reducing their corresponding  $|\hat{a}_j(\nu)|$  until either  $\hat{a}_j(\nu)$  or  $\lambda_j(\nu)$  changes sign, or the end of the path is reached (all  $\{\lambda_j(\nu) = 0\}_1^n$ ).

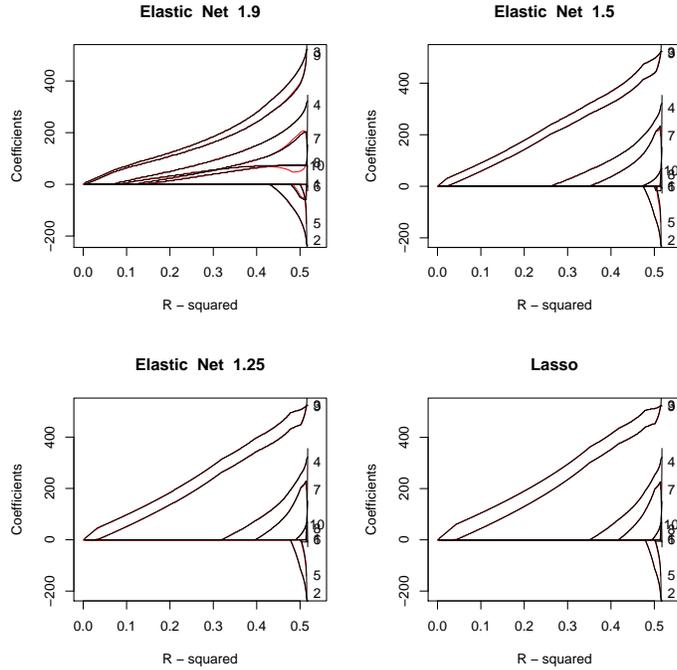


FIGURE 2. Exact (red) and GPS (black) paths for the diabetes data using convex elastic net penalties  $\beta \in \{1.9, 1.5, 1.25, 1.0\}$ .

## 4 Examples

In this section applications of the GPS algorithm to data using generalized elastic net penalties (19) (20) are presented, and compared to the exact paths for the convex members ( $\beta \geq 1$ ).

### 4.1 Least-squares regression: diabetes data

This data set, used in Efron *et al* 2004, consists of  $n = 10$  predictor variables and  $N = 442$  observations. The outcome variable is numeric so that squared-error loss (4) was employed.

Figure 2 shows the ten coefficient paths as a function of  $R^2$  (35) for  $\beta = 1.9$  (upper left),  $\beta = 1.5$  (upper right),  $\beta = 1.25$  (lower left), and  $\beta = 1$  (lasso, lower right). The red curves are the exact paths, whereas the black ones are the corresponding GPS paths. For  $\beta = 1.9$  slight differences are seen to occur for  $R^2 \gtrsim 0.45$  where one of the exact coefficient paths becomes non

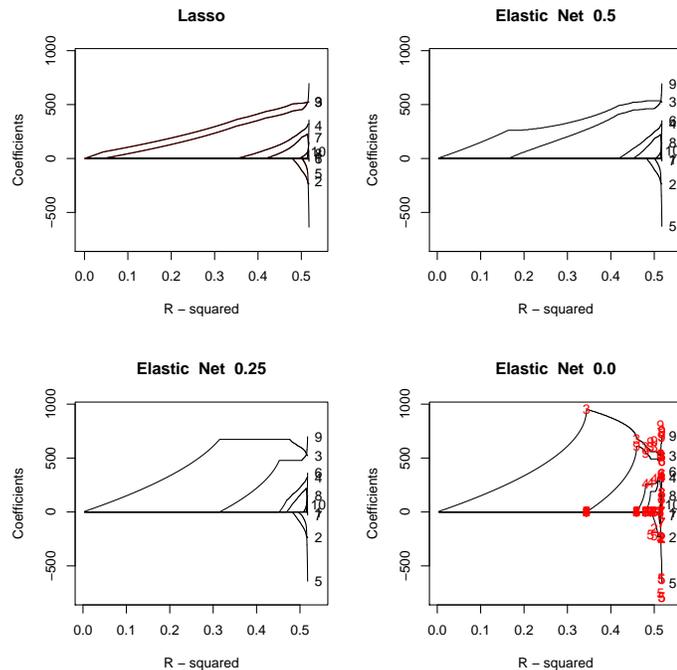


FIGURE 3. Paths for Lasso, and GPS non convex elastic net penalties  $\beta \in \{0.5, 0.25, 0.0\}$ , for the diabetes data. The red numbers indicate the forward stepwise solutions.

monotonic. For the other penalties the differences are seen to be smaller. As  $\beta$  decreases the solutions become sparser in that for the same degree of data fit as measured by  $R^2$  there tend to be fewer non zero coefficients. That is

$$S(\hat{\mathbf{a}}_{\beta}(R^2)) \geq S(\hat{\mathbf{a}}_{\beta'}(R^2)), \quad \beta < \beta' \quad (38)$$

with  $S(\mathbf{a})$  given by (15) for  $\eta = 0$ .

Figure 3 shows the GPS paths for the lasso (upper left) and for several non convex generalized elastic net penalties,  $\beta = 0.5$  (upper right),  $\beta = 0.25$  (lower left), and  $\beta = 0$  (lower right), plotted on the same vertical scale. Here one sees a similar pattern of further increasing sparsity (38) as  $\beta < 1$  decreases.

The red numbers in the lower right panel of Fig. 3 represent the (discrete) path points for forward stepwise regression. Here one sees that the  $\beta = 0$  GPS and stepwise paths coincide at the stepwise solutions. At other points

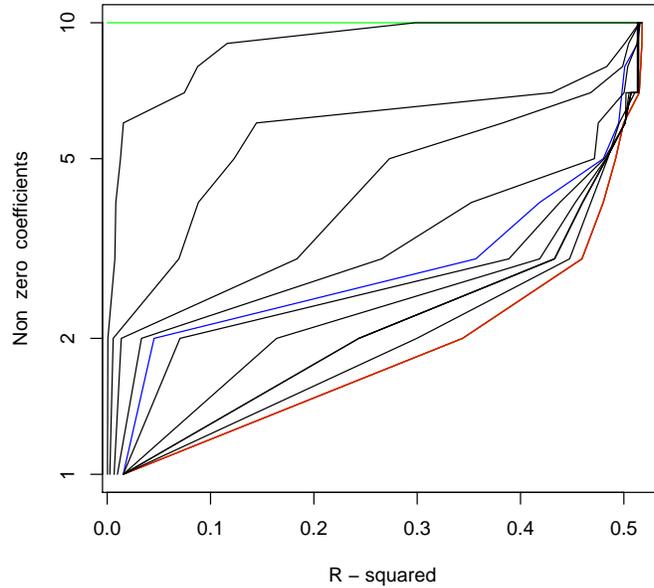


FIGURE 4. Number of non zero coefficient estimates along respective paths for diabetes data using elastic net penalties  $\beta \in \{2.0(\text{ridge, green}), 1.99, 1.9, 1.7, 1.5, 1.0(\text{lasso, blue}), 0.7, 0.5, 0.4, 0.3, 0.0\}$  and stepwise (red).

the GPS paths are continuous, interpolating between the stepwise solutions. This is not always the case. For  $\beta = 0$  the GPS paths interpolate the discrete path points generated by “statewise” regression. At each step, statewise regression successively selects the variable not in the model that is most correlated with the current residuals to next include in the model. It then performs a full multiple regression on the current variable set to obtain the solution coefficients. As a variable selection technique this can be slightly less aggressive than forward *stepwise* regression which selects each successive variable that gives the best multiple regression fit, given the variables that have already entered. In many situations the two procedures give identical results (as here), but this is not always the case. However, the results of the two procedures are seldom very different especially for the larger estimated coefficients.

Figure 4 shows the number of non zero coefficients as a function  $R^2$  along the path for a larger set of generalized elastic net penalties. This number is

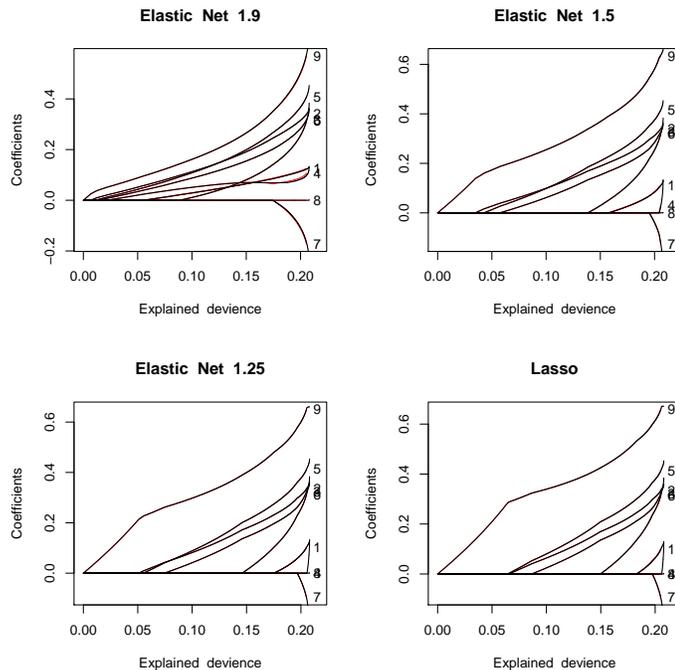


FIGURE 5. Exact (red) and GPS (black) paths for the heart transplant data using convex elastic net penalties  $\beta \in \{1.9, 1.5, 1.25, 1.0\}$ .

inversely related to sparsity (15) for  $\eta = 0$ . The results for each penalty are connected by straight lines to aid visualization. Results for forward stepwise regression (red) are also included, which here are identical to that of  $\beta = 0$  GPS. Results for  $\beta = 1$  (lasso) and  $\beta = 2$  (ridge-regression) are highlighted as well (blue and green). From Fig. 4 one sees that at  $R^2 \simeq 0.45$ , stepwise regression enters 3 variables, the lasso 4 variables, and ridge-regression all 10 variables. Using these curves as an inverse measure of sparsity, one sees a strict monotonicity among the members of this family. Smaller  $\beta$  produces sparser solutions at every point on the path as indexed by degree of data fit ( $R^2$ ).

#### 4.2 Logistic regression: South African heart transplant data

This data set was presented in Hastie, Tibshirani and Friedman 2001. It has  $n = 9$  predictor variables and  $N = 462$  observations. The outcome variable is binary so logistic loss (5) is appropriate.

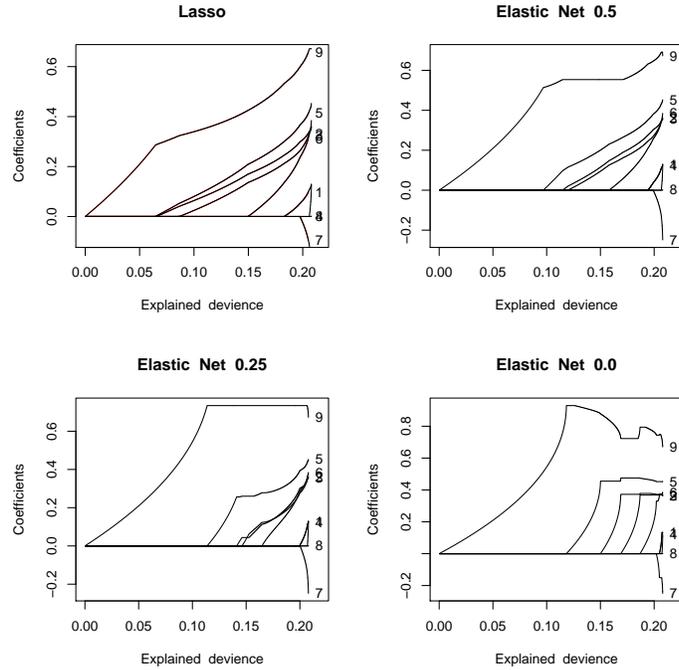


FIGURE 6. Paths for lasso, and GPS elastic net non convex penalties  $\beta \in \{0.5, 0.25, 0.0\}$ , for the heart transplant data.

Figure 5 compares the exact and GPS coefficient paths for selected convex members of the generalized elastic net for this data set. The paths are here indexed by fraction of explained deviance (5) (8) (35). As with squared-error loss (Fig. 2) the GPS paths closely track those for the exact solutions and become sparser for smaller  $\beta$ .

Figure 6 repeats the lasso for comparison, and shows the results for the same non convex penalties as in Fig. 3. Again sparsity is seen to continue to increase with decreasing  $\beta < 1$ . Figure 7 shows the number of non zero coefficients as a function of explained deviance along the path. As in the squared-error loss case (Fig. 4) there is a strict monotonicity; smaller  $\beta$  produces sparser solutions at every point on the path as indexed by degree of data fit, here measured by explained deviance.

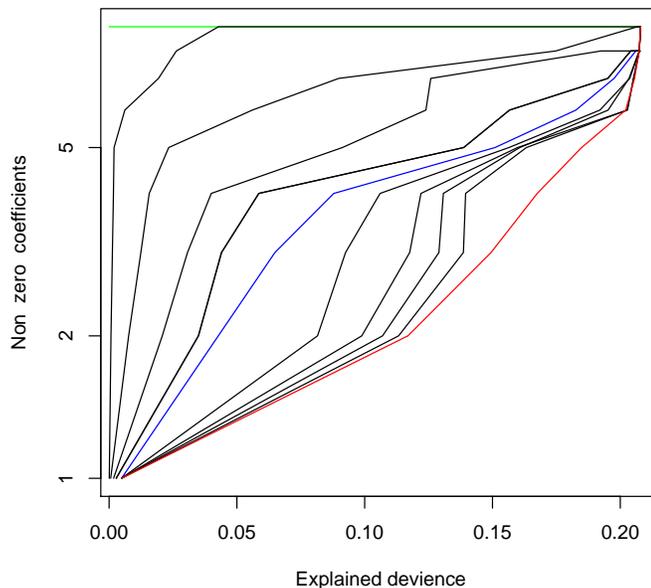


FIGURE 7. Number of non zero coefficient estimates along respective paths for heart transplant data using elastic net penalties  $\beta \in \{2.0(\text{ridge, green}), 1.99, 1.9, 1.7, 1.5, 1.0 (\text{lasso, blue}), 0.7, 0.5, 0.4, 0.3, 0.0 (\text{red})\}$ .

### 4.3 Least-squares regression: under-determined problem

The above two examples are highly over-determined in that the number of observations  $N$  is much larger than the number of predictor variables  $n$ . In such cases regularization is much less important than it is for highly under-determined problems where  $N \ll n$ . In this section a highly under-determined regression problem is considered. There are  $N = 200$  observations and  $n = 10000$  predictor variables. The data are simulated from the model

$$y_i = \sum_{j=1}^n a_j^* x_{ij} + \varepsilon_i \quad (39)$$

where the predictor variables are randomly drawn from a normal distribution  $\mathbf{x}_i \sim N(0, \mathbf{C})$ ; with covariance matrix elements  $C_{jj} = 1$ ,  $C_{jk} = 0.4$ ,  $j \neq k$ . The random error is also normally distributed  $\varepsilon_i \sim N(0, \sigma^2)$ , with the value of  $\sigma$  set to produce a 3/1 signal to noise ratio. The optimal

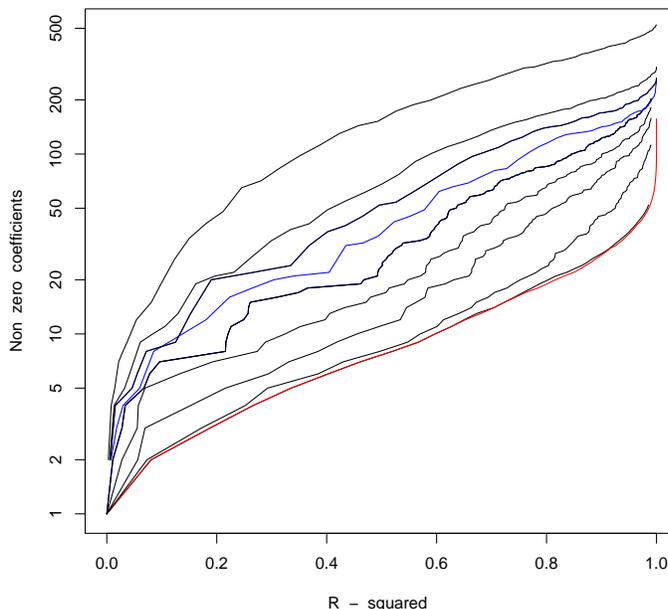


FIGURE 8. Number of non zero coefficient estimates along respective paths for the under-determined regression example ( $n = 10000, N = 200$ ) using elastic net penalties  $\beta \in \{1.9, 1.7, 1.5, 1.0$  (lasso, blue),  $0.5, 0.3, 0.2, 0.1, 0.0\}$ , and stepwise (red).

coefficient vector  $\mathbf{a}^*$  (6) has 30 non zero coefficients with uniformly distributed absolute values  $|a_j^*| = [31-j]_+$ , and alternating signs  $sign(a_{j+1}^*) = -sign(a_j^*)$ ,  $1 \leq j \leq 29$ .

Figure 8 shows the number of non zero coefficients as a function of  $R^2$  along the path for forward stepwise regression (red) and a selected set of generalized elastic net penalties. The  $\beta = 1$  (lasso) penalty is colored blue. One sees the same monotonic relation between the value of  $\beta$  and the sparsity of the induced path. At  $R^2 = 0.9$  on the training data, stepwise and  $\beta = 0$  GPS have 15 non zero coefficients, the lasso has 120, and  $\beta = 1.9$  elastic net has almost 400.

Thus, by varying  $\beta$  one can exercise sharp control over the sparsity of the induced solutions. Note that the  $\beta = 0$  and stepwise results are here slightly different for  $R^2 > 0.75$ .

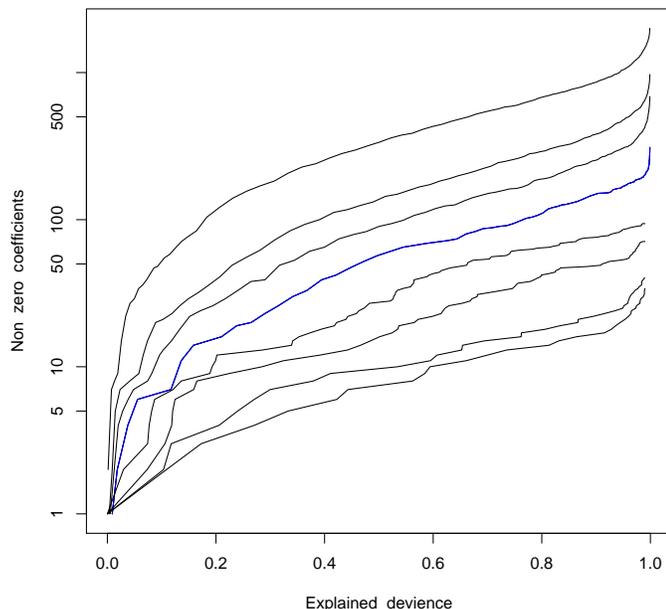


FIGURE 9. Number of non zero coefficient estimates along respective paths for the under-determined logistic regression example ( $n = 10000$ ,  $N = 200$ ) using elastic net penalties  $\beta \in \{1.9, 1.7, 1.5, 1.0$  (lasso, blue),  $0.7, 0.5, 0.3, 0.0\}$ .

#### 4.4 Logistic regression: under-determined problem

The data for this problem are similar to that in the previous section. There are  $N = 200$  observations and  $n = 10000$  predictor variables generated from the same model. The outcome variable has two values  $y \in \{-1, 1\}$  with the log-odds given by

$$\log[\Pr(y = 1)/\Pr(y = -1)] = s \cdot \sum_{j=1}^n a_j^* x_{ij}. \quad (40)$$

The value for  $s$  was chosen to produce a Bayes error rate of 0.05. Here the optimal coefficient vector has 15 non zero coefficients with uniformly distributed absolute values and alternating signs.

Figure 9 shows the number of non zero coefficients as a function of fraction of deviance explained along the path for the same selected set of generalized

elastic net penalties as in Fig. 8, with  $\beta = 1$  (lasso) colored blue. Again the sparsity of the solutions along the path is monotonically related the value of  $\beta$ . At 95% explained deviance, the  $\beta = 0$  path includes 10 non zero coefficients, the lasso has 120, and the  $\beta = 1.9$  elastic net path has almost 1000.

## 5 Speed

As the examples illustrate, one can control the sparsity of regularized regression solutions (9) (10) by appropriately selecting penalties  $P(\mathbf{a})$ . Since the sparsity (15) of the optimal coefficients (6) is generally unknown, bridge-regression (16) (17) can be used to estimate the best penalty, provided the resulting computational burden is not excessive.

**Table 1**

Time in seconds for computing 500 path points for GPS and exact convex algorithms with  $n = 10000$ ,  $N = 200$ .

Penalty $\beta$	Sq-err GPS	Sq-err exact	Logistic GPS	Logistic exact
0.0	0.37	1.82*	1.43	
0.1	0.47		1.43	
0.2	0.55		1.44	
0.5	0.62		1.44	
1.0	0.56	3.58	1.34	6.16
1.5	0.69	2.97	1.35	5.42

\* forward stepwise

Table 1 shows the computation time in seconds (column 2) required by the GPS algorithm to generate paths (500 points) for the  $n = 10000$ ,  $N = 200$  problem described in Section 4.3, for several generalized elastic net penalties (19) (20) as indexed by  $\beta$  (column 1). The third column (rows 5 and 6) show corresponding exact path times (500 path points) for the convex penalties ( $\beta \geq 1$ ) using the fastest known convex optimization methods for these particular problems (Friedman *et al* 2007). The entry in the first row of column 3 is for the (approximate) stepwise path.

The corresponding entries in the last two columns of Table 1 are for the logistic regression problem of Section 4.4. To facilitate comparison with the squared-error results (columns 2 and 3) the optimal coefficient vector  $\mathbf{a}^*$  used in (40) was here taken to have 30 non zero coefficients with uniformly distributed absolute values and alternating signs. Again the entries in column 6 are the exact path times using the fastest convex optimization

method for elastic net logistic regression (Friedman, Hastie and Tibshirani 2008).

As see from Table 1 bridge-regression is quite feasible for problems of this size. Performing 10-fold cross-validation to evaluate 500 path points for each of these six penalties would require 35 seconds for squared-error loss regression and 84 seconds for logistic regression. This is equivalent to solving 30000 optimization problems in (16), most of which are non convex. For  $n \gg N$  the computation for GPS scales roughly as  $n \cdot N$  so that bridge-regression with much larger problems is still quite feasible. For the convex elastic net (including the lasso) special fast exact algorithms are available that are competitive in speed with GPS as seen in Table 1. However, there are no such competitive algorithms for the non convex members (20). And even in the convex realm there are many loss-penalty combinations for which special fast exact algorithms competitive with GPS do not exist.

## 6 Utility

The results presented in Section 5 show that bridge-regression using GPS is computationally tractable for fairly large problems. In this section its potential statistical advantages are investigated. For regression the (lack of) quality of a particular coefficient path  $\hat{\mathbf{a}}(\rho)$ , as indexed by its path points  $\rho$ , can be measured by

$$\min_{\rho} [R(\hat{\mathbf{a}}(\rho)) - R(\mathbf{a}^*)] / R(\mathbf{a}^*) \quad (41)$$

where  $R(\mathbf{a}^*)$  is the minimum possible risk associated with the problem (6). This quantity is the minimal distance (11) between points on the path and the optimal solution  $\mathbf{a}^*$ , scaled by  $1/R(\mathbf{a}^*)$ . As discussed in Section 2.1, paths  $\hat{\mathbf{a}}(\rho)$  that produce smaller values for (41) have the potential for producing more accurate predictions given a model selection procedure such as cross-validation.

Figure 10 shows the distribution of (41) (boxplots) for paths produced by several squared-error loss (4) regression methods, based on 50 data sets randomly drawn from the model described in Section 4.3. The methods are (left to right) forward stepwise regression, GPS using (non convex) generalized elastic net penalties  $\beta \in \{0.0, 0.1, 0.2, 0.5\}$  (20), and the exact paths produced by the lasso ( $\beta = 1$ ) and elastic net with  $\beta = 1.5$ .

From Fig. 10 one sees that the lasso and  $\beta = 1.5$  elastic net consistently yield inferior paths on this very sparse problem (30 out of 10000 true non zero coefficients). The forward stepwise procedure yields paths of similar accuracy (41) to the lasso on average, but with much more variability. Stepwise paths based on some data sets are considerably better than those produced by the lasso, and some others are considerably worse. For the GPS paths variability decreases with increasing  $\beta$ . Expected performance is best for  $\beta = 0.1$  or  $\beta = 0.2$ , with the latter having less variability.

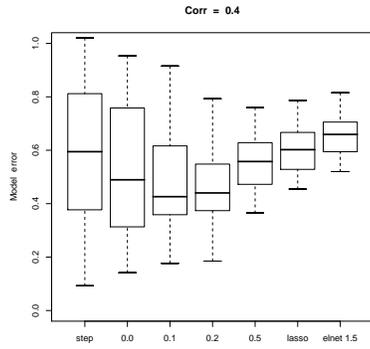


FIGURE 10. Inaccuracy of stepwise, several non convex elastic net GPS, and convex exact paths, over 50 simulated regression data sets with  $n = 10000$ ,  $N = 200$  (Section 6.1).

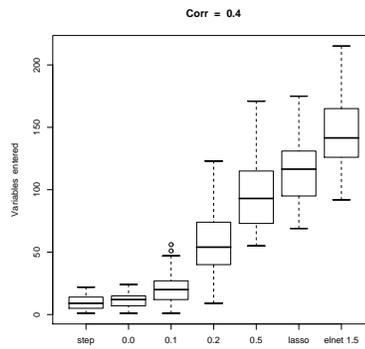


FIGURE 11. Number of non zero coefficients at optimal solutions for stepwise, several non convex elastic net GPS, and convex exact paths, over the 50 simulated data sets.

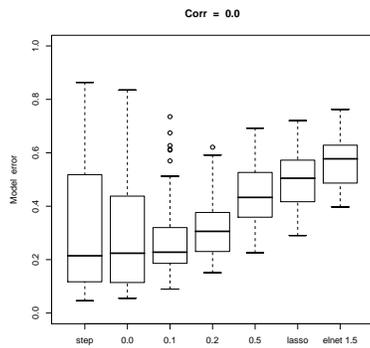


FIGURE 12. Inaccuracy of stepwise, several non convex elastic net GPS, and convex exact paths over 50 simulated data sets with population uncorrelated predictors.

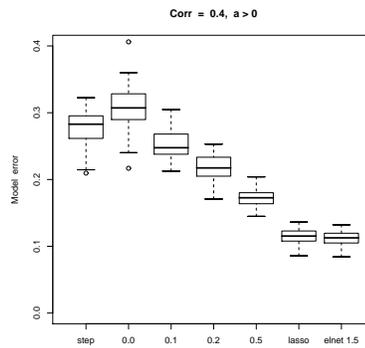


FIGURE 13. Inaccuracy of stepwise, several non convex elastic net GPS, and convex exact paths, over 50 simulated data sets, all optimal coefficients having the same sign.

Figure 11 shows the distribution of the number of non zero coefficients at the optimal path points minimizing (41) for each of the respective methods. Here one sees that forward stepwise regression typically has around 12 non zero coefficients at its optimal solutions. The slightly less aggressive  $\beta = 0$  GPS paths average 14. As  $\beta$  increases, the optimal points on GPS paths tend towards less sparsity. The lasso and  $\beta = 1.5$  elastic net produce even denser optimal solutions, typically involving 125 and 150 non zero coefficients respectively. Here one sees that penalty choice closely controls the sparsity of the solutions with sparse ( $\beta = 0.1$  or  $0.2$ ), but not the sparsest ( $\beta = 0$  or stepwise), being the best.

Figure 12 show results analogous to those in Fig. 10 for a slightly modified problem. Here for each of the 50 data sets, the predictor variables are drawn from a standard normal distribution,  $\mathbf{x}_i \sim N(0, \mathbf{I})$ . That is, the variables are uncorrelated with respect to their (population) joint distribution. All other aspects of the generating model are the same. For this problem all methods produce better paths, as measured by (41), with the sparsest procedures improving the most. Again the forward stepwise procedure produces the least stable paths in terms of variability, with the  $\beta = 0$  GPS path being almost as unstable. The stability of the paths produced by the other procedures are all about the same. The results for this problem are qualitatively similar to those shown in Fig. 10, with the best stability–expected performance trade–off appearing to be for the  $\beta = 0.1$  GPS path. The distributions of the number of non zero coefficients for the optimal solutions of each of the methods (not shown) is quite similar to that shown in Fig. 11.

Figure 13 shows analogous results to those in Fig. 10 for a slightly different modification of the problem. Here the joint distribution of the predictor variables is as described in Section 4.3, as is all other aspects of the model, except that the signs of the optimal non zero coefficients are taken to be the same ( $sign(a_{j+1}^*) = sign(a_j^*)$ ), instead of alternating. Here one sees a very different pattern of results. All methods produce much more stable paths. However their relative quality (41) is reversed; the worst methods in the previous two problems are here the best and vice versa. The lasso and the  $\beta = 1.5$  elastic net paths here dramatically out–perform methods producing sparser solutions. The distributions of the number of (optimal solution) non zero coefficients for each of the methods (not shown) is for this problem again quite similar to that shown in Fig. 11. The optimal  $\beta = 1.5$  elastic net solutions, typically involving 150 non zero coefficients, are far more accurate than methods producing much sparser solutions, even though the population optimal coefficient vector  $\mathbf{a}^*$  has only 30 non zero entries.

The results shown in Figs. 10–12 show that the accuracy of a given method for the same population joint distribution can strongly depend on the particular training data set realized from that distribution. This is especially the case for methods that induce very sparse paths. In Figs. 10 and 12 one

sees that the sparsest methods often produce much better solutions than denser methods on particular data sets, and much worse on others. Thus, comparisons based on one or a small number of data sets (simulated or real) can be highly misleading.

A somewhat surprising result from the above examples is that the optimal sparsity of the *estimated* coefficients depends upon more than just the sparsity of the optimal coefficients  $\mathbf{a}^*$  (6) characterizing the problem. In all three examples the sparsity (15) of  $\mathbf{a}^*$  was the same; 30 out of 10000 non zero coefficients. In fact, all  $\{|a_j^*|\}_1^n$  were identical. For the situation shown in Fig. 10 the best solutions typically involved 50 non zero coefficients whereas for that shown in Fig. 12 penalties producing around 20 were best. For the situation shown in Fig. 13 the densest method being considered produced the most accurate solutions with 150 non zero coefficients on average.

The penalties used here depend only on the absolute values of the coefficients. One might then expect that the best penalty would depend mainly on the relative absolute values of the optimal coefficients  $\{|a_j^*|\}_1^n$  (6). Thus, as discussed in Section 2.3, this knowledge (if available) would drive penalty choice. The examples presented here show that this is not the case. The relative signs of the optimal coefficients as well as the correlational structure of the predictor variable distribution also influence which such penalty is best. For example, methods that induce sparser solutions than the lasso are not always better than the lasso solutions, even in sparse situations as characterized by the optimal coefficients  $\mathbf{a}^*$ . Even with knowledge of the latter, the other aspects of the problem that influence choice of a good penalty are likely to be unknown. Thus, using bridge-regression to aid penalty choice can be helpful.

## 7 Post-processing selectors

As illustrated in Section 6, there are some applications where methods producing sparser solutions than is possible with convex penalties achieve higher prediction accuracy. One reason for this is the shrinkage of the absolute values of the coefficient estimates inherent in most regularized regression procedures (10).

For convex penalties ( $\beta \geq 1$ ) high sparsity solutions tend to involve heavy shrinkage of their non zero coefficient estimates. This can induce large bias if the optimal coefficients  $\mathbf{a}^*$  are also very sparse. In order to reduce this bias, the optimal solutions (41) for these convex penalties trade decreased sparsity for decreased shrinkage in an attempt to overcome this bias. Non convex penalties ( $\beta < 1$ ) shrink less for the same sparsity thereby producing sparser less biased optimal solutions (41) as illustrated in Fig. 11. This can sometimes improve performance as illustrated in Figs. 10 and 12.

These considerations suggest that the performance of a convex method such as the lasso might be improved by using it as a “selector”. At each

path point  $\lambda$ , the “active” variables corresponding to the non zero coefficients  $A(\lambda) = \{j \mid \hat{a}_j(\lambda) \neq 0\}$  are identified. Then a different regression procedure, producing less shrinkage, is used to estimate the coefficient values  $\{\tilde{a}_j(\lambda)\}_{j \in A(\lambda)}$  of these active variables. The overall solution at the path point  $\lambda$  is then taken to be  $\{\hat{a}_j(\lambda) = \tilde{a}_j(\lambda)\}_{j \in A(\lambda)}$  and  $\{\hat{a}_j(\lambda) = 0\}_{j \notin A(\lambda)}$ . In this way the selector identifies the non zero coefficients and the post-regression procedure determines their values, at each path point. Often an unregularized regression (7) (8) is used for this the second “post-processing” step.

Generally, convex selectors are employed due to the unattractive computational aspects associated with non convex optimization. With GPS, paths corresponding to non convex penalties ( $\beta < 1$ ) can be obtained with computation similar to that of convex ones ( $\beta \geq 1$ ) (Table 1), thereby expanding the pool of eligible selectors. In this section the utility of selectors based on non convex generalized elastic net penalties ( $\beta < 1$ ) is examined, and compared to that of convex selectors ( $\beta \geq 1$ ), using unregularized regression (7) (8) as the post-processor.

Figures 14–17 show the corresponding results of using the various GPS and exact convex procedures as selectors in the same set of situations represented in Figs. 10–13, respectively. The dashed (red) lines on each boxplot represents the medians of the corresponding distributions in Figs. 10–13. The results for forward stepwise regression are (by construction) identical and repeated for comparison. The results for  $\beta = 0$  GPS are very similar since this GPS path interpolates the unregularized *statewise* regression solutions at path points where each successive variable enters (Section 4.1). From Fig. 15 one sees that the selector based optimal solutions (41) are sparser than those for the corresponding direct solutions (Fig. 11) for all  $\beta > 0$ , with this effect being more pronounced as  $\beta$  increases. For the convex procedures ( $\beta \geq 1$ ) the optimal selector solutions typically involve 25 non zero coefficients rather than around 120 for their corresponding direct methods. This is seen to increase accuracy in those situations (Figs. 14, 16) where the sparser direct methods ( $\beta < 1$ ) provide superior results. Again, this accuracy increase is more pronounced for larger  $\beta$ , improving the convex methods the most. However, using these convex methods as selectors does not result in enough improvement to be competitive with the best non convex methods, especially when the latter are themselves used as selectors.

For the situation in Fig. 13 where the direct convex methods were seen to provide the best performance, using them as selectors *decreases* their accuracy (Fig. 17). In this situation using the  $\beta < 1$  GPS procedures as selectors improves their performance, but not enough to compete with the direct convex methods.

As with the direct methods, the success of the selector strategy in improving performance is seen to depend on more than just the sparsity of the optimal coefficients  $\mathbf{a}^*$ . Other factors including their signs and the correlational

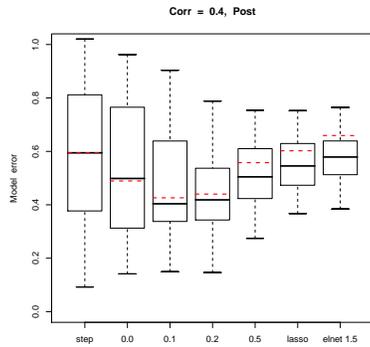


FIGURE 14. Results for Fig. 10 data when methods are used as selectors. Red lines are medians from Fig. 10 distributions.

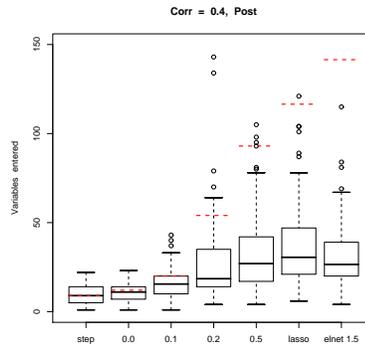


FIGURE 15. Results for Fig. 11 data when methods are used as selectors. Red lines are medians from Fig. 11 distributions.

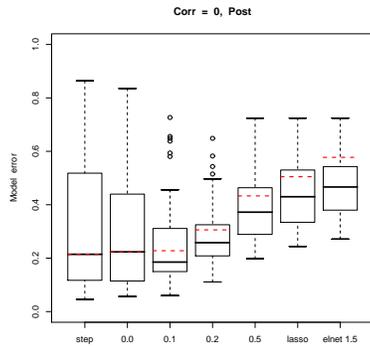


FIGURE 16. Results for Fig. 12 data when methods are used as selectors. Red lines are medians from Fig. 12 distributions.

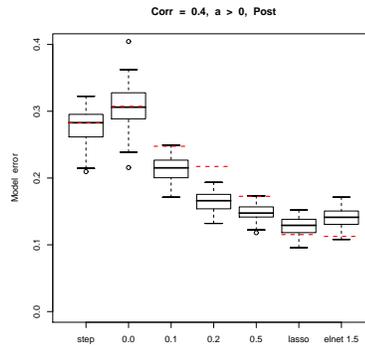


FIGURE 17. Results for Fig. 13 data when methods are used as selectors. Red lines are medians from Fig. 13 distributions.

structure of the predictor variable distribution are seen to be relevant. The results shown in Figs. 14–17 indicate that the best direct methods give rise to the best selectors. Namely, when the sparser direct methods are superior it is because they are better selectors as well as better shrinkers. This also suggests that one can reduce computation by estimating the best direct method through bridge-regression (16) (17) and then use its path as the selector to ascertain whether or not accuracy is improved.

## 8 Logistic regression

The results presented in Sections 6 and 7 were based on squared-error loss (4). Analogous results for logistic regression  $y \in \{-1, 1\}$  using (5) are quite similar and are not shown here due to space limitations. They are presented in Friedman 2008. As seen there the respective conclusions drawn from the squared-error results in Figs. 10–13 and Figs. 14–17 directly carry over to the logistic regression setting as well.

## 9 Related work

There is a large literature pertaining to regularized regression and classification. Most work involves the use of convex loss functions with convex penalties so that the overall criterion (10) is convex. Standard algorithms for convex optimization problems (Boyd and Vandenberghe 2004) can then be employed to repeatedly solve (10) for a sequence of  $\lambda$ -values. For squared-error loss and the lasso penalty, special one-at-a-time coordinate descent algorithms have been developed (Daubechines, DeFrise and De Mol 2004, Wu and Lang 2008) that are much faster than general convex optimizers for this special case. The method was extended to the full convex elastic net family of penalties by Var der Kooij 2007 and Friedman *et al* 2007. The one-at-a-time coordinate descent strategy was applied to regularized logistic and multinomial regression by Balaaji *et al* 2005 and Genkin, Lewis, and Madigan 2007, and was further generalized to the convex elastic net family by Friedman, Hastie, and Tibshirani 2008. These one-at-a-time coordinate descent algorithms are currently the fastest methods for these particular convex problems and, as seen in Table 1, their speed can rival that of GPS for those special loss-penalty combinations.

In order to obtain sparser solutions than the lasso with squared-error loss Fan and Li 2001 proposed the non convex SCAD penalty. They use an iterative approximate Newton–Raphson method for solving (10) at each  $\lambda$ -value. Lin and Wu 2007 proposed a family of non convex penalties bridging subset selection and the lasso consisting of an adjustable mixture of those two penalties. They use a mixed integer programming technique to solve (10) for square-error loss at each path point. Neither of these methods are

speed competitive with GPS. However, the GPS algorithm can be used to approximate paths based on these penalties, for any convex loss.

Direct path seeking algorithms for producing paths sparser than the lasso have been proposed by Buhlmann and Yu 2006 based on a modification of squared-error loss boosting (Friedman 2001). This procedure has connections to the sparsity inducing non negative garrote (Breiman 1995).

Again for squared-error loss, a variety of post-processing strategies using convex selectors have been proposed to create sparser paths. The relaxed lasso (Meinshausen 2007) and VISA (Radchenko and James 2008) use the lasso as the basic selector. The Dantzig selector (Candes and Tao 2007) uses a different convex constrained procedure for variable selection along its path. Although generally faster than exact methods based on non convex penalties, these selectors are still considerably slower than GPS based on those penalties. Also GPS is not limited to squared-error loss, and itself can be used to produce (non convex) selectors, as illustrated in Section 7. Rosset 2003 proposed a direct path seeking algorithm for the convex members of the power family (18) ( $\gamma \geq 1$ ) based on boosting, and illustrated its use in approximating ridge penalty ( $\gamma = 2$ ) solutions. The GPS method is a generalization of Rosset's proposal that more closely approximates exact paths for convex penalties, and extends application to non convex penalties.

## 10 Discussion

The principal advantages of using GPS to generate paths based on chosen loss-penalty combinations are simplicity, generality, and speed. The same basic algorithm can be used with a wide variety of penalty and loss criteria without the need to develop specialized search strategies for each such combination. One can concentrate on the statistical merits of the resulting regularized procedure with less concern for computational complexity. The speed of GPS extends the application of regularized regression to very large problems using any convex loss with any penalty satisfying (23) (24).

As seen in Section 6, the best penalty for any given application can strongly depend on various different aspects of the particular problem. These include the actual sparsity of the optimal coefficients  $\mathbf{a}^*$  (6), their relative signs, and the correlational structure of the predictor variable distribution. Since some or all of these properties are usually unknown, bridge-regression (16) (17) can aid in penalty choice. Again the speed of GPS makes this possible for large problems.

## 11 Acknowledgements

Helpful discussions with Trevor Hastie, Rob Tibshirani and Saharon Rosset are gratefully acknowledged. This work was partially supported by the National Science Foundation under grant DMS-97-64431.

**References**

- Balaji, K., Carlin, L., Figueiredo, M. A. T., Hartemink, A. J. (2005). Sparse multinomial logistic regression: fast algorithms and general bounds. *IEEE Trans. Pattern Analysis and Machine Intelligence* **27**, 957.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.
- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.
- Bühlmann, P. and Yu, B. (2006). Sparse Boosting. *Journal of Machine Learning Research* **7**, 1001–1024).
- Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$  (with discussion). *Annals of Statistics* **35**, 2313–2351.
- Daubechines, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics* **57**, 1413–1457.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499 (with discussion).
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.
- Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* **29**, 1189–1232.
- Friedman, J. H. (2008). Fast sparse regression and classification (long version). Stanford University, Dept. of Statistics technical report.
- Friedman, J. H., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1**, 302–332.
- Friedman, J. H., Hastie, T. and Tibshirani, R. (2008). Regularized paths for generalized linear models via coordinate descent. Stanford University, Dept. of Statistics technical report.
- Genkin, A., Lewis, D., Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**, 291–304.

- Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York.
- Hastie, T., Taylor, J., Tibshirani, R. and Walther, G. (2007). Forward stage-wise regression and the monotone lasso. *Electronic Journal of Statistics* **1**, 1–29.
- Horel, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.
- Lin, Y. and Wu, Y. (2007). Variable selection via a combination of the  $L_0$  and  $L_1$  penalties. *J. Computational and Graphical Statistics*, **16**, 782–798.
- Meinshausen, N. (2007). Relaxed lasso. *Computational statistics and Data Analysis* **52**, 374–393.
- Radchenko, P. and James, G. (2008). Variable Inclusion and Shrinkage Algorithms. *J. Amer. Statist. Assoc.* (To appear).
- Rosset, S. (2003). Topics in regularization and boosting. Ph. D. Thesis, Dept. of Statistics, Stanford University.
- Tibshirani, R. (1996). Regularization shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B* **58**, 267–288.
- Van der Kooij, A. (2007). Prediction accuracy and stability of regression with optimal scaling transformations. Ph. D Thesis, Dept. of Data Theory, Leiden University.
- Wold, S., Ruhe, A., Wold, H. and W. J. Dunn III (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal of Scientific and Statistical Computing* **5**, 735–742.
- Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* **2**, 224–244.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. B* **67**, 301–320.



# Random effects meta-analysis in the framework of the general(ized) linear mixed model

Theo Stijnen<sup>1</sup> and Taye H. Hamza<sup>2</sup>

<sup>1</sup> Dept. of Medical Statistics and Bioinformatics, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands

<sup>2</sup> Dept. of Biostatistics, Erasmus University Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

**Abstract:** In this paper we review advanced meta-analysis methods. We discuss univariate, bivariate and multivariate (regression) methods, all put into the framework of the (generalized) linear mixed model. In particular we pay attention to particular cases where an exact within study likelihood can be used. We show that the advent of the flexible (generalized) linear mixed model programs in the widely available statistical packages has made it feasible to fit advanced meta-analysis models relatively easily in practice.

**Keywords:** meta-analysis; random effects; exact likelihood; multivariate; dichotomous outcome

## 1 Introduction

Meta-analysis is used for combining findings from independently performed studies that address the same question. In the medical field it is especially used to combine all published evidence on the effect of a treatment. In this paper we consider meta-analysis of summary data, i.e. each study provides an estimate and standard error of one or more outcome parameters. Mostly these summary data are extracted from published study reports. During the last decades many statistical methods have been developed and many special purpose programs for meta-analysis have been written. In this paper we put these methods into the framework of the (generalized) linear mixed model ((G)LMM). Almost all methods used in practice make employ a normal approximation of the within studies likelihood. In that case the methods can be fitted in the LMM modules of widely available statistical packages. We pay particular attention to dichotomous outcome, as is very common in medical research, and emphasize that in that case often the exact likelihood can be used. We will show that then the modern flexible GLMM modules of packages such as SAS, STATA and R/S-Plus can be used. In section 2 we review univariate meta-analysis and discuss cases

where a hypergeometric or binomial within study likelihood can be employed. In section 3 we review bivariate meta-analysis. In section 4 we discuss multivariate meta-analysis and discuss applications to survival curve meta-analysis and to meta-analysis of ROC curves of diagnostic tests. In section 5 we formulate some conclusions.

## 2 Univariate meta-analysis

### 2.1 General case

The most common situation in meta-analysis is that we are dealing with  $i = 1, \dots, N$  studies in which a parameter of interest  $\theta_i$  is estimated. In medical applications the parameter of interest is often a measure of difference in effect between two treatments. Each study provides an estimate  $\hat{\theta}_i$  of the true value  $\theta_i$  and a standard error  $s_i$ . In the standard random effects model approach (DerSimonian and Laird, 1986) one acts as if  $\hat{\theta}_i$  has a standard normal distribution with unknown mean  $\theta_i$  and known standard error  $s_i$ , that is

$$\hat{\theta}_i \sim N(\theta_i, s_i^2) \quad (1)$$

We call this the within studies model. The true values  $\theta_i$  may vary across studies because of differences between study populations, differences in study designs, minor differences in treatments, etc. A standard assumption is that the  $\theta_i$ 's vary according to a normal distribution

$$\theta_i \sim N(\theta, \sigma^2) \quad (2)$$

This is called the between studies model. The parameter  $\theta$  is the main parameter of interest and is called the overall effect. We will refer to model (1) and (2) as the normal-normal (NN) model. In most special purpose programs for meta-analysis  $\theta$  is estimated as the weighted mean

$$\hat{\theta} = \frac{\sum_{i=1}^N w_i \hat{\theta}_i}{\sum_{i=1}^N w_i} \quad \text{with} \quad w_i = \frac{1}{\hat{\sigma}^2 + s_i^2} \quad (3)$$

where  $\hat{\sigma}^2$  is the method of moments estimator proposed by DerSimonian and Laird (1986). Alternatively the inference can be based on the likelihood

$$\prod_{i=1}^N \int L_i(\theta_i) \frac{1}{\sigma} \varphi\left(\frac{\theta_i - \theta}{\sigma}\right) d\theta_i \quad \text{with} \quad L_i(\theta_i) = \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{1}{2s_i^2}(\hat{\theta}_i - \theta_i)^2\right) \quad (4)$$

with  $\varphi$  the standard normal density. Then the estimate of  $\theta$  can be written as in (3) with  $\hat{\sigma}^2$  the ML or the REML estimate. The latter is equal to the iterated version of DerSimonian's and Laird's estimate. The model is straightforwardly extended to the NN meta-regression model by letting  $\theta$

depend on covariates,  $\theta = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$ . Since the NN meta-regression model is a special case of a linear mixed model, it can be fitted in linear mixed model modules of widely used statistical packages, such SAS or R/S-Plus. Van Houwelingen et al (2002) provide syntax examples for SAS Proc Mixed.

## 2.2 Dichotomous outcome

The above method can be applied very generally. We pay some extra attention to the case where the outcome variable is dichotomous. Then the above method can be improved by replacing the approximate normal within study likelihood by the appropriate exact likelihood. Then the model becomes a generalized linear mixed model. We present some particular cases.

### *Log odds*

Suppose the outcome variable is the occurrence of some event, and the parameter of interest is a proportion  $\pi$ , for instance the incidence of a certain adverse event under a treatment. Each study reports the number of patients with the adverse event,  $Y_i$ , and the total number of treated patients,  $n_i$ . The standard approach is to work with the  $\theta_i = \text{logit}(\pi_i)$  as the effect parameter.  $\theta_i$  is estimated by  $\hat{\theta}_i = \log(Y_i/(n_i - Y_i))$  with standard error  $s_i = \sqrt{1/Y_i + 1/(n_i - Y_i)}$ . There are some possible problems here. The first is that estimate and standard error are correlated, negatively for  $\theta_i < 0$ , and positively for  $\theta_i > 0$ . That means that studies with more extreme values of  $Y_i$  get a too low weight in (3), causing to bias towards zero (Chang et al, 2001). The second problem is what to do when  $Y_i$  is zero. In practice one adds a continuity correction of 0.5 to  $Y_i$  and  $n_i - Y_i$ . There is quite some literature on this kind of continuity corrections in meta-analysis, see for instance Bradburn et al (2007) and Sweeting et al (2004). So it is questionable whether the normal within study model (1) is appropriate, especially if events are rare. Therefore we advocated (Hamza et al, 2008) to replace it by

$$Y_i \sim \text{Binomial} \left( n_i, \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} \right) \quad (5)$$

Then the normal within study likelihood in (4) is replaced by the binomial likelihood  $\exp(\theta_i)^{Y_i}/(1 + \exp(\theta_i))^{n_i}$ . We will refer to this model as the binomial-normal (BN) model. The model now has become a *generalized* linear mixed model. More specifically, the BN model is a random intercept logistic regression model. It can be fitted in the GLMM modules of for instance SAS, STATA or R/S-Plus. See Hamza et al (2008) for syntax examples of SAS Proc NLMIXED. In extended simulations studies Hamza et al (2008) have shown that the BN model always performed better than the NN, which turned out to behave quite poorly in many scenarios, with high bias and low coverage of confidence intervals, in particular when the

proportion was near to one or zero and events were relatively rare. The BN method resulted in unbiased estimates and reasonable coverage probabilities. For smaller number of studies  $N$ , they recommended profile likelihood based confidence intervals for  $\theta$ .

#### *Log odds ratio*

Now we consider the most common situation in medical research, where two treatments are compared with respect to the occurrence of some outcome event. Almost always the log odds ratio is chosen as the treatment effect parameter to compare the treatments. Let  $Y_{i0}$  and  $n_{i0}$  denote the number of events and the number of patients in the control group of study  $i$  and let  $Y_{i1}$  and  $n_{i1}$  be defined analogously for the treatment group. The estimated log odds ratio and corresponding standard error for study  $i$  are

$$\hat{\theta}_i = \log \left( \frac{Y_{i1}/(n_{i1} - Y_{i1})}{Y_{i0}/(n_{i0} - Y_{i0})} \right) \quad \text{and} \quad s_i = \sqrt{\frac{1}{Y_{i1}} + \frac{1}{n_{i1} - Y_{i1}} + \frac{1}{Y_{i0}} + \frac{1}{n_{i0} - Y_{i0}}} \quad (6)$$

The log odds ratio is not defined in case of zero number of events, then 0.5 is added to all four numbers in the 2x2 table. The bias problem as outlined above is expected to be less serious here because a difference of two log odds is estimated, and the bias will at least partly cancel out. However, the bias in one log odds is larger when the log odds is more extreme, and the bias will not cancel out completely when the log odds ratio is big. So, especially for small numbers of events, big effects or when the within study samples sizes are unbalanced, we advocate to replace the normal within study likelihood by the exact conditional likelihood given the total number of events in the study, i.e. the likelihood of the non-central hypergeometric distribution. The likelihood becomes

$$\prod_{i=1}^N \int L_i(\theta_i) \frac{1}{\sigma} \varphi\left(\frac{\theta_i - \theta}{\sigma}\right) d\theta_i \quad \text{with} \quad L_i(\theta_i) = \frac{\binom{n_{1i}}{Y_{1i}} \binom{n_{0i}}{Y_{0i}} \exp(\theta_i Y_{1i})}{\sum_y \binom{n_{1i}}{y} \binom{n_{0i}}{Y_{0i} + Y_{1i} - y} \exp(\theta_i y)} \quad (7)$$

The idea of using this hypergeometric-normal (HN) likelihood was already given in Van Houwelingen et al (1993), but seems to have passed unnoticed. To the best of our knowledge we do not know of any applications since then. This is probably explained by the fact that the HN likelihood could not be maximized in standard statistical programs and home-written software had to be used. Nowadays routine use of this model is practically feasible, using the GLMM modules of widely available packages. We use Proc NLMIXED of SAS to fit the HN model. See Niël-Weise (2007) for an example of application of this method in a meta-analysis with very rare events. The extension to meta-regression is straightforward and can be carried out by the same software.

#### *Log incidence rate*

Now we consider the situation where per study the number of events per group,  $Y_{0i}$  and  $Y_{1i}$  respectively for group 0 and 1, and the total follow-up

time,  $T_{0i}$  and  $T_{1i}$  respectively for group 0 and 1, is given. Let  $\lambda_{0i}$  and  $\lambda_{1i}$  be the incidence rates, estimated by  $Y_{0i}/T_{0i}$  and  $Y_{1i}/T_{1i}$ , respectively. The standard errors of the log rates are  $1/\sqrt{Y_{0i}}$  and  $1/\sqrt{Y_{1i}}$ . Suppose that the association parameter used in the meta-analysis is the log incidence rate ratio. Then the standard meta-analysis approach would take

$$\hat{\theta}_i = \log \left( \frac{Y_{i1}/T_{i1}}{Y_{i0}/T_{i0}} \right) \sim N(\theta_i, s_i^2) = N \left( \log \left( \frac{\lambda_{1i}}{\lambda_{0i}} \right), \frac{1}{Y_{i0}} + \frac{1}{Y_{i1}} \right) \quad (8)$$

as the within studies model, where the usual continuity correction is applied in case of zero events in one group. Again, the normal within study distribution might not always be appropriate. Therefore, analogous to what we did above, we recommend to replace it by the exact likelihood of  $Y_{1i}$  given the total number of events  $Y_i = Y_{0i} + Y_{1i}$ . If  $Y_{0i} \sim \text{Poisson}(\lambda_{0i}T_{0i})$  and  $Y_{1i} \sim \text{Poisson}(\lambda_{1i}T_{1i})$ , then this distribution is

$$Y_{i1} \sim \text{Binomial} \left( Y_i, \frac{\exp(\log(T_{i1}/T_{i0}) + \theta_i)}{1 + \exp(\log(T_{i1}/T_{i0}) + \theta_i)} \right) \quad (9)$$

Similarly as above, (2) and (9) specify a random intercept logistic regression, but now with an offset variable  $\log(T_{i1}/T_{i0})$ . Again SAS Proc NLMIXED can be used to fit this model. Of course,  $\theta$  is allowed to depend on covariates. See Niël-Weise (2008) for an example in a case of very rare events.

### 3 Bivariate meta-analysis

We consider the case that there are two parameter estimates of interest per study,  $\hat{\theta}_{i1}$  and  $\hat{\theta}_{i2}$ . In addition to the standard errors, we assume that also an estimate of the correlation between  $\hat{\theta}_{i1}$  and  $\hat{\theta}_{i2}$  is available. The univariate NN is readily generalized to the bivariate NN model.

The within study model is

$$\begin{pmatrix} \hat{\theta}_{i1} \\ \hat{\theta}_{i2} \end{pmatrix} \sim N \left( \begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix}, \begin{pmatrix} s_{i1}^2 & s_{i12} \\ s_{i12} & s_{i2}^2 \end{pmatrix} \right) \quad (10)$$

with the estimated covariance matrix assumed to be known and fixed. The between studies model is

$$\begin{pmatrix} \theta_{i1} \\ \theta_{i2} \end{pmatrix} \sim N \left( \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right) \quad (11)$$

This model was introduced already in 1993 by Van Houwelingen et al, but seems to be not applied in practice for a long time, probably because the model could not be fitted in general statistical packages. The advent of powerful linear mixed model modules in these packages has made this feasible in practice, and in recent years the application of the bivariate model

has become more popular. Van Houwelingen et al (2002) give extended syntax examples for SAS Proc Mixed. The model may be extended with outcome specific covariates.

There are many practical situations where this model might be applied. For instance, the two outcome parameters could correspond to two different evaluation criteria in the trials, for instance the effect of a drug on HDL and LDL cholesterol. Notice that the bivariate model allows MAR missing values, so studies reporting only one of the outcomes can be included in the analysis. Another example is meta-analysis where three treatments are compared in the individual studies. In practice, when two treatments A and B are compared, researchers restrict the meta-analysis to studies where A and B are directly compared, while there might be other studies in which either A or B is compared to another treatment C. These studies carry indirect information on the A to B comparison, and can be included in the framework of a bivariate meta-analysis. The bivariate model has become very popular in special areas, such as surrogate endpoints (Molenberghs, 2007), relation between treatment effect and baseline risk (Arends et al 2007), Van Houwelingen et al (2003), Ghidry et al (2007)) and diagnostic data (Reitsma et al (2005), Arends et al (2008), Hamza et al (2008)). For instance, in meta-analysis of the accuracy of a diagnostic test often the studies report estimates of the sensitivity and specificity of the test corresponding to a particular choice of the threshold of the test. Then  $\hat{\theta}_{i1} = \text{logit}(Y_{i1}/n_{i1})$ , with  $n_{i1}$  the number of patients in the "healthy" group and  $Y_{i1}$  the number of patients with a positive test, is the estimated logit specificity with estimated variance  $s_1^2 = 1/Y_{i1} + 1/(n_{i1} - Y_{i1})$ . Similarly  $\hat{\theta}_{i2} = \text{logit}(Y_{i2}/n_{i2})$  is the estimated logit specificity with variance  $s_2^2 = 1/Y_{i2} + 1/(n_{i2} - Y_{i2})$ . Instead of the logit transformation any other transformation to  $(-\infty, +\infty)$  might be used. For instance, the probit transformation is quite popular as well. The true  $\theta_{i1}$ 's and  $\theta_{i2}$ 's usually will vary between studies because of variability in the choice of the threshold and because of possibly many other differences between the studies. Since sensitivity and specificity are negatively correlated, the above bivariate model was introduced (Reitsma, 2005) as a natural model to analyze this type of data. An estimate of a ROC curve describing the trade-off between sensitivity and specificity can be based on the estimated between studies model (11).

*Exact within studies likelihood*

When  $\hat{\theta}_{i1}$  and  $\hat{\theta}_{i2}$  are estimated proportions, the approximate within studies likelihood can be replaced by the exact within study model where  $Y_{i1}$  and  $Y_{i2}$  are conditionally independent and binomially distributed:

$$\begin{aligned} Y_{i1} &\sim \text{Binomial}\left(n_{i1}, \frac{\exp(\theta_{i1})}{1+\exp(\theta_{i1})}\right) \\ Y_{i2} &\sim \text{Binomial}\left(n_{i2}, \frac{\exp(\theta_{i2})}{1+\exp(\theta_{i2})}\right) \end{aligned} \tag{12}$$

This model was recently proposed by Chu and Cole (2006) and Arends et

al (2008) in the context of diagnostic test meta-analysis. Again the model can be fitted in GLMM programs. See Arends et al (2008) for examples of syntax for SAS Proc Mixed. In extended simulation studies Hamza et al (2008) show that using (12) instead of (11) in general leads to better estimates of the overall ROC curve.

When  $\hat{\theta}_{i1}$  and  $\hat{\theta}_{i2}$  are estimated incidence rates, i.e. with  $Y_{ij}$  the observed number of events and  $T_{ij}$  the total duration of follow-up in group  $j$ , the approximate model may be replaced by a within study model where we act as if  $Y_{i1}$  and  $Y_{i2}$  are conditionally independent and Poisson distributed:

$$\begin{aligned} Y_{i1} &\sim \text{Poisson}(n_{i1} \exp(\theta_{i1})) \\ Y_{i2} &\sim \text{Poisson}(n_{i2} \exp(\theta_{i2})) \end{aligned} \tag{13}$$

We expect that this model will perform better especially in cases of rare events. See Niël-Weise (2008) for an example of application of this model in a meta-analysis where the events were very rare.

## 4 Multivariate meta-analysis

### 4.1 General approximate likelihood model

In general we have  $p$  outcomes per study, of which some may be missing, assumed missing at random. Let  $\hat{\theta}_i = (\hat{\theta}_{i1}, \dots, \hat{\theta}_{ip})$  denote the vector of parameter estimates, with estimated covariance matrix  $\hat{\Sigma}_i$ . The approximate within studies model specifies that  $\hat{\theta}_i$  has a multivariate normal distribution

$$\hat{\theta}_i \sim MVN(\theta, \hat{\Sigma}_i) \tag{14}$$

with  $\theta_i = (\theta_{i1}, \dots, \theta_{ip})$  the vector of true study specific parameter values and  $\hat{\Sigma}_i$  assumed to be known. The between studies model is

$$\theta_i \sim MVN(\theta, \Psi) \tag{15}$$

where  $\theta = X\beta$  may depend on (outcome specific) study level covariates. The multivariate NN model can be fitted in flexible LMM programs like SAS Proc Mixed. Applications of the multivariate model are scarce at the moment, but potentially manifold. For instance it might be used for multiple outcome variables, multiple treatment groups, or when outcome is repeatedly measured over time. Network meta-analysis (Lumley, 2002), where several treatments are compared on the basis of all studies reporting a comparison of at least two of them, would be a possible application too. Arends et al (2003) discuss an application of a tri-variate meta-analysis and provide syntax examples of SAS Proc Mixed. Arends et al (2008) propose special application to meta-analysis of survival curves. We briefly describe their approach.

Consider the situation where survival is compared between two treatments  $j=1,2$  and each study in the meta-analysis provides one or more estimated survival probabilities and accompanying standard errors for both treatments. Let  $\hat{S}_{ijk}$  and  $se_{ijk}$  be the estimate and standard error of the survival probability  $S_{ijk}$  of treatment group  $j$  of study  $i$  at time  $t_{ijk}$ . In practice the number of survival estimates and their timing might be different between studies. We transform the estimates to  $\hat{\theta}_{ijk} = \log(-\log(\hat{S}_{ijk}))$  and write model (14)-(15) in standard LMM form.

$$\hat{\theta}_i = X_i\beta + Z_i b_i + \epsilon_i \quad (16)$$

with

$$b_i \sim N(0, D), \epsilon_i \sim N(0, V_i) \quad (17)$$

and

$$V_i = \begin{pmatrix} V_{1i} & 0 \\ 0 & V_{2i} \end{pmatrix} \quad (18)$$

where

$$V_{ij}(k, k') = \frac{se_{ijk}}{\hat{S}_{ijk} \log(\hat{S}_{ijk})} \sqrt{\frac{S_{ijk}(1 - S_{ijk'})}{S_{ijk'}(1 - S_{ijk})}} \frac{se_{ijk'}}{\hat{S}_{ijk'} \log(\hat{S}_{ijk'})} \quad (19)$$

The fixed part of (16) models the two average survival curves over studies. The design matrix  $X_i$  may contain variables such as intercept, time and treatment. The vector of random effects  $b_i$  typically contains a random intercept and a random time effect, assumed to be independently normally distributed with expectation zero and between studies covariance matrix  $D$ , independent from  $\epsilon_i$ .  $Z_i$  is the design matrix of the random effects, typically containing intercept, time and treatment effect. Since the residual components across time are correlated across time within a trial arm (or survival curve) but independent between treatment arms, the covariance matrix  $V_i$  is a block diagonal matrix existing of two blocks corresponding to the treatment arms. Notice that the first and third factors in (19) are the standard errors of  $\hat{\theta}_{ijk}$  and  $\hat{\theta}_{ijk'}$ . The middle factor is their correlation. Arends et al (2008) fit the model as follows. First starting values are given to  $S_{ijk}$  and  $S_{ijk'}$  in (19). Then the model is a LMM, and can be fitted for instance with SAS Proc Mixed. Next  $S_{ijk}$  and  $S_{ijk'}$  in (19) are updated, and the model is fitted again. This is repeated until convergence.

#### 4.2 Exact within studies model

Similar as in the bivariate case, in many cases it is possible to replace the approximate within studies likelihood by the exact likelihood. Arends et al (2003) describe an example of a meta-analysis in which endarterectomy (an operation to remove stenosis in the carotid arteries) is compared with

conservative medical treatment in the prevention of stroke. In the surgical group there is peri-operative mortality. Per study, three outcome parameter estimates were considered:  $\hat{\theta}_1 =$  proportion of patients in the surgical group who died due to the operation;  $\hat{\theta}_2 =$  event rate (no. of events/total follow-up time) in surgical group;  $\hat{\theta}_3 =$  event rate (no. of events/total follow-up time) in conservative group. Arends et al used the normal within study model (14). Alternatively an exact within studies likelihood could employ the product of a binomial likelihood and two Poisson likelihoods. Such a model could be fitted for instance in SAS Proc Nlmixed.

As another example we consider diagnostic meta-analysis of studies in which a single diagnostic test is administered and the results are reported using  $J - 1$  thresholds or, equivalently, in  $J$  ordered categories (Hamza et al, 2008). Let the number of non-diseased and diseased patients with test result in category  $j$  from the  $i^{th}$  study be given by  $Y_{0ij}$  and  $Y_{1ij}$ , respectively. The total number of non-diseased and diseased patients for study  $i$  is denoted by  $n_{0i} = \sum_j Y_{0ij}$  and  $n_{1i} = \sum_j Y_{1ij}$ , respectively. The aim of the analysis is to estimate an overall receiver operating characteristic (ROC), which is a curve describing *sensitivity* as a function of  $1 - \textit{specificity}$ .

Let the true logit transformed  $1 - \textit{specificity}$  and *sensitivity* for a given threshold  $j$  be denoted by  $\xi_{ij}$  and  $\eta_{ij}$  respectively, where the  $\xi_{ij}$ 's and  $\eta_{ij}$ 's are ordered in the  $j$  index. Now  $(\xi_{i1}, \eta_{i1}, \dots, \xi_{i,J-1}, \eta_{i,J-1})$  is going to play the role of  $\theta_i$  in the above formulation of the model. We assume the following hierarchical model.

*Between studies model*

Within a study we assume a linear relation between  $\eta_{ij}$  and  $\xi_{ij}$  with common slope  $\beta$  and study specific intercept  $\alpha_i$ .

$$\eta_{ij} = \alpha_i + \beta\xi_{ij} \quad \text{with} \quad \alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2) \quad (20)$$

Furthermore we assume the following model for the  $\xi_{ij}$ 's:

$$\xi_{ij} = \bar{\xi}_j + \Delta_i + \delta_{ij} \quad (21)$$

Here  $\bar{\xi}_j$  is the mean  $\xi_{ij}$  over studies,  $\Delta_i$  represents the study specific systematic deviation of the  $\xi_{ij}$ 's from the overall means  $\bar{\xi}_j$ , and  $\delta_{ij}$  represents the random residual deviation. The  $\Delta_i$ 's are assumed to follow a normal distribution given by  $\Delta_i \sim N(0, \sigma_\Delta^2)$ . The  $\delta_{ij}$ 's are assumed to be independent identically normally distributed,  $\delta_{ij} \sim N(0, \sigma_\delta^2)$ . Furthermore, the  $\delta_{ij}$ 's are independent of the  $\Delta_i$ 's and  $\alpha_i$ 's. The covariance between  $\alpha_i$  and  $\Delta_i$  is denoted by  $\sigma_{\alpha\Delta}$ . A negative  $\sigma_{\alpha\Delta}$  for instance would mean that in studies with a relatively small  $\alpha_i$  the  $\xi_{ij}$  tend to be chosen relatively high.

These assumptions lead to the following marginal between studies model:

$$\begin{pmatrix} \alpha_i \\ \xi_{i1} \\ \vdots \\ \vdots \\ \xi_{i,J-1} \end{pmatrix} \sim N \left( \begin{pmatrix} \bar{\alpha} \\ \bar{\xi}_1 \\ \vdots \\ \vdots \\ \bar{\xi}_{J-1} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\Delta} & \cdots & \cdots & \sigma_{\alpha\Delta} \\ \sigma_{\alpha\Delta} & \sigma_\Delta^2 + \sigma_\delta^2 & \sigma_\Delta^2 & \cdots & \sigma_\Delta^2 \\ \vdots & \sigma_\Delta^2 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \sigma_\Delta^2 \\ \sigma_{\alpha\Delta} & \sigma_\Delta^2 & \cdots & \sigma_\Delta^2 & \sigma_\Delta^2 + \sigma_\delta^2 \end{pmatrix} \right) \quad (22)$$

Note that the covariance structure for the  $\xi_{ij}$ 's is of compound symmetry. However, another choice such as Toeplitz, auto-regressive or unstructured might be chosen as well.

*Within study model:*

Given  $(\alpha_i, \beta, \xi_{i1}, \dots, \xi_{i,J-1})$ , the observed number of subjects in the non-diseased  $(Y_{0i1}, \dots, Y_{0iJ})$  and diseased  $(Y_{1i1}, \dots, Y_{1iJ})$  groups have independent multinomial distributions with parameters  $(\pi_{0i1}, \dots, \pi_{0iJ})$  and  $(\pi_{1i1}, \dots, \pi_{1iJ})$ , where

$$\pi_{0ij} = \begin{cases} \frac{\exp\{\xi_{ij}\}}{1+\exp\{\xi_{ij}\}} & \text{for } j = 1 \\ \frac{\exp\{\xi_{ij}\}}{1+\exp\{\xi_{ij}\}} - \frac{\exp\{\xi_{i,j-1}\}}{1+\exp\{\xi_{i,j-1}\}} & \text{for } j = 2 \dots, J-1 \\ 1 - \frac{\exp\{\xi_{i,j-1}\}}{1+\exp\{\xi_{i,j-1}\}} & \text{for } j = J \end{cases} \quad (23)$$

$$\pi_{1ij} = \begin{cases} \frac{\exp\{\eta_{ij}\}}{1+\exp\{\eta_{ij}\}} & \text{for } j = 1 \\ \frac{\exp\{\eta_{ij}\}}{1+\exp\{\eta_{ij}\}} - \frac{\exp\{\eta_{i,j-1}\}}{1+\exp\{\eta_{i,j-1}\}} & \text{for } j = 2 \dots, J-1 \\ 1 - \frac{\exp\{\eta_{i,j-1}\}}{1+\exp\{\eta_{i,j-1}\}} & \text{for } j = J \end{cases} \quad (24)$$

Now the within study likelihood given the  $\pi_{0ij}$ 's and  $\pi_{1ij}$ 's of the observations of the  $i^{\text{th}}$  study is given by

$$\prod_{j=1}^J \pi_{0ij}^{Y_{0ij}} \pi_{1ij}^{Y_{1ij}} \quad (25)$$

up to factors not depending on the parameters.

*Example: Fine-needle aspiration cytologic examination*

Giard and Hermans (1992) present 29 studies evaluating the accuracy of fine-needle aspiration cytologic examination (FNAC) of the breast to assess the presence of breast cancer. FNAC provides a non-operative way of obtaining cells for the establishment of the nature of a breast lump and therefore plays a pivotal role in the preoperative diagnostic process. The selected FNAC results were classified in the following three cytologic categories: definitely malignant, suspect for malignancy, and benign (Table 1). Giard and Hermans (1992) provide the full data.

Giard and Hermans (1992) analysed sensitivity and specificity of the FNAC test by reducing the two-by-three table for each study into a two-by-two table. They classified malignant and suspect test results as test result positive,

TABLE 1. Two-by-three contingency table for study  $i$  for relating the FNAC outcome to the final diagnosis of breast lesion.

FNAC outcome	Malignant	Suspect	Benign	Total
Final diagnosis				
Malignant	$Y_{1i1}$	$Y_{1i2}$	$Y_{1i3}$	$n_{1i}$
Benign	$Y_{0i1}$	$Y_{0i2}$	$Y_{0i3}$	$n_{0i}$

and benign as test result negative. We fitted the above multivariate model with SAS Proc Nlmixed. The estimated mean  $\xi_j$ 's were  $\bar{\xi}_1 = -7.084(0.408)$  and  $\bar{\xi}_2 = -2.548(0.260)$ , and the estimated variances and covariances were  $\sigma_\alpha^2 = 0.363(0.117)$ ,  $\sigma_{\alpha\Delta} = -0.045(0.143)$ ,  $\sigma_\Delta = -0.042(0.443)$  and  $\sigma_\delta^2 = 1.841(0.615)$ . The test for the significance of the covariance between  $\alpha_i$  and  $\xi_i$  was not significant (likelihood ratio  $\chi_1^2 = 1.00$ , p-value 0.317), and hence there is no indication for the choice of the  $\xi_i$ 's to depend on the level of individual curves. The estimates for the intercept and slope were 2.368(0.135) and 0.224(0.016). The area under the overall ROC curve was 0.902.

### 4.3 Conclusions

In this contribution we have put random effects meta-analysis into the framework of the (generalised) linear mixed model. We have shown that the powerful software provided by widely available general statistical packages makes it possible to fit advanced models relatively easily. In particular multivariate meta-analysis is feasible in practice now and should be applied more than it is today, since it potentially has many advantages. We have shown that in many cases the approximate within studies likelihood can be replaced by an exact likelihood, and that GLMM programs might be used to fit the resulting exact likelihood methods.

### References

- Arends, L.R., Voko, Z., Stijnen, T. (2003) Combining multiple outcome measures in a meta-analysis: an application. *Statistics in Medicine*, **22**, 1335-1353
- Arends, L. R., Hamza, T.H., Van Houwelingen, J.C., Heijenbrok-Kal, M.H., Hunink, M.G.M. and Stijnen, T. (2008) Meta-Analysis of ROC Curves Using a Bivariate Normal Distribution of Sensitivities and Specificities. *Medical Decision Making* (to appear)
- Arends, L.R., Hunink, M.G.M. and Stijnen, T. (2008) Meta-analysis of summary survival curve data. *Statistics in Medicine*, to appear.

- Chu, H. and Cole, S.R. (2006) Bivariate meta-analysis for sensitivity and specificity with sparse data: a generalized linear mixed model approach (letter to the Editor). *Journal of Clinical Epidemiology*, **59**, 1331-1331.
- Bradburn, M.J., Deeks, J.J., Berlin, J.A. and Localio, A.R. (2007) Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, **26**, 53-77.
- Chang, B.H., Waternaux, C. and Lipsitz, S. (2001). Meta-analysis of binary data: which study variance estimate to use? *Statistics in Medicine*, **20**, 1947-1956.
- DerSimonian, R., and Laird, N. (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177-188.
- Ghidey, W., Lesaffre, E. and Stijnen, T. (2007) Semi-parametric modelling of the distribution of the baseline risk in meta-analysis. *Statistics in Medicine 2007*, **26**, 5434-5444 .
- Giard, R.W.M. and Hermans, J. (1992) The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. *Cancer*, **69**, 2104-2110.
- Hamza, T.H., Van Houwelingen, J.C., Stijnen T. (2008) Random effects meta-analysis of proportions: The binomial distribution should be used to model the within study variability. *Journal of Clinical Epidemiology*, **61**, 41-51.
- Hamza, T.H., Reitsma, J.B., Stijnen, T. (2008) Random effects meta-analysis of diagnostic tests: a comparison of random intercept, normal-normal and binomial-normal bivariate random effects SROC approaches. *Medical Decision Making*, to appear.
- Hamza, T.H., Van Houwelingen, J.C. and Stijnen, T. Multivariate random effects meta-analysis of diagnostic tests with multiple thresholds. (Submitted)
- Lumley, T (2002) Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**, 2313-2324.
- Molenberghs, G., Burzykowski, T., Alonso, A., Assam, P., Tilahun, A. and Buyse M. (2007) The Meta-analytic Framework for the Evaluation of Surrogate Endpoints in Clinical Trials. In: *Proceedings of the IWSM 21*, 21-27.
- Niël-Weise, B.S., Stijnen, T. and Van den Broek, P.J. (2007) Anti-infective-treated central venous catheters: a systematic review of randomized controlled trials. *Intensive Care Medicine*, **33**, 2058-2068.

- Niël-Weise, B.S., Stijnen, T. and Van den Broek, P.J. (2008) Anti-infective-treated central venous catheters for total parenteral nutrition or chemotherapy? A systematic review. *Journal of Hospital Infection* (to appear)
- Reitsma, J.B., Glas, A.S., Rutjes, A.W.S., Scholten, R.J.P.M., Bossuyt, P.M., Zwinderman, A.H. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology* **58**, 982-990.
- Sweeting MJ, Sutton AJ, Lambert PC. (2004) What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine*, **23**, 1351-75.
- Van Houwelingen, H.C., Zwinderman, K.H., and Stijnen, T. (1993) A bivariate approach to meta-analysis. *Statistics in Medicine*, **12**, 2273-2284.
- Van Houwelingen, H.C., Arends, L.R. and Stijnen, T. (2002) Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, **21**, 589-624.



# Boosting Strategies in Semiparametrically Structured Regression

Gerhard Tutz<sup>1</sup>

<sup>1</sup> Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München

**Abstract:** Early boosting procedures were very successful in improving classification algorithms by applying reweighted versions of the input data. With the representation of the procedure as a functional optimization algorithm boosting has become a tool with strong potential in high dimensional regression problems. Here a general likelihood-based boosting procedure is considered that provides tools for model selection, regularization and feature selection in semiparametrically structured regression. From the wide range of applications we will select three: the structuring of the predictor in high-dimensional regression problems, constrained regression where the effect function is constrained to be monotonically decreasing or increasing and signal regression where predictors have an underlying metric but the number of predictors is in the hundreds.

**Keywords:** Boosting; semiparametrically structured regression; signal regression; constrained regression.

## 1 Introduction

Boosting has its origins in the machine learning community where boosting was introduced as a method to improve classifiers. Freund & Schapire (1997) introduced boosting as a method that converts a "weak" learning algorithm into one with high accuracy. The method was conceptualized in the spirit of ensemble schemes. When using ensemble methods one fits several models and lets them vote for the most popular class. In the widely used AdaBoost algorithm the ensemble is composed of the various classifiers that are built by putting different weights on observations. The statistical view of boosting was strongly stimulated by the seminal paper of Friedman et al (2000) who showed that earlier concepts of boosting perform stagewise additive model fitting, see also Breiman (1999) who showed for the first time a connection between boosting and numerical optimization techniques and Friedman (2001) who gave several algorithms which aim at the stagewise minimization of gradients. An up-to-date overview of the statistical view of boosting based on the least squares approximation of gradients was given by Bühlmann & Hothorn (2007).

The focus of our presentation of boosting methods is on generalized structured model building. We want to show how boosting may be used for

regularization, model choice and feature extraction in model building, in particular when the predictor space is high-dimensional. The semiparametrically structured models we have in mind have the form

$$\mu(\mathbf{x}) = h(\eta(\mathbf{x})) \quad \text{or} \quad g(\mu(\mathbf{x})) = \eta(\mathbf{x}),$$

where  $\mu(\mathbf{x})$  denotes the mean response  $\mu(\mathbf{x}) = E(y | \mathbf{x})$  which depends on covariate vector  $\mathbf{x}$ ,  $g$  is the link function ( $h = g^{-1}$  denotes the response function) and it is assumed that  $y | \mathbf{x}$  follows a simple exponential family. When the link function has been fixed it is especially important to find an appropriate structure of the predictor  $\eta(\mathbf{x})$ . In the simplest case of generalized linear models it is a linear term,  $\eta(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$ , in the more general generalized additive model the predictor has the form

$$\eta(\mathbf{x}) = f_{(1)}(x_1) + \cdots + f_{(p)}(x_p),$$

where  $f_{(j)}(x_j)$  are unspecified functions to be estimated. More generally, the additive predictor may also include spatial effects or varying coefficient terms. Very high-dimensional predictors occur in functional data analysis. For example in signal regression the predictor is a function which in discretized form yields a high-dimensional vector  $\mathbf{x}$  with some attached metric. As an example consider Figure 1 which shows signal regressors from near infrared spectroscopy applied to the compositional analysis of 32 marzipan samples (Christensen et al, 2004). Each "signal" consists of 600 digitizations along the wavelength axis. The objective of the analysis is to determine moisture and sugar content from these signals.

## 2 Functional Gradient Descent Boosting

A breakthrough for the statistical view of boosting was the representation of boosting as a steepest descent algorithm in function space. In the following the basic concept is given. For a new observation  $(y, \mathbf{x})$  one considers the problem of minimizing  $E[L(y, G(\mathbf{x}))]$  with respect to  $G(\mathbf{x})$  for some specified loss function  $L(\cdot, \cdot)$ , and an appropriately chosen value  $G(\mathbf{x})$ . A simple example is the squared error loss  $L_2(y, G(\mathbf{x})) = (y - G(\mathbf{x}))^2$ , where  $G(\mathbf{x})$  is considered to represent an approximation of  $y$ . In order to obtain a practical implementation for finite data sets, one minimizes the empirical version of the expected loss,  $\sum_i L(y_i, G(\mathbf{x}_i))/n$ . Minimization is obtained iteratively by utilizing a *steepest gradient descent approach*. An essential ingredient of the method is the fitting of a structured function as an approximation to  $G(\mathbf{x})$ . This fitting may be seen as a *base procedure*. Depending on the modeling problem one may for example fit a regression spline function or a tree. Thus in each iteration step  $G(\mathbf{x})$  is approximated by an (parameterized) estimate  $\hat{g}(\mathbf{x}, \{u_i, \mathbf{x}_i\})$  which is based on input data  $\{u_i, \mathbf{x}_i\}$ . The input data are not the original data but are generated during the fitting process by computing the derivatives

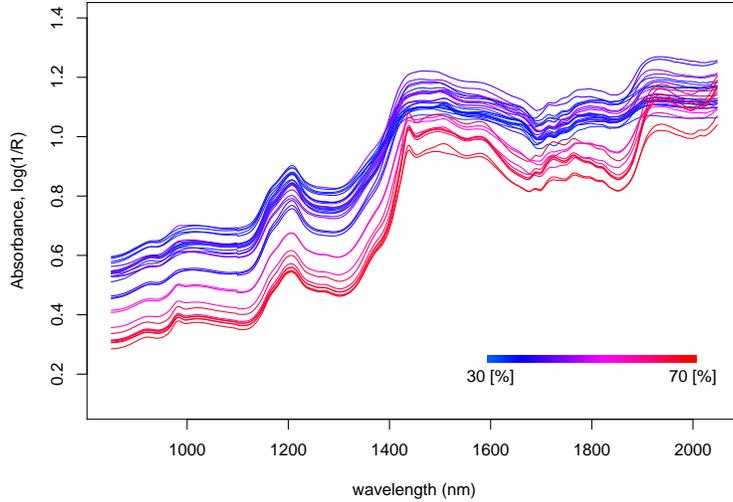


FIGURE 1. Near-infrared spectra of 32 marzipan samples; colors corresponding to sugar content.

$$u_i = -\frac{\partial L(y_i, G(\mathbf{x}))}{\partial G(\mathbf{x})} \Big|_{G(\mathbf{x})=G(\mathbf{x}_i)}, \quad i = 1, \dots, n. \quad (1)$$

Basically the functional gradient descent algorithm consists of iteratively refitting these pseudo-response values. Therefore the final solution is attained in a stage-wise manner. We give a version of this algorithm which is close to the Gradient Boost algorithm given in Friedman (2001).

Let  $\hat{g}(\mathbf{x}, \{u_i, \mathbf{x}_i\})$  denote the base procedure at value  $\mathbf{x}$  based on input data  $\{u_i, \mathbf{x}_i\}$ , which are not necessarily the original data  $\{y_i, \mathbf{x}_i\}$ .

---

### Functional Gradient Descent Boosting

#### *Step 1 (Initialization)*

Given data  $\{y_i, \mathbf{x}_i\}$  fit a base procedure for initialization, yielding the function estimate  $G^{(0)}(\mathbf{x}) = \hat{g}_0(\cdot, \{y_i, \mathbf{x}_i\})$ . For example a constant  $c$  is fitted, yielding  $G^{(0)}(\mathbf{x}) \equiv \arg \min_c \frac{1}{n} \sum_{i=1}^n L(y_i, c)$ .

*Step 2 (Iteration)* For  $l = 0, 1, \dots$

1. *Fitting step*

Compute the values of the negative gradient  $u_i$  as given in (1), evaluated at  $G^{(l)}(\mathbf{x}_i)$ . Fit a base procedure to the current data  $\{u_i, \mathbf{x}_i\}$ . The fit  $\hat{g}(\cdot, \{u_i, \mathbf{x}_i\})$  is an estimate based on the original predictor variables and the current negative gradient vector.

2. *Update step*

The improved fit is obtained by the update

$$G^{(l+1)}(\cdot) = G^{(l)}(\cdot) + \nu \hat{g}(\cdot, \{u_i, \mathbf{x}_i\}),$$

where  $\nu \in (0, 1]$  is a fixed shrinkage parameter which should be sufficiently small.

*Step 3 (Final estimator)*

Obtain the final estimator after an optimized number of iterations  $l_{\text{opt}}$ , i.e.  $\hat{G}^{(l_{\text{opt}})}(\mathbf{x})$ .

An example for which the negative gradient has a very simple form is the squared error loss  $L_2(y, G(\mathbf{x})) = (y - G(\mathbf{x}))^2$ . Then the negative gradient vector consists of simple residuals  $u_i = -\partial L(y_i, G_i)/\partial G = 2(y_i - G_i)$ . Therefore, in the fitting step a model is fit with the original responses replaced by the current residuals. Essentially the same procedure with just one boosting step was proposed already by Tukey (1977). In contrast to the original version of GradientBoost, the algorithm renounces an additional line search step between the fitting and update step, which calibrates the shrinkage parameter within each step  $\nu$ . For obtaining an accurate estimate  $\hat{G}^{(l_{\text{opt}})}(\mathbf{x})$  constant  $\nu$  seems to do well (see Bühlmann & Hothorn (2007)). However, the shrinkage parameter  $\nu$  plays an important role. It makes the learner a weak one and avoids early overfitting of the procedure. Therefore it should be chosen rather small (e.g.  $\nu = 0.1$ ).

Boosting by gradient descent is a stagewise strategy that is different from a stepwise approach that readjusts previously entered terms. Since previously entered terms are not adjusted it is a *greedy function approximation* which performs *forward stagewise additive modelling*. The additive fit becomes obvious from the final estimator which has the form

$$\hat{G}^{(l_{\text{opt}})}(\cdot) = \hat{g}_0(\cdot, \{y_i, \mathbf{x}_i\}) + \sum_{j=1}^{l_{\text{opt}}} \nu \hat{g}(\cdot, \{u_i^{(j-1)}, \mathbf{x}_i\}),$$

where  $u_i^{(j)} = -\partial L(y_i, G^{(j)}(\mathbf{x}_i))/\partial G(\mathbf{x})$  are the negative gradient values from step  $j$ .

For the squared error loss,  $G_i = G(\mathbf{x}_i)$  is estimated as an approximation to  $y_i$ . However, when the negative likelihood is used as loss function it is advisable to choose  $G(\mathbf{x}_i)$  as a value that is connected to the response. For

example when the response is binary one usually uses the half of the logits, thus implicitly fitting a logistic regression model (see Friedman et al, 2000). When viewing boosting as functional gradient descent approach it is possible to derive some interesting properties of the resulting estimates. Bühlmann & Yu (2003) showed that one obtains for the  $L_2$ -loss a bias-variance trade-off where the variance increases exponentially small with the number of iterations. In addition, they show that when a smoothing spline is used as base procedure, the algorithm reaches the optimal rate of convergence for one-dimensional function estimation. Furthermore, the procedure captures a higher degree of smoothness than the smoothing spline. Bühlmann (2006) investigated also componentwise least squares estimates as base procedure and showed that the algorithm provides a consistent estimator for high-dimensional models where the number of covariates is allowed to grow exponentially with the sample size under some sparseness assumptions.

### 3 Likelihood-based boosting

Within the functional gradient descent approach one may define the negative likelihood as a loss function and specify an appropriate value  $G(\mathbf{x}_i)$  that is estimated. When fitting semiparametrically structured regression models we prefer an alternative representation of the boosting procedure which is outlined in the following.

#### 3.1 Basic concept

Let  $y_i$  be from an exponential family distribution with mean  $\mu_i = E(y_i|\mathbf{x}_i)$  and the link between the mean and the structuring term specified in the usual form

$$\mu_i = h(\eta_i) \quad \text{or} \quad g(\mu_i) = \eta_i$$

with link function  $g$ . Then the *likelihood-based generalized model boosting* (GenBoost) tries to improve the predictor  $\eta(\mathbf{x})$  by greedy forward additive fitting. It may be given in the following form:

---

#### Likelihood GenBoost

*Step 1 (Initialization)*

For given data  $(y_i, \mathbf{x}_i), i = 1, \dots, n$ , fit the intercept model  $\mu^{(0)}(\mathbf{x}) = h(\eta_0)$  by maximizing the likelihood, yielding  $\eta^{(0)} = \hat{\eta}_0, \hat{\mu}^{(0)} = h(\hat{\eta}_0)$ .

*Step 1 (Iteration)* For  $l = 0, 1, 2, \dots$ ,

Fit the model

$$\mu_i = h(\hat{\eta}^{(l)}(\mathbf{x}_i) + \eta(\mathbf{x}_i, \gamma))$$

to data  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , where  $\hat{\eta}^{(l)}(\mathbf{x}_i)$  is treated as an offset and the predictor is estimated by fitting the parametrically structured term  $\eta(\mathbf{x}_i, \boldsymbol{\gamma})$ , obtaining  $\hat{\boldsymbol{\gamma}}$ .

The improved fit is obtained by

$$\hat{\eta}^{(l+1)}(\mathbf{x}_i) = \hat{\eta}^{(l)}(\mathbf{x}_i) + \hat{\eta}(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}), \quad \hat{\mu}_i^{(l+1)} = h(\hat{\eta}^{(l+1)}(\mathbf{x}_i))$$

One candidate for fitting is shrunk Fisher scoring which is familiar from generalized linear model fitting. One has to compute the pseudo-responses and weights

$$\tilde{\eta}_i = \frac{y_i - \hat{\mu}_i^{(l)}}{\partial h(\hat{\eta}_i^{(l)})/\partial \eta}, \quad w_i = \frac{(\partial h(\hat{\eta}_i^{(l)})/\partial \eta)^2}{\sigma_i^2},$$

and then compute weighted regression with weights  $w_i$  and dependent variables  $\tilde{\eta}_i$  in order to obtain  $\hat{\boldsymbol{\gamma}}_l$ . For the logit model one has  $\partial h(\hat{\eta}_i)/\partial \eta = h(\eta_i)/(1 - h(\eta_i))$  and therefore pseudo-responses and weights simplify to

$$\tilde{\eta}_i = \frac{y_i - \hat{\mu}_i^{(l)}}{\hat{\pi}^{(l)}(1 - \hat{\pi}^{(l)})}, \quad w_i = \hat{\pi}^{(l)}(1 - \hat{\pi}^{(l)}).$$

In the case of the logit model one obtains the *LogitBoost* algorithm for two classes which was proposed by Friedman et al (2000) and which may be derived as a gradient descent procedure. For early concepts of likelihood-based boosting procedures see also Ridgeway (1999).

### 3.2 Componentwise Boosting and Generalized Semiparametric Structuring

The advantage of GenBoost is that it applies for all kinds of link functions and exponential family responses. The crucial part in GenBoost, as in all boosting procedures, is the choice of the learner, i.e. the term  $\eta(\mathbf{x}_i, \boldsymbol{\gamma})$ . In earlier statistical boosting literature (Breiman (1998), Friedman et al (2000), Friedman (2001)) CARTs were recommended. Then the performance depends on the tree depth and higher tree depths may yield superior error rates; however, results are hardly interpretable. When the focus is on structured regression it is preferable to use so-called *componentwise boosting methods* where in each step the contribution of only one variable is updated. Within the fitting step a selection step is included that determines which of the variables is refitted. In the simplest linear case one fits all one-covariate models  $\eta(\mathbf{x}_i, \boldsymbol{\gamma}) = \gamma_0 + \gamma_j x_{ij}$ ,  $j = 1, \dots, p$ , obtaining  $\hat{\gamma}_j$ , and then selects the variable that has the strongest impact on the improvement of the fit. A criterion is the improvement in deviance

$$Dev(\hat{\eta}^{(l)}) - Dev(\hat{\eta}_{new(j)}),$$

where  $\tilde{\eta}_{new(j)}$  is based on the parameter vector in which only the intercept and the  $j$ th component are updated to  $\hat{\gamma}_j^{(l)} + \hat{\gamma}_j$ . When the  $s$ th variable is selected the new parameter vector is  $\hat{\gamma}^{(l+1)} = (\hat{\gamma}_0^{(l)} + \hat{\gamma}_0, \hat{\gamma}_1^{(l)}, \dots, \hat{\gamma}_s^{(l)} + \hat{\gamma}_s, \dots)^T$ .

For the general case of semiparametrically structured models let the candidate structuring be given by

$$\eta(\mathbf{x}) = \gamma_0 + \sum_j f_j(\mathbf{x}),$$

where the functions  $f_j$  denote generic representations of different types of covariate effects, usually depending only on few components of the predictor vector  $\mathbf{x}$ . Examples of  $f_j$  are

- $f(\mathbf{x}) = x_r \gamma$ , which specifies the linear effect of one covariate
- $f(\mathbf{x}) = \mathbf{x}_r^T \boldsymbol{\gamma}$  where  $\mathbf{x}_r$  is a vector of dummy variables corresponding to a categorical variable
- $f(\mathbf{x}) = f_{(r)}(x_r)$  where a smooth function of one covariate is assumed
- $f(\mathbf{x}) = f_{(r,s)}(x_r, x_s)$  representing an interaction surface depending on two covariate (including spatial effects)
- $f(\mathbf{x}) = x_r f(x_s)$  where the effect of  $x_r$  is modified (smoothly) by covariate  $x_s$ .

In addition one may also specify cluster-specific effects where observations are given in clusters.

Componentwise boosting means that in the fitting step just one of the functions  $f_j$  is updated. For example a smooth additive term  $f_{(r)}(x_r)$  may be fitted by P-splines (Eilers & Marx, 1996). Let the function be expanded in basis functions  $f(x_r) = \sum_s \gamma_{rs} \phi_s(x_r)$  and  $\mathbf{X}_r$  denote the corresponding design matrix composed from the evaluations of the basis functions at observations. Then the update of the vector of evaluations based on one-step Fisher scoring has the form

$$\hat{\mathbf{f}}_r = \mathbf{X}_r (\mathbf{X}_r^T \hat{\mathbf{W}} \mathbf{X}_r + \lambda \mathbf{A})^{-1} \mathbf{X}_r^T \hat{\mathbf{W}} \hat{\mathbf{D}}^{-1} (\mathbf{y} - \hat{\boldsymbol{\mu}}),$$

where  $\hat{\mathbf{W}}$  is the corresponding weight matrix,  $\hat{\mathbf{D}}$  contains the derivations, and  $\mathbf{y}, \hat{\boldsymbol{\mu}}$  are the vectors of observations and previously fitted values. The form of the penalty term  $\lambda \mathbf{A}$  which includes the smoothing parameter,  $\lambda$  that steers the weakness of the learner, is determined by the basis functions that have been chosen. The algorithm minimizes a penalized likelihood with penalty term  $(\lambda/2) \boldsymbol{\gamma}_r^T \mathbf{A} \boldsymbol{\gamma}_r$ .

When all the candidate functions are univariate smooth effects of just one variable one fits a generalized additive model (GAM). Then the stepwise algorithm with fixed (large) smoothing parameter  $\lambda$  adapts automatically to

the amount of smoothing needed for the single components of the additive predictor by selecting functions that have stronger curvature more often. In conventional algorithms for the fitting of GAMs appropriate smoothing of single components may be obtained by selecting one smoothing parameter for one variable which requires a simultaneous search in  $p$ -dimensional space and usually fails when  $p$  is large. Tutz & Binder (2006), Binder & Tutz (2008) show that the fitting of GAMs by boosting techniques works very well in high dimensions (up to  $p = 50$ ) and is able to select the relevant variables.

In the case of mixed terms in the candidate predictor the smoothing parameter that is used in the update step corresponds to different types of functions, for example to linear terms for categorical variables or to smooth additive terms for continuous variables. Then the smoothing should be adapted to the type of function since otherwise smooth functions are preferred because the underlying continuous variable contains more information than a categorical variable. An adaptive scheme was proposed by Kneib et al (2007).

It is important that the fitting step uses a weak learner, that means one does not want to optimize the closeness of the fit to the data, but improve only slightly on the fit. Otherwise the resistance to overfitting which is an attractive feature of boosting approaches is lost. That means that the smoothing parameter  $\lambda$  which essentially is a shrinkage parameter has to be chosen very large, with the effect that the change within each step is small. The crucial tuning parameter in boosting procedure is not the  $\lambda$  but the stopping criterion. While performance is rather insensitive to the choice of the smoothing parameter (given it is large) the stopping criterion determines how smooth single components are fitted and how many covariates are included - it represents the regularization of the estimate.

A prediction-based stopping criterion is cross-validation. Alternatively one may use the likelihood-based AIC or BIC criterion which seems more appropriate for regression model fitting. One strategy is to stop when AIC or BIC gets worse, or one performs a large number of boosting steps and then obtains the number of steps that shows best performance in terms of AIC or BIC.

### 3.3 Constrained Regression

In many studies it is known that the effect of one or more of the covariates is constrained to a special type of function. For example in some economic theories one knows that the relationship between input and output is concave and nondecreasing. In biometrical applications, when modelling the effect of air pollution on mortality or illness, it is also sensible to assume that the effect is monotonically nondecreasing.

Although there is a large body of literature on isotonic regression usually only the one-dimensional predictor case is considered. When several co-

variates are given within a GAM modelling approach the predictor has the form

$$\eta(x_j) = \gamma_0 + f_{(1)}(x_1) + \dots + f_{(p)}(x_p)$$

where it is assumed that  $f(x) \geq f(z)$ , if  $x > z$ , for one or more than one regression function  $f_{(j)}$ . Boosting as a stepwise learner offers a simple way to handle monotonicity constraints. How these are implemented depends on the basis functions that are used in the expansion  $f_{(r)}(x) = \sum_s \gamma_{rs} \phi_s(x)$ . One approach is to use monotonically increasing basis functions, for example integrated splines (so-called I-splines, see Ramsay, 1988) or sigmoidal functions like the logistic functions  $\phi_s(x) = \{1/[1 + \exp(-(x - t_j)/\delta)]\}$  with knots  $t_1 < t_2 \dots < t_m$  and control in the fitting step for the monotonicity constraint  $\gamma_{rs} \geq 0$  for  $s = 1, \dots, m$ . This may be accomplished in the spirit of componentwise boosting by fitting (with strong shrinkage) just one of the parameters  $\gamma_{rs}$  at a time and then select from the resulting set of updated parameters which maintain the monotonicity constraint that parameter that improves the fit the most (for details see Tutz & Leitenstorfer, 2007).

Alternatively, one may use the familiar B-splines which are not monotonic by themselves and control for the monotonicity constraint  $\gamma_{r,s+1} \geq \gamma_{r,s}$ ,  $s = 1, \dots, m$ . The basic concept is to split the expansion of function  $f_{(r)}$  within the fitting step into two components

$$\gamma_{r1(c)} \sum_{s=1}^c B_s(x_r) + \gamma_{r2(c)} \sum_{j=c+1}^m B_s(x_r)$$

and consider updates  $\hat{\gamma}_{rs}^{\text{new}} = \hat{\gamma}_r^{\text{old}} + \hat{\gamma}_{r1(c)}$ ,  $r = 1, \dots, c$ ,  $\hat{\gamma}_{rs}^{\text{new}} = \hat{\gamma}_r^{\text{old}} + \gamma_{r2(c)}$ ,  $r = c + 1, \dots, m$  only for strongly shrunk fits which fulfill  $\hat{\gamma}_{r1(c)} \leq \hat{\gamma}_{r2(c)}$ . The embedding into boosting algorithm and examples are given in Leitenstorfer & Tutz (2007).

## 4 Signal Regression

An area where regularization and feature extraction is indispensable is signal regression where the predictors itself are functions. For example the functions in Figure 1 may be seen as high-dimensional predictors for the response sugar content. Common methods for signal regression assume that the effect of the signal is a smooth function that is estimated by some regularization technique. However, by specifying a smooth function the whole signal is used without reduction of the signal to the relevant parts. The boosting procedure proposed in the following is able to do feature extraction by selecting areas from the range of the signal that have an effect on the response.

In general, in signal regression data are given by  $(y_i, x_i(t))$ ,  $i = 1, \dots, n$ , where  $y_i$  is the (metric) response variable and  $x(t)$ ,  $t \in I$  denotes a function

defined on an interval  $I \subset \mathbf{R}$ , also called the signal. A functional linear model for scalar responses has the form

$$y_i = \beta_0 + \int x_i(t)\beta(t)dt + \varepsilon_i,$$

where  $\beta(\cdot)$  is a parameter function and  $\varepsilon_i$  with  $E(\varepsilon_i) = 0$  represents a noise variable, cf. Ramsay & Silverman (2005).

The discretized form of the functional linear model is given by

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i \quad (2)$$

where  $x_{ij} = x_i(t_j)$ ,  $\beta_j = \beta(t_j)$  for values  $t_1 < \dots < t_p$ ,  $t_j \in I$ . For simplicity we take the values  $t_1, \dots, t_p$  as equidistant,  $t_{j+1} - t_j = \Delta$ . For the (original) marzipan data the digitization along the wavelength axis yields  $p = 600$ , where  $\Delta = 2$  nm has been chosen.

The naive approach, fitting by least squares frequently yields perfect fit of the data with poor predictive value. Common fitting procedures are principal components regression, partial least squares (Wold, 1975), ridge regression, the functional data approach suggested by Ramsay & Silverman (2005) or P-spline signal regression (Marx & Eilers, 1999). All of these methods fit smooth functions without interpretable feature extraction. In contrast, the lasso (Tibshirani, 1996) selects predictors and the fused lasso (Tibshirani et al, 2005) creates piecewise constant fits.

We suggest a blockwise boosting technique that fits smooth functions  $\beta(t)$  on selected areas of the predictor. The method differs from componentwise boosting in the way features are selected. Rather than selecting single variables the approach selects groups of variables where the grouping of variables is based on a metric. Since a signal  $x_i(\cdot)$  may be seen as a mapping  $x_i : I \rightarrow \mathbf{R}$  one utilizes that a metric is available on  $I$ . A potentially relevant part of the signal  $x_i(\cdot)$  may be characterized by  $\{x_{i,U}(t) | t \in U(t_0)\}$  where  $U(t_0)$  is defined as a neighborhood of  $t_0 \in I$ , i.e.  $U(t_0) = \{t | \|t - t_0\| \leq \delta\}$  for some metric  $\|\cdot\|$  on  $I$ . For the digitized signal the potentially relevant signal part turns into the group of variables  $\{x_{ij} | t_j \in U(t_0)\}$ . When the Euclidean metric is used,  $U(t_0)$  may be identified as a sub-interval from  $I$ . Blockwise boosting aims at updating groups of adjacent variables  $\{x_{ij} | t_j \in U\}$  for alternative sets  $U$ . For simplicity, we use in the updating procedure the subsets  $U_s = U_k(t_s) = [t_s, t_s + (k - 1)\Delta]$ ,  $k \in \{1, 2, \dots\}$ . Thus for  $k = 1$  one obtains the limiting case of single variables  $\{x_i(t_1)\}$ ,  $\{x_i(t_2)\}$ , for  $k = 2$  one gets pairs of variables  $\{x_i(t_1), x_i(t_2)\}$ ,  $\{x_i(t_2), x_i(t_3)\}$ , etc. In the following  $k$  is considered as fixed and the index  $k$  is suppressed.

Let  $X^{(s)}$  denote the design matrix of variables from  $U_s$ , i.e.  $X^{(s)}$  has rows  $(x_i(t_s), \dots, x_i(t_{s+k-1})) = (x_{is}, \dots, x_{i,s+k-1})$ . An update step of blockwise boosting will be based on estimating the vector  $b^{(s)} = (b(t_s), \dots, b(t_{s+k-1}))^T$  from data  $(u, X^{(s)})$  where  $u = (u_1, \dots, u_n)^T$  denotes the current residual.

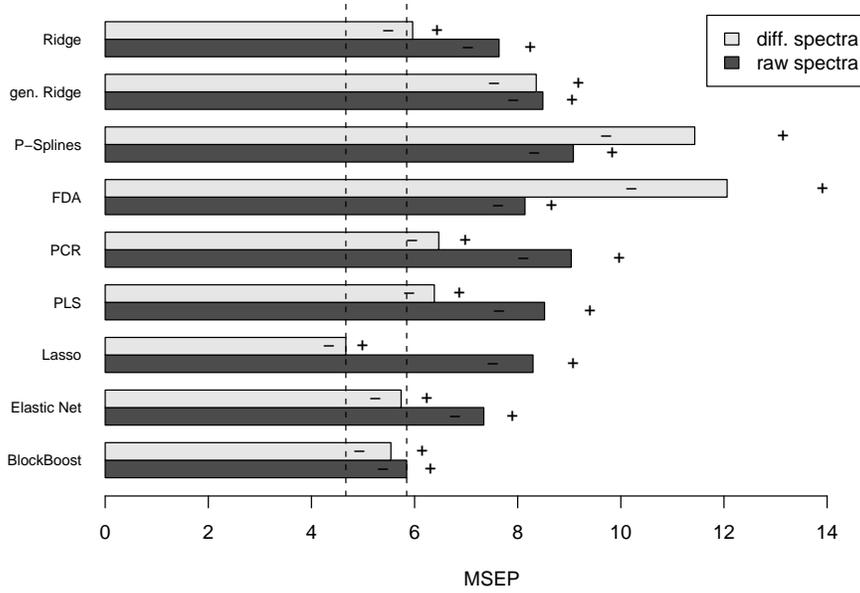


FIGURE 2. Mean Squared Error of Prediction for the considered methods averaged over all 200 random splits,  $\pm 2$  (estimated) standard errors, minima marked by dashed lines; raw and first-difference NIR spectra as predictors, sugar content as response.

Least squares fitting cannot be recommended, since variables in  $X^{(s)}$  tend to be highly correlated. Therefore the parameter estimate we use is the generalized ridge estimator

$$\hat{b}^{(s)} = (X^{(s)T}X^{(s)} + \lambda\Omega)^{-1}X^{(s)T}u$$

with the penalty matrix

$$\Omega = D^T D, \quad D = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 1 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

In addition to penalizing differences the coefficients at the boundaries of the blocks are penalized in order to obtain smoother transitions when two or more blocks (selected in different boosting iterations) are overlapping. In the fitting step of the boosting algorithm the estimate is computed for all neighborhoods  $U_s$  and then the neighborhood (block) is selected that improves the fit maximally. The procedure is stopped by deriving the corresponding AIC criterion.

## 5 Example: NearInfrared Spectroscopy

We consider 200 random splits of the sample data into two independent data sets consisting of 22 training and 10 test observations respectively and compare several methods in terms of mean squared error of prediction evaluated on the test data sets. In the P-spline and the functional data approach the regression function is expanded in 22 equally-spaced basis functions. The rationale is to use the number of basis functions corresponding to the number of observations in the training data set, since without penalty this would yield perfect fit. In the BlockBoost algorithm we chose  $\lambda = 10^{-1}$  and  $\lambda = 10^{-3}$  for raw and difference spectra respectively. Figure 2 shows the performance for the original signal and the difference spectra for various methods, including the elastic net (Zou & Hastie, 2005). BlockBoost performs very well for original spectra as well as for difference spectra. In particular P-splines and the functional data approach, which do not select features, show rather inferior performance.

### References

- Binder, H. and Tutz, G. (2008). Fitting generalized additive models: a comparison of methods. *Statistics and Computing*, **18**, 87-99.
- Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, **11**, 1493-1517.
- Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, **34**, 559-583.
- Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science*, **22**, 477-505.
- Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, **98**, 324-339.
- Christensen, J., Norgaard, L., Heimdal, H., Pedersen, J. G. and Engelsen, S. B. (2004). Rapid spectroscopic analysis of marzipan - comparative instrumentation. *Journal of Near Infrared Spectroscopy*, **12**, 63-75.

- Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*, **11**, 89–121.
- Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119–139.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, **29**, 1189–1232.
- Friedman, J. H. and Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, **28**, 337–407.
- Kneib, T., Hothorn, T. and Tutz, G. (2007). Variable Selection and Model Choice in Geoadditive Regression Models. *Technical Report 3, Department of Statistics LMU Munich*.
- Leitenstorfer, F., Tutz, G. (2007). Generalized Monotonic Regression Based on B-Splines with an Application to Air Pollution Data. *Biostatistics*, **8**, 654–673.
- Marx, B. D., Eilers, P. H. C. (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics*, **41**, 1–13.
- Ramsay, J.O. (1988). Monotone splines in action (with discussion). *Statistical Science*, **3**, 425–461.
- Ramsey, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. (2nd ed.) *Springer, New York*.
- Ridgeway, G. (1999). The state of boosting. *Computing Science and Statistics*, **31**, 172–181.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of Royal Statistical Society*, **B 58**, 267–288.
- Tibshirani, R. and Saunders, M. and Rosset, S. and Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society*, **B 67**, 91–108.
- Tukey, J. (1977). *Exploratory Data Analysis*. *Addison Wesley, New York*.
- Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood based boosting. *Biometrics*, **62**, 961–971.
- Tutz, G. and Leitenstorfer, F. (2007). Generalized smooth monotonic regression in additive modeling. *Journal of Computational and Graphical Statistics*, **16**, 165–188.

- Wold, H. (1975). Soft modelling by latent variables: The nonlinear partial least squares approach. In J. Gani (Ed.), *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett*. London: Academic Press.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society*, **B 67**, 301–320.

**Part 2**  
**Contributed papers**



# Semi-parametric regression to identify farms with high *Salmonella* infection burden

Marc Aerts<sup>1</sup>, Kaatje Bollaerts<sup>1</sup>, Stefaan Ribbens<sup>2</sup>, Yves Van der Stede<sup>3</sup>, Ides Boone<sup>3</sup>, Koen Mintiens<sup>3</sup>

<sup>1</sup> Center for Statistics, Hasselt University, Belgium

<sup>2</sup> Department of Obstetrics, Reproduction and Herd Health, Ghent University, Belgium

<sup>3</sup> Veterinary and Agrochemical Research Center, Brussels, Belgium

**Abstract:** Consumption of pork contaminated with *Salmonella* is an important source of human Salmonellosis worldwide. To control and prevent Salmonellosis, the 10% Belgian pig herds with the highest *Salmonella* infection burden are encouraged to take part in a control programme supporting the implementation of control measures. To identify these herds, serological data reported as SP-ratios are collected. However, SP-ratios have an extremely skewed distribution and are heavily subject to confounding seasonal and animal age effects. Therefore, we propose to identify the 10% high risk herds using semi-parametric quantile regression with P-splines. In addition, a risk factor analysis is conducted to identify potential control measures.

**Keywords:** *Salmonella*, pigs herds, risk factors, semi-parametric quantile regression, Generalized Linear Mixed Models.

## 1 Introduction

Worldwide Salmonellosis, the illness from *Salmonella* infection, is the most frequently occurring zoonoses having a significant economic impact. In the EU Regulation on the control of *Salmonella* and other zoonotic agents it is stated that proper and effective measures are to be taken to detect and control *Salmonella* and other zoonotic agents at all relevant stages of the food production chain, particularly at the level of primary production (that is, at herd level). In line with this regulation, the Belgian government started with a national *Salmonella* surveillance programme in professional pig herds in January 2005. Objective of this programme is to identify pig herds with high *Salmonella* infection burden, which are then (financially) encouraged to take part in a control programme supporting the implementation of control measures to reduce the infection burden.

In order to quantify the *Salmonella* infection burden in pig herds, serological data are collected. In particular, blood samples of breeding and fattening pigs are taken which are tested for *Salmonella*-specific antibodies. The

results are normalized as Sample to Positive ratios (SP-ratios). Currently, identification is based on a simple calculation of herd average SP-ratios. But SP-ratios have an extremely skewed distribution and are heavily subject to confounding seasonal and animal age effects. Therefore, we propose an alternative method to identify risk farms stressing the importance of high SP-ratios since the latter are indicative for recent infection and as such, might better reflect the *Salmonella* infection status at the time of sampling. In particular, we propose to base identification on the number of animals in a herd having very high SP-ratios (called high risk animals for convenience) with a very high SP-ratios being defined as SP-ratios above a specific upper quantile  $\theta$ . To correct for confounding seasonal and animal age effects, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal weight (proxy for animal age). To identify *Salmonella* risk farms, the serological data are linked with data from a survey on biosecurity in Belgian pigs farms (Ribbens et al, 2005). A risk factor analysis is conducted using Generalized Linear (Mixed) Models.

## 2 Methodology

In order to identify high risk animals,  $\theta \times 100\%$  quantile curves of animal SP-ratios are estimated while accounting for confounding seasonal and animal age effects. In particular, the following semi-parametric model is used for each pig  $i$  of herd  $k$  measured at time  $j$

$$\widehat{SP}_{\theta,ijk} = h(\text{time})_{ijk} + I(\text{weight})_{ijk} \quad (1)$$

with  $h(\cdot)$  being a smooth P-splines (Eilers & Marx, 1996) function and  $I$  being an indicator matrix. An extensive treatment of quantile regression with P-splines is given in Bollaerts et al. (2006). In this analysis, seasonal effects are modelled in a flexible way using a basis of 25 B-splines of degree  $q = 3$  with smoothness parameter  $\lambda$  being optimally chosen using cross-validation whereas age effects are modeled in an additive way. Confidence intervals are obtained using residual bootstrap. A graphical representation of the fitted conditional quantile curves  $\theta = 0.90$  and corresponding 95% confidence intervals are given in Figure 1a and 1b, respectively.

High risk animals are then defined as pigs for which the observed SP-ratio is higher than the corresponding estimated 90% quantile, denoted as  $Z_{ijk} = 1$ , otherwise  $Z_{ijk} = 0$ . Then, under the null hypothesis that *Salmonella* infection is equal in all pig herds, it follows that the number of high risk animals in herd  $k$  are beta-binomially distributed as

$$Y_k \sim BB(n_k, 1 - \theta, \rho)$$

with  $n_k$  being the total number of sampled pigs in herd  $k$ , with  $1 - \theta = 0.10$  the probability of being a high risk animal and with  $\rho$  being the intra-herd

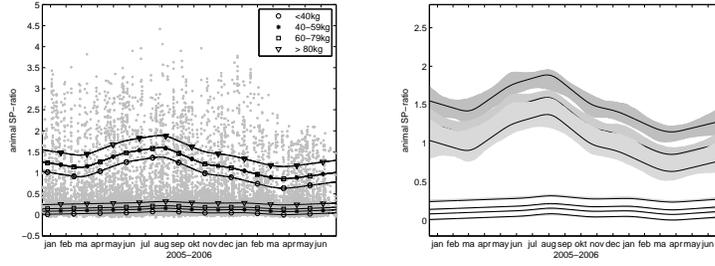


FIGURE 1. Estimated quantile curves  $\theta = 0.90$  of animal SP-ratios conditional on sampling time and animal weight + 95% confidence intervals. The results for  $\theta = 0.50$  are given by means of comparison.

correlation. The p-value corresponding to the alternative hypothesis that *Salmonella* infection burden is higher in herd  $k$  compared to the other herds equals

$$p_k = P\{Y_k \geq y_k | Y_k \sim BB(n_k, 1 - \theta, \rho)\},$$

based on which high risk herds can be selected. The results are graphically displayed in Figure 2. In this figure, herds are ordered along the X-axis following increasing proportion high risk animals. Large differences between herds can be observed with the percentage of herds for which  $p_k < \delta = 0.001$  being equal to 11.8%, being close to the target of 10%.

Of course, other upper conditional quantile curves could be considered as well. To investigate the effect of the choice of  $\theta$  on the selection of the 10% high risk herds, a small sensitivity analysis is conducted comparing the results for  $\theta = 0.85$ ,  $\theta = 0.90$  and  $\theta = 0.95$ . The results (not shown here) indicate minor effects of the choice of threshold. Finally, since high risk herds are encouraged to implement control measures, risk factor anal-

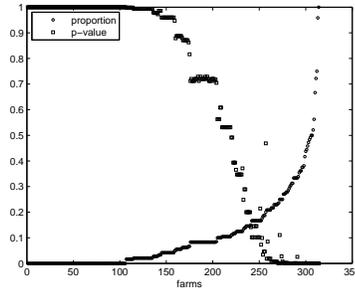


FIGURE 2. Proportion risk animals per herd and corresponding beta-binomial p-values.

yses using Generalized Linear Mixed Models are conducted. Nose contact between pigs from different pens seems to be most consistent risk factor.

### 3 Conclusions

In the current study, an alternative method is proposed to identify pig herds with high *Salmonella* infection burden based on serological data. The alternative method stresses the importance of high SP-ratios, which are indicative for recent infection. To deal with confounding effects and issues of skewness, semi-parametric quantile regression with P-splines is used. In particular, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal age. Then, pigs are classified into low and high risk animals with high risk animals having an SP-ratio larger than the corresponding estimated upper quantile  $\theta$ . Finally, for each herd, the number of high risk animals is calculated as well as the beta-binomial p-value reflecting the hypothesis that the *Salmonella* infection burden is higher in that herd compared to the other herds with the lower the p-value the higher the infection burden. As such, herds with high *Salmonella* infection burden can be identified within a proper inferential framework. In addition, since these herds are encouraged to implement control measures, a risk factor analysis is conducted as well.

### References

- Bollaerts, K., Eilers, P., Aerts, M. (2006). Quantile regression with monotonicity constraints using P-splines and the L1-norm. *Statistical Modelling*, 6, 189-207.
- Bollaerts, K., Aerts, M., Ribbens, S., Van der Stede, Y., Boone, I. & Mintiens, K. (2008). Identification of *Salmonella* high risk pig-herds in Belgium using semi-parametric quantile regression. *Journal of Royal Statistical Society, Series A*, 171 (2), 1–16.
- Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11(2),89-121.
- Ribbens, S., Dewulf, J., Maes, D., Koenen, F., Mintiens, K., Desadeleer, L. & Kruif, A. (2006). A survey on biosecurity in Belgian pig herds. *Preventive Veterinary Medicine*, doi:10.1016/j.prevetmed.2007.07.009

# Modelling the Duration of Necessary and Non-necessary Activities in Daily Life: Fitting Mixture Models to Japanese Time Use Survey Data

Tatsuhiko Anzai<sup>1</sup>, Hideyasu Shimadzu<sup>2</sup> and Toshiki Endo<sup>1</sup>

<sup>1</sup> Jiyu Gakuen College, 1-8-15 Gakuen-cho, Higashi-kurume, Tokyo 203-8521 Japan

E-mail: 99104@std.jiyu.ac.jp (T. Anzai; communicating author);

end@prf.jiyu.ac.jp (T. Endo)

<sup>2</sup> Geoscience Australia, GPO Box 378, Canberra ACT 2601, Australia

E-mail: shimadzu@stat.math.keio.ac.jp (H. Shimadzu)

**Abstract:** Mixture models which consist of normal and exponential distributions for the duration of daily activities are proposed. We assume that daily activities can be classified into two classes: *Necessary Activity* and *Non-necessary Activity*, whose duration has normal or exponential distribution respectively. The parameters are estimated from the data provided by a women's social organisation in Japan. The model fits the data well and also provides good description that the necessity of each activity largely depends on the life stage of the individuals as expected.

**Keywords:** Japanese time use survey, mixture model, Necessary Activity, Non-necessary Activity

## 1 Introduction

Many time use surveys have been conducted sequentially by national organisations. For example, Canada, Holland, Japan, Korea and Norway are well known as having a long history of conducting such surveys (see Harvey and Pentland, 1999; NHK, 2002). However, there appears to have been few that model the duration of daily activities. We therefore propose a new model constructed on a suitable stochastic framework. Our key idea in modelling is to assume that daily activities can be classified into at least two classes: *Necessary Activity* and *Non-necessary Activity* whose duration has normal or exponential distribution respectively. The data used here were provided by a women's social organisation in Japan. The proposed mixture model fits well to the data. As a consequence, the proposed model provides a good description that the necessity of each activity largely depends on the life stage of the individuals that can usually be characterised

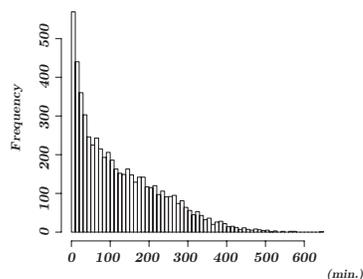


FIGURE 1. Histogram of the duration for time spent with their children excepting the cases of having no child or living separately.

by their attributes.

## 2 Data

The data analysed here were 14,670 records from the time use survey conducted by a women's social organisation, Zenkoku Tomonokai in Japan in 2004. The survey method adopted was the time-diary method (Robinson, 1999). The participants aged from 20s to 90s were all women and members of the organisation. They reported their attributes (eg. Number of family member, Employment, Age of the youngest child etc.) and "when and what they did" for full seven days as well. In this survey, daily activities are placed into 13 categories (eg. Sleep, Work, Meal preparation, etc.). It is worthy of noting that the duration of each activity is not allowed to overlap with each other; in other words, the summation of such durations over a day should be 24 hours.

## 3 Exploratory data analysis

In this survey, the activities are classified into two classes: *Necessary Activity* and *Non-necessary Activity*. The former represents essential behaviours (eg. Sleep, Eat) to keep their life and the latter represents optional behaviours (eg. Hobby). However, it is obvious that the necessity of activities in life would not remain the same but change according to their life stage. For example, we consider here the time spent with their children. Figure 1 shows the distribution of the whole data excepting the cases of having no child or living separately. And Figure 2 shows the histograms (A) through (E) for the data split by the age of their youngest child. Clearly, these five distributions are different from one another. It is easily interpreted that

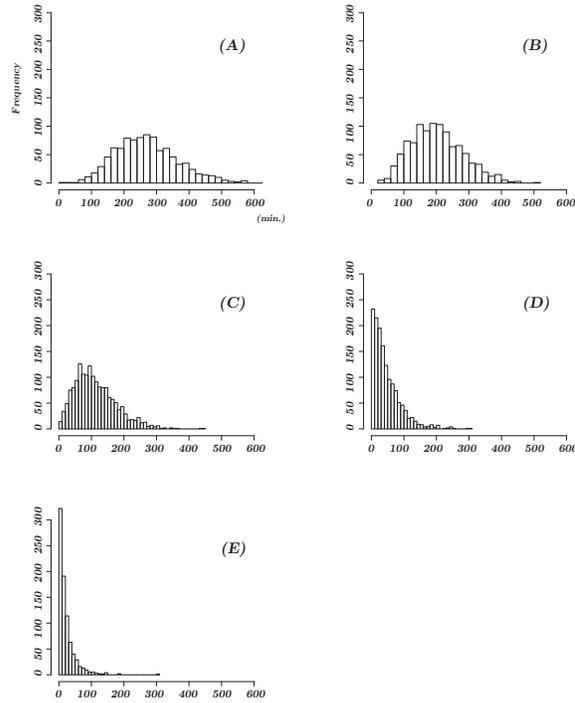


FIGURE 2. Histograms of data split by the age of the youngest child: (A) 0-3; (B) 4-5; (C) 6-12; (D) 13-18; (E) over 19.

when their children are young they spend a longer time for their children to give necessary care so the duration would have distribution as close as normal distribution. By contrast, as their children grow, it is not necessary to spend a long time for them. So the distribution of the duration can be close to exponential distribution. In fact, it is interesting to note that these typical aspects hold true not only for the time with the children but also for the time for work, etc. This implies that the necessity of activities largely depends on their life stage.

#### 4 Model

As discussed, we assume that activities in daily life can be classified into two classes: *Necessary Activity* and *Non-necessary Activity*, and the duration of activities is shown to have normal and exponential distribution respectively, reflecting the necessity of activities. This leads us to consider that the

TABLE 1. Estimated parameters.

	Age	$\hat{p}$	$\hat{q}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\lambda}$
A	0-3	0.000	1.0	269.95	95.81	-
B	4-5	0.000	1.0	203.12	79.76	-
C	6-12	0.000	0.8	113.02	54.09	0.009
D	13-18	0.066	0.0	-	-	0.021
E	19-	0.597	0.0	-	-	0.043

Ⓐ

duration of activities can be modelled as a mixture distribution (McLachlan and Peel, 2000) which mainly consists of three components. More formally, the distribution function of duration  $T$  is given by

$$f(t; \boldsymbol{\theta}) = p I_{\{t=0\}}(t) + (1-p) \{q \mathcal{N}(t; \mu, \sigma^2) + (1-q) \mathcal{E}(t; \lambda)\},$$

where

$$\begin{aligned} \mathcal{N}(t; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}, \\ \mathcal{E}(t; \lambda) &= \lambda \exp\{-\lambda t\}, \end{aligned}$$

and  $\boldsymbol{\theta} = (p, q, \mu, \sigma, \lambda)$  is the vector of parameters. An indicator function  $I = 1$  if and only if the duration  $t = 0$ , which represents the excess zeros. The parameter  $p$  represents the proportion of excess zeros and  $q$  is also the mixture proportion of normal distribution. It is obvious that  $0 \leq p, q \leq 1$ . The parameters are estimated by the maximum likelihood method. Here, the expected value and variance of this mixture model are given by

$$\begin{aligned} \text{E}[T] &= (1-p) \left\{ q\mu + (1-q)\frac{1}{\lambda} \right\}, \\ \text{Var}[T] &= (1-p) \left\{ q(\sigma^2 + \mu^2) + (1-q)\frac{2}{\lambda^2} \right\} - (1-p)^2 \left\{ q\mu + (1-q)\frac{1}{\lambda} \right\}^2. \end{aligned}$$

## 5 Result

Here we will, as an example, revisit the case of the time spent with their children. Table 1 shows the estimated parameters. It is clearly shown that the parameters  $p$  and  $q$  which respectively represent the proportion of excess zeros and normal distribution largely depends on the age of their youngest child. For instance, the parameter  $p$  is getting high in value as the children grow and it is shown that about 60 % of the families having children over 19 have no chance to communicate with their children. By contrast, the parameter  $q$  is getting low in value. This means that the proportion

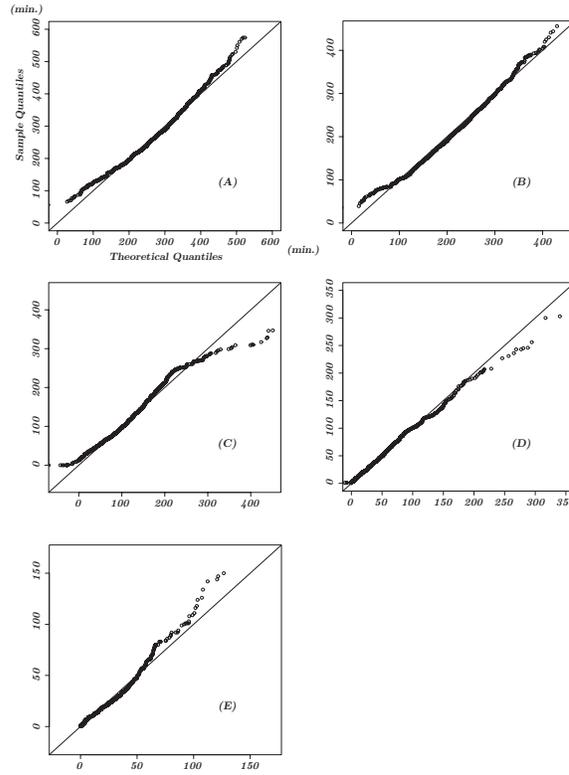


FIGURE 3. Q-Q plots for checking the goodness of fitting the model.

of exponential distribution is getting high. From the view point of our assumption, getting high proportion of exponential distribution implies that the necessity of such activity is changed. Further the parameters  $\mu$  and  $\lambda$  representing the average of duration also depend on the growth of the children. By combining these, our model clearly shows that the duration of the time spent with the children is getting short as the children grow.

Figure 3 shows Q-Q plots for checking the goodness of fitting our mixture model. These plots relatively show the liner relationship which indicates that the models adopted are well fitted to the data.

The mixture model is also applied to other activities (eg. Work, Sleep etc.) and succeeds in describing several aspects of each activity. The proposed model is simple but explains well the change of necessity of activities in our life.

## 6 Concluding remarks

A mixture model which consists of normal and exponential distribution has been developed for the duration of daily activities. A key assumption of this model is that daily activities can be classified into two classes: *Necessary* and *Non-necessary Activities*, whose duration has normal or exponential distribution respectively. The model fits the data well, is easy to understand and interpret, and also provides a good description for changing necessity of daily activities in their life, which largely depends on the life stage of the individuals.

As an outcome of our modelling, it is clearly shown that the duration spent with their children is getting shorter as their children grow. However, our model describes not only such shortening of the duration but also explains the reason lying behind as the estimated parameter that represents the proportion of normal distribution is getting low in value. This leads us to consider that as their children grow it is not necessary to spend a long time for them to give care, which is a quite reasonable conclusion.

This model is fitted to the duration of other activities and proves to be a good fit. As a consequence, the model should be of use to analyse and quantitatively evaluate the individual life in terms of time use.

## Acknowledgements

The authors are grateful to Mrs Midori Yamazaki (Zenkoku Tomonokai, Japan) and Mr Yasuhiro Yano (Jiyu Gakuen, Japan) for their help and suggestions.

## References

- European commission (2003). *Time use at different stages of life: Results from 13 European countries*. Luxembourg, Office for Official Publications of the European Communities.
- Harvey, A.S. and Pentland W.E (1999). Time use research. In: W. E. Pentland, A. S. Harvey, M. P. Lawton and M. A. McColl (eds), *Time Use Research in the Social Sciences*. 3–18, New York, Kluwer Academic.
- McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. New York, Wiley.
- NHK Broadcasting Culture Research Institute (2002). *Japanese Time Use in 2000*. Tokyo, Japan Broadcast Publishing.
- Robinson, J.P. (1999). The time-diary method: structure and uses. In: W. E. Pentland, A. S. Harvey, M. P. Lawton and M. A. McColl (eds), *Time Use Research in the Social Sciences*. 47–89, New York, Kluwer Academic.

# Developing an expert system for predicting Legionella outbreaks in evaporative installations by using Bayesian hierarchical models

Carmen Armero<sup>1</sup>, Alejandro Artacho<sup>1</sup> and Antonio López-Quílez<sup>1</sup>

<sup>1</sup> Departament d'Estadística i I.O.,  
Grup d'Estadística espacial i temporal en Epidemiologia i Medi ambient,  
Universitat de València,  
C/ Doctor Moliner, 50, 46100 Bujassot (València), Spain

**Abstract:** Quick detection in installations where there is risk is one of the keys to fighting against the bacterium *Legionella*. The main target of this paper is the use of the Bayesian hierarchical modelling approach to take advantage of minimal and remote information about the quality of the water in evaporative installations. We have obtained real-time predictions of intermediate, non-immediately observable variables involved in the Bayesian network in order to predict the risk of *Legionella* in real time in spite of the absence of hard historical evidence of this variable.

**Keywords:** Hierarchical Bayesian Networks; Markov Chain Monte Carlo methods.

## 1 Introduction

Developing monitoring tools in order to prevent and control outbreaks of Legionnaires' disease in real-time is a real issue. It is a special type of pneumonia originating from the bacterium *Legionella*, which usually lives in environmental water sources from where they colonize urban installations, such as evaporative condensers and cooling towers (Praveen Verma, 2000). In the period 1987-1997, the European Working Group for Legionella Infections (EWGLI, [www.ewgli.org](http://www.ewgli.org)) reported a total of 1365 affected travellers among various European countries, 28% of them associated with infections acquired in Spanish facilities. It is not hard to appreciate that the detection of infection outbreaks is a very serious public health issue. Although current regulations in Spain establish different hygienic-sanitary criteria for prevention, there are many difficulties in homogenizing them and some essential deficiencies related to bacteria detection methods were observed.

This paper presents the methodology, modelling and results of a project devoted to constructing an expert system (ES from now on) capable of monitoring information on water quality from remote evaporative installations in real-time, analyzing the risk of *Legionella* and providing control and immediate warnings mechanisms for the companies responsible for such installations. The model is implemented through the hierarchical Bayesian networks paradigm (Gelman *et al.*, 2004). Markov Chain Monte Carlo, MCMC, algorithms are applied for estimation by simulating the posterior distribution of parameters and hyperparameters of the model through the free software WinBUGS (Spiegelhalter *et al.*, 2003).

## 2 The data

The available information consists of historical data from a multiparametric panel developed by the companies participating in the project. This panel was located in six different cooling towers and provided real-time monitoring of five physical and chemical water variables (conductivity, hardness,  $pH$ , temperature and turbidity). In addition, we have data of other relevant variables (alkalinity,  $Cl$  and  $Fe$  levels, suspended solids, salinity and aerobic and *Legionella* levels) which cannot be measured immediately because laboratory tests are needed. Since the ES is currently working in alert mode, both new evidence being propagated through the network and predicted outcome values are stored in the knowledge base.

## 3 Statistical modelling

The ES combines deductive with inductive inference. Immediately the parameters of the model are adjusted from the data, the system will be capable of operating in alert mode and propagate all new daily evidence received from the panel installed in the cooling tower. At that point, this new information is stored for subsequent parametric adjustments allowing the ES to learn in a Bayesian way.

We use Bayesian Networks (BN from now on) as the inference engine, in particular Hierarchical Bayesian Networks (HBN) implemented through the BUGS language. HBN are generalizations of BN where the nodes may be an aggregate data type (Lauritzen, 1996). This hierarchical framework with multilevel parametric families allows us to reduce the number of network parameters without any loss of complexity. In addition, the parameters in BUGS are explicitly represented as nodes in the network, which are associated with a final distribution after adjustment and prediction. In the classic BN approach it is not clear how to incorporate the associated parameters error into the propagation algorithm. Figure 1 shows a flow chart of the expert system.

The appropriate graphical structure has been provided by the expert knowledge of the system we have acquired (Verma, 2000). It is realistic to admit that any structure suggested can only be tentative, and it ought to be subject to revision in the light of future data. If the appropriate knowledge is not initially available, we need to look for information in order to suggest appropriate conditional independence assumptions (See Buntine (1994) for a general review on structural learning in graphical models). Also, due the presence of non-observable variables, we will have to evaluate our assumptions later on. For example, if we consider the following  $pH$  transformation:

$$pH^* = \begin{cases} 0, & \text{if } 5 \leq pH \leq 9 \\ 1, & \text{other cases} \end{cases}$$

the weight sign for the link  $pH^* \rightarrow \text{aerobian}$ , is expected to be negative since the optimum growth interval of the aerobian lies in the  $pH$  interval [5, 9]. Hence, if our model cannot yield a negative estimation of this weight (it is a model parameter), we will have to remove the link.

The use of subjective static relationships between nodes has been combined with local linear regressions defined for some of the network families. The modelling for the second class of family may be summarized in the expressions below for both parametric and propagation phases. In particular, for a family  $i$  with children node  $Y_i$ , parents  $(x_1, \dots, x_s)$  and link weights  $(w_1, \dots, w_s)$ :

1. PARAMETRIC LEARNING:

$$\begin{array}{l} W = (w_1, \dots, w_s) \\ X = (x_1, \dots, x_s) \\ Y_i \sim \text{Beta}(a_i, b_i), a_i = lw_i, b_i = l(1 - w_i) \\ w_j \sim N(\mu = 0, \tau = 0.001), j = 1, \dots, s \\ \text{logit}(w_i) = WX \end{array} \rightarrow \begin{array}{l} \text{coda} \\ \text{(MCMC-sampling result)} \\ (w_j^1, \dots, w_j^{5000}), j = 1, \dots, s \\ \tilde{W} = (\tilde{w}_1, \dots, \tilde{w}_s) \\ \tilde{w}_j = \sum_{k=1}^{5000} w_j^k / 5000 \end{array}$$

2. PROPAGATION:

$$\begin{array}{l} W^k = (w_1^k, \dots, w_s^k) \\ w_i^k = e^{W^k X} / (1 + e^{W^k X}) \\ Y_i^k \sim \text{Beta}(a_i^k, b_i^k), a_i^k = lw_i^k, b_i^k = l(1 - w_i^k) \\ k = 1, \dots, 5000 \end{array} \rightarrow \begin{array}{l} \text{the estimation of node } i: \\ \tilde{Y}_i = \sum_{k=1}^{5000} Y_i^k / 5000 \end{array}$$

where  $l$  is the length of the historical data and 5000 is the length of the stored MCMC-chain. This perspective allows us to synchronize the parametric learning with biochemical hypotheses that regularize the process we want to model.

In our network, we can find families where the link weights are completely estimated from the data but also families where the child node is a non-observable variable. In these cases, with a total absence of data, estimation is only based on the opinion of the different human experts in the project.

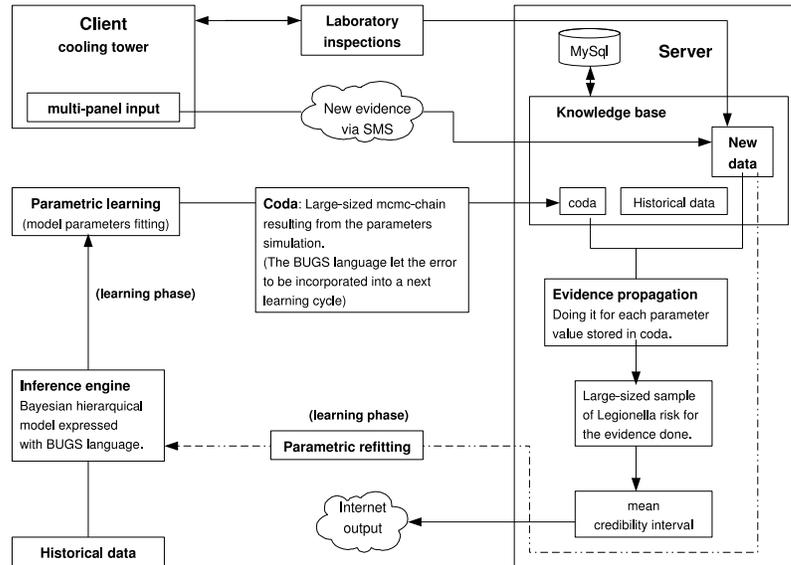


FIGURE 1. Expert system flow chart. The Bayesian-based learning is reflected in the cyclical structure. Outcomes will improve as the knowledge base grows. Customers can consult the resulting *Legionella* risk prediction via Internet.

This is a key point because the specific choice of weights has a sensitive effect on the resulting *Legionella* risk predictions. In this sense, it has been useful to compare the simulated outcomes with real *Legionella* observations to choose the best-fitted set of weights.

Some of the families implicitly include a parent node which sends information to the child related to the current season of the year (fixed seasonal effect) and a parent node which sends information about the specific installation where the data come from (random installation effect). Consequently, we can learn about the child node, but always bearing in mind that each installation operates in a different manner and its performance will change throughout the year.

The network structure reflects an ancestral scheme composed of four levels. The first one gathers variables that do not need to be explained (without parents) since each multi-panel measures them in real time. These nodes are the parents of the second level variables which are not immediately measurable. At the third level there are three intermediate and non-observable

variables (corrosion, incrustation, and biofilm levels) that transform information from the upper levels by imposing relationships based on knowledge of the theoretical system. The response variable, the risk of *Legionella* infection, is placed in the final level.

It has been verified that the network structure allows the variables not immediately measured in the laboratory to be well explained by the others measured in real time. Thus, the whole set of variables can be considered in real time to evaluate infection risk and it is not necessary to wait for laboratory confirmation. It was not possible to model the response variable family in a complete stochastic way because of the presence of multiple missing data and a lack of positive *Legionella* observations. However, the risk levels derived from the network are quite consistent with the available observations. The convergence of the whole set of parameters involved was achieved (by applying the Gelman-Rubin diagnostic) at a low computational cost and with quite reasonable associated error.

**Acknowledgments:** Special thanks to the Ministerio de Educación y Ciencia grant MTMT2007-61554, Conselleria Empresa, Universitat i Ciència de la Generalitat Valenciana grant GV/2007/079 and Ministerio de Industria, Turismo y Comercio FIT-350300-2006-87

## References

- Buntine, W. L. (1994). *Operations for learning with graphical models*. Journal of Artificial Intelligence Research, 2, 159-225.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford: Oxford University Press.
- Spiegelhalter, D.J., Thomas, A., Best, N. and Lunn, D. (2003). WinBUGS. User Manual. [www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf](http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf)
- Verma, P. (2000). *Cooling Water Treatment Hand Book*. New Delhi: Albartross.

# Bayesian hierarchical model for the prediction of football results

Gianluca Baio<sup>1</sup> and Marta Blangiardo<sup>2</sup>

<sup>1</sup> University College London (UK)

<sup>2</sup> Imperial College London (UK)

**Abstract:** The problem of modelling football data has become increasingly popular in the last few years and many different models have been proposed with the aim of estimating the characteristics that bring a team to lose or win a game, or to predict the score of a particular match. We propose a Bayesian hierarchical model to address both these aims and test its predictive strength on data about the Italian Serie A championship 2005-2006.

**Keywords:** Bayesian hierarchical models; Football data; Bivariate Poisson.

## 1 Introduction

Statistical modelling of sport data is a popular topic and much research has been produced, specifically about the distributional form associated with the number of goals scored in a single game by the two opponents. Although the Binomial or Negative Binomial have been proposed in the late 1970s (Pollard et al 1977), the Poisson distribution has been widely accepted as a suitable model for these quantities, with the simplifying assumption of independence between the goals scored by the home and the away team often used (Maher 1982).

Despite this, some authors have shown empirical (although relatively low) levels of correlation (Lee 1997, Karlis and Ntzoufras 2003). Consequently, the use of more sophisticated models have been proposed, for instance in Dixon and Cole (1997), who applied a correction factor to the independent Poisson model to improve the performance in terms of prediction.

More recently Karlis and Ntzoufras (2003) advocated the use of a Bivariate Poisson (BP) distribution that has a more complicated formulation for the likelihood function, and includes an additional parameter explicitly accounting for the covariance between the goals scored by the two competing teams. They specify the model in a frequentist framework (although extensions using the Bayesian approach have been described by Tsionas 2001), and their main purpose is the estimation of the effects used to explain the number of goals scored.

We propose in this paper a Bayesian hierarchical model for the number of goals scored by the two teams in each match. Hierarchical models are widely

used in many different fields as they are a natural way of taking into account relations between variables, by assuming a common distribution for a set of relevant parameters, thought to underlay the outcomes of interest. Within the Bayesian framework, which naturally accommodates hierarchical models, using a BP model is not fundamental. We show here that, assuming two conditionally independent Poisson variables for the number of goals scored, correlation is taken into account since the observable variables are mixed at an upper level.

## 2 The model

The league (we consider Italian ‘Serie A’ for the season 2005-2006) is made by a total of  $T = 20$  teams, playing each other twice in a season. We indicate the number of goals scored by the home and by the away team in the  $g$ -th game of the season ( $g = 1, \dots, G = 380$ ) as  $y_{g1}$  and  $y_{g2}$  respectively. The vector of observed counts  $\mathbf{y} = (y_{g1}, y_{g2})$  is modelled as independent Poisson:

$$y_{gj} \mid \theta_{gj} \sim \text{Poisson}(\theta_{gj}),$$

where the parameters  $\boldsymbol{\theta} = (\theta_{g1}, \theta_{g2})$  represent the scoring intensity for the  $g$ -th game and the team playing at home ( $j = 1$ ) and away ( $j = 2$ ), respectively.

We model these parameters according to a formulation that has been used widely in the statistical literature (see Karlis and Ntzoufras 2003, and the reference therein), assuming a log-linear random effect model:

$$\begin{aligned} \log \theta_{g1} &= \textit{home} + \textit{att}_{h(g)} + \textit{def}_{a(g)} \\ \log \theta_{g2} &= \textit{att}_{a(g)} + \textit{def}_{h(g)}. \end{aligned}$$

The parameter *home* represents the advantage for the team hosting the game and we assume that this effect is constant for all the teams and throughout the season. In addition, the scoring intensity is determined jointly by the attack and defense ability of the two teams involved, represented by the parameters *att* and *def*, respectively. The nested indexes  $h(g), a(g) = 1, \dots, T$  are uniquely associated with one of the 20 teams, and identify the team that is playing at home (away) in the  $g$ -th game of the season. The data consist of the name and code of the teams, and the number of goals scored for each game of the season.

The prior distributions for all the random parameters are specified as follows. The variable *home* is modelled as a fixed effect, assuming a standard flat prior distribution.

The team-specific effects are modelled as exchangeable from a common distribution:

$$\textit{att}_t \sim \text{Normal}(\mu_{\textit{att}}, \tau_{\textit{att}}), \quad \textit{def}_t \sim \text{Normal}(\mu_{\textit{def}}, \tau_{\textit{def}}),$$

for  $t = 1, \dots, T$ . As suggested by various works, we need to impose some identifiability constraints on the team-specific parameters. Although Karlis and Ntzoufras (2003) use a sum-to-zero constraint, we noted that the convergence of the model is improved if a corner-constraint is used instead.

Finally, the hyper-priors of the attack and defense effects are modelled independently using again a flat prior distribution.

The inherent hierarchical nature implies a form of correlation between the observable variables  $y_{g1}$  and  $y_{g2}$  by means of the unobservable hyper-parameters  $\boldsymbol{\eta} = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$ . In fact, the components of  $\boldsymbol{\eta}$  represent a latent structure that is assumed to be common for all the games played in a season and that determines the average scoring rate.

Each game contributes to the estimation of these parameters, which in turn generate the main effects that explain the variations in the parameters  $\boldsymbol{\theta}$  therefore implying a form of correlation on the observed counts  $\mathbf{y}$ .

### 3 Results

We estimated the parameters of the model by means of a standard MCMC-based procedure. Similarly to what found in other works, the home effect is positive (the posterior mean and 95% CI are 0.2894 and [0.1637; 0.4148], respectively). A.C. Milan have the highest propensity to score (as suggested by the posterior mean of 0.5930 for the effect *att*), followed by Juventus, the league winner in that year (0.4237). Juventus also performed very well in terms of defence showing the lowest value for the parameter *def* (posterior mean  $-0.4754$ ), while Treviso (who finished bottom) showed the highest propensity to concede goals (0.0417). These effects are incremental with respect to the baseline, which we defined as Ascoli.

Moreover, we produced a set of  $G$  games from the predictive distribution of  $\mathbf{y}$ , which we used for model checking (see Figure 1). Our model seems to predict well the dynamics throughout the season and even when there are some discrepancies between the observed and the predicted values, the overall prediction at the end of the season seems to be accurate. However, for teams showing “extreme” performance (particularly Treviso, who ended bottom of the table with an unusual low number of points) the predicted results are too shrunk by the hierarchical model and hence not too accurate. Also, since our model simulates the overall season at once, it is not sensitive enough to fit well teams who are associated with periods of consecutive particularly bad results in the observed campaign (i.e. Sampdoria or Lecce). Finally, we replicated our analysis using the data on Italian Serie A 1991-1992, to allow direct comparison with the model used by Karlis and Ntzoufras (2003). As one can see from Figure 2, the predictive distributions from the Bayesian Hierarchical model are generally closer to the observed results.

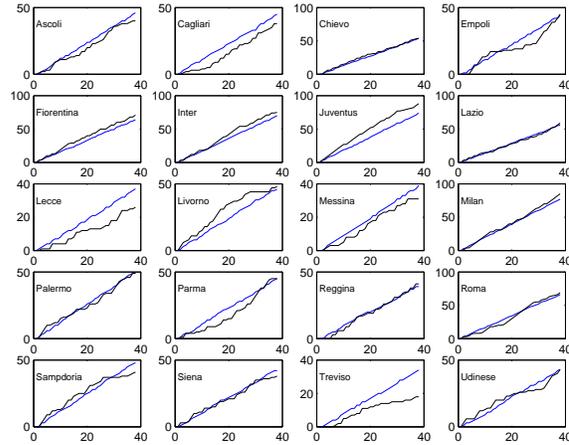


FIGURE 1. Posterior predictive validation of the model. For each team, the dark and the light curves represent the observed and predicted cumulative points through the season, respectively

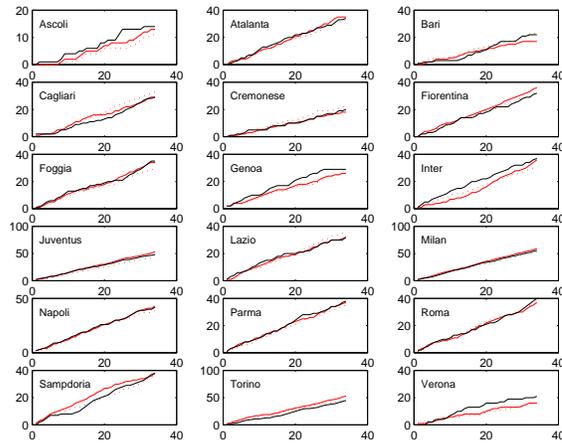


FIGURE 2. Posterior predictive validation of the model in comparison with Karlis and Ntzoufras (2003). Data are for the Italian Serie A 1991-92. For each team, the dark line represents the observed cumulative points through the season, while the light solid and the light dotted line represent predictions for the Bayesian Hierarchical and the Bivariate Poisson model, respectively

For clubs like Ascoli, Foggia and Inter the predictions from our model seem to fit the observed curves much better than the Bivariate Poisson model, which is probably better only for Genoa and Fiorentina. For all the other

clubs (excluding Verona), the two models seem to provide similar results, although generally ours are slightly closer to the realised values.

## 4 Conclusions

Using data from the Italian Serie A 1991-92 and 2005-2006, we showed in this paper that the use of a hierarchical structure produces results that are not inferior to the ones obtained with a BP model, which requires more complicated estimation methods. This methodology can be fruitfully applied in other applications where the interest is on two or more correlated observations (e.g. number of deaths in different vehicles involved in an incident).

## References

- Dixon, M. and Coles, S. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society, Series C*, **46**.
- Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society Series D*, **52**.
- Lee, A. (1997). Modeling scores in the Premier League: is Manchester United really the best? *Chance* **10**.
- Maher, M. (1982). Modelling association in football scores. *Statistica Neerlandica*, **36**.
- Pollard, R., Benjamin, P. and Reep, C. (1977). Sport and the negative binomial distribution. In: *Optimal Strategies in Sports*. New York: North Holland.
- Tsionas, E. (2001). Bayesian Multivariate Poisson Regression. *Communications in Statistics — Theory and Methodology*, **30**.

# A multidimensional latent Markov IRT model

F. Bartolucci<sup>1</sup>, I. L. Solis-Trapala<sup>2</sup>

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

<sup>2</sup> Department of Medicine, Faraday Building, Lancaster University, Lancaster, LA1 4YB, UK.

**Abstract:** We introduce a multidimensional extension of the latent Markov model which is suitable for the analysis of longitudinal binary data collected by the administration of batteries of tests which measure more than one latent trait. The model is based on the one-parameter or two-parameter logistic parametrization of the conditional distribution of the response variables given the latent process. The latter is assumed to follow a first-order Markov chain with transition matrices parameterized according to the application of interest. We outline an EM algorithm for the maximum likelihood estimation of the model parameters and we focus on testing the dimensionality of the problem through a likelihood ratio test on these parameters. To illustrate the approach we analyse data from a study aimed at assessing cognitive development in early childhood.

**Keywords:** Cognitive development; Item Response Theory; Latent class model.

## 1 Introduction

Multidimensional models represent an important class of Item Response Theory (IRT) models which may be used to analyse data deriving from the administration of a set of dichotomously-scored items when the probability of responding correctly to each item depends on more than one latent trait; for a recent contribution in this field and a detailed description of the related literature see Bartolucci (2007). Items with these features arise in many contexts and, especially, in educational assessment and psychometrics. Obviously, comparing the fit of a multidimensional IRT model with that of its unidimensional version, using for instance a likelihood ratio (LR) statistic, offers the possibility to assess if the structure of the data at hand is indeed multidimensional. This is a relevant aspect since rejecting unidimensionality means that standard IRT models based on this assumption may lead to misleading conclusions.

Even if the problem of assessing the dimensionality of the problem has found satisfactory solutions in the IRT literature when the data are collected by a single questionnaire, no approaches have been proposed to specifically address this problem in a repeated measurement or longitudinal context. We are referring, in particular, to cases in which the same

set of items is administered to the same subjects at a certain number of occasions which are separated by a suitable interval of time. This scheme is very common in psychological applications as the one we deal with for illustrative purposes. The particular feature of this scheme is that a subject may evolve in his/her latent characteristics between occasions and this is not taken into account in the multidimensional IRT models mentioned above.

In this paper, we propose a multidimensional model for the analysis of data deriving from a repeated measurement scheme. The basic tool is the latent Markov (LM) model of Wiggins (1973), which may be seen as an extension for longitudinal data of the latent class model (Lazarsfeld and Henry, 1968) in which each subject is allowed to move between latent classes during the period of observation. For a detailed description of the LM model see Langeheine and van de Pol (2002) and Bartolucci (2006); for a description with an IRT perspective see Bartolucci *et al.* (2008). In our formulation, the probability of responding in a certain way to each item given the latent state is formulated as in a multidimensional IRT model. This implies that a different level of every latent trait under consideration is associated to each latent state. Evolution of the subjects with respect to these latent traits depends on the transition matrix of the model and, on the basis of this matrix, we can test the hypothesis that the subjects always remain in the same state. Obviously, by exploiting the proposed model, we can also perform an LR test for the hypothesis of unidimensionality against that of multidimensionality. In contrast to more standard approaches, this test takes into account the longitudinal structure of the data. We give particular consideration to these aspects and to the maximum likelihood estimation of the model on the basis of the EM algorithm (Dempster *et al.*, 1977).

In the following, we first describe the proposed model in more detail and then we describe an application based on a dataset deriving from a psychological experiment.

## 2 The model

Let  $Y_{jt}$  denote the binary response variable corresponding to the  $j$ -th item administered to a subject at the  $t$ -th occasion, with  $j = 1, \dots, J$  and  $t = 1, \dots, T$ . The proposed model assumes that all the  $J \times T$  response variables are independent given a latent process  $X_1, \dots, X_T$  which follows a first-order Markov chain with state space  $\{1, \dots, k\}$ , initial probabilities  $\pi_c$ ,  $c = 1, \dots, k$ , and transition probabilities  $\pi_{d|c}$ ,  $c, d = 1, \dots, k$ . These initial and transition probabilities may also depend on individual time-varying covariates, such as age, so that the resulting process is not homogeneous. In order to give an interpretation of the model as a multidimensional IRT model for longitudinal data, we also assume that

$$\text{logit}\{p(Y_{jt} = 1|X_t = c)\} = \sum_a \delta_{aj} \theta_{ac} - \beta_j, \quad (1)$$

where  $\theta_{ac}$  is the level of the latent trait of type  $a$  for the subjects in latent state  $c$  and  $\delta_{aj}$  is a dummy indicator equal to 1 if the latent trait of type  $a$  is involved in responding to item  $j$ , with  $a = 1, \dots, b$  and  $b$  denoting the dimension of the model. Moreover,  $\beta_j$  is a parameter which measures the average effect of item  $j$  and that is usually interpreted as its *difficulty level*. In our approach, we also consider an extended parametrization based on the assumption

$$\text{logit}\{p(Y_{jt} = 1|X_t = c)\} = \alpha_j \left( \sum_a \delta_{aj} \theta_{ac} - \beta_j \right),$$

where  $\alpha_j$  is the discriminant level of item  $j$ . This is referred to as *two-parameter logistic* parametrization in contrast to the former one which is referred to as *one-parameter logistic* parametrization. We also consider an extension of the model in which transition between latent states is possible even within the subsequence of items  $Y_{1t}, \dots, Y_{Jt}$  administered at the same occasion on the basis of a latent process which follows a first-order Markov chain.

In order to estimate the parameters of the proposed model we make use of an EM algorithm (Dempster *et al.*, 1977). The E-step of this algorithm consists of computing, for every subject in the sample, the conditional probability of each latent state at each time occasion given the responses he/she provided. This is done by exploiting certain recursions which are well known in the hidden Markov literature (MacDonald and Zucchini, 1997). The M-step consists of updating the model parameters on the basis of simple iterative algorithms.

We also deal with testing hypotheses on the latent process and hypotheses concerning the conditional distribution of the response variables given this process. Among the hypotheses of the first type, of particular interest is that the transition matrix  $\mathbf{\Pi}$ , with elements  $\pi_{cd}$ ,  $c, d = 1, \dots, k$ , is diagonal and then each subject always remains in the same latent state during the period of observation. By exploiting the approach proposed by Bartolucci (2006), this hypothesis may be tested by an LR statistic between the model in which  $\mathbf{\Pi}$  is unconstrained and that in which it is diagonal. Among the hypotheses on the conditional distribution of the response variables, of particular interest is that of unidimensionality which may again be tested on the basis of the LR statistic between a model based on  $b > 1$  latent traits and its unidimensional version based on only one latent trait. By repeating a test of this type for increasing values of  $b$  we can also assess the number of latent traits in a way similar to that applied by Bartolucci (2007) for the case of items administered at a single occasion.

### 3 An application to experimental psychology

We illustrate the applicability of the model developed in the previous section by examining data from a study conducted by Shimmom (2004), which aims to investigate some aspects of the development of the construct *executive function* in early childhood. The study comprises data collected from the administration of a battery of tests to 115 young children during a *single* testing session. This session was subsequently replicated *over* two 6-month periods when it was believed that key changes in child cognition might have occurred. In this example we focus our attention on two components of executive function, namely *inhibitory control* and *attentional flexibility*. These are two abstract concepts defining two closely related psychological constructs. We address two methodological issues: i) we investigate the nature of the interrelationship between the two constructs and ii) we assess developmental trends in task performance and scalability of various tasks at various ages.

The response data collected for each participant consist of a sequence of correlated binary outcomes with each component indicating a success or failure for each trial on *four tasks* administered at each of three time periods. Inhibitory control was measured by the *day/night* (DN) and the *abstract pattern* (AP) tasks; each of these tasks was administered in blocks of 16 trials. Similarly, attentional flexibility was measured by two versions of the *Dimensional change card-sort* (DCCS) tasks: the DCCS face-up and the DCCS face-down tests. Each of these tests was administered in blocks of 6 trials. Participants were randomly allocated to one of two testing orderings. Half of the group performed the tasks in the following order: DN, DCCS face-down, AP, DCCS face-up, whereas the other half of the group followed the order: AP, DCCS face-up, DN, DCCS face-down. Thus the length of the response sequence for each participant is 44 (16+6+16+6) at each time period.

For illustration we restrict our attention to the analysis of the first sequence of 44 trials. We initially fitted a model formulated as described in Section 2 with  $k = 2$  latent states. The model, indicated in the following by *Model 1*, is based on assumption (1) with the first ability representing inhibitory control and the second attentional flexibility. We allow initial and transition probabilities of the latent process to depend on age through a logit parametrization and specify different sets of parameters for the transition probabilities within sessions and between sessions.

We also considered other models which are nested in *Model 1*:

*Model 2*: unidimensional version of *Model 1* in which common ability parameters are assumed for all the items;

*Model 3*: a probability matrix of independence is assumed for the transition between blocks of trials, with parameters equal to those for the initial probabilities;

*Model 4*: a diagonal transition matrix is assumed within blocks of trials.

Table 1 compares the fit of the models specified above, with respect to *Model 1*, on the basis of the deviance, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

TABLE 1. Maximum log-likelihood, number of parameters, deviance with respect to *Model 1*, AIC and BIC for each fitted model.

Model	max. log-lik.	n. par.	deviance	AIC	BIC
1	-2003.7	24	-	4055.5	4121.3
2	-2019.0	23	30.5	4084.0	4147.1
3	-2023.2	16	39.0	4078.5	4122.4
4	-2179.0	16	350.6	4390.1	4434.0

It is clear from Table 1 that all the models have an inadequate fit in comparison with *Model 1*. Consequently, we reject the hypothesis of unidimensionality, which is incorporated into *Model 2*, and we conclude that two different abilities underlie the response process; the first represents inhibitory control as measured by the DN and AP tasks and the second represents attentional flexibility which is measured through the DCCS tasks. We also have to reject the hypothesis, incorporated into *Model 3*, that the two abilities are independent. Finally, we conclude that trend in performance exists, since the hypothesis of absence of this effect, on which *Model 4* is based, must definitely be rejected.

Under *Model 1*, which we adopt as our final model, we obtained the initial probabilities for the latent process equal to 0.3602 for the first state and to 0.6398 for the second one. The second latent state identifies those children who show better task performance. Table 2 displays the within- and between-sequence transition probability matrices. Both the initial and the transition probabilities are averaged over all the sample.

TABLE 2. Estimated transition probability matrices under *Model 1*.

latent state	within		between 16-item	
	16-item sequence		and 6-item sequences	
1	0.9850	0.0150	0.6127	0.3873
2	0.0514	0.9486	0.5728	0.4272
latent state	within		between 6-item	
	6-item sequence		and 16-item sequences	
1	0.9642	0.0358	0.2626	0.7374
2	0.1090	0.8910	0.3313	0.6688

Briefly, both within transition probability matrices in Table 2 show some evidence of a tiring effect as reflected by some tendency of participants to move from latent state 2 to 1 within the sequences. Importantly, the

between transition probabilities show a positive impact of building experience on attentional flexibility tasks over performance on inhibitory control tasks. Finally, estimates of age effects (not reported here) show that older children are less susceptible to experiencing tiring effects and, in general, older children tend to move to the second latent state.

**Acknowledgments:** Solis-Trapala acknowledges support from the UK Economic and Social Research Council (RES-576-25-5020).

## 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. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, **72**, 141-157.
- Bartolucci, F., Pennoni, F., and Lupporelli, M. (2008). Likelihood inference for the latent Markov Rasch model. In: C. Huber, N. Limnios, M. Mesbah and M. Nikulin (Eds.), *Mathematical Methods for Survival Analysis, Reliability and Quality of Life*. 239-254, Wiley.
- 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. Hagenars and A.L. McCutcheon (Eds.), *Applied Latent Class Analysis*. 304-341, Cambridge University Press.
- Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.
- MacDonald, I.L., and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. London: Chapman & Hall.
- 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.
- Wiggins, L.M. (1973). *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. Amsterdam: Elsevier.

# Complex additive penalties for generalized structured additive regression

Christiane Belitz<sup>1</sup> and Stefan Lang<sup>2</sup>

<sup>1</sup> Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany

<sup>2</sup> Department of Statistics, University of Innsbruck, Universitätsstr. 15, A-6020 Innsbruck, Austria

**Abstract:** Models with structured additive predictor provide a very broad and rich framework for complex regression modeling. In this talk we discuss inference for structured additive regression based on complex additive penalties. The penalties are an additive combination of quadratic penalties common in spline smoothing, spatial statistics and random effects modeling. The additive combination provides an enormous enhancement of the modelers toolbox for flexible modeling of complex phenomena. The proposed penalties also allow for built in model and variable selection within the estimation process. Extensive simulations show that the covariate effects and the predictor are usually estimated with more precision compared to inference based on standard penalties.

**Keywords:** ANOVA-decomposition, geoadditive models, P-splines, spatial smoothing, variable selection

## 1 Models with structured additive predictor

Suppose that observations  $(y_i, \mathbf{z}_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , are given, where  $y_i$  is a response variable, and  $\mathbf{z}_i = (z_{i1}, \dots, z_{iq})'$  and  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  are vectors of covariates. For the variables in  $\mathbf{z}$  possibly nonlinear effects are assumed whereas the variables in  $\mathbf{x}$  are modeled in the usual linear way. The components of  $\mathbf{z}$  are not necessarily continuous covariates. A component may also indicate a time scale, a spatial index denoting the region or district a certain observations pertains to, or a unit- or cluster index denoting the unit (e.g. community) a certain observation pertains to. Moreover, the components of  $\mathbf{z}$  may be two- or even three dimensional in order to model interactions between covariates.

Generalized structured additive regression (STAR) models (Fahrmeir, Kneib and Lang, 2004) assume that, given  $\mathbf{z}_i$  and  $\mathbf{x}_i$  the distribution of  $y_i$  belongs to an exponential family. The mean  $\mu_i = E(y_i | \mathbf{z}_i, \mathbf{x}_i)$  is linked to a structured additive predictor  $\eta_i$  by

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(z_{i1}) + \dots + f_q(z_{iq}) + \mathbf{x}_i' \boldsymbol{\gamma} + \varepsilon_i. \quad (1)$$

Here,  $h$  is a known response function,  $f_1 - f_q$  are nonlinear functions of the covariates  $z_{ij}$  and  $\mathbf{x}'_i \gamma$  is the usual linear part of the model.

The nonlinear functions  $f_j$  are modeled by a basis functions approach, i.e. a particular nonlinear function  $f$  is approximated by a linear combination of basis (or indicator) functions:

$$f(z) = \sum_{k=1}^K \beta_k B_k(z)$$

The  $B_k$  are known basis (or sometimes indicator) functions and  $\beta = (\beta_1, \dots, \beta_K)'$  is a vector of unknown regression coefficients to be estimated. To ensure enough flexibility, typically a large number of basis functions is defined. To avoid overfitting a roughness penalty on the regression coefficients is additionally specified. We use quadratic penalties of the form  $\beta' \mathbf{K}(\lambda) \beta$  where

$$\mathbf{K}(\lambda) = \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 + \dots + \lambda_M \mathbf{P}_M$$

is a complex penalty matrix which is effectively the sum of (simpler penalty) matrices  $\mathbf{P}_1, \mathbf{P}_2, \dots$ . The number  $M$  of individual penalties used to construct  $\mathbf{K}$  is typically between 1 and 4. The penalty depends on a vector of smoothing parameters  $\lambda = (\lambda_1, \dots, \lambda_M)'$  that govern the weight of the individual penalty matrices  $\mathbf{P}_m$ ,  $m = 1, \dots, M$  and as a result the amount of smoothness imposed on the function  $f$ . As we will see below, a suitable chosen combination of individual penalty matrices provides an enormous enhancement of the modelers toolbox for flexible modeling of complex phenomena.

Defining the  $n \times K$  design matrix  $\mathbf{Z}$  with elements  $\mathbf{Z}[i, k] = B_k(z_i)$  the vector  $\mathbf{f} = (f(z_1), \dots, f(z_n))'$  of function evaluations can be written in matrix notation as  $\mathbf{f} = \mathbf{Z}\beta$ . Accordingly, for model (1) we obtain

$$\eta = \mathbf{Z}_1 \beta_1 + \dots + \mathbf{Z}_q \beta_q + \mathbf{X}\gamma,$$

where  $\mathbf{X}$  is the design matrix for linear effects,  $\gamma$  is the vector of regression coefficients for linear effects, and  $\eta$  is the predictor vector. In the next section we will give specific examples for modeling the unknown functions  $f_j$  or in other words for the choice of basis functions and penalty matrices.

## 2 Examples for additive penalties

### 2.1 Continuous covariates

Suppose first that a particular component  $z$  of  $\mathbf{z}$  is univariate and continuous.

The P-splines approach proposed by Eilers and Marx (1996) assumes that the unknown functions can be approximated by a polynomial spline of degree  $l$  and with equally spaced knots over the domain of  $z$ . The spline can

be written in terms of a linear combination of  $K = m + l$  B-spline basis functions. The columns of the design matrix  $\mathbf{Z}$  are given by the B-spline basis functions evaluated at the observations  $z_i$ . To overcome the difficulties involved with regression splines, Eilers and Marx (1996) suggest a relatively large number of knots (usually between 20 to 40) to ensure enough flexibility, and to introduce a roughness penalty on adjacent regression coefficients based on squared  $r$ -th order differences, i.e.

$$\lambda\beta'\mathbf{P}\beta = \lambda \sum_{k=r+1}^K (\Delta^r \beta_k)^2.$$

The penalty matrix is given by  $\mathbf{P} = \mathbf{D}'_r \mathbf{D}_r$  where  $\mathbf{D}_r$  is a  $r$ -th order difference matrix.

Usually, the difference order used for the penalty is not discussed further and second or third order differences are typical choices in practice. However, extensive simulations and some theoretical considerations suggest that even for relatively simple functions the choice of the difference order is crucial. Wiggly functions or functions with peaks are best estimated using first order differences. For linear effects a second order difference penalty is the optimal choice because a degree one polynomial is the limit of the spline for large smoothing parameters. Third order difference penalties are best suited for quadratic functions and most functions with moderate curvature (e.g. a sine functions). In this paper we propose to use the additive penalty

$$\mathbf{K}(\lambda) = \lambda_1 \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 + \lambda_3 \mathbf{P}_3, \quad (2)$$

where  $\mathbf{P}_1$ ,  $\mathbf{P}_2$  and  $\mathbf{P}_3$  are penalty matrices based on first, second and third order differences. The penalty has the following properties:

- The limit  $\lambda_j \rightarrow 0$ ,  $j = 1, 2, 3$  results in an unpenalized fit.
- The limit  $\lambda_1 \rightarrow \infty$  results in a constant fit, i.e. the covariate  $z$  is removed from the model. This limiting behavior is irrespective of the values of the remaining smoothing parameters  $\lambda_2$  and  $\lambda_3$ .
- The limit  $\lambda_2 \rightarrow \infty$  results in a linear fit (irrespective of the value for  $\lambda_3$ ). An additional shrinkage effect is obtained if  $\lambda_1 > 0$ . The higher the value of  $\lambda_1$  the more shrinkage is induced. The combination  $\lambda_1 \rightarrow 0$ ,  $\lambda_2 \rightarrow \infty$  results in a linear fit with no additional shrinkage effect.
- The limit  $\lambda_3 \rightarrow \infty$  results in a quadratic fit. Depending on the values of the remaining smoothing parameters  $\lambda_1$  and  $\lambda_2$  additional shrinkage effects towards a linear fit (for increasing  $\lambda_2$ ) and/or a constant fit (for increasing  $\lambda_1$ ) are obtained.

The proposed penalty provides built in model choice and variable selection since a constant fit (i.e. covariate  $z$  removed from the model), a linear fit and a quadratic fit are nested.

## 2.2 Surface fitting

Assume now that  $z$  is two-dimensional, i.e.  $z = (z^{(1)}, z^{(2)})'$  with continuous components  $z^{(1)}$  and  $z^{(2)}$ . A common approach for surface fitting is to approximate the surface  $f(z)$  by the tensor product of one dimensional B-splines, i.e.

$$f(z^{(1)}, z^{(2)}) = \sum_{k=1}^{K_1} \sum_{s=1}^{K_2} \beta_{ks} B_{1,k}(z^{(1)}) B_{2,s}(z^{(2)}), \quad (3)$$

where  $B_{11}, \dots, B_{1K_1}$  are the basis functions in  $z^{(1)}$  direction and  $B_{21}, \dots, B_{2K_2}$  in  $z^{(2)}$  direction. The  $n \times K = n \times K_1 K_2$  design matrix  $\mathbf{Z}$  now consists of products of basis functions.

Belitz and Lang (2008) discuss several alternatives for the penalty matrix  $\mathbf{K}(\lambda)$ . Of particular interest is the additive penalty

$$\mathbf{K}(\lambda) = \frac{\lambda_1}{K_1} \mathbf{I}_{K_2} \otimes \mathbf{K}_1 + \frac{\lambda_2}{K_2} \mathbf{K}_2 \otimes \mathbf{I}_{K_1} + \lambda_3 \tilde{\mathbf{K}}_1 \otimes \tilde{\mathbf{K}}_2, \quad (4)$$

where  $\mathbf{K}_1$  and  $\mathbf{K}_2$  are penalty matrices corresponding to one dimensional P-splines based on first or second order differences. The matrices  $\tilde{\mathbf{K}}_1$  and  $\tilde{\mathbf{K}}_2$  are penalty matrices of P-splines based on first order differences. This penalty has the following nice properties:

- The limit  $\lambda_3 \rightarrow \infty$  results in a mere main effects model. The main effects are one dimensional P-splines with smoothing parameters  $\lambda_1$  and  $\lambda_2$ .
- The limit  $\lambda_1 \rightarrow \infty$ ,  $\lambda_2 \rightarrow \infty$  and  $\lambda_3 \rightarrow \infty$  results in a main effects model with linear or constant main effects depending on the difference order used to construct  $\mathbf{K}_1$  and  $\mathbf{K}_2$ .

## 2.3 Spatial heterogeneity

In this subsection we assume that  $z$  represents the location a particular observation pertains to. The location is typically given in two ways. If exact locations are available  $z = (z^{(1)}, z^{(2)})'$  is two-dimensional and the components  $z^{(1)}$  and  $z^{(2)}$  correspond to the coordinates of the location. In this case the spatial effect  $f(z^{(1)}, z^{(2)})$  could be modeled by two-dimensional surface estimators as described in the preceding section.

In many applications, however, exact locations are not available. Typically, a geographical map is available and  $z \in \{1, \dots, K\}$  is an index that denotes the region (e.g. district) an observation pertains to. A common approach is to assume  $f(z) = \beta_z$ , i.e. separate parameters  $\beta_1, \dots, \beta_K$  for each region are estimated. The  $n \times K$  design matrix  $\mathbf{Z}$  is an incidence matrix whose entry in the  $i$ -th row and  $k$ -th column is equal to one if observation  $i$  has been

observed at location  $k$  and zero otherwise. To prevent overfitting a penalty based on squared differences is defined that guarantees that parameters of neighboring regions are similar. In most applications two regions are assumed to be neighbors if they share a common boundary although other neighborhood definitions are possible, see below. The penalty is defined as

$$\lambda\beta'\mathbf{P}\beta = \lambda \sum_{k=2}^K \sum_{s \in N(k), s < k} (\beta_k - \beta_s)^2, \quad (5)$$

where  $N(k)$  denotes all sites that are neighbors of site  $k$ .

In some situations a smooth spatial effect is not justified because of local, spatial heterogeneity. In this case, the assumption of spatial dependence of neighboring parameters is not meaningful. Instead, a simple ridge type penalty

$$\lambda\beta'\mathbf{P}\beta = \lambda\beta'\beta = \lambda \sum_{k=1}^K \beta_k^2 \quad (6)$$

with penalty matrix  $\mathbf{P}(\lambda) = \lambda\mathbf{I}$  may be defined. This penalty does not assume any spatial dependence but prevents highly variable estimates induced by small samples for some regions or sites.

In most applications the choice of the appropriate penalty is not clear a priori. A remedy is to use an additive composition of the two penalties (5) and (6) to obtain the complex penalty matrix

$$\mathbf{K}(\lambda) = \lambda_1\mathbf{P} + \lambda_2\mathbf{I}.$$

Here  $\mathbf{P}$  corresponds to the spatial penalty (5).

In some applications the choice of the appropriate neighborhood system for modeling spatial heterogeneity is not obvious. Two sites may not only be 'neighbors' because they share a common boundary but also because they share certain other characteristics. For instance in ecological modeling sites may be defined neighbors because of similar environmental conditions. In this case we could combine two or more spatial penalties based on different neighborhood systems to obtain the complex penalty

$$\mathbf{K}(\lambda) = \lambda_1\mathbf{P}_1 + \lambda_2\mathbf{P}_2 + \dots$$

where  $\mathbf{P}_1, \mathbf{P}_2, \dots$  are spatial penalties of the form (5).

### 3 Inference

Inference (including confidence intervals) is similar to Belitz and Lang (2008) sections 3 and 4. More details will be given in the talk.

## 4 Simulation results

A simulation study has been carried out to test the proposed additive penalty (2) for smooth effects of continuous covariates. A model with Gaussian responses and predictor  $\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$  is assumed. Here  $f_1(x_1) = 0.33x_1$  is linear,  $f_2(x_2) = 0.2 \cdot x_2^2 - 0.1 \cdot x_2 - 0.63$  is quadratic,  $f_3(x_3) = \sin(x_3)$  is sinusoidal and  $f_4(x_4) = (\sin(4/3x_4) + 2 \exp(-(16^2) \cdot (x_4/6)^2) - 0.217)/6$  is an example for a function with peaks. Eight possible covariates  $x_1 - x_8$  have been simulated, the latter 4 with no effect on the response. Covariates  $x_1$  and  $x_2$  as well as  $x_5$  and  $x_6$  are correlated with correlation around 0.5. Inference is based on the estimation techniques of Belitz and Lang (2008). We tested four penalties based on first differences (rw1), second differences (rw2), third differences (rw3) and the additive penalty (2) (rw1rw2rw3).

Box plots of  $\log(MSE)$  based on 250 replications of the model are shown in figure 1. The new additive penalty performs quite favorable compared to the simpler penalties.

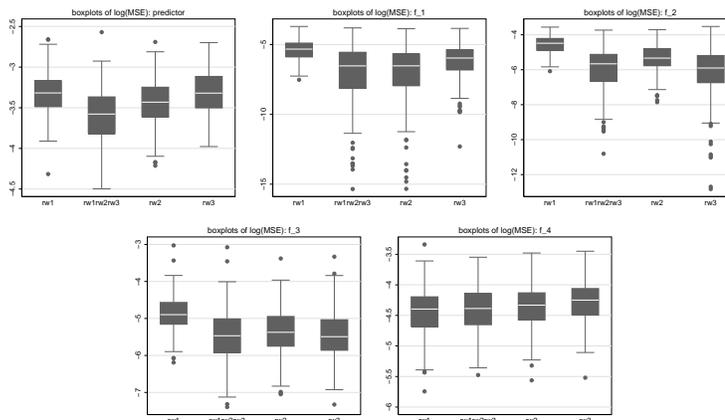


FIGURE 1. *Boxplots of  $\log(MSE)$  for the predictor and  $f_1 - f_4$ .*

## References

- Belitz C. and Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis*, to appear.
- Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, 14, 731-761.
- Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, 11, 89-121.

# Hierarchical Generalised Linear Models Analysis of Bovine Tuberculosis on Milk Data

Fiona Boland<sup>1</sup>

<sup>1</sup> UCD School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Ireland. Email: fiona.boland@ucd.ie

**Abstract:** This paper describes the modelling of bovine tuberculosis (TB) and milk yield data using Hierarchical Generalised Linear Models (HGLM's). In broad terms HGLM's extend generalised linear mixed models (GLMM's) by allowing more flexibility in the choice of the distribution for the random effects and also modelling of the dispersion of the random effects and error term. In this paper HGLM's are applied to a random sample of Irish dairy herds restricted from trading because of TB between the 1st June 2004 and the 31st May 2005 to model the relationship between milk production and TB infection. There is an inherent hierarchical structure in the data, lactations are nested within animals and animals within herds and this entails the use of two random effects in the model. The methods for testing hypotheses about the fixed effects, random effects and dispersion parameters are presented here and the most appropriate model is found and interpreted.

**Keywords:** Hierarchical Generalised Linear Models (HGLM's); h-likelihood; random effects; Bovine Tuberculosis (TB); milk production.

## 1 Introduction

Despite control measures the number of bovine TB infected cattle in Ireland has been approximately 24,000 animals per year for 2004-2006. TB outbreak in a dairy herd causes monetary losses to the exchequer as the infected animals are slaughtered and farmers compensated but it also affects the farmer due to trade restrictions imposed on the whole herd. A random sample of Irish dairy herds restricted from trading between the 1st June 2004 and the 31st May 2005 is examined. Milk variables belonging to all lactations on an animal in the study are considered. The primary question is to ascertain whether TB infection affects milk yield. Since observations relating to lactations within an animal will be correlated and in addition animals in the same herd will probably be correlated any model will need to incorporate these effects. HGLM's using the estimation method of h-likelihood as outlined in Lee, Nelder and Pawitan (2006) provide a useful class of models that do this. In addition the variances of the random effects are also modelled. Likelihood criteria were used for both inclusion/exclusion

of random and fixed effects to find the best model. Diagnostic plots were examined and the implications of the results of the model for TB were interpreted.

## 2 HGLM's

The standard linear mixed model specifies:

$$y = X\beta + Zv + e \quad (1)$$

where  $y$  is a vector of responses,  $X$  is a known design matrix for the fixed effects,  $\beta$  is vector of unknown fixed-effect parameters,  $Z$  is a known design matrix for the random effects,  $v$  is  $N(0, \sigma_v^2)$  and  $e$  is  $N(0, \sigma_e^2)$ . Let  $\tau = (\sigma_e^2, \sigma_v^2)$ . The extended ( $l_e$ ) or h-likelihood of all unknown parameters is given by:  $h = l_e(\beta, \tau, v) = \log f(y, v) = \log f(y|v) + \log f(v)$

$$= -\frac{1}{2} \{ \log |2\pi\sigma_e^2| - (y - X\beta - Zv)^t (\sigma_e^2)^{-1} (y - X\beta - Zv) - \log |2\pi\sigma_v^2| - v^t (\sigma_v^2)^{-1} v \} \quad (2)$$

The marginal distribution for the fixed effects is obtained via integration, so that:

$$L(\beta; y) \equiv f_\beta(y) = \int f_\beta(v, y) dv \quad (3)$$

Lee, Nelder and Pawitan (1996) propose three deviances for carrying out tests on various components of HGLM's. For inferences about the random effects they propose to use -2h-likelihood and for the fixed effects the difference in adjusted profile h-likelihood,  $p_v(h)$  as a  $\chi^2$  test with degrees of freedom equal to the number of extra parameters. For the dispersion parameters the adjusted profile h-likelihood,  $p_{\beta, v}(h)$  is used.  $p_v(h)$  for example denotes the adjusted profile likelihood  $h$  after eliminating the nuisance parameter  $v$ . The dispersion parameters can have structures where they are defined to be related to a set of covariates. This brings together the joint modelling of the mean and the dispersion. For the milk data there are two random effects, herd ( $v$ ) and animal within herd effect ( $a$ ) and the variance of the random herd effect is investigated to see if it increases with herd size. The model for the dispersion parameter  $\sigma_v^2$  of the random herd effect  $v$  is  $E(\sigma_v^2) = \omega_0 + \omega_1 z$  where  $z$  is the covariate herd size. Finally if there are two random effects  $v$  and  $a$ , together with  $\hat{e}$  two sets of residuals  $\hat{v}$  and  $\hat{a}$  are produced. Thus, assumptions about these three random components can be checked separately and in addition three sets of (deviance) residuals are available for checking the dispersion model (Lee, Nelder and Pawitan, 2006).

## 3 Results

The data consists of TB infected and non-infected animals from a random sample of 80 dairy herds in Ireland restricted from trading between the

TABLE 1. Testing the Fixed Terms in the Model

Parameter	Y*P*T	Y*P	T*Y	Y	T*P	T	P
$-2p_v(h)$	11.15	24.88	11.87	198.06	13.17	24.83	411.72
df	20	20	5	5	4	1	4
p-value	0.9422	0.204	0.0366	0.0001	0.0105	0.0001	0.0001

Parameters: Y=Year, T=TB and P=Parity

1st June 2004 and the 31st May 2005. The TB data were obtained from the Department of Agriculture, Fisheries and Food (DAFF) and the milk production data comes from the Irish Cattle Breeding Foundation (ICBF). The model fitted (saturated model) is as follows:

- ▷Response: Milk Yield for each lactation of an animal
- ▷Fixed effects: Parity (lactation no.), TB result (positive or negative for TB), year (of parity/lactation record) and all possible interactions between these three variables.
- ▷Random effects: Herd (unique herd identifier),  $N(0, \sigma_v^2)$  and animal (unique animal identifier) within herd,  $N(0, \sigma_a^2)$ .

The random terms were tested first (Lee, Nelder and Pawitan, 2006) and both herd and animal within herd were found to be needed in the model. Subsequently the fixed effects were tested and Table 1 shows the difference in  $-2P_v(h)$  for testing the fixed effects. From Table 1 it can be seen that all terms are significant and should be left in of the model except for Year\*Parity\*TB and Year\*Parity. Finally a HGLM with structured dispersion was modelled by adding herd size to model the random herd effect. The deviance difference in  $-2p_{v,\beta}(h)$  is  $0.273$ . Using the chi-squared test with 1 degree of freedom and  $\alpha=0.05$  the structured dispersion is not significant. Figure 1(a) shows the residual plot of the random herd effect in the dispersion model when herd size was included to model this effect. There was an effect of TB in all years and in all parities except for parity 5. For parity 5 only the year 2001 showed a significant difference between TB and non-TB animals. The herd variance estimate  $\sigma_v^2$  (0.4518) is much larger than the animal within herd estimate  $\sigma_a^2$  (0.1593) indicating more variability between herds than between animals within a herd. Figure 1(b) shows the plot of the residuals versus the fitted values for the error component in the mean model and it can be seen that the points are fluctuating randomly around zero except perhaps at large fitted values. This requires further examination.

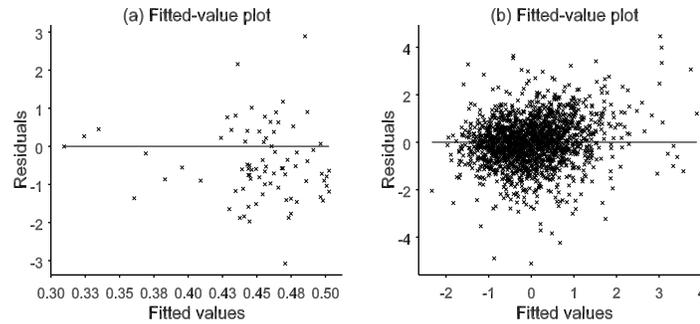


FIGURE 1. Fitted-value plots of (a) the residuals from fitting the random herd effect in the dispersion model using the covariate herd size and (b) the  $\hat{\epsilon}_{ijk}$  component in the mean model.

## 4 Discussion

There was an effect of TB in all years and in all parities except for parity 5. For parity 5 only the year 2001 showed a significant difference between TB and non-TB animals. In addition both the random animal within herd and herd effect are important. The dispersion of the random herd effect was modelled using herd size but there was no evidence to indicate that the variance of the herds changes with herd size (i.e. large herds are no more variable than small herds). The model allows simultaneous estimation of all parameters and in addition there is a suitable likelihood-based criteria for the model selection.

**Acknowledgments:** Special thanks to my supervisor Dr. Gabrielle Kelly for her ongoing support, the Department of Agriculture, Fisheries and Food (DAFF) for supplying the TB data and the Irish Cattle Breeding Foundation (ICBF) for supplying the milk data. This research was supported by a grant from DAFF.

## References

- Lee, Y., and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**.
- Lee, Y., Nelder J.A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman and Hall.

# Dose-Illness Models for Human Salmonellosis Based on Outbreak Data

Kaatje Bollaerts<sup>1</sup>, Marc Aerts<sup>1</sup>, Christel Faes<sup>1</sup>, Koen Grijspeerdt<sup>2</sup>, Jeroen Dewulf<sup>3</sup>, Koen Mintiens<sup>4</sup>

<sup>1</sup> Center for Statistics, Hasselt University, Belgium

<sup>2</sup> Agricultural Research Centre-Ghent, Department of Animal Product Quality, Melle, Belgium

<sup>3</sup> Department of Obstetrics, Reproduction and Herd Health, Ghent University, Belgium

<sup>4</sup> Veterinary and Agrochemical Research Center, Brussels, Belgium

**Abstract:** The quantification of the relationship between the amount of microbial organisms ingested and a specific outcome such as infection, illness or mortality is a key aspect of quantitative risk assessment. In this paper, we model the dose-illness relationship based on data of 20 *Salmonella* outbreaks, as discussed by the *World Health Organization*. In particular, we model the dose-illness relationship using Generalized Linear Mixed Models and modified fractional polynomials of dose.

**Keywords:** Human Salmonellosis; Dose-illness; Fractional polynomials; Generalized Linear Mixed Models; Data uncertainty.

## 1 Introduction

Salmonellosis, the illness from *Salmonella* infection, is one of the most frequently occurring foodborne diseases worldwide. Global estimations vary between 14 and 120 per 100 000 people. The majority of the cases is due to *Salmonella* Enteritidis and *Salmonella* Typhimurium infections, which comprised almost 80% of the total number of *Salmonella* infections in Belgium in 2005. Salmonellosis is characterized by fever, stomach cramps, and diarrhea. The degree to which a person becomes sick depends on his or her health status and the number and virulence of *Salmonella* spp. ingested.

An important aspect of quantifying microbial risk is the assessment of the dose-response relationship, which is the relationship between the amount of microbial organisms ingested and a specific outcome, like infection, illness or even mortality. Different sources of heterogeneity in dose-response are known to exist. A first important source of heterogeneity are differences in host susceptibility, with the YOPI-group (Young children, Older persons, Pregnant women and Immunocompromised) being typically highly susceptible. A second source of heterogeneity are differences in serovar types and

associated pathogen virulence. In addition, differences in food matrix cause heterogeneity in dose-response relationships as well.

According to the *World Health Organization* outbreak data are considered to be more valuable in order to model dose-response compared to experimental data (WHO, 2003). Whereas the latter are typically limited to young healthy volunteers, high doses and one specific combination of serovar type and food-matrix, outbreak data refer to real-life situations (including low doses). However, outbreak data are heavily subject to data-uncertainty. An example of modeling dose-response using outbreak data can be found in a report on risk assessment of *Salmonella* from the WHO. In this study, dose-illness is modeled using a beta-poisson model and the effect of data uncertainty is investigated by sampling from specific uncertainty distributions and refitting the beta-poisson model. However, beta-poisson models are developed to reflect the biological process of infection, not illness. Furthermore, heterogeneity is not accounted for. Bollaerts *et al.* (2008) re-analysed the data from the outbreak studies reported by the WHO, based on modified fractional polynomials satisfying the properties of the different types of dose-illness models proposed by Teunis *et al.* (1999). Heterogeneity due to differences in host susceptibility, serovar type and food matrix is taken into account as well as data-uncertainty.

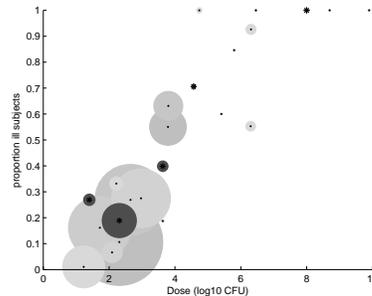


FIGURE 1. Bubble plot of proportion of ill subjects as function of dose, for 20 outbreak studies as reported in WHO, 2003. The area of the bubbles is proportional to the number of exposed subjects. Observations on normal subjects are indicated with a dot and light gray colored bubbles and observations on susceptible subjects are indicated with a star and dark gray colored bubbles.

## 2 Dose-Illness Models

Within dose-response modeling, most attention is given to dose-infection models with the most popular one being the beta-poisson model, developed to reflect the biological process of infection. Although commonly practice,

Teunis *et al.* (1999) advocate that the use of the beta-poisson model to model illness as a function of dose is questionable and introduce a multiple stage model instead. Following this model, exposure to pathogens might lead to infection (inf) and infection might lead to illness (ill). Hence, the probability of illness given dose (d) equals

$$\pi(\text{ill}|\text{dose}) = \pi(\text{ill}|\text{inf},d)\pi(\text{inf}|d). \quad (1)$$

Whereas  $\pi(\text{inf}|d)$  is typically assumed to be a monotonically increasing function of dose bounded between zero and one (e.g. beta-poisson model), some experimental evidence seems to indicate different relationships for  $\pi(\text{ill}|d)$ . Teunis et al (1999) argue that the length of the infection period may be dose-dependent. Starting from the assumption that during infection the host has a certain hazard of becoming ill and using a Gamma distribution for the duration of infection  $\tau$ , Teunis et al (1999) derive that the probability of illness given infection equals

$$\pi(\text{ill}|\text{inf}, d) = 1 - (1 + \lambda)^{-r} \quad (2)$$

with  $r > 0$  being the shape parameter of the Gamma distribution and with  $\lambda$  being the integral over duration time  $t$  of the hazard function for illness in an infected person. Assuming that  $\lambda$  varies with dose, Teunis et al (1999) explore three different alternatives (see also Figure 2):

1.  $\lambda$  increases linearly with dose  $D$  or  $\lambda = \eta D$  implying that  $\pi(\text{ill}|\text{inf},d)$  is a monotonically increasing function of dose bounded between zero and one. As such, given the common assumption that  $\pi(\text{inf}|d)$  is a monotonically increasing function of dose bounded between zero and one,  $\pi(\text{ill}|d)$  is also monotonically increasing as a function of dose with the same boundaries.
2.  $\lambda$  decreases with dose  $D$  or  $\lambda = \eta/D$  implying that  $\pi(\text{ill}|\text{inf},d)$  is a monotonically decreasing function of dose bounded between zero and one. Given the common assumptions for  $\pi(\text{inf}|d)$ , it follows that  $\pi(\text{ill}|d)$  is monotonically unconstrained with the probability of illness being zero for zero dose and infinitely large dose levels.
3.  $\lambda$  is dose-independent or  $\lambda = \eta$  such that  $\pi(\text{ill}|\text{inf},d) = \pi(\text{ill}|\text{inf})$ . Hence, given the common assumptions for  $\pi(\text{inf}|d)$ , it follows that  $\pi(\text{ill}|d)$  is monotonically increasing function of dose that is bounded by zero and reaches the asymptote of  $\pi(\text{ill}|\text{inf}) < 1$  for infinitely large dose levels.

Clearly, the three alternatives for the dose-dependency of  $\lambda$  results in three different types of dose-illness models, being the monotonically increasing dose-illness model bounded between zero and one (M1), the monotonically unconstrained dose-illness model with the probability of illness being zero

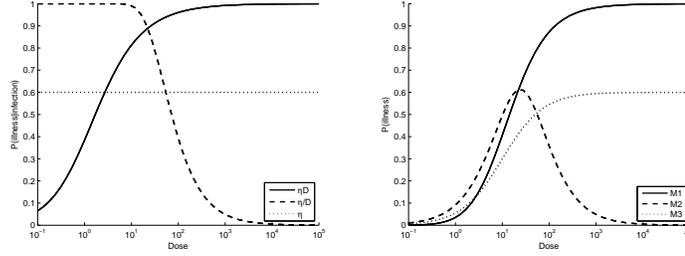


FIGURE 2. Three different probability models of illness given infection (left) and (b) the corresponding dose-illness models (right).

for zero dose and infinitely large dose levels (M2) and the monotonically increasing dose-illness model bounded between zero and  $\pi(\text{ill}|\text{inf})$  being some constant  $c < 1$  (M3).

### 3 GLMMs and Fractional Polynomial of Dose

Most dose-response models fit within the framework of Generalized Linear Models (GLMs). The linear predictor  $\eta$  for such GLMs contains indicator variables for discrete covariates and conventional polynomials, mostly of linear or quadratic order, for continuous covariates. To enhance flexibility, Royston and Altman (1994) introduced fractional polynomials, which are a set of parametric models offering a wide range of functional forms including the conventional polynomials. Fractional polynomials of degree  $m$  for a continuous covariate  $\mathbf{x}$  subject to the constraint  $\mathbf{x} > 0$ , are defined as

$$f(x; \boldsymbol{\beta}, \mathbf{p}, m) = \sum_{r=0}^m \beta_r H_r(x) \tag{3}$$

with  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)$  being a vector of coefficients and  $\mathbf{p} = (p_0, p_1, \dots, p_m)$  a vector of powers with  $p_0 \equiv 0$  and  $H_0(x) \equiv 1$  representing the intercept. The powers  $p_1 \leq p_2 \leq \dots \leq p_m$  can be positive or negative integers or fractional powers.  $H_r(x)$  is a transformation on a continuous variable  $x$  defined as

$$H_r(x) = \begin{cases} x^{p_r} & \text{if } p_r \neq p_{r-1} \\ H_{r-1}(x) \times \ln(x) & \text{if } p_r = p_{r-1} \end{cases}$$

and with  $\mathbf{x}^0 \equiv \ln(x)$ . Following Royston and Altman (1994), models of degree  $m = 2$  or lower with powers  $\mathbf{p}$  selected from a fixed set  $\mathcal{R}^m$  with  $\mathcal{R} = \{-2, -1, -0.5, 0, 0.5, 1, 2, \dots, \max(3, m)\}$  are sufficiently flexible to cover most practical cases adequately. Then, given a degree  $m$  and a fixed set  $\mathcal{R}^m$ , all possible models are fitted using maximum likelihood estimation and the model with the lowest deviance is selected.

However, fractional polynomial models do not inherently fit into the three above-mentioned dose-illness models. Bollaerts *et al.* (2008) proposed the following modified fractional polynomials. For M1, a fractional polynomial of dose  $D$  of degree  $m = 2$  can be easily modified to satisfy these properties as follows (with  $g$  a link function)

$$g(\pi(\text{ill}|\text{dose} = D)) = \beta_0 + \beta_1 D^{p_1} + \beta_2 D^{p_2} \quad (4)$$

with  $p_1 < 0$ ,  $p_2 \geq 0$ ,  $\beta_1 < 0$  and  $\beta_2 > 0$ . M2 is covered by (4) with  $p_1 < 0$ ,  $p_2 \geq 0$ ,  $\beta_1 < 0$  and  $\beta_2 < 0$ . Finally, a fractional polynomial of degree  $m = 1$  can be modified to cover M3 as follows

$$g(\pi(\text{ill}|\text{dose} = D)) = \beta_0 + \beta_1 D^{p_1} \quad (5)$$

with  $p_1 < 0$  and  $\beta_1 < 0$ . Using fixed effects and random intercepts this model can further be extended to account for different sources of heterogeneity. The three different types of dose-illness models extended with random effects are fitted to the WHO outbreak data. The best fitting model is a monotonically increasing dose-illness model bounded between zero and some constant  $c < 1$  (M3) suggesting that length of the infection period is dose-independent. The estimated dose-illness model is given in Figure 3 (restricting attention to *Salmonella* Enteritidis combined with the chicken and sauce food matrices).

Finally, the uncertainty in estimated dose-illness curves is assessed by means of a 2-stage bootstrap procedure taking into account both stochastic variability as well as data uncertainty. For each outbreak study, the WHO (2003) report defines the uncertainty on dose  $D$ , on the total number of exposed subjects  $N$  and on the number of ill subjects  $Y$  by means of uncertainty distributions. We incorporate data uncertainty in a bootstrap procedure using the same uncertainty distributions. Given the original sample,  $\{(D_j, S_j, N_j, Y_j, t_j, m_j)\}_{j=1}^J$  with  $J = 23$ , generate a new ‘pseudo-original’ sample  $\{(\tilde{D}_j, S_j, \tilde{N}_j, \tilde{Y}_j, t_j, m_j)(b)\}_{j=1}^J$ ,  $b = 1, \dots, B$ , according to the uncertainty distributions as detailed in WHO (2003). Given this ‘pseudo-original’ sample, generate a bootstrap sample  $\{(\tilde{D}_j, S_j, \tilde{N}_j, \tilde{Y}_j^*, t_j, m_j)(b)\}_{j=1}^J$ , by sampling a binomial observation  $\tilde{Y}_j^* \sim \text{Binomial}(\tilde{N}_j, \tilde{\pi}_j)$  with  $\tilde{\pi}_j = \tilde{Y}_j / \tilde{N}_j$ , for  $j = 1, \dots, J$ . In this way  $B = 500$  bootstrap samples are generated, on which the M3 model is refitted. In order to estimate the  $(1 - \alpha/2)100\%$  confidence intervals, percentile intervals are calculated.

## 4 Conclusions

In this paper, dose-illness is modeled using GLMMs and fractional polynomials of dose. The fractional polynomial models are modified in order to satisfy the properties of three different types of dose-illness models that are proposed by Teunis *et al.* (1999). Within these models, heterogeneity due to differences in host susceptibility are modeled using fixed effects

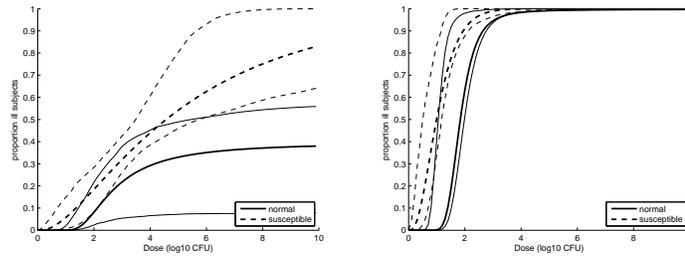


FIGURE 3. (Serovar  $\times$  food matrix)-specific dose-illness curves based on GLMM with fractional polynomial of dose (thick lines) + 90% confidence intervals incorporating stochastic variability and data uncertainty (thin lines). Left panel: Enteritidis  $\times$  chicken; right panel: Enteritidis  $\times$  sauce.

whereas heterogeneity due to differences in serovar type and food matrix are modeled using random effects that are defined for unique combinations of serovar type  $\times$  food matrix.

Based on confidence intervals that incorporate both stochastic variability and data uncertainty, it is concluded that the susceptible population has a higher probability of illness at low dose levels when the combination pathogen-food matrix is extremely virulent and at high dose levels when the combination is less virulent. Furthermore, the analyses suggest that immunity exists in the normal population but not in the susceptible population.

## References

- Bollaerts, K., Aerts, M., Faes, C., Grijspeerd, K., Dewulf, J., Mintiens, K. (2008). Human Salmonellosis: estimation of dose-illness from outbreak data. *Risk Analysis*, to appear.
- Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, **43**, 429-467.
- Teunis, P. F. M., Nagelkerke, N. J. D. and Havelaar, C. N. (1999). Dose Response Models for Infectious Gastroenteritis. *Risk Analysis*, **19**, 1251-1260.
- World Health Organization. (2003). Risk assessments of *Salmonella* in eggs and broiler chickens. Microbial risk assessment series, Nr. 2. World Health Organization (WHO), Geneva, Switzerland.

# Empirical random-effects models to predict the amount individuals withdraw at cash machines

Adam R. Brentnall<sup>1</sup>, Martin J. Crowder<sup>2</sup> and David J. Hand<sup>1,2</sup>

<sup>1</sup> Institute for Mathematical Sciences, Imperial College London, 53 Princes Gate, SW7 2AZ, UK

<sup>2</sup> Department of Mathematics, Imperial College London

**Abstract:** Retail finance organisations use data on past behaviour to make predictions for customer value management strategies. Random-effects models, where each individual has a behavioural pattern drawn from an overall population distribution, are a natural statistical form in this context. This paper develops, using real data, three multinomial random-effects models to predict the amounts customers withdraw from automated teller machines. The use of an empirical distribution for the random-effects is compared to a Dirichlet distribution in two prediction tests. The Dirichlet distribution is found to inflate bin probabilities where no previous withdrawal has been observed more than is appropriate.

**Keywords:** Cash machine; Customer; Random-effects; Prediction.

## 1 Introduction

A range of statistical models is used to help manage risk in the retail financial services sector. However, as noted in Hand and Crowder (2005), most of them seek out past correlations between a set of covariates and a measure of individual risk, and the performance of this type of model can degrade rapidly with changes in conditions such as the economy, technology or competition. Customer management strategies based on predictions about the deeper processes of how individuals behave may be less affected by such changes. In this paper we develop three random-effect models for this purpose in the context of automated teller machine withdrawals. They are based on a random sample (not stratified) of 5 000 accounts with a UK high-street bank, where the time and amount of each withdrawal was recorded over a four-month period in 2005.

## 2 Approach

The general modelling approach taken is as follows. Let  $p(\mathbf{y}|\mathbf{u})$  be a likelihood function, where  $p$  denotes a probability density or mass function,

$\mathbf{y}$  the data set and  $\mathbf{u}$  the parameter set. We take the  $\mathbf{u}$  to vary over individuals  $i = 1, \dots, n$ , i.e. as random effects, with  $p(\mathbf{u}; \theta)$ , where  $\theta$  is the underlying parameter vector. Parametric assumptions might be made for  $p(\mathbf{u}; \theta)$ . Alternatively, one can use the empirical distribution function of the estimated random effects from maximum-likelihood fits to many individual accounts. That is, to replace  $p(\mathbf{u}; \theta)$  by a discrete probability mass function  $\hat{\pi}(\mathbf{u})$ , with atoms  $1/n$  at each  $\hat{\mathbf{u}}_i$ . In addition to avoiding parametric specification of  $p(\mathbf{u}; \theta)$  there may be a computational advantage from not fitting the random-effects distribution (Brentnall *et al.* 2008). The posterior probability distribution of the random effects for each individual  $i$  may then be estimated as

$$\hat{p}(\mathbf{u}|\mathbf{y}_i) = \begin{cases} \frac{p(\mathbf{y}_i|\mathbf{u})}{\sum_u p(\mathbf{y}_i|\mathbf{u})} & \text{if } u \in (\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_n) \\ 0 & \text{otherwise} \end{cases}$$

where  $\sum_u$  denotes summation over the observed  $\hat{\mathbf{u}}_i$ -values. This may be used to predict some future aspect, say  $C$ , of behaviour:

$$p(C|\mathbf{y}) = \sum_u p(C|\mathbf{y}, \mathbf{u})\hat{p}(\mathbf{u}|\mathbf{y}).$$

The approach will be demonstrated on models for the amount withdrawn by individuals at cash machines by running some prediction tests.

### 3 Models

This section describes the models. We choose to bin the withdrawal distributions because around 4 in 5 withdrawals are multiples of £5 or £10, and there are a wide variety of individual distributions. For example, 9% of accounts always withdraw the same amount, but others have a preferred amount of £10 to £50 and a long tail into hundreds of pounds.

#### 3.1 Model 1

The first model uses multinomial probabilities for binned-withdrawal ranges. Suppose that the set of distinct withdrawal amounts in the complete data is  $(a_1, \dots, a_M)$ , where  $M$  is the number of different withdrawal amounts. Then the same number  $m < M$  of bins are chosen for each account and the withdrawal-amount record of account  $i$  is represented as  $\mathbf{y}_i = (y_{i1}, \dots, y_{im})$ , where  $y_{ij}$  is the number of amounts falling in the  $j$ -th bin. The distribution of  $\mathbf{y}_i$  is taken to be multinomial with probability vector denoted by  $\mathbf{q}_i = (q_{i1}, \dots, q_{im})$ , and we take the  $\mathbf{q}_i$ -distribution as Dirichlet (Evans *et al.*, 2000) with parameter set  $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_m\}$ .

### 3.2 Model 2

In model 2 the Dirichlet random-effects distribution of model 1 is replaced by an empirical distribution  $(\hat{\mathbf{q}}_1, \dots, \hat{\mathbf{q}}_n)$ .

### 3.3 Model 3

The last model incorporates a form of dependence observed for some accounts in an exploratory data analysis, where withdrawals made a relatively short-time apart are of similar amounts. Let the withdrawal-amount record of account  $i$  to time  $t$  be represented as  $\mathbf{z}(t) = \{z_1, \dots, z_{n(t)}\}$  where  $z_k \in (1, \dots, m)$ ,  $m$  is the number of bins chosen for the distribution of amounts and  $n(t)$  is the number of withdrawals made to time  $t$ . Denote by  $\mathbf{x}(t) = \{x_1, \dots, x_{n(t)}\}$  the withdrawal times. For  $j = 1, \dots, m$  a model for the probability of the amount bin  $b(t)$  chosen by an individual  $i$  at time  $t > t_p$  is:

$$P\{b(t) = j | \mathbf{v}, \mathbf{w}, b_p, t_p\} = \frac{v_j + I(j = b_p) \exp\{w_1 - w_2(t - t_p)\}}{\mathbf{v}_+ + \exp\{w_1 - w_2(t - t_p)\}}$$

where  $I(\cdot)$  is the indicator function,  $b_p$  is the previous withdrawal bin,  $t_p$  is the previous withdrawal time,  $\mathbf{v} = (v_1, \dots, v_m)$  and  $\mathbf{v}_+ = \sum_{j=1}^m v_j$ . The parameters  $\mathbf{w} = (w_1, w_2)$  are constrained to be greater than or equal to zero. This model may be interpreted in the following way. Conditional upon a withdrawal occurring at time  $t$ , the probability of it being from each bin is linked to a set of parameters  $\mathbf{v}$ . If  $\exp(w_1)$  is zero then it is equivalent to a multinomial model. The  $\mathbf{w}$  terms have the effect of increasing the probability that the last withdrawal bin  $b_p$  is chosen again soon afterwards. The increased probability, determined by  $w_1$ , decays exponentially following the previous transaction at  $t_p$  at a rate determined by  $w_2$ . The other bin probabilities are proportionally adjusted so they all sum to unity. The distribution of random effects is unspecified and approximated by  $(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \dots, \hat{\mathbf{r}}_n)$  where  $\hat{\mathbf{r}}_i = (\hat{\mathbf{v}}_i, \hat{\mathbf{w}}_i)$  for each account  $i$ .

## 4 Prediction tests

The models were fitted to the first three months data by maximum likelihood, using a Nelder & Mead (1963) simplex algorithm to estimate  $\hat{\mathbf{a}}$  for model 1 and  $(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \dots, \hat{\mathbf{r}}_n)$  for model 3.

### 4.1 Predicting a binned range

The first set of results examines predictions about the probability that each individual account will withdraw in a bin containing £20, by using the bins  $[\text{£}0, \text{£}19)$ ,  $[\text{£}19, \text{£}29)$ ,  $[\text{£}29, \text{£}99)$  and  $[\text{£}99, \infty)$ .

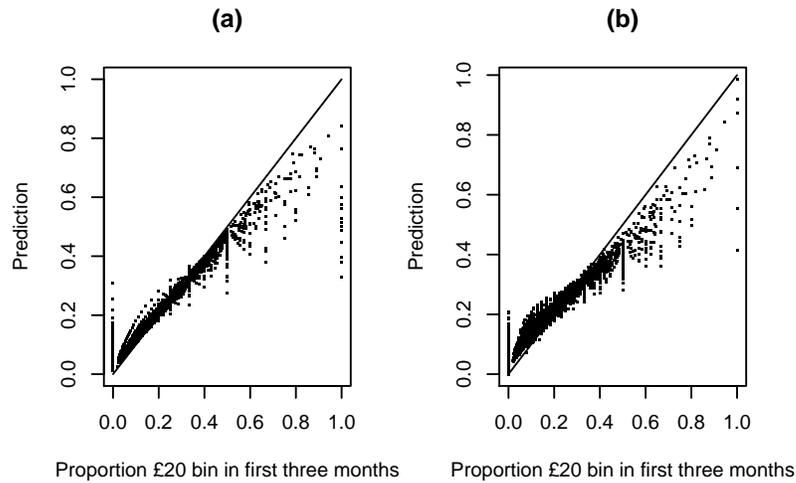


FIGURE 1. Shrinkage in (a) model 1 (Dirichlet) and (b) model 2 (Empirical) multinomial random-effects model prediction

The charts in Figure 1 show that models 1 and 2 have similar shrinkage, but the former shrinks predictions more for those who only withdrew amounts in the £20 bin during the first two months, or made no withdrawals in the £20 bin. That is, the accounts that did not make any withdrawals in the £20 bin during the first three months, are predicted higher probabilities under the Dirichlet model; the accounts that only made withdrawals in the £20 bin during the first three months, are predicted lower probabilities under the Dirichlet model.

Predictions on the probability that a withdrawal is in the £20 bin are compared by using the approach recommended by Copas (1983). Here we set  $b = 1$  if the next withdrawal after the fitting period is in the range [£19, £29) and  $b = 0$  if not, and plot  $b$  against the predicted probability using (Normal) kernel-smoother lines. The lines in Figure 2 show that the predictions are ordered sensibly, and the random-effects models show slightly better performance to projecting past behaviour forwards for predictions between 0.4 and 0.6 when few withdrawals have been observed.

#### 4.2 Predicting extreme percentile points

The next set of results use 11 bins to test predictions on extreme percentile points, which are also of interest to banks. Define  $x_\gamma$  by  $P(W > x_\gamma) = \gamma$  for some specified (small) probability  $\gamma$ , where  $W$  is the withdrawal amount;

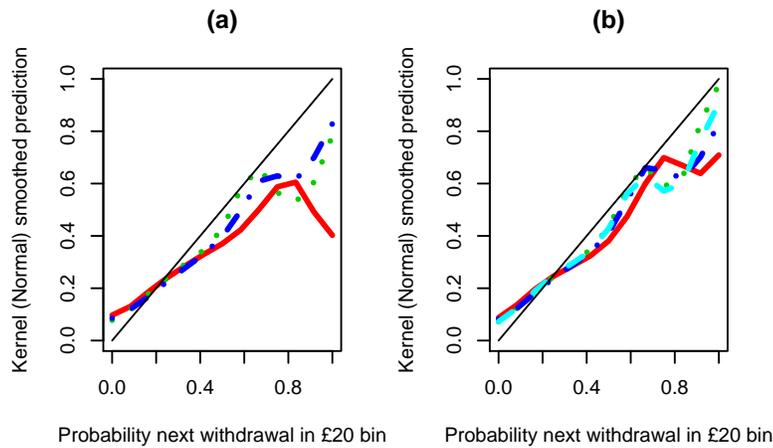


FIGURE 2. Predictive performance of a multinomial projection (—), model 1 (— · —), model 2 (· · ·) and model 3 (— · ·) for those with (a) at least 1 withdrawal in the first 3 months (3 778 accounts), and (b) at least 8 (2 753 accounts).

so  $x_\gamma$  represents an unusually large withdrawal amount. Then, predictions may be made based on an estimate of this probability and then used within a fraud-detection strategy. A test of the random-effects model is to choose different values of  $\gamma$ , and compare the proportion of accounts whose next withdrawal is beyond the upper  $\gamma$  percentile. If the model is accurate, the observed proportion should equal  $\gamma$ .

The results are presented in Table 1. In this comparison model 2 performs noticeably better than use of individual empirical distribution functions (EDFs), especially for those with not many transactions. Model 1 does worse than model 2 for the higher  $\gamma$  levels, but better for the lower ones. The difference is related to the Dirichlet distribution used by model 1. Many accounts do not make withdrawals in all the bins, so there may be a step at zero in the marginal bin distributions. This is smoothed out by the Dirichlet model, increasing such bins' predicted probabilities, but the mismatch at the higher levels of  $\gamma$  shows that they may be inflated too much. However, the Dirichlet model does better at the lower  $\gamma$  levels, and may be more useful for predictions in this range.

TABLE 1. Percentage of predictions that fall outside upper  $\gamma$  confidence levels for accounts with a transaction in the last month and (a) at least one transaction in the first three months (3 778 accounts), and (b) more than 13 transactions in the first three months (2 100 accounts).

		Level $\gamma$ (%)				
	Model	10.0	5.0	2.5	1.0	0.1
(a)	EDF	10.8	8.3	7.5	7.2	7.1
	1	4.3	2.5	1.7	1.2	1.0
	2	8.9	5.3	3.8	2.7	1.7
(b)	EDF	9.3	5.7	4.2	3.7	3.6
	1	5.1	2.6	1.6	0.9	0.7
	2	9.0	5.3	4.0	3.0	1.7
	3	8.9	5.3	4.0	2.7	1.5

## 5 Conclusion

In this paper three models to predict individual accounts' ATM withdrawal amount have been developed using a sample of 5 000 accounts with a UK high-street bank. Empirical random-effects models were compared to a Dirichlet form and shown, in several cases, to predict more accurately. This is probably because the Dirichlet distribution inflates bin probabilities where no previous withdrawal has been observed more than is appropriate.

**Acknowledgments:** Adam Brentnall's work on this project was supported by EPSRC grant number EP/D505380/1 and the work of David Hand was partially supported by a Royal Society Wolfson Merit Award.

## References

- Brentnall, A.R., Crowder, M.J., and Hand, D.J. (2008). A statistical model for the temporal pattern of individual automated teller machine withdrawals. *Applied Statistics* **57**, 43-59.
- Copas, J. B. (1983). Plotting  $p$  against  $x$ . *Applied Statistics* **32**, 25-31.
- Evans, M., Hastings, N., and Peacock B. (2000). *Statistical Distributions*. New York: Wiley-Interscience.
- Hand, D.J., and Crowder, M.J. (2005). Measuring customer quality in retail banking. *Statistical Modelling*, **5**, 145-158.
- Nelder, J.A., and Mead, R. (1965). A Simplex Method for Function Minimization. *Computer Journal*, **7**, 308-313.

# Student monitoring using Chess ratings

Matthieu J.S. Brinkhuis<sup>1</sup>, Gunter Maris<sup>1,2</sup>

<sup>1</sup> Cito, Psychometric Research and Expertise Center. P.O. Box 1034, 6801 MG Arnhem. E-mail: matthieu.brinkhuis@cito.nl

<sup>2</sup> University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam.

**Abstract:** The Elo rating system has its origins in chess ability estimation. The system is proposed as a means of analyzing data from student monitoring systems, such as the *Rekentuin*. The data consists of varying amounts of measurements at varying time points. Applying the Elo rating system to this data has several advantages. There is no need for a growth model, both person ability and item difficulty are estimated simultaneously, ratings are updated after each response and the algorithm requires no computationally intensive calculations. Since bias was found in the logistic Elo rating system, a correct variant is presented in order to obtain a stationary distribution.

**Keywords:** Dynamic paired comparisons data; chess rating; dynamic ability monitoring; Elo rating system.

## 1 Introduction

In educational measurement, one increasingly popular demand is to follow student ability over time. Such systems are known as *student monitoring systems*. Examples include progress testing at the University of Maastricht's medical faculty, and at the psychology program at the Erasmus University. Also, progress is monitored by several products from the *Dutch national institute for educational measurement* (Cito), tracking several abilities throughout primary and special education. Clear reasons for the interest in this topic are the possibilities to compare students and for example provide remedial teaching or accelerated programs.

In progress testing both the frequency of test administration and the accuracy at each administration are important. Frequency of administration is important in many applications because it allows for quick intervention when deviant growth patterns are observed. Accuracy of results is usually obtained by creating longer tests and is especially important in high stake progress testing. A concern in progress testing is to balance these two interests, the frequency at which tests are administered versus the test length with the response burden as the main constraint.

One specific field in which short tests are used to track ability are sports. A player, or a team, plays a match against an opponent which leads to a

win, loss or draw. Data from such matches are known as *paired comparisons data*. In the field of chess a particular ability estimation system has been developed by Arpad Elo, the *Elo Rating System* (ERS) (e.g., Elo, 1978; Glickman, 1999). The ERS has several large implementations, of which the most well known are the ratings of the *World Chess Federation* (FIDE). An advantage of the ERS that it allows for continuous ability monitoring of an individual, instead of occasional snapshots that traditional testing provides. While in chess two players compete, one can regard a player as a respondent and its opponent as an item. This approach allows for updating an ability estimate after each item administered, instead of after a more lengthy test.

Since the ERS is already functioning for a long time in several large applications, it seems to work in practice. However, does it also work in theory? Several properties of the Elo rating system are evaluated for this purpose.

## 2 Data

Data have been acquired from the *Rekentuin*. The *Rekentuin* is an arithmetic testing website from the University of Amsterdam. Pupils from primary education can answer arithmetic questions, where growth in ability translates in growth of their virtual garden. Pupils use the website at their schools, but can login from their homes. Applying Elo estimates in this configuration provides several advantages. First, an ability estimate is updated after every answer without changing other pupils' ability estimates. Second, tracking changes over time is quite easy with no need for a growth model. Third, both pupil ability and item difficulty are estimated simultaneously. And fourth, the algorithm requires light calculations.

A more traditional approach to the data might be calculating *maximum likelihood* (ML) estimates. However, ML provides several disadvantages over the ERS in this context. ML estimates are updated for all persons and items after a single question is answered by a single person, which is difficult to explain to individual players and computationally burdensome in larger applications. Furthermore, assuming a growth model is required with the risk of introducing bias.

## 3 Methods

A representation of the Elo algorithm for estimating ability is

$$\begin{aligned}\theta_v &= \hat{\theta}_v + K [x_{vi} - E(X_{vi})] \\ \delta_i &= \hat{\delta}_i - K [x_{vi} - E(X_{vi})]\end{aligned}$$

where  $\hat{\theta}_v$  represents the current ability estimate of pupil  $v$ , and  $\theta_v$  the updated ability estimate. The dichotomous answer under evaluation is  $x_{vi}$ .

The parameter  $\hat{\delta}_i$  is the current estimated item difficulty and the factor  $K$  is a multiplicand of the update step size. The update step is determined by the difference between the answer and the expected value of the answer, which can for example be determined by a Rasch model (Rasch, 1960)

$$E(X_{vi}) = \frac{e^{\hat{\theta}_v - \hat{\delta}_i}}{1 + e^{\hat{\theta}_v - \hat{\delta}_i}}.$$

One should note the relation to the *Bradley Terry Luce* (BTL) model here (e.g. Bradley & Terry, 1952), where players compete against each other instead of answering items.

Due to the design of the ERS, an update corrects the algorithm in the right direction with a certain step size. The behavior of the estimates is therefore quite erratic, jumping up and down between updates. One can for example influence the step size  $K$ , or use a smoother to obtain less erratic ability estimates. In figure 1, the ability estimation of pupil 598 from the *Rekentuin* is displayed. The vertical positions of the triangles illustrates

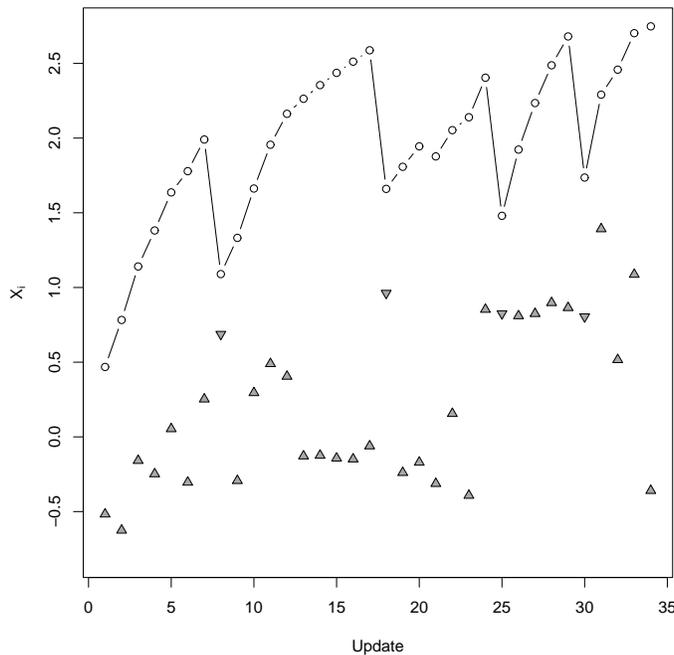


FIGURE 1. Ability estimates for pupil 598.

the difficulties of the items answered, pointing upwards indicates correct answers, pointing downwards incorrect answers. One should note that with

correct answers the estimate increases. The amount of increase is larger if the item was relatively difficult and smaller if the item was relatively easy. The reverse holds for incorrect answers. Note how generally quite easy items are selected in order to keep the younger pupils motivated. Investigating several properties of the logistic ERS, it is found that the logistic ERS does not converge to its true value. While the technicalities are not worked out here, figure 2 demonstrates the bias using a simulated example with an exaggerated scaling factor. Using a symmetrical response

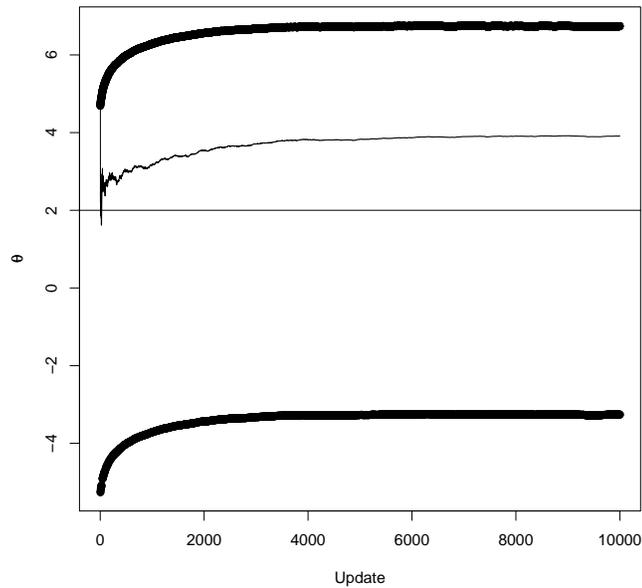


FIGURE 2. Simulation illustrating bias.

function as suggested by Batchelder and Simpson (1979) combined with adaptive opponent or item selection, one finds an algorithm that converges to the correct value and holds the correct variance. However, one still does not exactly obtain a stationary distribution. Using Sapiro (1991), one can see that the *Mixtures of Multivariate Normal Distributions* (MMND) models imply the logit model, and thus that regarding the ERS as a mixture is the solution of obtaining its correct distribution. One should note that to obtain the MMND distribution, the obtained normal distribution needs to be shifted occasionally. A step-wise illustration is provided in figure 3. The mixture distribution in step 1 is obtained by the two distributions of

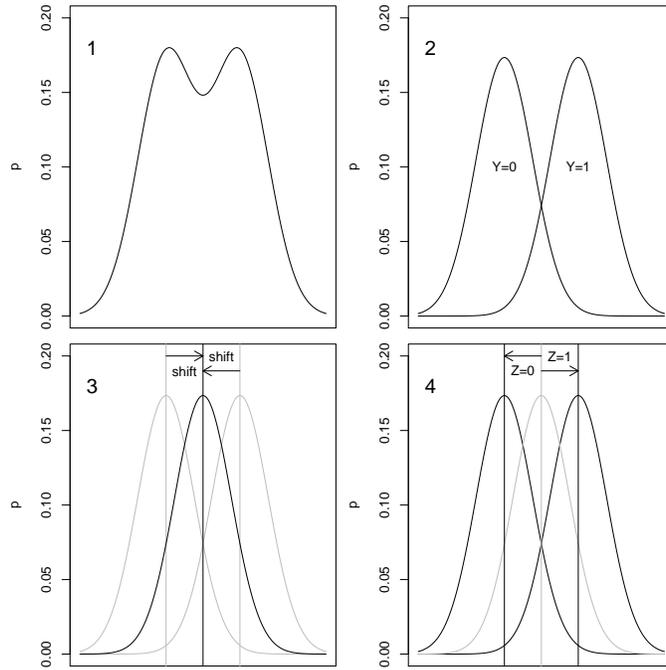


FIGURE 3. Step-wise composition of the mixture distribution.

correct and incorrect answers  $Y$  in step 2. Those are shifted to obtain the single normal distribution in step 3. Using a Bernoulli trial  $Z$ , the distributions are again shifted to obtain the correct mixture. The outcome of the Bernoulli trial determines whether the ability estimate is updated or not, in which one can recognize a similarity to the Metropolis-Hastings algorithms (Hastings, 1970) where an update is accepted or not. The advantage of these steps is that the correct mixture distribution is obtained, the disadvantage that the ability estimate is not always updated after a match, despite of the outcome of the match.

## 4 Results

A quite remarkable result is found when investigating the properties of the logistic ERS, the algorithm does not converge to its true ability estimate. While this is shown by means of a simulation, it can also be shown more formally. A solution to obtain a correct stationary distribution is found using a symmetrical response function and adaptive matching between persons or persons and items. Since these measures are not implemented in for ex-

ample chess ability estimation nor in the *Rekentuin*, one might expect their estimates to be biased. However, matching in chess and sports in general is adaptive to some degree. Both in practice and in tournaments players compete with opponents at an approximate equal ability level. Also the *Rekentuin* is designed to be adaptive to some degree, in order not to pose questions that are too difficult. These adaptive elements limit the amount of bias introduced by the standard ERS. However, to obtain a stationary distribution the adaptive mixture ERS can be implemented. In the *Rekentuin*, these changes can be made quite easily by slightly changing the adaptive matching and to change the update rules to including the mixture distributions.

**Acknowledgments:** Special thanks to Timo Bechger of Cito for expressing thoughts and sharing insights.

### References

- Batchelder, W.H., & Simpson, R.S. (1979). Rating systems for human abilities: The case of rating chess skill. In *Modules in undergraduate mathematics and its applications: Module 698*, 1–22. Arlington, MA: COMAP.
- Bradley, R.A., & Terry, M.E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*, 324–345.
- Elo, A.E. (1978). *The rating of chess players past and present*. New York, NY: Arco Publishing.
- Glickman, M. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, *48*, 377–394.
- Hastings, W.K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)
- Sapra, S.K. (1991). A connection between the logit model, normal discriminant analysis, and multivariate normal mixtures. *The American Statistician*, *45*, 265–268.

# Do pollution time-series studies contain residual confounding by personal risk factors for acute health events?

Sabit Cakmak<sup>1</sup>, Vladislav Brion<sup>1</sup>, Mark Goldberg<sup>2</sup>, Timur Gultekin<sup>3</sup>, Ranjeeta Mallick<sup>1</sup> and Rick Burnett<sup>1</sup>

<sup>1</sup> Health Canada, 137 EHC, Tunneys Pasture A.L 0801A, Ottawa, ON. Canada.  
e-mail: [sabit\\_cakmak@hc-sc.gc.ca](mailto:sabit_cakmak@hc-sc.gc.ca)

<sup>2</sup> McGill University, Montreal, Canada

<sup>3</sup> Ankara University, Ankara, Turkey

**Abstract:** Short-term exposure to ambient air pollution has become a public health concern, largely due to the number of time series studies showing a positive association between daily variations in ambient air pollution concentrations and in the number of hospitalization and deaths.

One of the limitations of time series studies is that they do not provide information on longitudinal changes in health status and whether or not health status is modified by air pollution. Dewanji and Moolgavkar<sup>1</sup> have developed an alternative estimation approach to this problem using a Poisson counting process. We extended the Poisson point process model of Dewanji and Moolgavkar to include multiple diseases and interactions. We illustrate our method through an analysis of data on multiple hospital admissions for congestive heart failure in Montreal to investigate whether air pollution has a differential effect on these event-disease outcomes.

**Keywords:** air pollution; morbidity; poisson process, time series.

## 1 Introduction

Short-term exposure to ambient concentrations of combustion-related pollution has become a public health concern over the last decade largely due to a number of studies linking fluctuations in the daily number of hospitalization and deaths with daily variations in ambient air pollutants ( Pope et al, Spix et al, Cakmak et al, Stieb et al. ). The time series studies consistently show positive associations across diverse study areas having different levels and mixes of pollution and different weather patterns. The fact that only those (few) factors that co-vary temporally with pollution and mortality could act as possible confounding variables made one to think that the association between exposure short term air pollution and health are insensitive to any residual confounding by personal risk factors. However, these studies make use of grouped data, provide little or no information on

individual risk factors (e.g., for assessing effect modification), they provide only a snapshot of acute health events (do not include information on concurrent comorbid conditions or medical history), do not assess individual characteristics or clinical conditions that may vary on short time scales (possible confounding of acute effects), and, as indicated above.

In this paper, we conduct a dynamic population study that takes advantage of knowledge of the recurrent event nature of disease processes. Individuals can have multiple events of a single type (e.g., emergency rooms visits for congestive heart disease), or can experience multiple diseases. The statistical methodology that we propose allow the examination of the effects of short-term exposure to ambient air pollution on the individual's underlying process that generates a potentially complex temporal pattern of disease progression, incorporating the modifying influence of current disease condition on predicting future events related to air pollution.

## 2 Dynamic population study

Consider a hazard model for recurrent events. The relationship between the risk factors  $x_i(t)$  and survival is modeled by the hazard function  $\lambda_i(t)$  for the  $i$ th subject

$$\lambda_i(t) = \lambda_{i0}(t) \exp(x'_i(t)\beta)$$

where  $\lambda_{i0}(t)$  is the baseline hazard function. We have  $n$  subjects under observation with at least one event during the time period  $(0, \tau_I]$ . Estimates of the baseline hazard function and regression parameter vector  $\beta$  can be obtained by finding the value of  $\beta$  that maximizes the likelihood for the time to a recurrent event. This likelihood can be written as a function of the risk factors for the subjects that died on day  $t$  and those that are at risk of dying on day  $t$ . There are almost two million people living in Montreal on any given day and thus the likelihood would be a complex function of a large number of individuals' risk information, making parameter estimating numerically challenging.

Dewanji and Moolgavkar have developed an alternative estimation approach to this problem using a Poisson counting process. Their approach incorporates time dependent covariates and does not use information on those subjects who did not have an event during the observation period. Thus the numerical burden of estimation is greatly reduced. The baseline hazard is defined as a (possibly subject dependent) piecewise constant function in time which define strata (i.e., year, month).

$$\lambda_{i0}(t) = \lambda_{i0l} \text{ for } t \in I_{il} = (\tau_{i,l-1}, \tau_{il}]$$

for  $l = 1, \dots, K_i$  with  $0 = \tau_{i0} < \dots < \tau_{iK_i} = \tau_I$  being  $K_i$  prespecified times for the  $i^{th}$  subject. The Poisson model assumes that the events are

independent within a subject so that the log-likelihood for this process is given by

$$l = \sum_{i=1}^n \left\{ \left( \sum_{j=1}^{d_i} \ln \lambda_i(t_j) \right) - \int_0^{\tau_i} \lambda_i(t) dt \right\}$$

where  $d_i$  denotes the number of events for the  $i^{th}$  subject. Now further denote  $d_{il}$  as the number of events for the  $i^{th}$  subject in the time interval or strata  $I_{il}$  with  $t_{ilj}$  denoting  $j^{th}$  event time. Dewanji and Moolgavkar estimate the parameter vector  $\beta$  using the profile likelihood

$$l_\beta = \sum_{i=1}^n \sum_{l=1}^{K_i} \left\{ \left( \sum_{j=1}^{d_{il}} x'_{ilj}(t) \beta \right) - d_{il} \log \left( \int_{I_{il}} \exp\{x'_i(t) \beta\} dt \right) \right\}$$

which is obtained by substituting the maximum likelihood estimate of the  $\lambda_{i0l}$  into the full log-likelihood,  $l$ .

We extended Dewanji and Moolgavkar's Poisson process approach to include multiple diseases and interactions, thus enabling the examination as to whether air pollution has a differential effect on these event-disease outcomes. Consider the hazard function for both multiple types and diseases of the form

$$\lambda_{ijk}(t) = \lambda_{ijk0}(t) \exp(x'_i(t) \beta_{jk})$$

for the  $j^{th}$  of  $J$  types of events (ER visits, HA, death) and the  $k^{th}$  of  $K$  diseases (IHD, CHF, COPD, pneumonia). The baseline hazard,  $\lambda_{ijk0}(t)$ , can vary by subject, type, and disease while the relation between air pollution and time to event can vary by type of event and disease,  $\beta_{jk}$ . Focus of the analysis is on examining if air pollution has a differential effect on both type of event and disease. We consider reduced models for  $\beta_{jk}$  such as

$$\beta_{jk} = \theta_j + \phi_k$$

that assume no interaction of the air pollution effect on time to event between type and disease. Further models considered are:  $\beta_{jk} = \theta + \phi_k$  (air pollution effect on time to event is independent of type of event);  $\beta_{jk} = \theta_j + \phi$  (air pollution effect is independent of disease); and  $\beta_{jk} = \beta$  (air pollution effect is common to all types and diseases). The baseline hazard is modeled in a similar manner to that of the air pollution effect (e.g.  $\lambda_{ijk0}(t) = \lambda_{i0}(t) e^{\gamma_{jk}}$ ) so that the baseline hazard varies in a proportional manner with type of event and disease.

## 2.1 Example: Air pollution and hospital admissions for congestive heart failure (CHF) in Montreal, Canada

We illustrate the method by applying it to the analysis of hospitalization for congestive heart failure in Montreal, Canada over the period 1991-2002. We

considered the cohort of individuals admitted at least once in 1989-1990 to a hospital in Montreal with a primary diagnosis of congestive heart failure. For this cohort of individuals, we constructed a history of hospitalizations over the entire period 1991-2002. Our data consists of approximately 70,000 admissions for 12,000 individuals. We obtained air pollution and weather information on a daily basis over this period of time from fixed-site monitors on the Island of Montreal. Data was analyzed using four distinct stratification of the time period: each of the years as a stratum, every three months, each month and every single week as a strata. A program designed to analyze multiple events of a single type was provided by Drs. Dewanji and Moolgavkar. The program is written in S-Plus and run time for a simple case excluding personal information and interactions took about 10 hours. We wrote a new program that perform multiple events and interactions. Our program is written in Fortran and our analysis that included personal information and interactions took about 20 minutes to run, which shows a considerable improvement in run time.

### 3 Results

Our results are sensitive to the particular stratification we used. It is clear that temporal trends and seasonal variations in the covariates affect the results of analyses. And it is not clear what the optimal stratification scheme should be chosen. Persons with certain medical conditions, such as congestive heart failure, are more susceptible to air pollution related hospitalization than is the general population. The sensitivity of  $NO_2$  effect and the interactive effect of  $NO_2$  and time dependent variables to the model specification is given in Table 1. The  $NO_2$  effect based on the standard time series regression model (Model 1) assuming daily count of hospitalizations are independent was similar to the corresponding estimate based on "Poisson Process" model ( Model 2) where no personal information is included in the model, indicating that the Poisson process estimation approach gave similar results to the time series regression model for this example. This was most likely due to the large number of subjects and cases of heart disease in the sample. The estimates of the standard errors of the  $NO_2$  effect for the two approaches were also similar (Table 1). There existed strong statistical evidence of interaction between the disease state indicator and  $NO_2$  based on the likelihood ratio test comparing Models 2 and 3 ( $p < 0.0001$ ). The inclusion of a health status variable ( Model 3) at least doubled the estimate of the  $NO_2$  effect but almost did not change the estimated standard error compared to the error obtained from a model including no personal information (i.e. Model 2) and time series regression model (Model 1). This suggests that there was not more variation in heart disease between patients than can be fully explained by the personal information and  $NO_2$ . The  $NO_2$  effect estimate was sensitive to inclusion of health status variable.

Table 1: Effect of Unit Change in  $NO_2$  on the Logarithm of the Congestive Heart Failure in Montreal, Canada by Model specification.

Model	Personal variable	Time/strata			
		Year	3 months	month	week
Time Series (1)	None	0.0022	0.001	0.00093	0.00133
	Model	(0.000369)	(0.000368)	(0.000423)	(0.000482)
Poisson Process (2)	None	0.00223	0.00173	0.00107	0.00133
	Model	(0.000339)	(0.000363)	(0.000463)	(0.000460)
Poisson Process (3)	Health Status	0.0047	0.0073	0.00712	0.0042
	Model	(0.000409)	(0.000460)	(0.000507)	(0.000611)
	Interactive Effect	-0.00167	-0.00283	-0.00276	-0.00126
		(0.000142)	(0.000166)	(0.000182)	(0.000208)

Note: Standard errors are given in parenthesis and Interactive effects are related to interaction between pollution & personal information specified in the model.

## 4 Discussion

A challenge of time series studies is the lack of a clear-cut method to choose the smooth time function to eliminate "long-term" and seasonal trends in the data, and different estimation methods can lead to different results. It is usually suggested that the smooth function be selected so that the residual time series is consistent with a white noise process. It seems clear now that estimates of the air pollution effect are sensitive to the method of modeling time and weather, although this sensitivity can vary by location and season depending on how these variables are correlated. A fundamental assumption in time series studies is that the relative risks estimated in the study population can be applied uniformly to all members of the population. (namely, there is no effect modification between individual characteristics and ambient air pollution). This assumption may not be valid as evidenced, for example, from the findings of Cakmak et al. in which an interaction was found between attained education and income (a measure of socioeconomic status) and level of air pollution. In addition, Information on disease status can be incorporated into synthetic longitudinal cohort model (as in this study) by defining an individual-level covariate as an indicator function of the presence/absence of a disease, which would vary with time. The interaction between the disease state indicator and air pollution would provide a means of assessing the effect modification of host conditions on air pollution related hospitalizations.

## 5 Summary and Conclusions

37 ppb increase in daily mean  $NO_2$  is associated with 5 % (95%CI:1.4-8.9) increase in total hospitalization based on time series analysis (Model 1). The  $NO_2$  effect based on poisson process model (Model 2) where no personal information is included in the model is similar to the corresponding estimates based on time series model (Model 1). The addition of interaction

of air pollution and health status indices into the model resulted in an increasing of the estimate of  $NO_2$  effect to 16.8% (95%CI:11.5-22.1).

We conclude that there exists the effect modification of host conditions on air pollution related hospitalizations.

## References

- Dewanji, A. and Moolgavkar, S. H. (2000). A Poisson process approach for recurrent event data with environmental covariates. *Environmetrics*, **11**, 665-673.
- Pope CA. (1989). Respiratory disease associated with community air pollution and a steel mill, Utah Valley. *American Journal of Public Health*, **79**(5), 623-628.
- Spix C, Anderson HR, Schwartz J et al. (1998). A Short-term effects of air pollution on hospital admissions of respiratory diseases in Europe: a quantitative summary of APHEA study results. *Arch Environ Health*, **53**(1), 54-64.
- Cakmak S, Dales RE, Blanco Vidal C. (2007). Air Pollution and Mortality in Chile: susceptibility among the elderly. *Environ Health Perspect*, **115**(4), 524-527.
- Stieb DM, Judek S, Burnett RT. (2002). Meta-analysis of time-series studies of air pollution and mortality: effects of gases and particles and the influence of cause of death, age, and season. *J Air & Waste Manage Assoc*, **52**, 470-484.
- Cakmak S, Dales R.E. Judek S. (2006). Respiratory health effect of air pollution gases: Modification by Education and Income. *Archives of Environmental & Occupational Health*, **61**(1), 5-10.

# A Warped Failure Time Model for Human Mortality

Carlo Giovanni Camarda<sup>1</sup>, Paul H. C. Eilers<sup>2</sup> and Jutta Gampe<sup>1</sup>

<sup>1</sup> Max Planck Institute for Demographic Research, Rostock, Germany.  
camarda@demogr.mpg.de, gampe@demogr.mpg.de

<sup>2</sup> Methodology and Statistics, Faculty of Social and Behavioural Sciences,  
Utrecht University & Data Theory Group, Leiden University, The Netherlands.  
P.H.C.Eilers@uu.nl

**Abstract:** We present an extension of the accelerated failure time model for comparison of density functions. With this model one can study how the time axis would have to be transformed so that one density distribution conforms to another. Parametric and non-parametric estimates from actual data can be used as target distributions. The only assumption we make about the warping function is the smoothness and we represent it by a linear combination of  $B$ -splines. To estimate the warping function we use a penalized Poisson likelihood approach. The optimal smoothing parameter is found by minimizing the BIC. Actual demographic applications are presented.

**Keywords:** Accelerate failure time models; human mortality; penalized likelihood; smoothing; warping.

## 1 Introduction

Traditionally human mortality is studied by comparing hazard functions. However, lifespan distributions can also be characterized by their probability density. Thus, instead of modelling trends in the hazard functions, we can study how the age-axis would have to be transformed so that one age-at-death distribution conforms to another. In the simplest case the transformation is linear, leading to an accelerated failure time model. A uniform rescaling of the time-axis often is too simplistic, though, to adequately capture mortality dynamics.

## 2 Comparing Age-at-Death Distributions

Figure 1 shows the age-at-death distributions, as derived from period life-tables (Keyfitz & Caswell, 2005), for Japanese women above age 30 in 1947 and in 2006. (The data are derived from the Human Mortality Database,

2008.) To investigate the changes in mortality that lead to the different patterns we want to transform the age-axis such that the two densities coincide. More specifically, we define one distribution as the target, with density  $f(x)$ , and want to obtain the transformation function  $w(x)$  so that the density of the other distribution,  $g(x)$ , conforms to the target density on the warped axis, i.e.,

$$g(x) = f(w(x)) \cdot |w'(x)|. \quad (1)$$

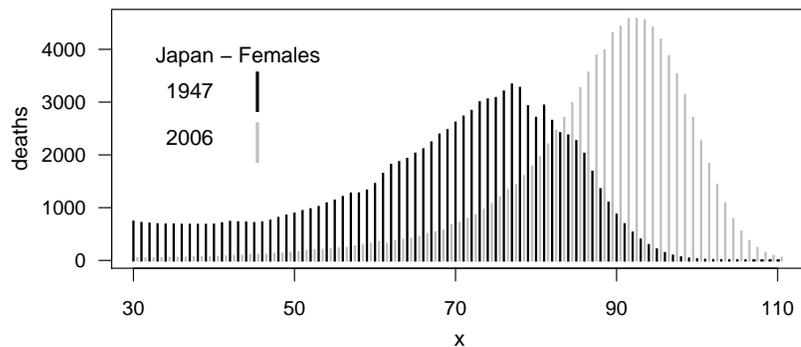


FIGURE 1. Life-table age-at-death distribution of Japanese females for the years 1947 and 2006.

We do not want to restrict the transformation function  $w(x)$  to any rigid shape. The only assumption will be the smoothness of  $w(x)$ .

Nonlinear transformation of the independent axis to achieve close alignment of functions is called ‘warping’ in functional data analysis (Ramsay & Silvermann, 2005). Hence we call this a Warped Failure Time (WaFT) model. Warping ideas have been introduced into mortality analysis in Eilers (2004). In this paper we extend this approach further.

### 3 The Warped Failure Time Model

Our model is not restricted to any particular target distribution,  $f(x; \boldsymbol{\theta})$ , and it will mostly be estimated from data. In the following we consider the parameters  $\boldsymbol{\theta}$  fixed.

The observed death counts at age  $x_i$  are denoted by  $y_i$  and are realizations from Poisson variables with  $E(y_i) = \mu_i$ . The values  $\mu_i$  derive from the density  $g(x)$  that generated the data. The proposed model is

$$\begin{aligned} \mu_i = E(y_i) &= \gamma \cdot f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta}) \cdot \left| \frac{\partial}{\partial x} w(x_i; \boldsymbol{\alpha}) \right| \\ &= \gamma \cdot f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta}) \cdot |v(x_i; \boldsymbol{\alpha})|, \end{aligned} \quad (2)$$

where  $\gamma$  is a normalizing constant and  $f(\cdot)$  is the target density. The warping function  $w(x; \boldsymbol{\alpha})$  is to be determined such that, after transforming the age-axis, the density matches the specified target. To allow for arbitrary shape of  $w(x)$ , we represent the warping function by a linear combination of  $K$   $B$ -splines of degree  $q$ . Their knots are equally spaced by a distance  $h$ :

$$w(x; \boldsymbol{\alpha}) = \sum_{k=1}^K B_k^q(x) \alpha_k.$$

Smoothness of the warping function is enforced by a difference penalty on neighbouring coefficients  $\alpha_k$ . The derivative then is

$$\frac{\partial}{\partial x} w(x; \boldsymbol{\alpha}) = v(x; \boldsymbol{\alpha}) = \frac{1}{h} \sum_k B_k^{q-1}(x) [\alpha_k - \alpha_{k+1}],$$

c.f. Eilers & Marx (1996).

#### 4 A Penalized Poisson Likelihood Approach

In order to estimate the coefficients  $\boldsymbol{\alpha}$  in (2) we use a penalized Poisson likelihood approach. The estimates of the solution to the first-order conditions are obtained by differentiating the log-likelihood with respect to the elements of  $\boldsymbol{\alpha}$ :

$$\sum_i (y_i - \mu_i) \frac{\partial \eta_i}{\partial \alpha_k} = 0, \quad (3)$$

where

$$\frac{\partial \eta_i}{\partial \alpha_k} = \frac{\partial \ln(\mu_i)}{\partial \alpha_k} = \frac{\frac{\partial f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta})}{\partial w(x_i; \boldsymbol{\alpha})}}{f(w(x_i; \boldsymbol{\alpha}), \boldsymbol{\theta})} \cdot \frac{\partial w(x_i; \boldsymbol{\alpha})}{\partial \alpha_k} + \frac{\frac{\partial v(x_i; \boldsymbol{\alpha})}{\partial \alpha_k}}{v(x_i; \boldsymbol{\alpha})}. \quad (4)$$

In matrix notation we can more succinctly write (4) in the following way:

$$\mathbf{Q} = \text{diag} \left( \frac{\mathbf{f}'}{\mathbf{f}} \right) \cdot \mathbf{B}(\mathbf{x}) + \text{diag} \left( \frac{1}{\mathbf{v}} \right) \cdot \mathbf{C}(\mathbf{x}). \quad (5)$$

Here  $\mathbf{B}(\mathbf{x}) = [B_k^q(x_i)]_{ik}$ ,  $\mathbf{C}(\mathbf{x}) = [C_k(x_i)]_{ik}$  with  $C_k(x_i) = \frac{1}{h} [B_k^{q-1}(x_i) - B_{k-1}^{q-1}(x_i)]$ . Given (3) and (5) we can adapt the iteratively reweighted least squares (IWLS) algorithm. Specifically, equation (5) gives the model matrix, which depends on the coefficients  $\boldsymbol{\alpha}$  and needs to be updated with each iteration. Moreover, the normalizing constant  $\gamma$  needs to be included in the algorithm. In matrix notation we have

$$(\tilde{\mathbf{X}}' \tilde{\mathbf{W}} \tilde{\mathbf{X}}) \boldsymbol{\beta} = \tilde{\mathbf{X}}' \tilde{\mathbf{W}} (\tilde{\mathbf{W}}^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}) + \tilde{\mathbf{X}} \tilde{\boldsymbol{\beta}}),$$

where the model matrix is  $\tilde{\mathbf{X}} = [\mathbf{1}, \tilde{\mathbf{Q}}]$ , the coefficient vector  $\boldsymbol{\beta}' = [\ln(\gamma), \boldsymbol{\alpha}']$  and the matrix  $\tilde{\mathbf{W}} = \text{diag}(\tilde{\boldsymbol{\mu}})$ .

Following Eilers & Marx (1996), the number of basis functions  $K$  for the warping function is chosen purposely high and a roughness penalty on the coefficient vector  $\alpha$  is used to ensure smoothness. If we introduce this penalty into the likelihood, we obtain the following system of equations

$$(\tilde{X}'\tilde{W}\tilde{X} + \lambda P)\beta = \tilde{X}'\tilde{W}(\tilde{W}^{-1}(y - \tilde{\mu}) + \tilde{X}\tilde{\beta}) \quad (6)$$

where  $P = \begin{pmatrix} 0 & 0 \\ 0 & \check{P} \end{pmatrix}$  and  $\check{P} = D_d'D_d$ . The matrix  $D_d$  calculates  $d$ -th order differences. Via the value of the parameter  $\lambda$  the smoothness the warping function can be controlled. To choose the optimal  $\lambda$  we minimize the Bayesian Information Criterion where the effective dimension is the trace of the hat-matrix from the estimated system in (6).

### 5 Applications

In the example of Figure 1 we use the year 2006 as the target density. Because of its prominent role in the study of adult human mortality, a Gompertz distribution was estimated. That is, the target density is  $f(x; \theta) = \theta_1 e^{\theta_2 x} \exp\{\frac{\theta_1}{\theta_2}[1 - e^{\theta_2 x}]\}$ , and we obtained  $\hat{\theta} = (2.93e^{-6}, 0.1149)'$ .

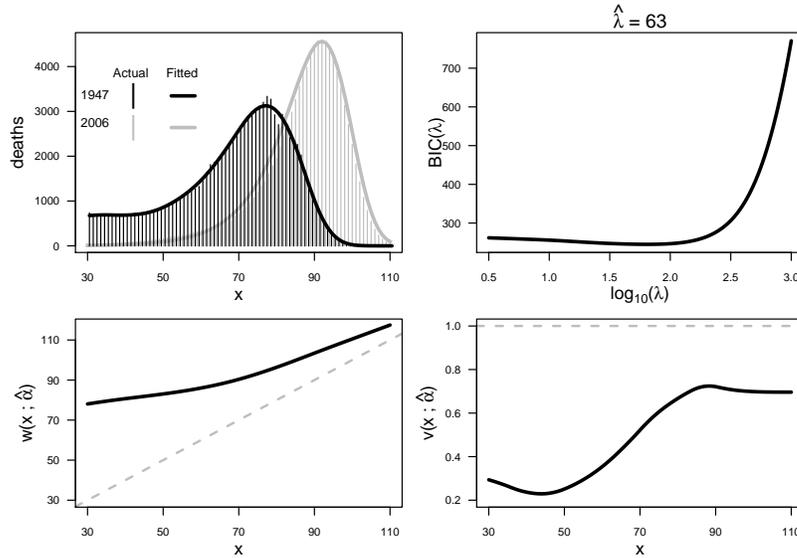


FIGURE 2. Life-table age-at-death distributions for the Japanese data. Data from 2006 are fitted with a Gompertz function and used as target distribution, data from 1947 are estimated with the WaFT model (top-left). BIC profile (top-right). Bottom row: estimated warping function  $w(x, \hat{\lambda})$  and its derivative  $v(x; \hat{\lambda})$ .

For the representation of the warping function  $K = 15$   $B$ -splines of degree  $q = 3$  were used. The order of the penalty was  $d = 2$ . As starting values we used the estimates from a warping function that only shifts the distribution so that the modes of the two densities coincides. The value of  $\lambda$  selected by minimizing the BIC is equal to 63.

The upper-left panel in Figure 2 shows the target distribution with its Gompertz estimates as well as the fitted values from the WaFT model. The change of the BIC with  $\log_{10}(\lambda)$  is given in the upper-right plot. The bottom images present the resulting transformation function  $w(x; \hat{\alpha})$  and its derivative. The identity is indicated by a dashed line. The warping function is nonlinear, that is, neither a simple shift nor a uniform scaling of the age-axis can map one density on to the other.

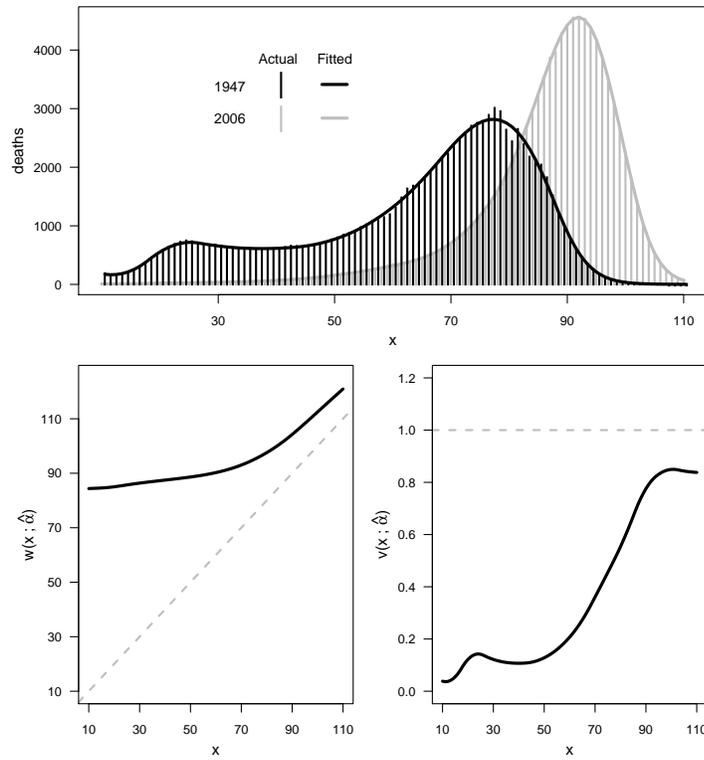


FIGURE 3. Life-table age-at-death distributions for the Japanese data. Non-parametric  $P$ -splines estimate for the target distribution (year 2006). Data from 1947 are estimated with the WaFT model (top). Bottom row: Estimated warping function  $w(x, \hat{\alpha})$  and its derivative  $v(x; \hat{\alpha})$ .

Often the Gompertz distribution with only two parameters cannot properly describe more complex patterns of adult mortality. Alternatively we can use a non-parametric estimate of the target distribution to improve the model. Again we choose a  $P$ -spline approach to obtain the estimated target density (Eilers & Marx, 1996).

Figure 3 shows outcomes for the Japanese women above age 10. We use the same specifications of the warping function ( $K = 15, q = 2, d = 2$ ). The smoothing parameter  $\lambda$  chosen by BIC is equal to 10. The warping function and its derivative depict again a non-linear transformation of the age-axis, especially under age 30.

## 6 Discussion

In this paper we presented a new approach for dealing with the estimation of a nonlinear transformation of densities. The proposed WaFT model is a rather general tool and brings together the ideas of warping and smoothing. By using a  $P$ -spline approach we may not only estimate the warping function, but also its derivative provides further useful details for interpretation of the model.

In this paper we presented applications from human mortality only, hence stressing the generalization of the more simple accelerated failure time models. However, the WaFT model is appropriate for comparison of any two densities. Further generalizations of the WaFT approach to a sequence of warping functions over time in a two-dimensional setting is in progress.

## 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. (2004). The Shifted Warped Normal Model for Mortality. *Proceedings of the 19th International Workshop of Statistical Modelling*, 159-163.
- Human Mortality Database (2008). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at [www.mortality.org](http://www.mortality.org).
- Keyfitz N. and Caswell H. (2005) Applied Mathematical Demography. 3rd ed. Springer.
- Ramsay J.O. and Silverman B.W. (2005) Functional Data Analysis. 2nd ed. Springer.

# Cd and Cu migration during electro dialysis: biregressional modeling

Paulo Canas Rodrigues<sup>1</sup> and Ana Teresa Lima<sup>2</sup>

<sup>1</sup> Faculty of Sciences and Technology, Nova University of Lisbon, 2829-516 Caparica, Portugal. Email: paulocanas@fct.unl.pt

<sup>2</sup> Superior School of Technology and Management of Oliveira do Hospital, Polytechnical Institute of Coimbra, 3400-124 Oliveira do Hospital, Portugal. Email: lima.at@gmail.com

**Abstract:** The electro dialytic process (EDR) is a remediation technique based on the principle of electrokinetics and dialysis for the removal of contaminants from contaminated solid media. The electro dialytic heavy metal extraction has been tried out in several matrices, including fly ash. This is a hazardous waste due to its enrichment on heavy metals. In the present study, a nonlinear biregressional methodology was used to model the heavy metal migration during the treatment of four different types of fly ash: from the combustion of straw (ST), municipal solid waste in Denmark (DK) and Portugal (PT) and the co-combustion of wood (CW). Firstly, 4th degree polynomial regressions were adjusted to Cd and Cu concentrations along time. After and since some variables are categorical, a categorical regression was adjusted considering “Ash type”, “Duration”, “Final pH” and “Dissolution” as explanatory variables and the coefficients from the polynomial regressions as dependent variable.

**Keywords:** Electro dialytic Process; Fly ash; Multiple regression design.

## 1 Introduction

Incineration is a feasible treatment for waste management, reducing the volume of waste up to 90%. In Europe, the use of wood, straw and certain crops for energy production represents the largest renewable energy source, where combustion of biomass represents an energy production with low CO<sub>2</sub> emission. This would increase incineration as the preferred treatment and the subsequent increase of its products, e.g. fly ash. Fly ash is enriched with volatile contaminants during incineration, among them heavy metals. However, depending on its source, fly ash may be reused if its heavy metal content would be reduced, for instances, as soil amendment, geotechnical, or concrete. The EDR is a remediation technique first described for heavy metal contaminated soil (Ottosen et al., 1997) which combines an electric DC field as cleaning agent with ion-exchange membranes, allowing the regulation of ion fluxes. There are different conditions affecting heavy metal

migration during EDR, e.g. pH and ash dissolution. A new statistical approach may support the understanding of heavy metals behaviour during the EDR. The linear biregessional method was first developed by Mexia (1990) and previously applied in the study of heavy metal migration in EDR soil remediation (Ribeiro and Mexia, 1997) and timber waste (Moreira et al., 2005). Here we introduce a nonlinear extension of this method for variables with different measurement levels.

## 2 Materials and methods

Ash dissolution was measured by putting fly ash in contact with distilled water. The electro dialytic cell used in this investigation consisted on a three compartment cell, two electrode compartments and one central compartment, where the contaminated waste was placed. Electrode compartments were separated from central compartment by an anion-exchange membrane and a cation-exchange membrane. Eight electro dialytic experiments were carried out with distilled water in a liquid/solid ratio of 4. The migration of the metals was controlled and quantified along remediation time. At the end of the experiments the electrolytes were analyzed for its content of Cd and Cu through atomic absorption spectrophotometry in flame.

Eight electro dialytic experiments were carried out with varying conditions. The considered variables were: “Ash type” - ST, CW, PT, DK; “Duration” - Remediation time (10, 11 and 14 days); “Final pH” - pH obtained in the ash after EDR treatment; and ‘Dissolution” - Ash dissolution, related to initial dry weight, occurring during EDR (%).

## 3 Results and discussion

The main goal of EDR is to remove the heavy metals present in the central compartment and transport them to electrode compartments, using electric current. Now we present the analysis of anolyte (electrolyte in circulation on the anode compartment) and catholyte (electrolyte in circulation on the cathode compartment) concentration profiles for Cd and Cu along experimental time. These concentration trends were analysed according to the biregessional methodology described in Moreira et al. (2005), introducing a nonlinear approach. In the first part of the nonlinear biregessional method, 4th degree polynomial regressions (1) were adjusted to each anolyte (AN) and catholyte (CAT) Cd and Cu concentration profiles along time.

$$y_i(t_i) = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 t_i^3 + \beta_4 t_i^4 + \varepsilon_i; \quad i = 1, \dots, n, \quad (1)$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $i=1, \dots, n$ , and  $n$  the number of experiments. In matrix notation (1) becomes  $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , with  $E(\boldsymbol{\varepsilon}) = \mathbf{0}$  and  $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$ .

To complete the analysis, a second series of regression were adjusted to each  $\beta_j$ ,  $j = 0, \dots, 4$  using the selected “Ash type”, “Duration”, “Final pH” and “Dissolution” as explanatory variables. However, since some variables are non numerical, a categorical regression (CATREG) was applied instead of the linear used in Moreira et al. (2005). The CATREG model fits the classical linear regression model with nonlinear transformations of the variables, written as

$$\varphi_r(\mathbf{y}) = \sum_{j=1}^J \beta_j \varphi_j(\mathbf{x}_j) + \varepsilon, \quad (2)$$

by minimizing the least squares loss function

$$L(\varphi_r; \varphi_1, \dots, \varphi_J; \beta_1, \dots, \beta_J) = N^{-1} \left\| \varphi_r(\mathbf{y}) - \sum_{j=1}^J \beta_j \varphi_j(\mathbf{x}_j) \right\|^2 \quad (3)$$

with  $N$  the number of observations,  $J$  the number of predict variables,  $\{\beta_j\}$ ,  $j = 1, \dots, J$ , the regression coefficients,  $\varphi_r \mathbf{y}$  the transformation for the response variable  $\mathbf{y}$ ,  $\varphi_j(\mathbf{x}_j)$  the transformation for the predict variables  $\{\mathbf{x}_j\}$ ,  $j = 1, \dots, J$  and  $\varepsilon$  the error vector, and where  $\|\cdot\|^2$  denotes the squared Euclidean norm. The loss function (3) is minimized over  $\{\beta_j\}$ ,  $\varphi_j(\mathbf{x}_j)$  and  $\varphi_r \mathbf{y}$  to maximize the least squares fit between  $\varphi_r \mathbf{y}$  and the linear combination  $\sum_{j=1}^J \beta_j \varphi_j \mathbf{x}_j$ . Because the transformed variables  $\varphi_r \mathbf{y}$  and  $\varphi_j \mathbf{x}_j$  are centered and normalized to have sum of squares equal to  $N$ , loss function (3) maximizes the (squared) multiple correlation (Van der Kooij and Meulman, 1997).

The parameters  $\beta_0$  (estimates the initial concentration at the beginning of the time series),  $\beta_1$  (measures the initial rate of migration, i.e. the velocity that the metal is entering the electrolyte compartments) and  $\beta_2$  (measures the rate that one object/metal changes its migration velocity) obtained with the first set of regressions (??) were then used in CATREG as dependent variables. “Ash type”, “Duration”, “Final pH” and “Dissolution” were the controlled variables. Table 1 shows the obtained coefficients and  $R^2$ .

Considering a significance level of 5% ( $\alpha = 0.05$ ), it is seen that all studied variables affect Cd and Cu migration to the catholyte (CAT). If we take Cd CAT migration, the interpretation of Table 1 may be as following: the lower the “Final pH”, the less changeable the initial concentration ( $\beta_0$ ) on CAT (Beta < 0); the higher the migration velocity ( $\beta_1$ ) (Beta > 0); the lower the rate of variation of velocity ( $\beta_2$ ) (Beta < 0). This interpretation is valid for any variable with  $\alpha < 0.05$ , with Beta value as positive or negative.

TABLE 1. CATREG to each  $\beta_0$ ,  $\beta_1$  and  $\beta_2$  of the linear regressions to Cd and Cu in anode and cathode compartments, considering each variable.

Variable	Cd AN		Cd CAT		Cu AN		Cu CAT		
	Beta	p	Beta	p	Beta	p	Beta	p	
$\beta_0$	Ash type	0.738	0.797	1.020	0.010	0.899	0.460	1.169	0.007
	Duration	-0.546	0.538	-0.334	0.019	-0.493	0.549	-0.007	0.534
	Final pH	1.409	0.584	-2.244	0.008	-0.208	0.908	-0.828	0.014
	Dissolution	1.663	0.485	-1.464	0.011	-0.261	0.888	-1.996	0.006
	R2		0.674		1.000		0.741		1.000
$\beta_1$	Ash type	-0.924	0.398	-0.629	0.008	0.735	0.376	-1.578	0.005
	Duration	0.586	0.449	-0.341	0.011	0.321	0.464	-0.146	0.021
	Final pH	0.071	0.965	1.807	0.005	-1.663	0.281	-1.295	0.007
	Dissolution	-0.038	0.980	1.859	0.005	-0.391	0.670	0.813	0.010
	R2		0.795		1.000		0.938		1.000
$\beta_2$	Ash type	0.814	0.362	-2.089	0.003	1.032	0.387	1.232	0.005
	Duration	0.249	0.467	-0.206	0.014	0.227	0.538	-0.233	0.023
	Final pH	-2.775	0.148	-1.241	0.007	-2.341	0.194	-0.648	0.018
	Dissolution	-2.304	0.146	0.788	0.010	-1.819	0.195	-0.793	0.015
	R2		0.960		1.000		0.952		1.000

## 4 Conclusion

The nonlinear biregessional study was found to be a useful tool when there is need of physical interpretation of the results and there are categorical variables in the study. “Ash type”, “Duration”, “Final pH” and “Dissolution” significantly influenced Cd and Cu migration towards the cathode.

## References

- Moreira, E.E., Ribeiro A.B., Mateus E.P. and Mexia J.T., Ottosen L.M. (2005). Regressional modeling of electro dialytic removal of Cu, Cr and As from CCA treated timber waste: application to sawdust. *Wood Science and Technology*, **39**, 291-309.
- Ottosen L.M., Hansen H.K., Laursen S. and Villumsen A. (1997). Electro dialytic remediation of soil polluted with copper from wood preservation industry. *Environmental Science and Technology*, **31**, 1711-1715.
- Mexia, J.T. (1990). Best linear unbiased estimators, duality of F tests and the Scheff multiple comparison method in presence of controlled heterocedasticity. *Computational Statistics and Data Analysis*, **10(3)**, 271-281.
- Ribeiro A.B. and Mexia J.T. (1997). A dynamic model for the electrokinetic removal of copper from a polluted soil. *Journal of Hazardous Materials*, **56**, 257-271.
- Van der Kooij A.J. and Meulman J.J. (1997). MURALS: Multiple regression and optimal scoring using alternating least squares. In: *Sofstat '97 Advances in Statistical Software 6*, Bandilla W, Faulbaum F (eds.). Stuttgart: Lucius & Lucius; 99-106

# Logistic Regression Model: continuous independent variables and linearity in the logit

Ana Isabel Carita<sup>1</sup>, Pedro Luís Marques<sup>2</sup> and Filomena Vieira<sup>3</sup>

<sup>1</sup> CIPER and Departamento de Métodos Matemáticos, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa, Estrada da Costa, 1495-688 Cruz Quebrada-Dafundo, Portugal; email:acarita@fmh.utl.pt

<sup>2</sup> Master Student, Departamento de Ciências do Desporto, F.M.H

<sup>3</sup> Departamento de Ciências da Motricidade, F.M.H.

**Abstract:** In logistic regression assuming linearity in the logit is correct for variable selection and is consistent with the goal of determining a set of significant independent variables for the outcome.

After fitting a model it is necessary and convenient to assess the significance of the estimated coefficients and interpret their values. When a logistic regression model included a continuous independent variable, the interpretation of the estimated coefficient assumed that the logit is linear in the variable. If the linearity condition is not verified, the continuous variable must be replaced by a most logical parametric shape for the scale of the variable. In this work, we approach the problem, analysing the effect in goodness-of-fit, of the replacement of continuous variable, not linear in logit, for categorical variables. We used different logistic regression models to predict pooled for national team basketball players, and we concluded that treat sitting height (cm), sum skinfold (cm) and lean body mass (kg) as continuous variables are better then treating this as categorical variables .

**Keywords:** multiple logistic regression; linearity in the logit; categorical variables; goodness-of-fit.

## 1 Multiple Logistic Regression Model

For a collection of  $p$  independent variables denoted by  $\mathbf{x}' = (x_1, x_2, \dots, x_p)$  were each variable  $x_i$  is at least interval scale, the conditional probability that the outcome is present is denoted by  $P(Y = 1|x) = \pi(x)$ . The logit of the multiple logistic regression model is given by the equation

$$g(\mathbf{x}) = \beta_0 + \beta_1 + \dots + \beta_p$$

in which case the logistic regression model is

$$\pi(\mathbf{x}) = \frac{\exp g(\mathbf{x})}{1 + \exp g(\mathbf{x})}$$

(Hosmer and Lemeshow, 2000)

### 1.1 Interpretation of the fitted logistic regression model

Assuming that a logistic regression model has been fit, its interpretation requires that we be able to draw practical inferences from the estimated coefficients in the model. The interpretation of the coefficients in the model require the knowledge of the functional relationship between the dependent and the independents variables and appropriate definition for the unit of change in the independents variables. In the logistic regression model the link function is the logit transformation

$$g(\mathbf{x}) = \ln \left( \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right) = \beta_0 + \beta_1 + \dots + \beta_p.$$

### 1.2 Linearity in the logit

Once we have a model that we feel contain the essential variables, the preliminary main effects model (Hosmer and Lemeshow, 2000), we should look carefully at the variables in the model. A easily implemented procedure is based on the following observation and is quite described in (Hosmer and Lemeshow, 2000). The difference, adjusted for other model covariates, between the logits for two different groups is equal to the value of an estimated coefficient from a fitted logistic regression model that treats the grouping variable as categorical. Summary, the linearity assumption for logistic regression is assessed by categorizing each continuous variable into a categorical variable, with four levels using three cutpoints based on the quartiles and plotting each variable's coefficients against midpoint of the groups. Visually inspect the plot and choose the most logical parametric shape for the scale of the variable. Refit the model with the correct scale.

### 1.3 Assessing the fit of the model

When the model building stage has completed, we would like to know how actually the model describes the outcome variable. The measures of goodness-of-fit are easily calculated and provided an overall indication of the fit of the model. As measures of goodness-of-fit we used the -2log-likelihood test and the Hosmer-Lemeshow test. A simple way to summarize the results of a fitted logistic regression model is a classification table. The area under the ROC curve measures the model's ability to discriminate between subjects according to the outcome variable.

## 2 Data analysis

Talent identification is crucial for the selection of high performance athletes in low ages. The more important factors for basketball players' success are morphologic, physiological and technical characteristics. (Bayios et

TABLE 1. Results for the fitted multiple regression model

Variable	Coeff.	S.E.	Wald	d.f.	$p$	$\widehat{OR}$
Constant	-53.812	15.94	11.396	1	0.001	
Sitting height	0.523	0.182	8.278	1	0.004	1.687(1.18, 2.41)
Sum skinfold	-0.029	0.010	9.122	1	0.003	0.971(0.95, 0.99)
Lean body mass	0.151	0.077	3.853	1	0.050	1.163(1.00, 1.35)

al.,2006) The sample data included a total of 83 male basketball players, with age between 14 and 16 years old (sub 16), divided into two groups, 48 pooled for national team athletes and 35 club athletes (not pooled for the national team).

## 2.1 Statistical analysis

Multiple logistic regression analysis is used to predict pooled for national team athletes. In the present work we only used morphological variables as independent variables. A step by step forward procedure was used to assess the multiple logistic regression model.

## 2.2 Results

Variables identified as significantly associated with pooled for national team athletes vs. not pooled for national team athletes in the logistic regression model were: sitting height, sum skinfold and lean body mass. The sitting height (cm) and sum skinfold (cm) variables were obtained according to ISAK norms described by Frago and Vieira (2005). Lean body mass (kg) variable was calculated as the difference between weight and body fat mass. The significance of the variables in the model was assessed by the Wald test (W) and CIs. Table 1 shows the results for the fitted multiple regression model. The Nagelkerke  $R^2$  for the logistic regression model was 0.66. The Nagelkerke  $R^2$  attempts to quantify the proportion of explained variance in the logistic regression model. The overall significance of the equation by -2log-likelihood test was 56.480 with  $p < 0.001$  and the Hosmer-Lemeshow goodness of fit test was 4.304 ( $df = 8, p = 0.829$ ). From the classification table we have 86.7% concordant pairs and 13.3% discordant pairs. The area under the ROC curve was 0.93 (S.E. 0.027).

Sitting height, sum skinfold and lean body mass were included as continuous variables. The linearity assumption for logistic regression was assessed by categorizing each continuous variable into a multiple categorical variable and plotting each variable's coefficients against midpoint of the group. For this purpose we create categorical variables using three cutpoints based on the quartiles. Figure 1 shows the plots of estimate logistic regression coefficients versus quartile midpoints of sitting height, sum skinfold and lean body mass. The results do not conclusively support linearity in the logit for

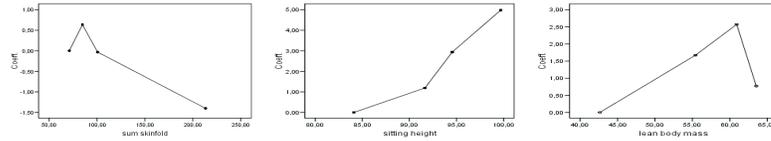


FIGURE 1. Plots of estimate logistic regression coefficients versus quartile mid-points.

TABLE 2. Goodness-of-fit results for the models (i), (ii) and (iii)

Model	$R^2$	-2log-likelihood	H-L test	conc./disc.pairs	ROC area
(i)	0.585	65.617 ( $p < 0.001$ )	0.427	77.1% / 22.9%	0.898
(ii)	0.471	77.225 ( $p < 0.001$ )	0.603	79.5% / 20.5%	0.825
(iii)	0.585	65.617 ( $p < 0.001$ )	0.956	78.3% / 21.7%	0.886

the three variables. The possible scaling suggested by these results is to create a dichotomous variable at the median for sum skinfold, a dichotomous variable at the third quartile for lean body mass, and a dichotomous variable at the median or a 4-level categorical variable based on the quartiles for sitting height. To explore the scale for the variables we used three different parameterizations, (i) sum skinfold as dichotomous variable, sitting height as continuous variable and lean body mass as dichotomous variable; (ii) sum skinfold as dichotomous variable, sitting height as dichotomous variable and lean body mass as dichotomous variable; (iii) sum skinfold as dichotomous variable, sitting height as 4-level categorical variable and lean body mass as dichotomous variable.

### 2.3 Conclusions

Results in table 2 are very similar for all models. The three parameterizations don't differ in goodness-of-fit measures. The choice between parameterizations should be based on the interpretability of their estimated coefficients, namely by the odds ratio. Results not show that modelling sitting height, sum skinfold and lean body mass as categorical variables provides a model that is better than one treating sitting height, sum skinfold and lean body mass as continuous and linear in the logit.

### References

- Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logist Regression (2nd edition)*. New York: John Wiley and Sons.
- Bayios, I.A., Bergeles, N.K., Apostolidis, N.G., Noutsos, K.S., and Koskolou, M.D. (2006). Anthropometric, body composition and somatotype differences of Greek elite female basketball, volleyball and handball players. *The Journal of sports medicine and physical fitness*, 46 (2), 271-280.
- Fragoso, I., and Vieira, F (2005). *Cin antropometria*. Cruz Quebrada: FMH-Serviço de Edições.

# On the Bayesian 2-stage procedure for parameter estimation in copula models

Silvia Cecere<sup>1</sup> and Emmanuel Lesaffre<sup>1,2</sup>

<sup>1</sup> Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, 3000 Leuven, Belgium

<sup>2</sup> Department of Epidemiology and Biostatistics, P.O. Box 2040 3000 CA Rotterdam, The Netherlands

**Abstract:** A copula model offers a flexible way to handle bivariate data when the marginals are non-Gaussian. In this paper we explore the Bayesian 2-stage procedure proposed by Romeo et al. (2006). We show that the results of the 2-stage procedure can be misleading both with respect to the estimated parameters as their variability. Therefore, we suggest to adapt the basic 2-stage approach by parallel sampling the marginal and the copula parameters in a first step and in a second step to adjust the parameter estimates obtained from the MCMC output by a reweighting procedure. These 2-stage procedures are applied to simulated and real data.

**Keywords:** Bivariate data; copula models; Bayesian analysis; Markov chain Monte Carlo; importance-sampling

## 1 Introduction

Copula models allow one to model separately the marginals from the association structure. Romeo et al. (2006) suggested a 2-stage Bayesian estimation procedure. In a first stage, the marginal parameters are estimated and in a second stage, the copula parameters are estimated whereby the estimated marginal parameters are plugged-in. This approach, however, underestimates the variability with which the parameters are estimated. We suggest to improve the above 2-stage procedure in two ways using parallel sampling and importance sampling (Rubin (1987)) in combination with MCMC techniques. We study the performance of our procedure in simulated data generated from the Gaussian and Clayton copula with different (semi-)parametric marginals. The methodology is illustrated with the Signal Tandmobiel<sup>®</sup> data to study the association structure of emergence times of permanent teeth. In this paper we will restrict our attention to the bivariate case.

## 2 Copulas Models

### 2.1 Introduction

The bivariate function  $C = C(u_1, u_2)$  is called a copula if it is a continuous distribution function on  $[0, 1]^2$  with uniform marginals. The copula framework can be used to construct bivariate distributions with given marginals. In fact, given  $F_1, F_2$  univariate cumulative distributions functions (cdf) and a copula  $C$ , the distribution  $F$  defined as  $F(y_1, y_2) = C(F_1(y_1), F_2(y_2))$ , has marginals  $F_1, F_2$ . If the densities of the marginals are  $f_1, f_2$ , (pdf) and  $c$  is the density of the copula  $C$  then the density of  $F$  is given by

$$f(y_1, y_2) = c(F_1(y_1), F_2(y_2))f_1(y_1)f_2(y_2). \quad (1)$$

Here, we will consider the Gaussian copula  $C_R$  and the Clayton copula  $C_\alpha$ . The Gaussian copula is defined by

$$C_R(u_1, u_2) = \Phi_R(\Phi^{-1}(u_1), \Phi^{-1}(u_2)), \quad (2)$$

where  $R$  is a  $2 \times 2$  correlation matrix with non-diagonal element  $\rho$ ,  $\Phi$  is the cumulative distribution function of the standard normal and  $\Phi_R$  is the cumulative distribution function of the bivariate normal distribution with mean vector zero and variance-covariance matrix  $R$ .

The Clayton copula is given by

$$C_\alpha(u, v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-\frac{1}{\alpha}}, \quad (3)$$

with  $\alpha \in (0, \infty)$ .

Copula-based models can be extended in various ways, for example, the marginals and/or the copula can depend on covariates. Such extensions does not pose theoretical difficulties but may complicate the estimation procedure considerably.

### 2.2 Bayesian copula models

Suppose there are  $n$  independent observations  $\mathbf{y}_i = (y_{i1}, y_{i2})'$ ,  $i = 1, \dots, n$ , coming from a bivariate distribution  $Y = (Y_1, Y_2)$  specified by a copula  $C_\alpha$  and its marginals with cdf's  $F_1(\cdot; \boldsymbol{\theta}_1), F_2(\cdot; \boldsymbol{\theta}_2)$  (and pdf's  $f_1(\cdot; \boldsymbol{\theta}_1), f_2(\cdot; \boldsymbol{\theta}_2)$ ) respectively with  $\boldsymbol{\theta}_j$  the total parameter vector associated with the  $j$ th-marginal. Let  $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2)'$  and  $\mathbf{Y} = \{\mathbf{y}_i\}_{1 \leq i \leq n}$ , then according to (1), the likelihood is given by

$$L(\mathbf{Y}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{i=1}^n c_\alpha(F_1(y_{i1}; \boldsymbol{\theta}_1), F_2(y_{i2}; \boldsymbol{\theta}_2))f_1(y_{i1}; \boldsymbol{\theta}_1)f_2(y_{i2}; \boldsymbol{\theta}_2). \quad (4)$$

Here we will use a Bayesian approach to estimate the parameters. Hence priors for all parameters are needed. The marginal and copula parameter

vectors are assumed to be a priori independent, so considering priors  $\pi_{\theta_j}(\cdot)$  and  $\pi_{\alpha}(\cdot)$  respectively, the joint posterior distribution is

$$p(\alpha, \theta | Y) \propto L(Y | \alpha, \theta) \times \pi_{\alpha}(\alpha) \times \pi_{\theta_1}(\theta_1) \times \pi_{\theta_2}(\theta_2). \quad (5)$$

### 3 Estimation procedures

In general, sampling from a copula model is not straightforward and things become quite complicated when the marginals are non-standard. We consider three approaches to sample from (5), namely: Basic 2-stage, Parallel 2-stage and Reweighted 2-stage.

#### 3.1 Basic 2-stage procedure

Since marginal distributions can be chosen independently from the association part in a copula model, it makes sense to consider the estimation of the marginals and the dependence parameters separately leading to the basic 2-stage procedure proposed by Romeo et al. (2006). The first stage consists of estimating the marginal parameters separately from

$$p(\theta_j | y_{1j}, \dots, y_{nj}) \propto \prod_{i=1}^n f_j(y_{ij}; \theta_j) \times \pi_{\theta_j}(\theta_j), \quad j = 1, 2. \quad (6)$$

In the second-stage the posterior means  $\hat{\theta}_j$  are plugged-in in the copula part of expression (5) yielding the pseudo-posterior distribution of  $\alpha$  given by:

$$p(\alpha | Y, \hat{\theta}_1, \hat{\theta}_2) \propto \prod_{i=1}^n c_{\alpha}(F_1(y_{i1}, \hat{\theta}_1), F_2(y_{i2}, \hat{\theta}_2)) \times \pi_{\alpha}(\alpha). \quad (7)$$

Posterior summary statistics for  $\alpha$  are obtained using an appropriate MCMC algorithm.

However, there are at least 2 potential problems with this procedure. Firstly, the posterior estimates of the parameters are not obtained from the true posterior distribution but from an approximation, whereby it is assumed that the marginal parameters are independent of the copula part. This could lead to inefficient parameter estimates. Secondly, the posterior variability of the parameter estimates obtained by this procedure underestimates the true variability.

#### 3.2 Parallel 2-stage procedure

To address the second problem we propose to sample the marginal parameters in parallel with the copula parameters as follows. While we are sampling from (6) at each iteration  $t$  of the samplers, we impute the current values  $\theta_j^t$  in the copula part of (5) so that  $\alpha^t$  will be sampled from (7) replacing  $\hat{\theta}_j$ ,  $j = 1, 2$  by  $\theta_j^t$ ,  $j = 1, 2$  respectively.

### 3.3 Importance Sampling: reweighting

The above procedures do not sample from the correct joint posterior, but rather from an approximation. A posterior sample of  $\boldsymbol{\alpha}$ , say  $\{\boldsymbol{\alpha}^t\}_{t=1}^T$  obtained by the parallel 2-stage procedure can be reweighted using the importance weights,  $w$ , to yield an approximate sample from the joint posterior given in (5). The weights  $w$  are given by the quotient between the target (5) and the approximation,  $p_{\text{approx}}(\boldsymbol{\alpha}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{Y})$ , where  $p_{\text{approx}}(\cdot, \cdot, \cdot | \mathbf{Y})$  is proportional to (6) multiplied by  $p(\boldsymbol{\alpha} | \boldsymbol{\theta}, \mathbf{Y})$ . After some algebra, it can be seen that the weights,  $w$ , satisfy

$$w \propto \int_{\Omega} \prod_{i=1}^n c_{\boldsymbol{\alpha}}(F_1(y_{i1}; \boldsymbol{\theta}_1), F_2(y_{i2}; \boldsymbol{\theta}_2)) \pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) d\boldsymbol{\alpha}. \quad (8)$$

where  $\Omega$  is the parameter space of  $\boldsymbol{\alpha}$ .

Thus, at each iteration  $t$  of the 2-stage procedure sampler, we calculate the weights  $w^t$  according to expression (8). Given the (MCMC-) sequences  $\{\boldsymbol{\alpha}^t\}$ , and  $\{w^t\}$ , we calculate the posterior mean of  $h(\boldsymbol{\alpha})$  as the reweighted average according to the following expression,

$$\widehat{h(\boldsymbol{\alpha})}_{\text{rw}} = \frac{\sum_t h(\boldsymbol{\alpha}^t) w^t}{\sum_t w^t}. \quad (9)$$

Following Rubin (1987), the importance weights can be used to get a sequence of draws that approximate the target distribution by the method of importance re-sampling (also called sampling-importance re-sampling, SIR).

## 4 Application

### 4.1 Simulations and applications

A small simulation study illustrates the performance of the 2-stage procedures in the following settings: (a) Normal Margins-Gaussian Copula ( $\rho = 0.2, 0.6$ , sample size = 100, 500), (b) Normal Margins-Clayton Copula ( $\alpha = 0.5, 8$ , sample size = 100, 500), (c) Mixture of two log-normals, (complete/ interval-censored data-Gaussian Copula ( $\rho = 0.5, 0.8$ , sample size = 100). In setting (c), the marginals were estimated using an AFT model with flexible distributional assumptions as proposed by Komárek et al. (2008). We consider a  $U[-1, 1]$  prior for the dependence parameter  $\rho$  in the settings with a Gaussian copula and in the Clayton copula settings we consider a  $\mathcal{G}(1, 100)$  for the dependence parameter  $\alpha$ . We will present the results only for the dependence parameter.

As an application, we estimated the association between emergence times of permanent teeth based on the Signal Tandmobiel<sup>®</sup> data, a 6 year longitudinal oral health study started in Flanders (Belgium) in 1996 in which

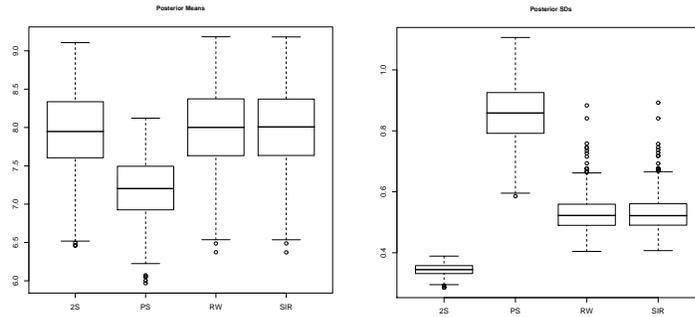


FIGURE 1. Normal Margins-Clayton Copula- $\alpha = 8$ . 2S: basic 2-stage, PS: parallel sampler, RW: reweighted posterior mean, SIR: resampling estimate

4468 children were annually examined on pre-scheduled visits by sixteen trained dental examiners in a mobile dental clinic. At each visit, the tooth emergence status was recorded. From a dental point of view, it was of interest to study the association structure of emergence times of permanent teeth. Due to the study design, the emergence times were recorded in an interval-censored manner. The copula framework allows us to model the marginals in a flexible way using the AFT models with flexible distribution assumptions introduced by Komárek et al. (2008) We use a Gaussian copula to handle the dependence.

## 4.2 Results

Figure 1 shows the posterior means and the posterior SDs for each data set with normal margins combined with the Clayton copula. While the point estimate of the dependence parameter obtained by the basic 2-stage (2S) has the smallest bias, see Table 1, its posterior SDs are underestimated. For the Signal Tandmobiel<sup>®</sup> study, we estimated the correlation coefficient of the emergence times for pairs of teeth (results not shown) based on copula models combined with flexible marginals. These correlation coefficients measure the dependence of the emergence times based on a monotone transformation, i.e.,  $\rho = \text{corr}(\Phi^{-1}(F_1(Y_1)), \Phi^{-1}(F_2(Y_2)))$ , where  $F_j(\cdot)$  is the cdf of the  $j$ th-margin corresponding to the Bayesian penalized AFT model. We envisage to extend the methodology to higher dimensions, so that we can model jointly the emergence times of permanent teeth.

**Acknowledgments:** The authors have been partly supported by the Research Grants OT/00/35 and OT/05/60, Catholic University Leuven. The

TABLE 1. Bayesian Penalized mixture AFT Margins-Normal copula-Interval-Censored  $N = 100$ - $\rho = 0.5$ 

Sampler	$\hat{\rho}$	SD	Bias	MSE
2S	0.4962	0.0796	0.0628	0.0103
PS	0.4429	0.0859	-0.0571	0.0106
RW	0.5644	0.0797	0.0644	0.0105
SIR	0.5644	0.0797	0.0644	0.0105

authors also acknowledge the partial support from the Interuniversity Attraction Poles Program P5/24 – Belgian State – Federal Office for Scientific, Technical and Cultural Affairs. Data collection was supported by Unilever, Belgium. The Signal Tandmobiel<sup>®</sup> study comprises following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

## References

- Cecere, S., Jara, A., Lesaffre, E. (2006). Analyzing the emergence times of permanent teeth: an example of modelling the covariance matrix with interval-censored data. *Statistical Modelling*, **6**, 352-372. appear.
- Komárek, A., Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly-interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association*. To appear.
- Romeo, J., Tanaka, N., Pedroso-de-Lima, A. (2006). Bivariate survival modeling: a Bayesian approach based on copulas. *Lifetime Data Analysis*, **12**, 205-222
- Rubin, D. (1987). The Calculation of Posterior Distributions by Data Augmentation: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, **82**, 398, 543-546

# Simulation-based results for bioequivalence studies using the 2x2 crossover design

Paula Rocha Chellini<sup>1</sup> and Arminda Lucia Siqueira<sup>2</sup>

<sup>1</sup> Centro de Pesquisa em Biotecnologia Ltda. (CEBIO), Belo Horizonte, MG, Brazil. E-mail: paulachellini@cebio.med.br

<sup>2</sup> Departamento de Estatística, ICEX, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil. E-mail: arminda@est.ufmg.br

**Abstract:** A simulation study was performed in order to discuss practical issues related to the planning and analysis of data from a bioequivalence study using the 2x2 crossover design. This paper focuses on the conclusion of bioequivalence considering problematic conditions, such as presence of period, formulation and residual effects, high variability and uncertainty in variability estimation.

**Keywords:** Bioequivalence study; 2x2 crossover design; simulation study.

## 1 Introduction

Bioequivalence studies aim at demonstrating that two drugs of the same active ingredient have similar bioavailability, that is, a similar rate and extent of drug absorption. The 2x2 crossover design is often used, with the administration of two formulations ( $T = \text{test}$  and  $R = \text{reference}$ ) to healthy volunteers. Following a pre-established schedule, blood samples are collected before and after drug administration, and then the profile of the blood concentration-time curve is studied by means of two pharmacokinetic parameters: the area under the blood concentration-time curve ( $AUC$ ) and maximum concentration ( $C_{max}$ ). Formulation  $T$  is said to be bioequivalent to  $R$  if, for both  $AUC$  and  $C_{max}$ , the 90% confidence interval for the ratio of the means lies within the interval (0.80, 1.25), or equivalently, the 90% confidence interval for the difference of means on the natural log scale is within the interval (-0.2231, 0.2231).

The model commonly used to describe pharmacokinetic data ( $Y$ ), which generally follows the log-normal distribution, with the assumption of no carry-over effect, obtained through a study conducted under the 2x2 crossover design is  $Y_{ijk} = \mu + S_{ik} + P_j + F_{(j,k)} + e_{ijk}$ ,  $i = 1, 2, \dots, n_k$ ;  $j = 1, 2$ ;  $k = 1, 2$ , respectively for subject, period and sequence. In the model, involving fixed and random terms,  $\mu$  represents the overall mean,  $S_{ik}$  is the random effect related to subject,  $P_j$  and  $F_{(j,k)}$  are the fixed period and formulation effects, with the constraint that  $\sum F_{(j,k)} = 0$ , and the last term  $e_{ijk}$  is the intrasubject random error related to the outcome  $Y_{ijk}$ . The usual

assumptions are: (i)  $\{S_{ik}\} \sim N(0, \sigma_s^2)$ ; (ii)  $\{e_{ijk}\} \sim N(0, \sigma_e^2)$ ; (iii)  $\{S_{ik}\}$  and  $\{e_{ijk}\}$  are mutually independent. Further details for bioequivalence studies can be found for instance in Chow and Liu (2000) and Patterson and Jones (2006), and on the Regulatory Agencies web sites.

This paper investigates some factors that can influence the bioequivalence conclusion, such as inaccurate sample size. Moreover, we evaluated the influence of the presence of residual, formulation and period effects on the conclusion of bioequivalence, as well as the consequences of a high inter-subject and/or intrasubject variability.

## 2 A simulation study

Using the model given in Section 1, the pharmacokinetics measure  $AUC$  was generated according to a log-normal distribution with  $\mu = 4.37$ ,  $\sigma_e = 0.19$  (corresponding to  $CV_d = 13.50\%$ ) and  $\sigma_s = 0.20$ . The parameters were chosen using data from Chow and Liu (2000), page 73. Several sample sizes ( $2n$ ) were considered. Residual, formulation and period effects were tested, the 90% confidence interval for the difference between the two means ( $T$  and  $R$ ) was built, and the Schuirmann's two one-sided tests were performed the significance level of 5%. The computational implementation was done in C language with 10,000 repetitions in all cases. Selected results are summarized next.

### 2.1 Conclusion of bioequivalence

Residual, formulation and period effects come up in approximately 5% of the tests for all  $2n$  values, which was already expected since this is the significance level of the test. To test the influence of the period effect on the conclusion of bioequivalence, the parameter  $P$  was set from 0.00 to 0.30 (with increment of 0.01), without modifying the other parameters. The same was done with  $F$  to test the formulation effect on the bioequivalence conclusion. The conclusion of bioequivalence are exactly the same with or without period effect, i.e., the presence of the period effect does not affect the bioequivalence conclusion (results not shown). However, the presence of formulation effect alters the conclusion of bioequivalence, i.e., there is a great reduction in the percentage.

Table 1 shows selected results for several values of intrasubject variation  $\sigma_e$  (0.19 to 0.65 with increase of 0.01) and the intersubject variation  $\sigma_s$  (0.20 to 1.00 with an increase of 0.10). In general, an increase in  $\sigma_e$  due to inadequate planning may undermine the study. In other words, a larger number of volunteers or higher order crossover designs are required for high variability conditions. Even for high values of  $\sigma_s$ , the bioequivalence conclusion did not alter very much. This can be explained by the fact that each individual is control of himself/herself and then the assessment of bioequivalence is based on the intrasubject variability.

Some combinations of period and formulation effects for various intrasubject and intersubject values were analyzed. Even in the presence of period and/or formulation effects, an increase in the intersubject variation does not change the bioequivalence conclusion. The same can be said for the period effect in the presence of a high intrasubject variability. However, the formulation effect-higher intraindividual variability combination leads to a higher drop in the bioequivalence conclusion.

TABLE 1. Percentage of conclusion of bioequivalence (BE) and presence of the formulation effect (Ef)

$2n$	$F = 0,00$		$F = 0,05$		$F = 0,00$		$F = 0,05$	
	$\sigma_e = 0,19$		$\sigma_e = 0,19$		$\sigma_e = 0,24$		$\sigma_e = 0,24$	
	BE	Ef	BE	Ef	BE	Ef	BE	Ef
12	69.83	4.79	67.37	9.16	37.27	4.79	34.69	7.38
14	80.12	4.87	71.10	9.56	50.51	4.87	45.43	8.06
16	86.75	4.70	78.18	10.17	60.86	4.70	54.23	8.26
18	91.49	4.72	82.93	11.42	69.73	4.72	61.66	8.87
20	94.73	4.91	86.80	12.17	76.17	4.91	67.75	9.60
22	96.73	4.73	89.58	13.41	81.90	4.73	72.85	9.96
24	97.86	5.01	91.95	14.29	86.07	5.01	76.79	10.77
26	98.56	4.79	94.16	14.43	89.44	4.79	80.94	10.74
28	99.16	5.32	95.05	16.41	91.97	5.32	82.48	12.15
30	99.48	5.10	96.11	16.94	93.69	5.10	85.60	12.41

$2n$  = sample size (16 is need to achieve 80% of power)

$F$ : difference of formulation effect

$\sigma_e = 0,19 \Leftrightarrow CV_d = 13,50\%$ ,  $\sigma_e = 0,24 \Leftrightarrow CV_d = 17,09\%$

The presence of formulation effect does not invalidate the study, because the test of equal averages of the two formulations is more rigorous than the test used in the assessment of bioequivalence, which allows the average difference lying within the bioequivalence interval.

## 2.2 Uncertainty of estimating variability

To investigate the impact of misspecifying the coefficient of variation ( $CV$ ) on the bioequivalence conclusion, for each value of  $CV$  taken as real we calculated how many times it is concluded by bioequivalence when it is incorrectly estimated, that the value of  $CV$  is less or greater than the true one. The simulation included several  $CV$  values and three values for the difference between means ( $\theta_\gamma$ ): 0.00, 0.05 and 0.10.

Table 2 presents the results for selected values of  $CV$ , for the case in which  $\theta_\gamma = 0.00$  and power is set at 80%. The higher the  $CV$  used, the higher the percentage of bioequivalence conclusion. For example, if the true  $CV$  is 0.18 and the value used to calculate the sample size was 0.12, the number of participants will be less than necessary, and the risk of not getting an outcome favourable to bioequivalence is increased. Finally, while a conservative position of using  $CV$  equal to 0.50 can ensure 100% of bioequivalence conclusion, but this greatly increases the cost and endangers the feasibility of the study.

TABLE 2. Percentage of bioequivalence for misspecified  $CV$ 

$CV$ (true)	$2n$	$CV$ used on sample size calculation							
		0,12	0,15	0,18	0,20	0,25	0,30	0,40	0,50
0,12	12	82,68	96,77	99,40	99,91	100,00	100,00	100,00	100,00
0,15	18	56,33	83,30	94,40	98,07	99,84	100,00	100,00	100,00
0,18	24	32,05	62,41	81,94	90,78	98,54	99,85	100,00	100,00
0,20	30	19,56	48,92	70,39	82,41	95,88	99,46	100,00	100,00
0,25	44	6,69	19,73	39,91	56,25	81,80	94,02	99,61	99,99
0,30	62	1,73	5,76	16,10	30,91	59,28	81,02	97,25	99,72
0,40	104	0,19	0,43	1,23	4,22	20,24	46,05	80,24	94,66
0,50	156	0,03	0,02	0,07	0,20	2,58	16,68	54,42	79,50

 $2n$  = sample size

### 3 Conclusion

Period effect does not affect the bioequivalence conclusion and the residual effect can be avoided with an adequate washout period. The presence of formulation effect does not invalidate the study but there is a damaging effect on the conclusion of bioequivalence.

Whenever possible, a larger sample than the one calculated for the drug's  $CV$  would be recommended in cases of uncertainty of the variability in bioequivalence studies, as smaller than necessary values may seriously undermine the study. However, taking a very conservative position can cause unnecessary increase in the number of participants which may be unfeasible due to cost and practical reasons. Therefore, with careful planning and appropriated conduction at the analysis, if the drugs are indeed bioequivalent, a favorable outcome is expected, that is, that the formulations ( $T$  and  $R$ ) will be declared bioequivalent.

**Acknowledgments:** This work was partly sponsored by FAPEMIG and CNPq, Brazilian Research Committees.

### References

- Chow, S.C., and Liu, J.P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies*. Marcel Dekker.
- Julious, S.A. (2004). Designing clinical trials with uncertain estimates of variability. *Pharmaceutical Statistics*, **23(2)**, 261-268.
- Patterson, S., and Jones, B. (2006). *Bioequivalence and Statistics in Clinical Pharmacology*. London: Chapman & Hall.
- Siqueira, A. L., Whitehead, A., Todd, S., and Lucini, M. M. (2005). Comparison of sample size formulae for 2x2 cross-over designs applied to bioequivalence studies. *Pharmaceutical Statistics*, **4(4)**, 233-243.

# The Reliability of Type II Censored Reliability Analyses for Weibull Data

S. J. Chua<sup>1</sup> and A. J. Watkins<sup>1</sup>

<sup>1</sup> School of Business and Economics, Swansea University, Singleton Park, Swansea SA2 8PP, United Kingdom. E-mail: a.watkins@swansea.ac.uk

**Abstract:** This paper considers the analysis of reliability data drawn from a Weibull distribution and subject to Type II censoring. We aim to determine the smallest number of failures at which the experiment can be reasonably or safely terminated with the interim analysis still providing a close and reliable guide to the final analysis. We present asymptotic results on the correlation between two estimates of percentiles; this, in turn, yields 95% confidence limits for the final estimate given the interim estimate. We illustrate our results using published data set, and also present results from simulation experiments, indicating the extent to which the asymptotic results apply in samples of finite size.

**Keywords:** Correlation; Interim analysis; Percentile estimation; Type II censoring; Weibull.

## 1 Introduction and Background

From a statistical viewpoint, the analysis of the complete data set is to be preferred, but, in practice, some censoring - such as Type I or Type II - is often inevitable; some systems are expensive to test; some failures may take years to observe; and some experiments may be hazardous to run for prolonged periods. In practice, an experimenter may wish to know the smallest number of failures at which the experiment can be reasonably or safely terminated but with the interim analysis still providing a close and reliable guide to the analysis of complete data. This paper gives some insight into the roles of censoring number  $r$  and sample size  $n$  in a Type II censoring setting.

It is often relevant in practical applications to make inferences on either the running time for the experiment or some percentile of lifetimes, since time is often directly linked to costs; for example, estimating the 10<sup>th</sup> percentile of failure times. Throughout, we assume that the data follows the Weibull distribution, with percentile function ( $0 < p < 1$ )

$$B_p = \theta \{-\ln(1-p)\}^{\frac{1}{\beta}}, \quad (1)$$

where  $\theta > 0$  and  $\beta > 0$  are, respectively, scale and shape parameters. We illustrate this experimental set-up using this distribution to model the

TABLE 1. Maximum likelihood estimates of Weibull parameters and 10<sup>th</sup> percentile calculated for various  $r$  for the ball-bearings data.

$r$	8	12	16	20	23
$\hat{\theta}_r$	67.6415	75.2168	76.6960	78.9674	81.8783
$\hat{\beta}_r$	3.2280	2.6241	2.4695	2.3539	2.1021
$\hat{B}_{0.1,r}$	33.6860	31.9063	30.8329	30.3563	28.0694

classic ball-bearings data (Kalbfleisch, 1979). Table 1 shows the maximum likelihood estimates calculated for various  $r$ , and we note that  $\hat{B}_{0.1,r} = \hat{\theta}_r (-\ln 0.9)^{\frac{1}{\beta_r}}$  converges to its complete counterpart as  $r \rightarrow n = 23$ .

We can now assess the difference in censoring at  $r = 8$  and  $r = 16$ . For  $r = 8$ , testing stops after 51.84 million revolutions, while, with  $r = 16$ , we need to wait roughly 30 million revolutions longer. We can also assess the effect of the final few failures by taking  $r = 20$ , when we intuitively expect estimates to be more consistent with final values than with  $r = 8$  or 16. More generally, we consider the precision with which we can make statements on final estimates, based on interim estimates. This approach requires the study of the relationship between  $\hat{B}_{0.1}$  ( $\equiv \hat{B}_{0.1,n}$ ) and  $\hat{B}_{0.1,r}$ , and the extent to which  $\hat{B}_{0.1,r}$  can be regarded as a reliable guide to  $\hat{B}_{0.1}$ . Since  $B_{0.1}$  is a non-linear function of model parameters, we consider a first order Taylor series expansion of (1) (for which the asymptotic mean and variance can be computed) to estimate  $B_{0.1}$ ; this can be written as

$$\hat{B}_{0.1,r} \simeq B_{0.1} + \mathbf{c}' \begin{pmatrix} \hat{\theta}_r - \theta \\ \hat{\beta}_r - \beta \end{pmatrix}, \quad (2)$$

where

$$\mathbf{c}' = \left( (-\ln 0.9)^{\frac{1}{\beta}}, -\theta\beta^{-2} (-\ln 0.9)^{\frac{1}{\beta}} \ln(-\ln 0.9) \right) \quad (3)$$

so that the asymptotic distribution for  $\hat{B}_{0.1,r}$  is  $N(B_{0.1}, \mathbf{c}' \mathbf{A}_r^{-1} \mathbf{c})$ , where  $\mathbf{A}_r$  is the Type II expected Fisher information matrix; see Watkins and John (2006).

## 2 Link between $\hat{B}_{0.1}$ and $\hat{B}_{0.1,r}$

We are interested in the agreement between  $\hat{B}_{0.1,r}$  and  $\hat{B}_{0.1}$ ; from (2), we see that this depends on the relationship between the two sets of maximum likelihood estimators of parameters. From Chua and Watkins (2007), we

TABLE 2.  $Corr(\widehat{B}_{0.1}, \widehat{B}_{0.1,r})$  calculated for various  $r, n$  using Weibull data generated with  $\theta = 100, \beta = 2$ .

$r$	$n$					
	25	50	100	1000	2500	5000
0.2n	.8451	.8511	.8539	.8564	.8565	.8566
	.8634	.8587	.8563	.8578	.8557	.8558
0.4n	.8692	.8668	.8654	.8639	.8638	.8638
	.8781	.8696	.8657	.8645	.8638	.8634
0.6n	.8961	.8930	.8913	.8897	.8896	.8896
	.8996	.8923	.8906	.8907	.8908	.8899
0.8n	.9356	.9330	.9317	.9304	.9304	.9303
	.9369	.9326	.9322	.9307	.9309	.9314

have, for large  $n$ ,

$$Corr(\widehat{B}_{0.1}, \widehat{B}_{0.1,r}) \simeq \sqrt{\frac{\mathbf{c}' \mathbf{A}_n^{-1} \mathbf{c}}{\mathbf{c}' \mathbf{A}_r^{-1} \mathbf{c}}}. \quad (4)$$

## 2.1 Some Numerical Results

We consider this result for finite samples. We take  $\theta = 100, \beta = 2$ , and, for combinations of  $r$  and  $n$ , replicate  $10^4$  sets of Weibull data; these yield  $10^4$  values from the sampling distribution of  $(\widehat{\theta}_r, \widehat{\beta}_r)$ , and hence  $\widehat{B}_{0.1,r}$ . Table 2 summarises the theoretical (upper) and observed (lower) values of  $Corr(\widehat{B}_{0.1}, \widehat{B}_{0.1,r})$  for these  $10^4$  estimates, and we observe good agreement between theory and practice for varying  $r, n$ .

Some indication of the precision with which we can make statements on final estimate, given the interim estimate, may also be useful. In particular, we can compute the 95% confidence limits for  $\widehat{B}_{0.1}$  based on  $\widehat{B}_{0.1,r}$ . The asymptotic Normality of maximum likelihood estimators implies that, for large samples,  $\widehat{B}_{0.1} - \widehat{B}_{0.1,r}$  is also asymptotically Normal, with zero mean and variance involving  $Var(\widehat{\theta} - \widehat{\theta}_r)$ ,  $Var(\widehat{\beta} - \widehat{\beta}_r)$  and  $Cov(\widehat{\theta} - \widehat{\theta}_r, \widehat{\beta} - \widehat{\beta}_r)$ ; see Chua and Watkins (2007).

We can now write down approximate 95% confidence intervals for  $\widehat{B}_{0.1}$  given  $\widehat{B}_{0.1,r}$ , and Table 3 shows these limits for various  $r$  for the ball-bearings data. Again, we are interested in the extent to which these limits apply in finite samples; we expect, in our simulations, to find 95% of the  $10^4$  simulated values of  $\widehat{B}_{0.1}$  within these limits based on  $\widehat{B}_{0.1,r}$ . Again, Table 4 shows generally good agreement between theory and practice.

TABLE 3. Standard deviation of  $\widehat{B}_{0.1} - \widehat{B}_{0.1,r}$  for the ball-bearings data.

$r$	8	12	16	20	23
$\widehat{B}_{0.1,r}$	33.6860	31.9063	30.8329	30.3563	28.0694
$sd(\widehat{B}_{0.1} - \widehat{B}_{0.1,r})$	2.9231	3.0910	2.6749	1.9673	0

TABLE 4. Number of  $\widehat{B}_{0.1}$  within the 95% confidence limits for  $10^4$  sets of Weibull data generated with  $\theta = 100, \beta = 2$ .

$r$	$n$					
	25	50	100	1000	2500	5000
$0.2n$	9613	9543	9538	9498	9459	9479
$0.3n$	9317	9417	9531	9475	9510	9497
$0.4n$	9483	9476	9505	9475	9456	9463
$0.5n$	9437	9550	9501	9505	9499	9524
$0.6n$	9444	9463	9488	9512	9471	9457
$0.7n$	9413	9492	9475	9501	9508	9514
$0.8n$	9442	9510	9524	9487	9492	9488
$0.9n$	9385	9471	9482	9493	9533	9512

### 3 Practical Implications

Table 3 shows that these limits increase then decrease to 0 as  $r \rightarrow n$ . In addition, the last three failures have a considerable effect on final estimates. Consequently, ultimate convergence relies heavily on the last few failures. We also note that the precision levels associated with  $\widehat{B}_{0.1,8}$  and  $\widehat{B}_{0.1,16}$  are quite similar, which provides partial answers to questions posed above. In real life scenarios, censoring often leads to earlier termination of a life test; for a given tolerance level, an experimenter may, in practice, terminate the test sooner than might have been thought.

### 4 Conclusion

We have obtained an expression for the correlation between two maximum likelihood estimators of a particular Weibull percentile. This, in turn, yields approximate 95% confidence limits for the final estimate given interim estimate. We have also shown that these limits agree with behaviour observed in simulation experiments for various combinations of  $r, n$ . Overall, results are encouraging, suggesting that, for a specified level of precision, it may be possible to design experiments in which early stopping is a viable option.

### References

- Chua, S.J. and Watkins, A.J. (2007). The reliability of type II censored reliability analyses. In: *Proceedings of the 5th International Mathematical Methods in Reliability Conference, Glasgow, Scotland*.
- Kalbfleisch, J.G. (1979). *Probability and Statistical Inference II*. New York: Springer-Verlag.
- Watkins, A.J. and John, A.M. (2006). On the expected Fisher information for the Weibull distribution with type II censored data. *International Journal of Pure and Applied Mathematics*, **26**, 93-106.

# Estimating Functional Principal Components using the Linear Mixed Effects Model.

Norma Coffey<sup>1</sup>, Kevin Hayes<sup>1</sup>, Orna Donoghue<sup>2</sup> and Andrew J. Harrison<sup>2</sup>

<sup>1</sup> Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland

<sup>2</sup> Biomechanics Research Unit, University of Limerick, Limerick, Ireland

**Abstract:** We propose a new estimation procedure for functional principal components analysis based on the linear mixed effects model. We use penalized spline regression and the linear mixed effects model to smooth both the population mean function and the subject-specific deviations from this mean and estimate the eigenfunctions of the resulting covariance function. Simulation results exhibited reduced integrated squared bias values of the estimated functions. We have also applied our method to a biomechanical data set.

**Keywords:** Functional principal components; linear mixed effects model; P-splines; smoothing.

## 1 Methodology

In many situations, data arising from an experiment or observational study form a collection of  $N$  smooth functions  $y_i(t)$ ,  $i = 1, \dots, N$  and  $t \in \mathcal{T}$ . Ramsay and Dalzell (1991) called the analysis of such data *functional data analysis* (FDA) while Ramsay and Silverman (2002) and Ramsay and Silverman (2005) give a broad discussion on the method and its associated procedures. Many classical statistical methods (e.g. principal components analysis, regression analysis, etc. ) have been extended to the functional case. We aim to develop functional principal components analysis (FPCA) in the linear mixed effects (**lme**) model framework. We model the  $y_i(t)$  using the approach of Wu and Zhang (2007) as

$$y_i(t) = \eta(t) + \nu_i(t) + \varepsilon_i(t), \quad (1)$$

where  $\eta(t)$  models the population mean function,  $\nu_i(t)$  (the  $i$ th random effect function) models the deviation of the  $i$ th individual from the population mean function and  $\varepsilon_i(t)$  is an additive noise process. The  $\nu_i(t)$  are assumed to be i.i.d. realizations of an underlying smooth process  $\nu_i(t)$  with 0 mean and covariance function  $\gamma(s, t)$  and the  $\varepsilon_i(t)$  are i.i.d. realizations of an uncorrelated noise process  $\varepsilon(t)$  with variance  $\sigma^2(t)$ . The  $\nu_i(t)$  represent

variation around the mean and extracting the eigenfunctions of their associated covariance function is equivalent to finding the functional principal components of the data set. We estimate both  $\eta(t)$  and  $\nu_i(t)$  functions using P-splines (Eilers and Marx, 1996) and the connection between P-spline smoothing and the `lme` model. Equation (1) becomes

$$\mathbf{y}_i = \underbrace{\mathbf{X}\boldsymbol{\beta}_\eta + \mathbf{Z}\mathbf{u}_\eta}_{\eta(t)} + \underbrace{\mathbf{X}\boldsymbol{\beta}_{v_i} + \mathbf{Z}\mathbf{u}_{v_i}}_{v_i(t)} + \boldsymbol{\varepsilon}_i, \quad (2)$$

where  $\mathbf{u}_\eta \sim N(\mathbf{0}, \sigma_\eta^2 \mathbf{I})$ ,  $\mathbf{u}_{v_i} \sim N(\mathbf{0}, \sigma_{v_i}^2 \mathbf{I})$  and  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ . We then determine the functional principal components of the resulting covariance function. Representing the penalized smoothing problem using the `lme` model facilitates the inclusion of additional covariates, nesting/grouping structure and replicate measurements for each experimental unit. We aim to develop model (2) to incorporate such structure.

## 2 Results

### 2.1 Simulation Study

Data were simulated from the model  $y_i(t) = \eta(t) + \sum_{r=1}^{\infty} f_{ir}\xi_r(t) + \varepsilon_i(t)$ , where  $f_{ir}$  is the score of the  $i$ th individual on the  $r$ th principal component  $\xi_r(t)$ . The simulated curves had mean function  $\eta(t) = t + \sin(2\pi t)$  and two eigenfunctions  $-2^{\frac{1}{2}} \cos(\pi t)$  and  $2^{\frac{1}{2}} \sin(\pi t)$ ,  $t \in [0, 1]$  were specified, with corresponding eigenvalues  $\rho_r$ ,  $r = 1, 2$ . The remaining eigenvalues  $\rho_3, \dots$  were set to zero. The scores  $f_{ir}$  were generated from  $N(0, \rho_r)$  distributions, while uncorrelated measurement errors with variance  $\sigma^2(t)$  were generated from the  $N(0, 0.25)$  distribution. See Coffey *et al.* (2008) for further details. We present results for two cases that were investigated. Initially we set  $\rho_1 = 2$  and  $\rho_2 = 1$  implying that  $\xi_1(t) = -2^{\frac{1}{2}} \cos(\pi t)$  and  $\xi_2(t) = 2^{\frac{1}{2}} \sin(\pi t)$ . We then explored the effect of changing the order of the eigenfunctions by setting  $\rho_1 = 1$  and  $\rho_2 = 2$  so that now  $\xi_1(t) = 2^{\frac{1}{2}} \sin(\pi t)$  and  $\xi_2(t) = -2^{\frac{1}{2}} \cos(\pi t)$ . Table 1 displays IBIAS<sup>2</sup> values determined for each method.

TABLE 1. Simulation results.

$\rho_1$	$\rho_2$	FPC	FDA	LME
2	1	$\hat{\xi}_1(t)$	0.0350	0.1143
		$\hat{\xi}_2(t)$	0.5743	0.1302
1	2	$\hat{\xi}_1(t)$	0.9153	0.1853
		$\hat{\xi}_2(t)$	0.4156	0.1566

## 2.2 Biomechanical Data

We now present the results of applying our method to the biomechanical data set collected by Dr. Orna Donoghue and Dr. Andrew J. Harrison at the Biomechanics Research Unit, University of Limerick, Ireland. Thirteen subjects who displayed excessive pronation and had a history of Achilles tendon injury consented to participate in the study, as did 13 control subjects. Retro-reflective markers were placed on both lower extremities and these were used to define the angles described in Table 2. Eight Qualisys ProReflex MCU240 cameras obtained three-dimensional coordinates of the markers during treadmill running at self-selected speeds. The Achilles tendon group ran in two conditions, with and without customized orthoses. Control subjects did not wear orthoses. Angle-time series data for five footfalls for each subject and condition were calculated. Initially we focussed on individual angles of the Achilles tendon group without orthoses (i.e. AT(NO) group) and the control group. Only the first footfall for each subject was analyzed. The results presented here are for the ankle-dorsiflexion (ADF) angle. Coffey *et al.* (2008) provides further results for the Achilles tendon (EV) angle, while Donoghue *et al.* (2008) provides a standard FDA analysis of these data. Figure 1 displays the functional principal components of the control group, estimated using the standard FDA functions and the corresponding components estimated via the `lme` method. Figure 2 displays the functional principal components extracted for both the control group and the AT(NO) group. The individual covariance functions were determined by extending model (2) to include grouping structure, where  $\eta(t)$  now represents a group-specific mean function. The common functional principal components (Flury, 1984) determined for both groups simultaneously using the FG algorithm outlined in Flury and Constantine (1985) and Flury and Gautschi (1986) are also displayed.

## 3 Discussion

We have shown that the `lme` model framework can be used to estimate functional principal components. We used this model to simultaneously smooth the data and determine the functional principal components of the resulting covariance function. The simulation results presented in Table 1 shows reduced IBIAS<sup>2</sup> in almost all cases for the `lme` method. Finally, it can be seen from Figure 1 that both methods are identifying the same main sources of variation in the data. However, the `lme` method achieves these results without pre-processing of the data, automatic selection of smoothing parameters and no additional smoothing of the estimated functional principal components. Our method also has the capacity to incorporate grouping structure (as shown by the results in Figure 2), replicate footfalls for each subject and other levels of nesting evident in this data set.

TABLE 2. Angles measured.

Angle	Description
Achilles tendon (EV) angle	In/eversion position of rearfoot relative to the lower leg.
Leg abduction angle (ABD) angle	Angle between the lower leg and the ground on the inside as viewed from posterior.
Calcaneal angle	Angle between the rearfoot and the ground on the inside as viewed from posterior.
Knee flexion (KF) angle	Anatomical joint angle between the greater trochanter, fibular head and ankle.
Ankle dorsiflexion (ADF) angle	Anatomical joint angle between the fibular head, lateral malleolus and fifth metatarsal.

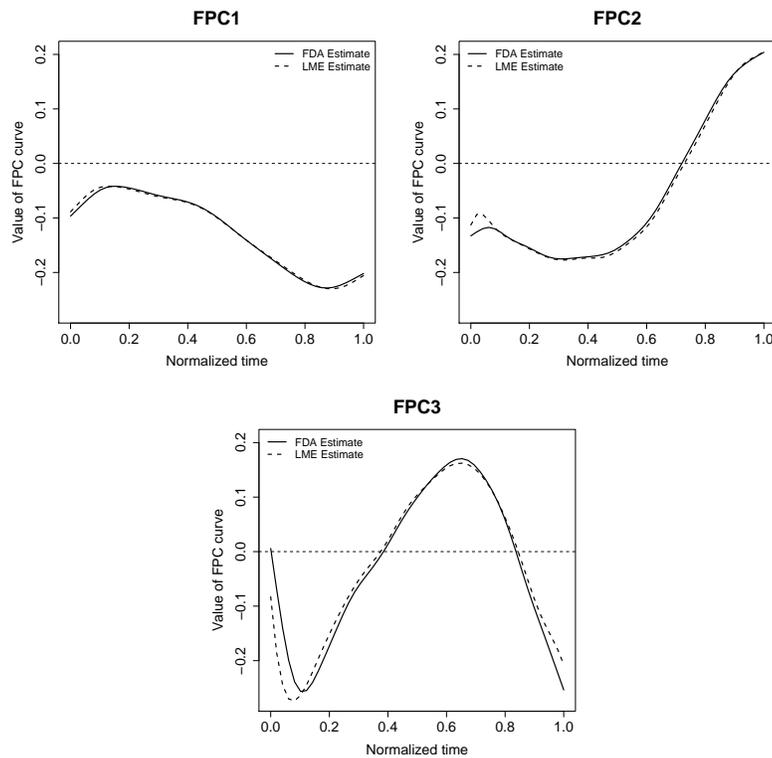


FIGURE 1. Estimates of FPC1-FPC3 calculated for the control group via the FDA method and the lme method.

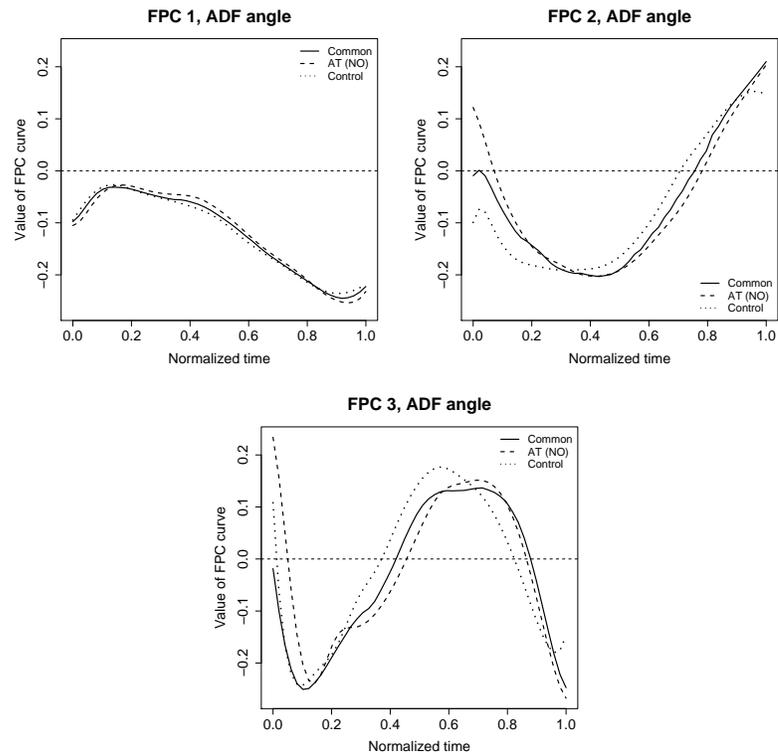


FIGURE 2. Estimates of common FPC1-FPC3 and corresponding estimates determined for each group individually.

**Acknowledgments:** Special Thanks to the Irish Research Council for Science, Engineering and Technology (IRCSET) who provided the funding for this work.

**References**

Coffey, N. and Hayes, K. and Donoghue, O. and Harrison, A.J. (2008) Functional Data Analysis via the Linear Mixed Effects Model. *Statistical Modelling*, Submitted.

Donoghue, O. and Harrison, A. J. and Coffey, N. and Hayes, K. (2008) Functional Data Analysis of Running Kinematics in subjects with Achilles Tendon Injury. *Medicine and Science in Sports and Exercise*, Accepted.

- Eilers, P.H.C. and Marx, B.D. (1996) Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11**, 89-121.
- Flury, B.N. (1984) Common Principal Components in k Groups. *Journal of the American Statistical Association*, **79**, 892-898.
- Flury, B.N. and Constantine, G. (1985) The F-G Diagonalization Algorithm. *Applied Statistics*, **34**, 177-183.
- Flury, B.N. and Gautschi, G. (1986) An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form. *SIAM Journal on Scientific and Statistical Computing*, **7**, 169-184.
- Ramsay, J.O. and Dalzell, C.J. (1991) Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society, Series B*, **53**, 539-572.
- Ramsay, J.O. and Silverman B.W. (2002) *Applied Functional Data Analysis*. New York: Springer.
- Ramsay, J.O. and Silverman B.W. (2005) *Functional Data Analysis*. New York: Springer.
- Wu, H. and Zhang, J.T. (2007) *Nonparametric Regression Methods for Longitudinal Data Analysis*. USA: Wiley.

# Search Algorithms for Log-Linear Models in Contingency Tables. Comorbidity Data

Susana Conde<sup>1</sup> and Gilbert MacKenzie<sup>1</sup>

<sup>1</sup> Centre of Biostatistics, Department of Mathematics and Statistics, University of Limerick, Ireland, [susana.conde@ul.ie](mailto:susana.conde@ul.ie) & [gilbert.mackenzie@ul.ie](mailto:gilbert.mackenzie@ul.ie)

**Abstract:** We have developed automatic search algorithms in R for finding optimal hierarchical log-linear models (HLLMs) in multi-dimensional contingency tables. The performance of the algorithms was compared using two complementary simulation strategies. In most of the cases the algorithms identified the correct model. We describe the algorithms and present some results.

**Keywords:** comorbidity; hierarchical log-linear model; search algorithms; simulation methods; composition.

## 1 Introduction

We adopt a log-linear model when we analyse contingency tables arising from  $p$  binary comorbidities (Conde and MacKenzie, 2007), with  $p$  any fixed integer  $\geq 2$ . Consider the  $p$ -dimensional contingency table with exactly  $n = 2^p$  cells. Let  $y_j$  be the observed count in the  $j$ th cell,  $j = 1, \dots, n$ ; assume the cells are ordered lexicographically in Fortran major order. Consider also the bijective mapping  $j \mapsto (i_1, \dots, i_p)$  where each  $i_1, \dots, i_p$  takes the value 0 (absent) or 1 (present) (O’Flaherty and MacKenzie, 1982).

Our model is then:

$$E(Y_j) = \mu_j = \exp(a'_j \theta)$$

where  $Y_j$  is the random variable denoting the frequency in the  $j$ th cell;  $a'_j$  is the  $j$ th row of the  $(n \times n)$  saturated design matrix  $A$ ; and  $\theta$  is the vector  $(n \times 1)$  of unknown parameters measuring the influence of the effects.

Our aim is to develop new search algorithms in R that efficiently identify *best* fitting HLLMs (Goodman, 1971), especially in multi-dimensional contingency tables, where software is lacking. In this paper our focus is methods of simulation of multi-dimensional contingency tables.

## 2 BE and other search algorithms

A backwards elimination search algorithm (BE) which mimics existing procedures in SPSS has been constructed in R (Conde and MacKenzie,

2007). Let  $model$  be an HLLM and  $modelI[i]$  a vector of HLLMs, where  $i = 1, \dots, J$  and  $J$  is the dimension of the generating set of  $model$ . All HLLMs are characterized by their generating sets. Let  $\alpha$  be the level of significance used in the likelihood ratio tests. And let  $pval$  be a vector with length  $J$ . The BE algorithm is:

- (A) [*Initialize model, modelI, J, and pval*]  $model \leftarrow saturated\ model;$   
 $modelI \leftarrow \emptyset; J \leftarrow 1; pval \leftarrow \vec{0}.$
- (B) *If*( $model \equiv null\ model$ ) *return*( $model$ ); *exit*.
- (C) *For*  $i = 1, J$  : *perform* (D)&(E).
- (D)  $ModelI[i] \leftarrow model \setminus \{ith\ effect\ in\ its\ generating\ set\}.$
- (E)  $Pval[i] \leftarrow compare(model, modelI[i]).$
- (F) *If*( $all\ pval \leq \alpha$ ) *return*( $model$ ); *exit*.  
*Else* [*update model & J*]  $model \leftarrow modelI[i_0] \mid pval[i_0] = max(pval);$   
 $J \leftarrow J(model);$  *go to* (B).

We have also constructed three new search algorithms. Firstly, another backwards elimination algorithm (BE2). It starts with the model with the  $m(\leq p)$ -way interactions such that this model fits the data and the model with all the  $(m - 1)$ -way interactions does not; thus bounding the *best* model above and below and thereby reducing the dimension of the model search space. Secondly, a forward selection algorithm (FS). It starts with the null model and adds effect(s) until a model that fits the data is found. These algorithms work by eliminating (BE2), or by adding (FS), one effect at a time: the proposed model is always compared to the previous model. Finally, another algorithm which uses the tests of partial associations (FS2): here, the saturated model is always the basis for the comparisons.

### 3 Simulation techniques

We have started to evaluate the performance of the new algorithms by means of a comprehensive simulation study which employs two different strategies, A & B. Let  $\Theta$  be the parameter space of dimension  $n$ . Let  $T_p^N \subset \mathbb{R}^n$  be the set of all possible contingency tables formed with  $p$  binary variables and with a total sample size  $N$ . Then  $n = 2^p$ . We consider:

#### 3.1 Strategy A - Compositions

A composition of  $N_1$  into  $n_1$  parts is  $N_1 = r_1 + r_2 + \dots + r_{n_1}$  where  $r_i \geq 0$  ( $i = 1, \dots, n_1$ ) and the order of the summands is important;  $N_1$  and  $n_1$  are fixed positive integers (see Nijenhuis & Wilf, 1978). Let  $C_{N,n}$  be the set of all the compositions of  $N$  into  $n$  parts.  $C_{N,n}$  maps bijectively onto

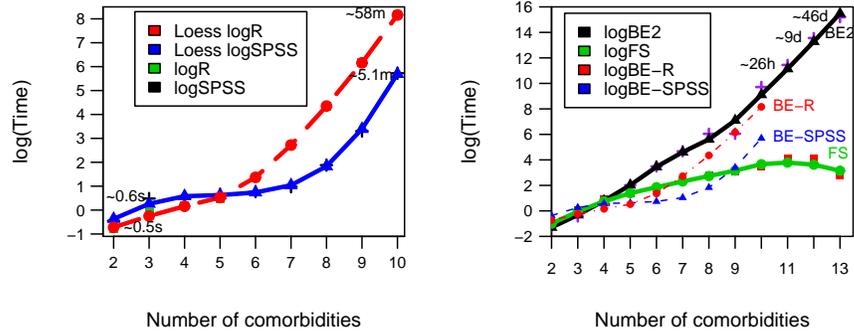


FIGURE 1. Comparison of times. (a) BE. (b) FS & BE2 (SPSS v15; R v2.6.1). (a) From 5 variables upwards the times become progressively longer in R: it is an interpreted language while SPSS uses compiled code in FORTRAN. (b) For large  $p$ , FS is the fastest algorithm. However, this algorithm often selected models unsatisfactory in terms of fit, as judged by the pattern of residuals. From this perspective, the backward elimination algorithms selected more satisfactory models, but they were always more complex.

$\mathcal{T}_p^N$ . Accordingly, the total number of contingency tables formed with  $p$  binary variables (i.e.  $n$  cells) and with a total sample size  $N$  is  $\binom{N+n-1}{N}$ . A contingency table is drawn as a random composition  $\mathcal{C}_{N,n}^* \in \mathcal{C}_{N,n}$ . The advantage of this approach is that the different tables generated are not “formulated” in advance.

### 3.2 Strategy B - Formulated Tables

We formulate the model a priori, by selecting a suitable starting  $\theta_0 = (\theta_{1_0}, \dots, \theta_{n_0})'$  which reflects the desired model structure. The  $Q$  function:

$$Q(\theta) = \left( \sum_{j=1}^n N^{-1} \exp(a'_j \theta) - 1 \right)^2 \tag{1}$$

is minimized starting from  $\theta_0$  and using `nlm` in R. All these quantities are as defined previously. This minimization yields a  $\hat{\theta}_Q$  defining a model satisfying our requirements selected from the space of all possible models. If necessary, we may re-adjust  $\theta_0$  at this stage in order to refine our desired model structure, e.g., by utilizing information of the standard errors (Agresti, 2002). Using  $\hat{\theta}_Q$  and  $N$  we generate a pseudo-contingency table and obtain the model mle  $\hat{\theta}$ , using `nlm` again. Next we check that the final formulated model (i.e.  $\hat{\theta}$ ) has the desired structure and then simulate contingency tables from it using the `rpois` function in R. This approach allows us to formulate a specific model *a priori*.

## 4 Results

BE was tested using real datasets comprising 10 comorbidities (Conde & MacKenzie, 2007) and 13 comorbidities. The solutions found by our algorithm in R were identical to those found using the analogous procedure in SPSS – Figure 1(a) shows their (*log*) times. For large values of  $p$  this combinatorial algorithm becomes infeasible and other approaches will be required. Figure 1(b) shows the (*log*) times for FS and BE2, which were obtained using a different set of comorbidities.

In addition to testing with real data sets, a comprehensive simulation study is being undertaken. A core scenario using strategy A involved generating  $m = 1000$  random compositions. In relation to strategy B, preliminary results in relation to formulated models are encouraging. In almost all the cases the final model given by the algorithms (BE, BE2 and FS) identifies the correct model. More details will be presented at the conference and in Conde & MacKenzie (2008) forthcoming.

## 5 Discussion

Testing new algorithms thoroughly via simulation is an essential step in their evaluation. We have designed stern tests for the algorithms by means of a two-pronged simulation strategy. In preliminary testing, these complementary simulation strategies have been useful and to date the new algorithms have performed well.

**Acknowledgments:** Special thanks to Dr. George Quartey, GSK, UK for supporting this work.

## References

- Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience.
- Conde, S. and MacKenzie, G. (2007). Modelling High Dimensional Sets of Binary Co-morbidities. In: *Proceedings of the 22nd International Workshop on Statistical Modelling*. 177-180, Barcelona, Spain.
- Conde, S. and MacKenzie, G. (2008). Evaluating Search Algorithms for Log-Linear Models in Multi-Dimensional Contingency Tables. In: *Proceedings of the 3rd International Workshop on Correlated Data Modelling*, Limerick, Ireland. Paper to appear.
- Goodman, L.A. (1971). The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications. *Technometrics*, **13**, 1, 33-61.
- Nijenhuis, A. and Wilf, H. S. (1978). *Combinatorial Algorithms*. Academic Press.
- O’Flaherty, M. and MacKenzie, G. (1982). Direct Simulation of Nested Fortran DO-LOOPS. *Algorithm AS 172*, **31**, 1, 71-74.

# Analyzing Randomized Response Data using a Doubly Zero-Inflated Poisson Model

Maarten J.L.F. Cruyff<sup>1</sup>, Ulf Böckenholt<sup>2</sup> and Peter G.M. van der Heijden<sup>1</sup>

<sup>1</sup> Utrecht University, Utrecht, The Netherlands

<sup>2</sup> McGill University, Montreal, Canada

**Abstract:** This paper presents a randomized response model for sensitive event counts that are assumed to follow a Poisson distribution. The model distinguishes two kinds of zero-inflation. The Poisson distribution of the sensitive event counts is assumed to be zero-inflated due to persons who are incapable of experiencing the sensitive event, and the observed response distribution is assumed to be zero-inflated due to respondents who do not comply with the randomized response design. The model is applied to randomized response data from a Dutch social security survey.

**Keywords:** regulatory noncompliance; sensitive event counts; self-protective response bias.

## 1 Introduction

Randomized response is an interview technique designed to eliminate response bias (Warner, 1965). Most randomized response designs have dichotomous questions that assess the presence or absence of a sensitive characteristic. A limited number of designs use questions with multiple answer categories. Examples are the multi-proportions randomized response design by Abul-Ela, Greenberg and Horvitz (1967), the unrelated question design with by Greenberg, Kuebler, Abernathy and Horvitz (1971), and a quantitative forced response design by Liu and Chow (1976). These design assume a nonparametric distribution for the sensitive categories.

A 2004 Dutch randomized response survey on social security fraud included randomized questions with multiple response categories. We found that nonparametric randomized response models do not fit the data. Therefore we in this paper a parametric model for the social security data that takes the possibility into account that other response generating processes are at work. The model assumes a Poisson distribution for the sensitive categories. In addition, we assume zero-inflation of both the observed response and the sensitive event count distribution. The sensitive events counts are potentially zero-inflated due to persons who are incapable of experiencing the sensitive event (one could for example think of a zero produced by a

nondrinker when the number of alcohol drinks is assessed). The observed count distribution is zero-inflated due to persons who do not follow the randomized-response design, and respond "zero" regardless the outcome of the randomization procedure. Böckenholt and van der Heijden (2007) call this *self-protective response behavior*.

The next section presents the data and randomized response design of the social security survey. The doubly zero-inflated Poisson regression is introduced in the model section. The results is presented in last section.

## 2 The Data

In 2004 the Dutch Department of Social Affairs conducts a nationwide survey to assess the level of noncompliance with social security regulations. The sample consists of 2.580 respondents receiving financial benefits. The survey includes two questions assessing the amount of illegally obtained income in addition to the welfare benefit:

*jobs* On average, how much money a month have you earned in the past 12 months in addition to your social security benefits by doing small jobs for friends or acquaintances (without reporting this to the social welfare agency)?

*work* 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 questions have the six response categories denoting the amount, ranging from 1 for '0 euros' to 6 for '251 euros or more'. A forced response design (Boruch, 1971) was specified in which participants give the truthful response if the sum of a pair of dice is 5, 6, 7, 8, 9, while otherwise they toss a new dice and answer with the number of eyes on that die. In order to obtain (pseudo) count data we interpret the amounts of money as units, and we recode the responses 1 to 6 into the pseudo counts "0" (no money earned) to "5" (250 euro or more). The following response frequencies were observed,

	0	1	2	3	4	5
<i>jobs</i>	1882	304	110	104	91	98
<i>work</i>	2014	245	108	74	72	67.

## 3 The Model

The doubly zero-inflated Poisson regression model (dZIP) is given by

$$\pi_{y_i^*} = (1 - \theta_i) \sum_{y_i=0}^K p_{y_i^*|y_i} \pi_{y_i}^{ZIP} + I_{(y_i^*=0)} \theta_i. \tag{1}$$

where  $Y_i^*$  is the observed response variable and  $Y_i$  denotes the true status of individual  $i$  (i.e. the amount of earned money), for  $y_i, y_i^* = 0, 1, \dots, K$  and  $i = 1, 2, \dots, n$ .

The parameter  $\theta_i$  in model (1) denotes the probability of zero-inflation in the distribution of variable  $Y_i^*$  due to a self-protective response, with the indicator variable  $I_{(y_i^*=0)}$  taking value one if  $y_i^* = 0$ , and zero otherwise. The parameter is modeled as a logistic function of the covariate vector  $\mathbf{u}_i$  and the parameter vector  $\psi$ .

The conditional misclassification probabilities of the randomized response design are denoted by  $p_{y_i^*|y_i}$ , which in the present forced response design are equal to

$$p_{y_i^*|y_i} = \begin{cases} p + q = 19/24 & \text{if } y_i^* = y_i \\ q = 1/24 & \text{if } y_i^* \neq y_i \end{cases}. \quad (2)$$

The variable  $Y_i$  is assumed to follow a censored Poisson distribution

$$\pi_{y_i}^P = \begin{cases} \exp(-\lambda_i)\lambda_i^{y_i}/y_i! & \text{for } y_i < K \\ 1 - \sum_{y_i=0}^{K-1} \exp(-\lambda_i)\lambda_i^{y_i}/y_i! & \text{for } y_i = K \end{cases}, \quad (3)$$

where the Poisson parameter  $\lambda_i$  is expressed as a logarithmic functions of the covariate vector  $\mathbf{x}_i$  and the parameter  $\beta$ . Due to individuals who are characterized by a deterministic zero on variable  $Y_i$  this distribution is also zero-inflated, so that

$$\pi_{y_i}^{ZIP} = (1 - \phi_i)\pi_{y_i}^P + I_{(y_i=0)}\phi_i. \quad (4)$$

The zero-inflation parameter  $\phi_i$  is modeled as a logistic function of the covariate vector  $\mathbf{z}_i$  and the parameter vector  $\gamma$ .

## 4 The Examples

We fitted a nonparametric, multinomial model without predictors and the doubly zero-inflated Poisson regression model (dZIP) with 17 predictors (including the intercepts); nine predictors were used for the Poisson parameter  $\lambda_i$ , four for the zero-inflation parameter  $\phi_i$  and four for the self-protective zero-inflation parameter  $\theta_i$ . Table 1 compares the fit statistics for the non-parametric model (the multinomial model without predictors) and the doubly zero-inflated Poisson regression model with all predictors included (dZIP model) for the respective dependent variables *jobs* and *work*. The positive  $X^2$  statistics (with zero degrees of freedom) show that the multinomial model does not fit, and a comparison of the AIC and BIC indicate that the doubly zero-inflated Poisson regression is a substantial improvement.

TABLE 1. Log-likelihood, parameters ( $k$ ), AIC, BIC and  $X^2$  statistic for the non-parametric and the dZIP model for the variables *jobs* and *work*.

	model	loglike	$k$	AIC	BIC	$X^2$ (df)
<i>jobs</i>	Multinomial	-2532.0	5	5074.0	5103.3	6.6 (0)
	dZIP	-2447.1	17	4928.3	5027.8	-
<i>work</i>	Multinomial	-2207.6	5	4425.2	4454.5	43.4 (0)
	dZIP	-2116.9	17	4267.7	4367.2	-

Table 2 reports the estimates of the regression parameters of the predictors that are included in the dZIP model. The upper panel show the results for the predictors of the Poisson parameter. In both models the strongest predictor of illegal earnings is perceived benefit of noncompliance. The Poisson parameter almost doubles if the perceived benefits of noncompliance with the rules increases by a standard deviation. Stronger social control significantly diminishes the amount of illegal earnings, in both models the Poisson parameters decreases by approximately 25% for each standard deviation increase in the predictor social control. Persons who perceive the costs to comply with the rules to be high have significantly higher illegal earnings from doing jobs for friends and relatives than persons who have no trouble complying with the rules. The sector in which the person was last employed is related to illegal earnings from working off the books; persons who worked in the construction sector have significantly higher earnings than persons from other sectors. There is no evidence that age and gender are related to the amount of illegal earnings.

The middle panel of Table 2 shows the parameter estimates for the predictors of zero-inflation of the Poisson distribution. Both models show a strong effect for type of Insurance Act; the odds of having a zero probability of earning illegal income for (partly) disabled persons who receive benefits under the Disability Act (reference category for Insurance Act) are about eight times higher than for persons who receive benefits under the the Assistance Act (although for *work* the estimate is not significant). The way people feel about abiding the rules is also a significant predictor; the odds of not having any illegal earnings double for each standard deviation increase on the predictor law conformity (although the estimate is again not significant for *work*).

The bottom panel of Table 2 shows the parameter estimates for the predictors of a self-protective response. In both models education is a significant predictor; the odds of a self-protective response decrease with 25% if the educational level increases with a standard deviation. The last two variables trust and understanding assess the degree to which the respondent understands the forced response method and trusts the privacy protection of the design. The latter predictor is significant in the model for *jobs*, the more trust the respondents has in the confidentiality of the design, the

lower the odds of a self-protective response.

TABLE 2. Parameter estimates, standard errors (se) and relative change in  $\lambda$  and the odds of  $\phi$  and  $\theta$  per unit change in the categorical predictors or per standard deviation change in the continuous predictors for the SP ZIP model with dependent variables *jobs* (left) and *work* (right).

covariate	<i>jobs</i>		<i>work</i>	
	$\hat{\beta}$ (se)	$\Delta\lambda$	$\hat{\beta}$ (se)	$\Delta\lambda$
<i>gender</i> (male)	0.37 (.20)	1.45	0.14 (.23)	1.15
<i>age</i> (31-50)	-0.14 (.18)	0.87	-0.33 (.24)	0.72
<i>age</i> (51 or older)	0.01 (.14)	1.00	-0.38 (.27)	0.69
<i>job sector</i> (food & drinks)	0.31 (.24)	1.36	0.38 (.28)	1.46
<i>job sector</i> (construction)	-0.41 (.38)	0.66	1.19 (.34)	3.29
<i>costs compliance</i>	0.47 (.17)	1.29	0.31 (.20)	1.18
<i>benefits noncompliance</i>	1.23 (.15)	2.00	1.04 (.18)	1.81
<i>social control</i>	-0.65 (.17)	0.75	-0.70 (.21)	0.73
	$\hat{\gamma}$ (se)	$\Delta\frac{\phi}{1-\phi}$	$\hat{\gamma}$ (se)	$\Delta\frac{\phi}{1-\phi}$
<i>Insurance Act</i> (UA)	-0.52 (.45)	0.59	-0.18 (.63)	0.84
<i>Insurance Act</i> (AA)	-1.97 (.92)	0.14	-2.02 (1.3)	0.13
<i>law conformity</i>	1.51 (.66)	1.92	2.05 (1.3)	2.46
	$\hat{\psi}$ (se)	$\Delta\frac{\theta}{1-\theta}$	$\hat{\psi}$ (se)	$\Delta\frac{\theta}{1-\theta}$
<i>education</i>	-0.34 (0.16)	0.71	-0.42 (.14)	0.75
<i>trust</i>	-0.74 (0.28)	0.72	-0.19 (.25)	0.92
<i>understanding</i>	-0.22 (0.43)	0.91	-0.32 (.20)	0.87

Table 3 presents the estimated joint distribution of self-protective zeros, the zero-inflated income counts and the Poisson distributed income counts under the dZIP model. The results show that the probability of a self-protective response is higher for working off working than for doing jobs for friends and relatives. This may be due to the fact that working off the books is considered the more serious offence. From Table 3 we can also compute the conditional probabilities of regulatory compliance given compliance with the randomized response design. Thus it is estimated that 78%  $(.231 + .324)/(1 - .289)$  does not have any earnings from doing jobs for friends and relatives, and that 80%  $(.167 + .314)/(1 - .401)$  does not have any earnings from working off the books. These estimates are only valid under the assumption that the probability self-protective response is independent of the amount of illegal earnings.

## 5 Discussion

The SP ZIP model presented in this paper allows for zero-inflation on the level of the sensitive event counts and on the level of the observed responses, thus stratifying the sample into self-protective respondents, respondents who always comply with the rules and respondents who are potentially

TABLE 3. Joint probability distribution of self-protective zeros ( $\hat{\theta}$ ), zero-inflated income counts ( $\hat{\pi}_0^{ZIP}$ ) and Poisson distributed illegal income counts ( $\hat{\pi}_0^P$  to  $\hat{\pi}_{5+}^P$ ) under the dZIP model.

	$\hat{\theta}$	$\hat{\pi}_0^{ZIP}$	$\hat{\pi}_0^P$	$\hat{\pi}_1^P$	$\hat{\pi}_2^P$	$\hat{\pi}_3^P$	$\hat{\pi}_4^P$	$\hat{\pi}_{5+}^P$
<i>jobs</i>	.289	.231	.324	.108	.032	.010	.004	.002
<i>work</i>	.401	.167	.314	.088	.021	.006	.002	.001

noncomplying. Although the model was originally meant for true count data, the examples from the social security survey show that the model can also be successfully applied to pseudo count data. It is probably true that imposing a Poisson distribution onto the pseudo counts may have induced bias in the prevalence estimates of regulatory noncompliance, but we feel that this problem is by far outweighed by the correction for self-protective response bias. The model fits the data significantly better than the existing multinomial randomized responses, and provides interesting information about the response generating processes.

**References**

Abul-Ela, A-L.A., Greenberg, G.B., and Horvitz, D.G. (1967). A Multi-Proportions Randomized Response Model, *Journal of the American Statistical Association* **62**, 990-1008.

Böckenholt, U., and van der Heijden, P.G.M. (2007). Item Randomized-Response Models for Measuring Noncompliance: Risk-Return Perceptions, Social Influences, and Self-Protective Responses. *Psychometrika* **72**, 245-262.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Econometric Monographs, 30, Cambridge University Press, Cambridge.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., and Horvitz, D.G. (1971). Application of the Randomized Response Technique in Obtaining Quantitative Data, *Journal of the American Statistical Association* **66**, 243-250.

Liu, P.T., and Chow, L.P. (1976). A New Discrete Quantitative Randomized Response Model, *Journal of the American Statistical Association* **71**, 72-73.

Warner, S. L. (1965). Randomized Response: a Survey Technique for Eliminating Answer Bias, *Journal of the American Statistical Association* **60**, 63-69.

# Smoothing overparameterized regression models

Iain Currie<sup>1</sup>

<sup>1</sup> 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:** Over-parameterized regression models occur throughout statistics and are often found, though not exclusively, when data are arranged in an array. Plots of fitted values for such models can suggest that smoothing is appropriate. Penalized splines are a very popular method of smoothing but their use in this setting is not straightforward. We discuss the difficulties of using penalized splines to smooth overparameterized regression models and suggest a new smoothing paradigm which gets round these difficulties. We call our method direct smoothing and illustrate it by smoothing the Lee-Carter model, an over-parameterized model used in the modelling and forecasting of human mortality. We illustrate our methods with male Swedish mortality data taken from the Human Mortality Database.

**Keywords:** Generalized linear array model; Lee-Carter model; mortality; penalized splines.

## 1 Introduction

We begin with an example which illustrates the difficulties of using penalized splines to smooth an overparameterized regression model. We suppose that we have mortality data, deaths and exposures to the risk of death, arranged in two matrices,  $\mathbf{D}$  and  $\mathbf{E}$ , each  $n_a \times n_y$ , whose rows and columns are classified by age at death,  $\mathbf{x}_a$ , and year of death,  $\mathbf{x}_y$ , respectively. The assumption in the Lee-Carter model (Lee and Carter, 1992) is that the log of the force of mortality or hazard rate,  $\theta_{ij}$ , at age  $i$  and in year  $j$  is given by

$$\theta_{ij} = \alpha_i + \beta_i \kappa_j, \quad i = 1, \dots, n_a, \quad j = 1, \dots, n_y. \quad (1)$$

This factor-type model has  $2n_a + n_y$  parameters but the parameters are not identifiable and so location and scale constraints, such as  $\sum \kappa_j = 0$  and  $\sum \kappa_j^2 = 1$ , are generally used to simplify estimation. If we assume that the number of deaths at each age in each year follows a Poisson distribution then we can fit the Lee-Carter model with these constraints by maximum likelihood. We fit the model to Swedish male mortality data taken from the Human Mortality Database with ages  $\mathbf{x}_a$  from 10 to 90 and years  $\mathbf{x}_y$

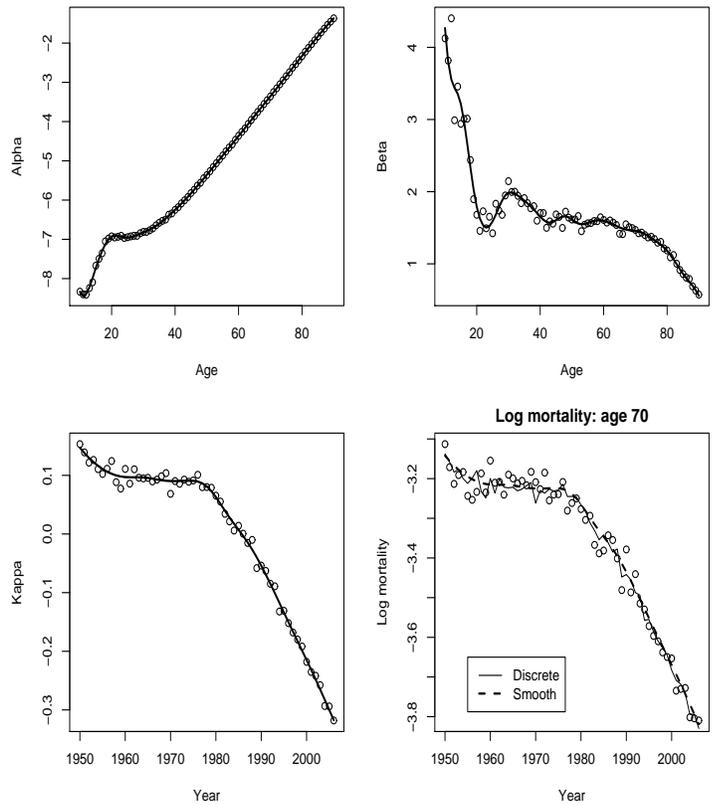


FIGURE 1. Fitted parameters from the Lee-Carter model (1) with — and without  $\circ$  direct smoothing; also observed  $\circ$  and fitted log mortality with — — and without — direct smoothing for age 70.

from 1950 to 2006. Figure 1 (which also includes smoothed values from section 3) displays the fitted parameter values,  $\hat{\alpha}$ ,  $\hat{\beta}$  and  $\hat{\kappa}$  together with the observed and fitted log mortality for age 70.

Figure 1 suggests that it is appropriate to smooth the fitted parameters. One approach is to fit model (1) by maximizing the penalized log likelihood,  $\ell_p(\theta)$ , where

$$\ell_p(\theta) = \ell(\theta) - \lambda_a \alpha' D_a' D_a \alpha - \lambda_b \beta' D_a' D_a \beta - \lambda_k \kappa' D_y' D_y \kappa; \quad (2)$$

here  $\ell(\theta)$  is the usual log likelihood for a generalized linear model (GLM),  $D_a$  and  $D_y$  are second order difference matrices of appropriate size, and  $\lambda_a$ ,  $\lambda_b$  and  $\lambda_k$  are smoothing parameters. Expression (2) corresponds to a block

diagonal penalty matrix. However, we now run into the following problem: model (1) is not identifiable and we can transform the fitted parameters

$$\hat{\boldsymbol{\alpha}} \rightarrow \hat{\boldsymbol{\alpha}}^* = \hat{\boldsymbol{\alpha}} - c\hat{\boldsymbol{\beta}}, \quad \hat{\boldsymbol{\beta}} \rightarrow \hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}, \quad \hat{\boldsymbol{\kappa}} \rightarrow \hat{\boldsymbol{\kappa}}^* = \hat{\boldsymbol{\kappa}} + c. \quad (3)$$

The fitted values from model (1) are invariant with respect to this transformation but it is a simple matter to check that, while the penalties on  $\boldsymbol{\beta}^*$  and  $\boldsymbol{\kappa}^*$  are unaltered from those on  $\boldsymbol{\beta}$  and  $\boldsymbol{\kappa}$ , the penalty on  $\boldsymbol{\alpha}^*$  tends to  $\infty$  as  $c \rightarrow \pm\infty$ . We conclude that the smoothed values are not invariant with respect to the choice of parameterization. These remarks apply equally when we attempt to implement a penalized spline system such as the  $P$ -spline system of Eilers and Marx (1996) by setting

$$\boldsymbol{\alpha} \rightarrow \mathbf{B}_a \mathbf{a}, \quad \boldsymbol{\beta} \rightarrow \mathbf{B}_a \mathbf{b}, \quad \boldsymbol{\kappa} \rightarrow \mathbf{B}_y \mathbf{k}, \quad (4)$$

where  $\mathbf{B}_a = \mathbf{B}_a(\mathbf{x}_a)$ ,  $n_a \times c_a$ , and  $\mathbf{B}_y = \mathbf{B}_y(\mathbf{x}_y)$ ,  $n_y \times c_y$ , are regression matrices of  $B$ -splines for age and year respectively; again, the penalty on the regression coefficients  $\mathbf{a}$  depends on the parameterization.

It is the purpose of this paper to propose a solution to this problem.

## 2 Direct smoothing

Let  $\mathbf{d}$  and  $\mathbf{e}$  be the observed numbers of deaths and the corresponding exposures to the risk of death for a single year. We assume that we can model  $\mathbf{d}$  by a penalized GLM with Poisson errors and linear predictor  $\log \mathbf{e} + \mathbf{B}_a \mathbf{a}$  where  $\mathbf{B}_a$  is a regression matrix of  $B$ -splines, as in the previous section. The  $P$ -spline system of smoothing uses the penalty  $\mathbf{a}' \mathbf{D}'_a \mathbf{D}_a \mathbf{a}$ , i.e., it penalizes differences in adjacent coefficients. There seems no *a priori* reason why the penalty should not be placed directly on adjacent values of the linear predictor; this gives the penalty  $\mathbf{a}' \mathbf{B}'_a \check{\mathbf{D}}'_a \check{\mathbf{D}}_a \mathbf{B}_a \mathbf{a}$  where  $\check{\mathbf{D}}_a$  is a second order difference matrix. We call smoothing with a  $B$ -spline basis and this new penalty *direct smoothing*. In summary, the difference between  $P$ -splines and direct smoothing is that the penalty changes as follows:

$$\mathbf{P} = \lambda \mathbf{D}'_a \mathbf{D}_a \rightarrow \check{\mathbf{P}} = \lambda \mathbf{B}'_a \check{\mathbf{D}}'_a \check{\mathbf{D}}_a \mathbf{B}_a. \quad (5)$$

We note that although  $\check{\mathbf{D}}_a$ ,  $(n_a - 2) \times n_a$ , is often much larger than  $\mathbf{D}_a$ ,  $(c_a - 2) \times c_a$ , both  $\mathbf{P}$  and  $\check{\mathbf{P}}$  are  $c_a \times c_a$ . Thus, both  $P$ -splines and direct smoothing are low rank methods.

We do not make any special claim for direct smoothing in one dimension except to say that in applications we have found it to give very similar results to  $P$ -splines. Our main reason for considering direct smoothing is that it applies easily to data distributed over an array, as in our mortality example. Penalizing the fitted values on the scale of the linear predictor in rows and columns gives the penalty

$$\check{\mathbf{P}} = \lambda_a \check{\mathbf{P}}_a + \lambda_y \check{\mathbf{P}}_y = \mathbf{X}' (\lambda_a \mathbf{I}_{n_y} \otimes \check{\boldsymbol{\Delta}}_a + \lambda_y \check{\boldsymbol{\Delta}}_y \otimes \mathbf{I}_{n_a}) \mathbf{X} \quad (6)$$

where  $\Delta_a = \check{D}'_a \check{D}_a$ ,  $\Delta_y = \check{D}'_y \check{D}_y$ ,  $\mathbf{X}$  is the regression matrix and  $\otimes$  denotes the Kronecker product. This penalty is a function of  $\mathbf{X}\boldsymbol{\theta}$  which is invariant with respect to the parameterization; here  $\boldsymbol{\theta}$  is the vector of regression coefficients. We conclude that direct smoothing is invariant with respect to the particular parameterization used to fit the model.

### 3 Lee-Carter model

We apply direct smoothing to the Lee-Carter model (1). First we make the transformation (4). We have a pair of coupled penalized GLMs, M1 and M2, with data vector  $\mathbf{d} = \text{vec}(\mathbf{D})$ , Poisson error and log link. The regression matrices, offsets and penalty matrices (calculated from expression (6)) for M1 and M2 are as follows.

- M1 Given current estimates of the age parameters  $\tilde{\mathbf{a}}$  and  $\tilde{\mathbf{b}}$  we have a GLM with regression matrix  $\mathbf{X} = \mathbf{B}_y \otimes \mathbf{B}_a \tilde{\mathbf{b}}$ , parameter vector  $\mathbf{k}$ , offset  $\text{vec}(\log \mathbf{E} + \mathbf{B}_a \tilde{\mathbf{a}} \mathbf{1}'_{n_y})$  and penalty matrices

$$\check{P}_a = \left( \tilde{\mathbf{b}}' \mathbf{B}'_a \Delta_a \mathbf{B}_a \tilde{\mathbf{b}} \right) \mathbf{B}'_y \mathbf{B}_y, \quad \check{P}_y = \left( \tilde{\mathbf{b}}' \mathbf{B}'_a \mathbf{B}_a \tilde{\mathbf{b}} \right) \mathbf{B}'_y \Delta_y \mathbf{B}_y. \quad (7)$$

- M2 Given current estimates of the year parameters  $\tilde{\mathbf{k}}$  we have a GLM with regression matrix  $\mathbf{X} = [\mathbf{1}_{n_y} : \mathbf{B}_y \tilde{\mathbf{k}}] \otimes \mathbf{B}_a$ , parameter vector  $(\mathbf{a}', \mathbf{b}')'$ , offset  $\text{vec}(\log \mathbf{E})$  and penalty matrices

$$\check{P}_a = \phi(\tilde{\mathbf{k}})' \phi(\tilde{\mathbf{k}}) \otimes \mathbf{B}'_a \Delta_a \mathbf{B}_a, \quad \check{P}_y = \psi(\tilde{\mathbf{k}})' \psi(\tilde{\mathbf{k}}) \otimes \mathbf{B}'_a \mathbf{B}_a \quad (8)$$

where  $\phi(\tilde{\mathbf{k}}) = [\mathbf{1}_{n_y} : \mathbf{B}_y \tilde{\mathbf{k}}]$  and  $\psi(\tilde{\mathbf{k}}) = [\mathbf{0}_{n_y} : \check{D}_y \mathbf{B}_y \tilde{\mathbf{k}}]$ .

The model can now be fitted by iterating back and forward between M1 and M2 until convergence.

We note that the penalty matrices are dynamic, i.e., in M1, they depend on the current estimate of  $\mathbf{b}$  while in M2 they depend on the current estimate of  $\mathbf{k}$ ; as a result, the fast computational forms given in (7) and (8) are much preferably to the general form in (6). Furthermore in both M1 and M2 the regression matrix has a Kronecker product structure. Thus, M1 and M2 can be regarded as a pair of coupled generalized linear array models or GLAMs (Currie *et al.*, 2006). A GLAM takes advantage of the Kronecker product structure of a regression matrix to give very fast fitting of a GLM. We chose smoothing parameters by minimizing the Bayesian Information Criterion, BIC. The smooth estimates of the parameters and the smooth estimate of the log mortality at age 70 have been added to Fig. 1; it appears that direct smoothing has been successful in smoothing both the estimated parameters and the fitted values. We emphasize that the object of direct smoothing has been achieved and the smoothed values are indeed invariant with respect to reparameterization.

TABLE 1. Summary statistics for Lee-Carter model.

Model	$\lambda_a$	$\lambda_y$	Trace	Deviance	$\phi$	BIC
Discrete LC	0	0	217	6227	1.4	8058
Smooth LC	5	2800	48	6737	1.5	7145

The discrete model (1) is a special case of direct smoothing and results when we set the regression matrices  $\mathbf{B}_a$  and  $\mathbf{B}_y$  to the identity matrices  $\mathbf{I}_{n_a}$  and  $\mathbf{I}_{n_y}$ , and the smoothing parameters  $\lambda_a$  and  $\lambda_y$  to zero. The results from both models are summarized in Table 1 ( $\phi$  denotes an estimate of the overdispersion parameter). On the BIC scale, there is a clear preference for the smooth Lee-Carter model.

#### 4 Forecasting with the Lee-Carter model

The principal use of the Lee-Carter model is as a forecasting tool. Lee and Carter (1992) originally proposed that forecasting of the mortality table be achieved by keeping the values of  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  fixed at their estimated values and forecasting the  $\boldsymbol{\kappa}$  values. Figure 2 shows the result of forecasting age 70 mortality to 2048 when  $\boldsymbol{\kappa}$  is forecast by an ARIMA(2,1,2) model. Forecasting is also possible with direct smoothing. We treat the forecast region of the table as missing data with zero weights. The penalty function allows forecasting to take place (see Currie *et al.*, (2004) for more detail). The forecast values from the two methods are very close; the 95% confidence interval with the time series method are wider.

#### 5 Concluding remarks

We have introduced a new method of penalized spline smoothing. Computationally, direct smoothing is very close to the  $P$ -spline method of Eilers and Marx (1996), the only difference being the change in penalty given by (5). Direct smoothing can also be regarded as a discrete approximation to O'Sullivan penalized splines (O'Sullivan, 1986), an observation which lends some support to our method. The method is designed for smoothing a GLM when data are distributed over an array but when the regression matrix can be written as a Kronecker product (as in the Lee-Carter model) then the GLM can be regarded as a GLAM. We have used direct smoothing successfully on the age-period-cohort model of mortality. More generally, the method is available for any overparameterized regression model where plots of fitted values suggest that smoothing is appropriate.

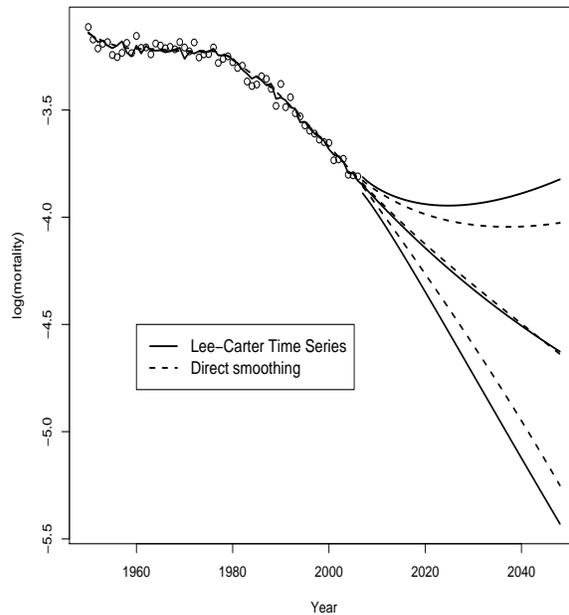


FIGURE 2. Forecast of age 70 log mortality with Lee-Carter model: original discrete model — and with direct smoothing - - -.

## 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.
- Lee, R.D., and Carter, L. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87**, 659-75.
- O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, **1**, 505-27.

# Joint Modeling of Progression Free Survival and Death

David Dejardin<sup>1</sup>, Emmanuel Lesaffre<sup>1,2</sup> and Geert Verbeke<sup>1</sup>

<sup>1</sup> Biostatistical Centre, Catholic University of Leuven, U.Z. St. Rafaël, Kapucijnenvoer 35, B3000 Leuven, Belgium , David.Dejardin@med.kuleuven.be

<sup>2</sup> Department of Epidemiology and Biostatistics, Erasmus MC, Rotterdam, The Netherlands

**Abstract:** Progression Free Survival ( $P$ ) and time to death ( $D$ ) are common endpoints in oncology clinical trials. It is of interest to study the joint distribution of these ordered endpoints ( $P$  will always be smaller than  $D$ ). Measures of association derived from the joint distribution (like Kendall's  $\tau$ ) are used to validate  $P$  as a surrogate for  $D$ . Current methods based on frailty models and copulas ignore the ordering of the endpoints. We propose a shared frailty model in which the conditional hazard of  $D$  given  $P$  includes an “at risk” function and is proportional to the hazard of  $P$ . This simple model allows to take the ordering into account. In addition, Kendall's  $\tau$  for this model depends only on the frailty distribution and the parameter of proportionality between the hazards of  $P$  and  $D$  given  $P$ . We assessed the performance of the model by simulations. The model will also be applied to clinical trial data.

**Keywords:** Survival; Progression Free Survival; Ordered Survival Times; Frailty Model; Kendall's  $\tau$

## 1 Introduction

Cancer clinical trials in the metastatic setting collect a number of efficacy endpoints. Among them, the time to first progression (Progression Free Survival or  $P$ ) and time to death ( $D$ ) are often used to test for treatment difference. Most often, these endpoints are analysed separately. Here, we are interested to study the relationship between those endpoints. For example, we would like to know whether subjects with a long  $P$  have a long  $D$ , which could be summarized with an association measure. This will highlight how good a surrogacy measure  $P$  is for  $D$ , which is important to assess the relevance of a clinical trial outcome based on  $P$ . On the other hand, it is also of interest to estimate the chances of success in terms of  $D$  using  $P$ . For these reasons, the joint distribution of  $P$  and  $D$  is worthwhile to study. The endpoints are ordered ( $P < D$ ), which needs to be taken into account in modeling the distribution. In the literature, frailty and copula models have been suggested (see Hougaard 2000). However, these models ignore the ordering of the variable.

In this paper, we propose a frailty model for the joint distribution of  $P$  and  $D$  in which the conditional hazard for  $D$  given  $P$  includes an “at risk” function and is proportional to the hazard of  $P$ . This model takes the ordering of the variables into account. Further, the association between  $P$  and  $D$  can be measured by Kendall’s  $\tau$  which depends on the frailty distribution and the parameter of proportionality between the hazards of  $P$  and  $D$  given  $P$ .

The model will be described in the next section. A simulation study was performed to investigate the bias induced by ignoring the ordering of the variables in previously proposed models. We also show that our approach has the desired characteristics. Results of the simulation study are presented in Section 3. In Section 4 we will apply the model to data from a clinical trial in ovarian cancer. In the last section, we indicate further generalizations of our approach.

## 2 Frailty Model with “at risk” Function

Let  $\lambda_0(\cdot)$  be the hazard of  $P$ . To take the restriction  $P < D$  into account, we propose to model the conditional hazard of  $D$  given  $P$  by introducing an “at risk” function that forces the hazard to be 0 for all death times below the progression time. In general, we can not assume that the hazard of  $P$  and  $D$  given  $P$  are the same, but we make the simplifying assumption that the two hazards are proportional with  $e^\beta$  as factor. This, together with the restriction on the possible values of  $D$ , induces some dependence between  $P$  and  $D$ . To allow us to express the surrogacy of  $P$  for  $D$ , we include a frailty term in the hazards. Thus the hazard of  $D$  given  $P$  is given by:

$$\lambda_{D|P,Z}(p, d) = ZR(p, d)e^\beta \lambda_0(d)$$

where  $R(p, d) = I(p < d)$  is the “at risk” function and the frailty term  $Z$  is a random variable with mean 1. In this context, the marginal density of  $P$  conditional on the frailty term is given by  $f_{P|Z}(p, Z) = Z\lambda_0(p) \exp(-Z \int_0^p \lambda_0(x) dx)$  and the conditional density of  $D$  given  $P$  and  $Z$  is given by

$$f_{D|Z,P}(p, d, Z) = ZR(p, d)e^\beta \lambda_0(d) \exp(-Ze^\beta \int_p^d \lambda_0(x) dx)$$

The cumulative joint distribution is given by:

$$\begin{aligned} S(p, d|Z) &= \Pr[P > p, D > d|Z] = \int_d^\infty \int_p^y f_{P|Z}(x, Z) f_{D|Z,P}(x, y, Z) dx dy \\ &= \frac{1}{1 - e^\beta} [\exp(-Z((1 - e^\beta) \int_0^p \lambda_0(x) dx + e^\beta \int_0^d \lambda_0(x) dx)) \\ &\quad - e^\beta \exp(-Z \int_0^d \lambda_0(x) dx)] \end{aligned}$$

For this model, it can be shown that Kendall's  $\tau$  depends only on the distribution of  $Z$  and  $\beta$  and not on  $\lambda_0(\cdot)$ .

Many possible choices for the baseline hazard function  $\lambda_0(\cdot)$  can be chosen from the literature. Simple parametric hazards (like Weibull hazards) allow the parameters to be estimated through maximizing the integrated likelihood (where the frailty term has been integrated out). Semi-parametric models are also available. They require more sophisticated methods for the estimation of the parameters but they also allow more flexibility to model the hazards and the inclusion of a covariate.

### 3 Simulation Study

We have performed a limited simulation study to assess the performance of our approach and the bias in the estimated value of  $\tau$  from models that do not take the ordering of the variables into account. The simulated data are taken from a gamma shared frailty model with Weibull marginal distribution. The progression time was generated first and the death time was generated such that the hazard of death was 0 for all death time below  $P$ . Results on the bias induced by ignoring the ordering are given in Table 1.

TABLE 1. Median of the estimated  $\tau$  over 100 simulated samples.  $P$  is simulated from a Weibull with parameters (2,2) and  $Z$  is drawn from a gamma distribution with both parameters equal. For model 1,  $e^\beta$  was 0.25, the gamma parameters were 4.5. For model 2,  $e^\beta$  was 1.6 and the gamma parameters were 5. For model 3,  $e^\beta$  was 2 and the gamma parameters were 1.2.

Model	True $\tau$	Model	
		Classical approach	Our approach
1	0.3	0.22	0.29
2	0.45	0.32	0.45
3	0.75	0.67	0.75

### 4 Analysis of an Ovarian Cancer Trial Dataset

We have applied our model to a clinical trial in metastatic advanced epithelial ovarian cancer. This phase III study compares the combination of taxol and platinum therapy (experimental treatment) with the combination of cyclophosphamide and platinum therapy (which was the standard of care) (see Piccard et al. (2000)). For all subjects,  $P$  and  $D$  were measured. In addition to these efficacy data, a number of prognostic factors were recorded, including stage of disease and pretreatment for their cancer.

For this dataset, a gamma shared frailty model with parametric models for the hazards were fitted with the purpose of assessing the association

between  $P$  and  $D$ . The conditional distribution of  $D$  given  $P$  was also derived. Covariates such as treatment and stage of disease were included in the model to assess whether the association differs if those covariates are taken into account.

The results show that the model with the “at risk” function provides a good alternative to model  $P$  and  $D$  compared to classical models that do not take the ordering into account.

## 5 Conclusions

Simulations showed that bias in the estimation of  $\tau$  is induced by ignoring the ordering of the variables. In applications like validation of  $P$  as a surrogate to  $D$  (see Burzykowski et al. 2005), the correction would be easy to implement. Further extensions under consideration: semi parametric estimation procedures based on the EM algorithm that do not require a parametric form for  $\lambda_0(\cdot)$  or estimators of the baseline hazard based on splines could be used to allow more flexibility in the estimation of the parameters. In addition,  $P$  is typically interval censored. The likelihood could be modified to take the interval censoring into account. The frailty models could also be modified to investigate the short term versus the long term dependence. This question is of interest in cancers that can almost be cured like breast cancers.

## References

- Burzykowski T., Molenberghs G. and Buyse M. (2005). *The Evaluation of Surrogate Endpoints*. Springer-Verlag, New York.
- Cook R. J. and Lawless J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer-Verlag, New York.
- Hougaard P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.
- Piccard M. et al. (2000) Randomized intergroup trial of cisplatin-paclitaxel versus cisplatin- cyclophosphamide in women with advanced epithelial ovarian cancer: Three-year results. *Journal of the National Cancer Institute* **92**, 699-708

# Smoothing zeros and small counts in meta-analysis of clinical trials

Johan J. de Rooi<sup>1</sup>, Paul H. C. Eilers<sup>1</sup>, Laurence E. Frank<sup>1</sup>

<sup>1</sup> Department of Methodology and Statistics, Utrecht University, The Netherlands

**Abstract:** In clinical trials on rare diseases one or both arms may contain low or even zero counts of events. This makes the computation of log-odds unreliable or even impossible. We present a pseudo-Bayes method of smoothing sparse data, using a prior based on the totals of all studies in a meta-analysis. To determine the optimal combination between prior and observed counts the use of an adjusted Akaike information criterion  $AIC_C$  is proposed.

**Keywords:** AIC; cross-validation; effective dimension; sparse tables

## 1 Introduction

In clinical trials on rare diseases one or both arms may contain low or even zero counts of events; in these cases we speak of sparse data. As a consequence log-odds can be unreliable or even impossible to calculate. A popular correction is to add a constant, like 0.1 or 0.5, to all counts. When a set of trials is being studied in a meta-analysis, we propose to compute pseudo-Bayes estimates (Fienberg and Holland, 1973) for each study, with a prior based on the sum of all studies. To optimize the weight of the prior we use a corrected AIC.

## 2 The method

Let  $X$  be a  $2 \times 2$  contingency table with cell counts  $x_{ij}$ , and the total number of observations in the study as  $n = \sum_{ij} x_{ij}$ . The row variable represents whether the event of concern did or did not occur, treatment and control group are defined in the columns. The maximum likelihood estimator of a cell probability is of low quality when  $x_{ij}$  is small. To smooth the data two priors (*overall prior* and *split prior*) based on information from all individual trials are proposed. The prior,  $\lambda_{ij}$  is calculated as the weighted mean expectation for each cell calculated over all tables in the array. The number of observations in corresponding cells of the distinctive studies are summed, and subsequently divided by the total number of observations  $N$  in the array:

$$\lambda_{ij} = \frac{\sum_{ij+} x_{ij+}}{N} \quad (1)$$

Where,  $N = \sum_{ijk} x_{ijk}$ , with  $i$  and  $j$  representing the rows and columns of the data matrix, and  $k$  as the number of dimensions of the array, or in other words the total number of studies in the meta-analysis. Combining all the priors from all individual cells in a matrix results in  $\Lambda = [\lambda_{ij}]$ , and constitutes the overall prior, which uses the structure of all four cells in relation to each other. For the split prior the same procedure holds, with the only difference that the relation between treatment and control group is disconnected. This results in two priors, and thus in two optimal combinations of prior and data; one for each arm of the study. The smoothed estimates based on raw data and the prior are calculated as follows:

$$x_{ij}^* = w\lambda_{ij}N + (1 - w)x_{ij}. \quad (2)$$

The smoothed estimate  $x_{ij}^*$  is a linear combination of the raw data and the prior. To determine the optimal weight  $w$  different methods are proposed. An early solution is presented by Fienberg and Holland (1973), who proposed Bayesian estimation with a conjugate prior. To determine the optimal weight of the prior they relied on the asymptotic properties of a risk function. The optimal combination of prior and data is the one that minimizes the expected mean square error, a routine that is currently hardly ever used in model selection. Eilers (1996) proposed leave-one-out cross-validation (LOOCV) as a criterion of model performance. This procedure is applicable as long as there are observations present in a cell. However if the data contain zeros there is nothing to leave out and no cross-validation can be performed.

Instead, the use of an improved Akaike information criterion (Hurvich et al., 1998) is proposed. The Akaike information criterion (AIC), introduced by Akaike (1974), is a fundamental concept in model selection. It is an entropic measure that uses the log likelihood ratio test in combination with a score that resembles the information gain of the modelling process (Ye, 1998). The information theoretic component makes model estimation possible, also in situations where not all cells are filled with observations. Instead of the standard AIC the adjusted AIC better suited for small datasets is used. This  $AIC_C$  presented by Hurvich et al. (1998) looks as follows;

$$AIC_C = 2 \sum_{ij} x_{ij} \ln \left( \frac{x_{ij}}{x_{ij}^*} \right) + 1 + \frac{2(\text{tr}(H) + 1)}{n - \text{tr}(H) - 2}, \quad (3)$$

where the first part of the equation represents the deviance of model  $X^*$  to the raw data  $X$  and is calculated with the likelihood ratio test. The  $\text{tr}(H)$  in the second part of equation (3) is the trace of the hat matrix. Ye (1998) defines the  $\text{tr}(H)$ , which is also conceived as the effective dimension  $ED$ ,

as;

$$\text{tr}(H) = ED = \sum_{ij} \frac{\partial x_{ij}^*}{\partial x_{ij}}, \quad (4)$$

which is a simple and elegant solution, also in comparison with calculating the trace of the hat matrix. Now adjusting equation (3) to our situation looks as follows:

$$\text{AIC}_C = 2 \sum_{ij} x_{ij} \ln \left( \frac{x_{ij}}{x_{ij}^*} \right) + 2 \frac{2 m(1-w)}{m - m(1-w)}, \quad (5)$$

with  $m$  the number of cells in the matrix, which is four in case of the overall prior and two in case of the split prior. Notice that in the case of the split prior two separate  $\text{AIC}_C$  scores are optimized.  $w$  is the weight given to the prior; it ranges from zero to one. The optimal weighted combination minimizes  $\text{AIC}_C$ .

### 3 An application

Table 1 shows data from seven different studies investigating whether electronic fetal heart rate monitoring (EFM) causes a drop in the perinatal mortality rate (Sweeting et al. 2004). Each study is presented on two lines; on the first line the counts are given for the raw data, on the second time the counts smoothed with the split-prior are shown. From the raw data we see that the studies are often not balanced in terms of group size of treated and controls. Moreover, the data seem to be heterogeneous when it comes to the treatment effect. Calculating the odds ratios is problematic due to the sparsity in most of the studies. For the studies two and six the problems are larger since in these studies no observations are present in one out of the four cells. Smoothing the data resolves this problem, as can be seen in table 1. After smoothing individual odds-ratios as well as a joint effect can be calculated.

The prior cell probabilities of the split-prior to which the data are smoothed are given in table 2. Clear are the very small event probabilities for both treatment and control group. However, the prior suggests a trend that assigns a positive effect of the EFM routine comparison with the situation where no EFM is used. In the last two columns of table 1 the weight assigned to the prior for respectively the treatment and control group are given. The assigned weight confirms the suggested heterogeneity among the studies especially in the control group. However in no instance the difference are such big that the prior is completely unwanted. In figure 1 the  $\text{AIC}_C$ -scores are plotted against the complete weight range. Picked is the split-prior in case of the control arm. With the different lines representing the different studies. Note that where the a line reaches its lowest point, this is the situation in which the trade-off between the most informative model (the

TABLE 1. Raw and smoothed data, investigating the effect of Routine induction on perinatal mortality

Study No.	EFM given		EFM not given		Weight	
	Event	No Event	Event	No Event	T	C
1	1160.00	2.00	5410.00	17.00		
$1_{split}$	1161.32	0.68	5415.99	11.01	0.78	0.61
2	150.00	0.00	6821.00	15.00		
$2_{split}$	149.96	0.04	6825.93	10.07	1.00	0.83
3	607.00	1.00	6142.00	37.00		
$3_{split}$	607.75	0.25	6153.36	25.64	0.89	0.39
4	4209.00	1.00	2914.00	9.00		
$4_{split}$	4208.90	1.10	2917.77	5.23	1.00	0.73
5	553.00	1.00	689.00	3.00		
$5_{split}$	553.75	0.25	690.74	1.26	0.88	0.83
6	4978.00	0.00	8632.00	2.00		
$6_{split}$	4976.70	1.30	8626.88	7.12	1.00	0.54
7	45870.00	10.00	66163.00	45.00		
$7_{split}$	45868.04	11.96	66145.55	62.45	1.00	0.41

prior) and closeness to the actual data is optimal, and thus results in the lowest  $AIC_C$ . The sketched example shows a situation with considerable amounts of heterogeneity between the studies. In situations where there is more agreement concerning the treatment effect the weight of the prior will more often go to one.

TABLE 2. Prior cell probabilities in case of the split prior.

EFM given		EFM not given	
Event	No Event	Event	No Event
.997	.00026	.998	.0013

## 4 Discussion

It is common practice to resolve problems with zero counts by simply adding a constant to all cells in the table. Our approach however uses information available in the data. A simulation study will be performed to determine the performance of the smoothing method in comparison with common applied corrections. The simulations will model different scenarios like for instance variable group ratios and between study heterogeneity. Next to this the consequences of either using or disconnecting the relation

between treated and control, or in other words using the overall- or split-prior, have to be further investigated. More sophisticated priors have to be considered as well. Although in this paper all attention is drawn towards the application of the method on clinical trials, extensions towards multinomial data are possible as previously shown by Eilers (1996).

## References

- Eilers P. H. C. (1996). Sparse contingency tables, pseudo-Bayes estimates and cross-validation. in: *Statistical modelling. Proceedings of the 11th international workshop on statistical modelling*, 402-405.
- Fienberg S.E., P.W. Holland (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of American Statistical Association*, Vol. 68 No. 343.
- Hurvich C. M., J. S. Simonoff (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, Vol. 60 part2, 271-293.
- Sweeting M. J., A. J. Sutton, P. C. Lambert (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; 23: 1351-1375.
- Ye J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of American Statistical Association*, Vol. 93, No. 441.

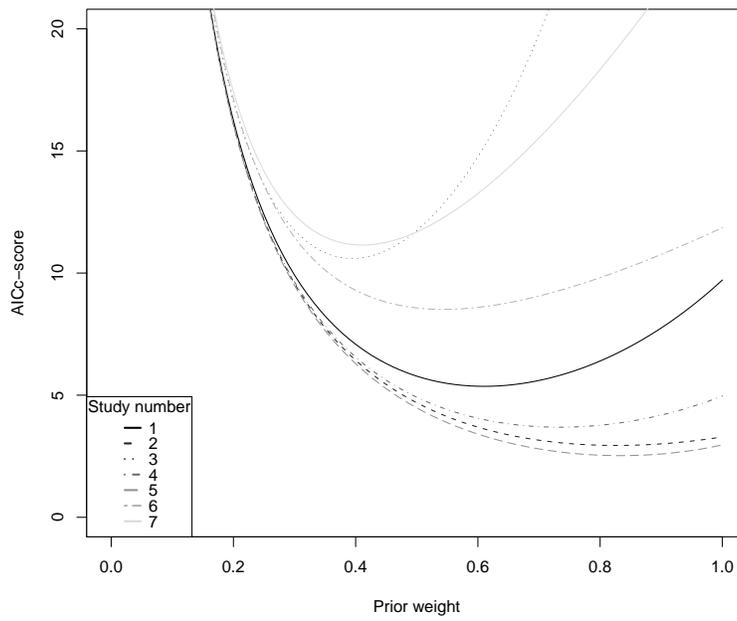


FIGURE 1.  $AIC_C$  scores of all seven studies with the split prior for the control group.

# Modelling space-time quantiles of ground-level ozone

Dana Draghicescu<sup>1</sup>

<sup>1</sup> Hunter College of the City University of New York, 695 Park Avenue, New York, NY 10065, USA, email: dana.draghicescu@hunter.cuny.edu

**Abstract:** Modelling effects of high order quantiles rather than mean effects is important for many environmental processes, since only higher values adversely affect human and environmental health. For this reason many air pollution standards issued by the United States Environmental Protection Agency are based on quantile estimates. In this paper we present a statistical framework for modelling quantiles of space-time processes assumed to be nearly stationary in time, isotropic in space, and space-time separable, and illustrate the proposed methodology on applications to ground-level ozone mixing ratios.

**Keywords:** kriging; nearly stationary processes; ozone; quantile maps; space-time modelling.

## 1 Introduction

There is an increasing interest in studying temporal and spatial variations of quantile functions. While there exists a large body of research on modelling and predicting trends (mean functions), comparatively little is known about spatio-temporal behavior of quantiles for environmental processes. For example, for air pollution, it is important to be able to model effects of high order quantiles rather than modelling mean effects, on the grounds that people are adversely affected only by very high levels of ozone, fine particles, or other pollutants. Maximum readings would give a more relevant statistic for monitoring than average values. However, high order quantiles are preferred to maximum values, in order to increase statistical stability. Medical information is typically collected at the ZIP-code level, whereas environmental data is mostly available at coarser spatial scales. For this reason, most studies focusing on the link between outdoor pollutants and acute or chronic respiratory illness use data aggregated spatially at city or county level for both pollution covariates and health outcomes. There is therefore a need for more accurate analyses, by using spatial scales at higher resolution. Guttorp (2000) recognizes the complexity of environmental data, and acknowledges the need for creating comprehensive descriptive tools and flexible statistical models to account for spatially heterogeneous

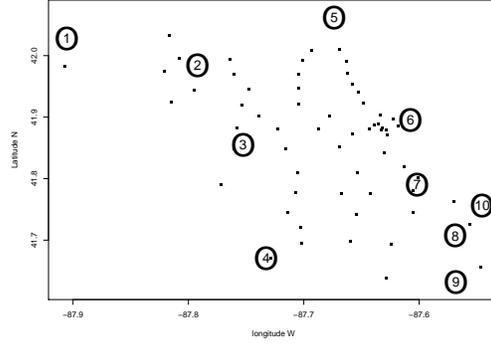


FIGURE 1. Locations of 10 air pollution monitors (ordered from East to West), and population centroids of 62 ZIP-codes in Chicago.

and temporally non-stationary processes. In this paper we propose a two-step procedure for modelling space-time quantiles in a wide class of processes. We combine smoothing in the time domain with spatial interpolation (universal kriging) to predict the quantile field of interest at locations with no observations, based on data with large temporal and relatively small spatial coverages. This method provides a fast, accurate, and informative exploratory tool, that can be used to describe various distributional characteristics, such as the center (medians), extremes (high or low quantiles), and spread (interquartile ranges).

## 2 Theoretical framework

Denote by  $X(t, s)$  a random field observed at  $n$  time points  $\{t_1, \dots, t_n\}$  and  $m$  spatial locations  $\{s_1, \dots, s_m\} \in D \subset \mathbf{R}^2$ . Typically  $m \ll n$ . For fixed location  $s \in D$ , and for rescaled time point  $t_i = \frac{i}{n}$ ,  $1 \leq i \leq n$ , we assume that

$$X(t_i, s) =: X_{i,n;s} = G_s(i/n; Z_i). \quad (1)$$

Here  $Z_i = (\varepsilon_i, \varepsilon_{i-1}, \dots)$ ,  $(\varepsilon_i)_{i \in \mathbf{Z}}$  are independent, identically distributed random variables, and  $G_s$  is a measurable function (unknown). Denote by

$$F_s(x; t) = P[G_s(t; Z_i) \leq x], \quad x \in \mathbf{R}, \quad 0 \leq t \leq 1 \quad (2)$$

the probability distribution function of the process  $(X_{i,n;s})_{i=1}^n$ , yielding the quantile function

$$q_\alpha(t, s) = \inf \{x : F_s(x; t) \geq \alpha\} \quad (3)$$

( $s \in D$  is fixed). If there exists a constant  $L < \infty$  such that for all  $0 \leq t, t' \leq 1$ ,

$$\sup_{x \in \mathbf{R}} |F_s(x; t) - F_s(x; t')| \leq L|t - t'|, \quad (4)$$

the process is called *nearly stationary* (Draghicescu, Guillas, Wu 2008). For rescaled time  $t_0$ , let  $n_1 = \lfloor n(t_0 - b_n) \rfloor$ ,  $n_2 = \lfloor n(t_0 + b_n) \rfloor$  and  $p = n_2 - n_1 + 1$ , where  $b_n$  is a sequence of bandwidths such that  $b_n \rightarrow 0$ ,  $nb_n \rightarrow \infty$  as  $n \rightarrow \infty$ . We can then estimate  $q_\alpha(t_0, s)$  by the sample quantile in this window,

$$\hat{q}_\alpha(t_0, s) = \inf \left\{ x : \frac{1}{p} \sum_{i=n_1}^{n_2} \mathbf{1}_{\{X_{i,n}; s \leq x\}} \geq \alpha \right\}. \quad (5)$$

As  $t$  changes from 0 to 1,  $\hat{q}_\alpha(t, s)$  may not be a continuous function of  $t$ , and thus an extra smoothing step is necessary. Consider the Nadaraya-Watson kernel estimator

$$\tilde{q}_\alpha(t, s) = \frac{\sum_{i=1}^n K\left(\frac{t - i/n}{h_n}\right) \hat{q}_\alpha(i/n, s)}{\sum_{i=1}^n K\left(\frac{t - i/n}{h_n}\right)}, \quad (6)$$

where  $h_n \rightarrow 0$  is another sequence of bandwidths such that  $nh_n \rightarrow \infty$ , and the kernel  $K$  is a nonnegative probability density function. Examples of nearly stationary processes, asymptotic properties of  $\tilde{q}$ , and a data-driven scheme for the selection of smoothing parameters are given in Draghicescu, Guillas, Wu (2008). The smooth physical evolution of many real life processes makes nearly stationarity a natural working assumption. In Figure 2 we display  $\tilde{q}_{0.9}(t, s)$  and  $iqr(t, s) := \tilde{q}_{0.75}(t, s) - \tilde{q}_{0.25}(t, s)$  for four ozone time series in Chicago (at sites 1, 2, 4, 7 showed in Figure 1). They were constructed based on formula (6), with the truncated Gaussian kernel for  $K$ . For details on kernel smoothing we refer to Simonoff (1998).

In the second step we use spatial interpolation (universal kriging), under the assumption of space-time separability of the observed process. We model the spatial covariances parametrically, by using the exponential covariance model

$$\text{cov}\left(\tilde{q}_\alpha(t, s_i), \tilde{q}_\alpha(t, s_j)\right) = \sigma_t^2 e^{-\theta_t \|s_i - s_j\|}, \quad (7)$$

where  $t, \alpha$  are fixed,  $\|s_i - s_j\|$  is the Euclidean distance between  $s_i$  and  $s_j$ ,  $\sigma_t^2$  is the variance, and  $\theta_t$  measures how fast the covariances decay with distance. Then, the best linear unbiased predictor (BLUP) of the quantile field at  $s_0 \in D$  is

$$q_\alpha^*(t, s_0) = \sum_{i=1}^m \lambda_i \tilde{q}_\alpha(t, s_i). \quad (8)$$

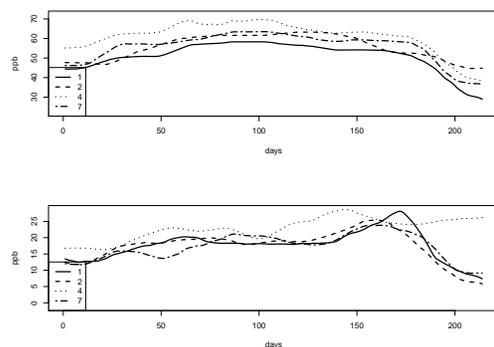


FIGURE 2. Moving window, smoothed 0.9 quantile curves (top), and moving window, smoothed interquartile range curves (bottom) of daily max-8hr-ave ozone (ppb) at monitoring sites 1, 2, 4, and 7, respectively; period April 1 – October 31, 1998.

The standard error of  $q_{\alpha}^*(t, s_0)$  can be also expressed in terms of the interpolation parameters  $\lambda_i$ ,  $1 \leq i \leq m$ , that are completely determined by  $\sigma_t^2$  and  $\theta_t$ . This procedure is known as *universal kriging*. For details we refer to Stein (1999). In practice the covariance parameters are estimated from the same data. To account for their uncertainty, the standard errors of  $q_{\alpha}^*(t, s_0)$  need to be adjusted. This can be done by using conditional simulation techniques (Stein 1999, Chapter 6) or resampling schemes (Lahiri 2003).

### 3 Quantile maps of ground level ozone

To illustrate this two-step procedure, we chose 10 monitoring sites in Chicago with complete hourly records of ground-level ozone mixing ratios (in parts per billion – ppb), period April 1 – October 31, 1998, source: Environmental Protection Agency (EPA). Their locations are displayed in Figure 1, together with the population centeroids of 62 ZIP-codes where we will predict the ozone quantile field (that can be then linked to time series of asthma-related outcomes). For each day we computed the 16 possible 8-hour averages (starting at 0:00, 1:00, ..., 15:00 for 8 consecutive hours), and retained the maximum of these 8-hour averages. This measure is part of the National Air Quality Standard (NAAQS) for ground-level ozone issued by the US EPA. Maps of estimated 0.9 quantiles and interquartile ranges are displayed in Figure 3, showing different spatio-temporal patterns. For instance, elevated air pollution as measured by the 0.9 quantiles of ozone

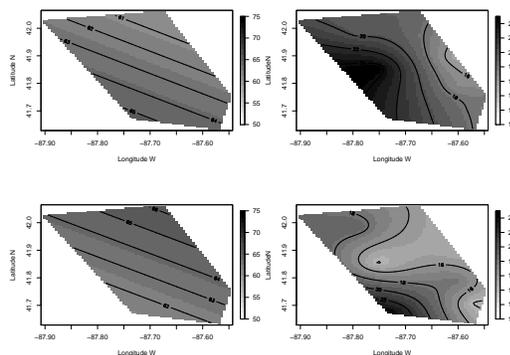


FIGURE 3. Map of 0.9 quantiles (left) and interquartile range (right) of ozone (in ppb) on June 1, 1998 (top), and on August 1, 1998 (bottom).

display an increasing trend towards South-East in June, whereas in August higher values tend towards North-West.

As mentioned in the previous section, the above spatial predictions are in fact EBLUP's (empirical or estimated BLUP's) since the covariance parameters are estimated from the same data. Therefore, the interpolation standard errors of  $q_{\alpha}^*(t, s_0)$  need to be corrected in order to account for this added uncertainty. Assuming Gaussian noise, we use a parametric simulation scheme (Stein 1999, Section 6.8) based on the exponential model (7), with 200 simulations carried out at each step. In Table 1 we display the interpolation standard errors, as well as the adjusted standard errors at the 10 ZIP-codes located at smallest distance from the monitoring sites, averaged over time (214 days from April 1 to October 31, 1998). In all cases the correction yielded slightly larger standard errors, as expected. The values of all the 0.9 ozone quantiles at the 10 monitoring sites are never higher than 75 ppb, with standard errors not exceeding 6 ppb. Chicago is known to be in attainment of the EPA regulations (one of the EPA thresholds for ground-level ozone is 80 ppb). Even though this application does not provide a breakthrough from a purely environmental or medical point of view, it shows the potential of quantile visualization as an exploratory tool. Additional information (such as urban setting, wind patterns) might give better insight on the different patterns in both higher quantiles, and variability of ground-level ozone. The computations were done in R, using the function `Krig`, library `Fields` (<http://www.image.ucar.edu/Software/Fields/>).

TABLE 1. Mean squared errors (averaged over time – 214 days) of the ozone 0.9 quantile estimates (ppb)

ZIP-code nearest site	interpolation error	adjusted error
1	4.47	5.73
2	2.81	2.97
3	3.95	4.29
4	5.02	5.64
5	4.95	6.29
6	3.02	3.24
7	2.95	3.49
8	1.02	3.24
9	1.93	2.21
10	2.07	3.48

## 4 Discussion

The increasing availability of high speed, inexpensive computing capabilities leads naturally to a growing demand for developing new flexible and informative statistical tools (such as summaries and graphs) for the exploration of large data sets with complex structures. The methodology proposed in this paper has a wide area of applicability in many other fields, such as atmospheric sciences, demography, ecology, epidemiology, finance, medicine, psychology.

**Acknowledgments:** This research is partially supported by the NSF Institutional Transformation grant number 40398-25 01.

## References

- Draghicescu, D., Guillas, S., and Wu, W.B. (2008). Quantile curve estimation and visualization for non-stationary time series. *Journal of Computational and Graphical Statistics* (to appear).
- Guttorp, P. (2000). Environmental statistics. *Journal of the American Statistical Association*, **95**(449), 289-292.
- Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer.
- Simonoff, J. S. (1998). *Smoothing Methods in Statistics*. Springer.
- Stein, M. L. (1999). *Interpolation of Spatial Data. Some Theory of Kriging*. Springer.

# A hidden semi-Markov model for the occurrences of water pipe bursts

T. Economou<sup>1</sup>, T.C. Bailey<sup>1</sup> and Z. Kapelan<sup>1</sup>

<sup>1</sup> School of Engineering, Computer Science and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, UK

**Abstract:** A frequently used approach when modelling the bursts of underground water pipes is to assume a non-homogeneous Poisson process (NHPP) for the occurrence of failures. This however does not account for possible serial dependence in the failures or that the occurrence of failures may also be affected by some temporal process other than ageing. This paper proposes a hidden semi-Markov model which is NHPP conditional on the states of the hidden process.

**Keywords:** Hidden semi-Markov; NHPP; Censoring; Truncation.

## 1 Introduction

Water companies (especially in the UK) have a need for proper and accurate estimations on water pipe bursts or blockages. In addition to the fact that processes driving the occurrences of pipe failures are complex and often unmeasurable, the available historical data is often scarce and unreliable. The statistical models employed would then need to be complex and flexible enough to capture the failure process. It is common practice to consider that the occurrences of pipe failures are generated by a point process in time, specifically a (possibly nonhomogeneous) Poisson process (Kleiner and Rajani, 2001). Recently Economou et al. (2007) considered an aggregated nonhomogeneous Poisson process (NHPP) model with an intensity function  $\lambda(t, \mathbf{x})$  (essentially the failure rate) dependent on time  $t$  and covariates  $\mathbf{x}$ . They further extended the model to account for possible zero-inflation in their pipe burst data. Although the models adequately captured the ageing process and accurately predicted the total number of failures in the network, they performed poorly in predicting at the individual pipe level. This could be because there was nothing in the proposed models to account for serial dependence and possible unobserved covariates or even processes that might have influenced the burst occurrence. In this paper we consider how aspects like these can be incorporated into such models.

Water pipes are degradable components of an ageing system whose failure mechanism may involve several ‘states’ relating to the ‘health’ of the pipe. This leads to the idea of incorporating a hidden Markov process to the

NHPP model such that the intensity function  $\lambda(t, \mathbf{x})$  is different according to which state the hidden process is in at time  $t$ . By doing this, the resulting process that generates the failures allows for both serial dependence and overdispersion (MacDonald and Zucchini, 1997). In a Markov process, the times between states are exponentially distributed and in the case of water pipe failures this may well be unrealistic. A semi-Markov process is essentially a Markov process with temporal structures and this not only makes the above model more flexible, but it also allows explicit modelling of the duration time between states (Dong and He, 2007). Following the thinking of Özekici (2003), the hidden process can be seen as an environmental process giving rise to variations in the parameters of the model. One could then say that model for the pipe failures includes time dependent random effects. The formulation of the model is presented in section 2 and model application in section 3.

## 2 Model Specification

Consider a single water pipe  $k$  which has been observed in the time interval  $[t_0, t_{end}]$  and has failed  $n$  times at  $t_1, t_2, \dots, t_n$  ( $t_0 = 0$  implies that observation started at the installation date of the pipe). Suppose that the occurrences of these failures occur as a NHPP that depends on a hidden state  $S_t$  of a semi-Markov process where  $S_t \in \{1, 2, \dots, M\}$  is the state space of the process. Here  $S_t$  is interpreted as the state of the process at time  $t$ . The semi-Markov process is basically defined by an initial distribution  $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_M)$ , a transition probability matrix  $\mathbf{P} = \{p_{i,m}\}$  and a matrix of holding times  $\mathbf{H}(\mathbf{t}) = \{f_{i,m}(t)\}$ . If at time  $t$  the process is in state  $i$ , it will decide to move to state  $m$  at time  $(t + s)$  with probability  $p_{i,m}$  and it will do so after holding for a time period with a density function  $f_{i,m}(s)$ . Here we assume that the resulting process occurs in discrete time, mainly due to the nature of pipe failure data but also because it somewhat reduces the complexity of the model.

For the characterization of the intensity function  $\lambda(t, \mathbf{x})$  of the NHPP we adopt a power function of time:

$$\lambda_k(t, \mathbf{x} | S_t) = \gamma_k \theta^{(S_t)} t^{(\theta^{(S_t)} - 1)} \exp \left\{ \boldsymbol{\beta}^{(S_t)} \mathbf{x} \right\}$$

So given the state  $S_t$  the process is NHPP with intensity function  $\lambda(t, \mathbf{x})$  which depends on state specific parameters  $\theta$  and  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_q)'$  corresponding respectively to the shape parameter and the parameters relating to possible explanatory variables  $\mathbf{x} = (x_1, x_2, \dots, x_q)$ . The scale parameter  $\gamma_k$  is pipe specific and thus constitutes a random effect to account for the between pipe variability. It is worth mentioning here that whilst the process conditional on  $S_t$  is a NHPP, the overall failure process risen from this model is not.

Note that from now on, the subscript  $k$  will be dropped for clarity until the results are generalized to more than one pipe. Now, in a NHPP the conditional distribution of  $t_j$ , the time of the  $j^{\text{th}}$  failure given the state and that the previous failure occurred at  $t_{j-1}$  is

$$\begin{aligned} h(t_j|t_{j-1}, S_{t_j}) &= \lambda(t_j|S_{t_j}) \exp \left[ - \int_{t_{j-1}}^{t_j} \lambda(u|S_{t_j}) du \right] \\ &= \lambda(t_j|S_{t_j}) e^{-\Lambda([t_{j-1}, t_j]|S_{t_j})} \end{aligned} \quad (1)$$

A second thing that needs to be considered in order to compute the likelihood of the model is the state sequence of the semi-Markov process  $\{S_{t_0}, S_{t_1}, \dots, S_{t_n}\}$ , i.e. the state at the start of the observation period, the state at the time of the first failure  $t_1$ , the state at  $t_2$  and so on. Consider a specific realization of this sequence  $\{S_{t_0} = z_0, S_{t_1} = z_1, \dots, S_{t_n} = z_n\}$  where  $z_0, \dots, z_n \in \{1, 2, \dots, M\}$ . The probability associated with this sequence then is

$$\begin{aligned} &\pi_{z_0} p_{z_0, z_1} f_{z_0, z_1}(t_1 - t_0) \cdots p_{z_{n-1}, z_n} f_{z_{n-1}, z_n}(t_n - t_{n-1}) \\ &= \pi_{z_0} \prod_{j=1}^n p_{z_{j-1}, z_j} f_{z_{j-1}, z_j}(t_j - t_{j-1}) \end{aligned} \quad (2)$$

Here,  $f_{z_{j-1}, z_j}$  is assumed to follow a negative binomial distribution. Conditional on (2), the joint probability distribution of the data can be computed using (1):

$$\begin{aligned} &\lambda(t_1|S_{t_1} = z_1) e^{-\Lambda([t_0, t_1]|S_{t_1} = z_1)} \cdots \lambda(t_n|S_{t_n} = z_n) e^{-\Lambda([t_{n-1}, t_n]|S_{t_n} = z_n)} \\ &= \prod_{j=1}^n h(t_j|t_{j-1}, S_{t_j} = z_j) \end{aligned} \quad (3)$$

The multiplication of equations (2) and (3) will result in the likelihood of the data conditional on the fact that  $\{S_{t_0} = z_0, \dots, S_{t_n} = z_n\}$  is known. Since this is not the case, to define the likelihood of the model we also need to sum over all possible values of the states. In other words, the likelihood of the model is:

$$L(\cdot) = \sum_{z_0=1}^M \sum_{z_1=1}^M \cdots \sum_{z_n=1}^M \left[ \pi_{z_0} \prod_{j=1}^n p_{z_{j-1}, z_j} f_{z_{j-1}, z_j}(t_j - t_{j-1}) h(t_j|t_{j-1}, S_{t_j}) \right]$$

Buried in the likelihood equation above is the assumption that the data is failure truncated, i.e.  $t_n = t_{end}$ . The situation where  $t_n < t_{end}$  is referred to as time truncation and if so, the probability of no failures occurring in  $[t_n, t_{end}]$  needs to be incorporated in the likelihood. In an NHPP, the number of events in a time period  $T$  is Poisson distributed (Meeker and

Escobar, 1998) with mean  $\int_T \lambda(t) dt$ . So to account for time truncation in  $L(\cdot)$  one also needs to also sum over  $\sum_{z_{end}=1}^M$  and multiply the product in the sums by  $\exp\{-\Lambda([t_n, t_{end}]|S_{t_{end}})\}$ .

The second modification to the likelihood that should be considered is the case when the data is left censored meaning that we started observing the pipe after its installation date, so  $t_0 = \tau > 0$ . In this case, the distribution of the time to the first failure  $h(t_1|t_0 = 0)$  should be replaced by  $h(t_1|t_0 > 0)$ . Formally, we need  $\Pr(\tau < t_1 \leq \infty)$  which can be computed by dividing the density function of  $t_1$  by  $1 - H(\tau)$  where  $H$  is the distribution function of  $t_1$ :

$$\begin{aligned} 1 - H(\tau) &= \int_{\tau}^{\infty} \lambda(t_1) e^{-\int_0^{t_1} \lambda(u) du} dt_1 = \int_{\tau}^{\infty} \gamma \theta t_1^{(\theta-1)} e^{\beta x} \exp\{-\gamma t_1^{\theta} e^{\beta x}\} dt_1 \\ &= [-\exp\{-\gamma t_1^{\theta}\}]_{\tau}^{\infty} = \exp\{-\gamma \tau^{\theta}\} = c(\tau) \end{aligned}$$

This implies that dividing (3) by  $c(t_0|S_{t_1} = z_1)$  one can obtain the 'left censored' likelihood  $L^*(\cdot)$ . Considering now the situation when no failures have been observed, it is fairly easy to see that the contribution to the likelihood will be

$$L^0 = \sum_{z_0=1}^M \sum_{z_{end}=1}^M \pi_{z_0} p_{z_0, z_{end}} f_{z_0, z_{end}} \exp\{-\Lambda([t_0, t_{end}]|S_{t_0} = z_{end})\}$$

Reintroducing the pipe subscript  $k$  and assuming that we have  $N$  independent pipes that failed  $n_k$  times each, the overall likelihood of the model is

$$\prod_{k=1}^N [\delta_k L_k^* + (1 - \delta_k) L_k^0] \quad (4)$$

where  $\delta_k = 0$  if  $n_k = 0$  and is equal to 1 otherwise.

### 3 Model Application

For the sake of neatly deriving and writing the likelihood in (4), we forced the transitions of the hidden chain to actually happen at each failure. This of course is not realistic as transitions could happen at any time, a thing which is taken into account when estimating the parameters. In addition, the likelihood in (4) is very complex to even compute which is why recursive algorithms were used to fit the model. These are not mentioned here due to lack of space.

The model is currently being applied to a Canadian distribution network of water pipes. This network consists of 1349 pipes with 5425 recorded failures in the period 1945-2003. The model is implemented within the Bayesian context using MCMC methods.

## References

- Dong, M. and He, D. (2007). A segmental hidden semi-markov model (hsmm)-based diagnostics and prognostics framework and methodology. *Mechanical systems and signal processing*, **21**, 2248-2266.
- Economou, T., Bailey, T. and Kapelan, Z. (2007). Bayesian modelling of time aggregated water pipe bursts with a zero-inflated, non-homogeneous Poisson process. *Proceedings of the 22nd international workshop on statistical modelling*, 227-232.
- Kleiner, Y. and Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water*, **3**, 131-150.
- MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall.
- Meeker, W. and Escobar, L. 1998. *Statistical methods for reliability data*. John Wiley and Sons Inc., New York.
- Ozekici, S. and Soyer, R. (2003). Bayesian analysis of Markov modulated Bernoulli processes. *Mathematical methods of operations research*, **57**, 125-140.

# On parsimonious higher-order binary Markov chain models

Anna Espinal<sup>1</sup> and Pedro Puig<sup>2</sup>

<sup>1</sup> Servei d'Estadística, Universitat Autònoma de Barcelona.

<sup>2</sup> Departament de Matemàtiques, Universitat Autònoma de Barcelona.

**Abstract:** We present two examples of higher-order binary Markov chain series where the number of parameters can be considerably reduced.

**Keywords:** Discrete Time Series; Mixture Transition Distribution models; Markov regression models

## 1 Higher-order Markov chains

The Markov chain is a well known probabilistic model used to explore dependences between successive observations. The usual first order Markov chain model considers that, given the present, the future is conditionally independent of the past. However, in some situations the future depends not only on the present but on the last observations.

Consider a discrete-time random variable  $X_t$  taking values in a finite set of states  $1, 2, \dots, m$ . We have a  $k$ -order Markov chain if

$$P(X_t = i_t | X_0 = i_0, X_1 = i_1, \dots, X_{t-1} = i_{t-1}) =$$

$$P(X_t = i_t | X_{t-k} = i_{t-k}, \dots, X_{t-2} = i_{t-2}, X_{t-1} = i_{t-1}),$$

where  $i_0, i_1, \dots, i_t \in \{1, 2, \dots, m\}$ . Each  $k$ -order Markov chain model is completely characterized by its transition probabilities. For instance, for  $k = 2$  and  $m = 3$  these transition probabilities can be arranged according to the following matrix form:

		$X_t$		
$X_{t-2}$	$X_{t-1}$	1	2	3
1	1	$p_{111}$	$p_{112}$	$1 - p_{111} - p_{112}$
1	2	$p_{121}$	$p_{122}$	$1 - p_{121} - p_{122}$
1	3	$p_{131}$	$p_{132}$	$1 - p_{131} - p_{132}$
2	1	$p_{211}$	$p_{212}$	$1 - p_{211} - p_{212}$
2	2	$p_{221}$	$p_{222}$	$1 - p_{221} - p_{222}$
2	3	$p_{231}$	$p_{232}$	$1 - p_{231} - p_{232}$
3	1	$p_{311}$	$p_{312}$	$1 - p_{311} - p_{312}$
3	2	$p_{321}$	$p_{322}$	$1 - p_{321} - p_{322}$
3	3	$p_{331}$	$p_{332}$	$1 - p_{331} - p_{332}$

Consequently this model has 18 parameters to be estimated. Unfortunately the number of parameters of these fully parameterized models increases very rapidly: a  $k$ -order Markov chain model with  $m$  states has  $m^k(m-1)$  parameters. The Mixture Transition Distribution (MTD) models (Raftery, 1985, Berchtold and Raftery, 2002) are higher order Markov chains with fewer parameters than the fully parameterized models. However the meaning of these parameters is not always clear and the MDT models can not be fitted by using standard statistical packages. The Markov regression models of Zeger and Qaqish (1988) are another way to reduce the number of parameters. In the following examples we show how the number of parameters can be reduced taking into account simple concepts of "short" and "long" time memory using a special kind of Markov regression models. We have used standard statistical packages for fitting these models.

## 2 Short-time + Long-time binary memory models

Given a zero-one  $k$ -order Markov chain model (here  $m = 2$ ), we define  $S_r(t) = \sum_{i=1}^r X_{t-i}$  and  $L_r(t) = \sum_{i=r+1}^k X_{t-i}$  when  $r < k$ , and  $S_k(t) = \sum_{i=1}^k X_{t-i}$  and  $L_k(t) = 0$ . A Short-Long-time memory Markov chain model of degree  $r$  (SLMC $_k(r)$ ) can be defined by the relation,

$$P(X_t = i_t | X_0 = i_0, X_1 = i_1, \dots, X_{t-1} = i_{t-1}) =$$

$$P(X_t = i_t | S_r(t) = j, L_r(t) = l)$$

where  $i_0, i_1, \dots, i_t \in \{0, 1\}$ . Here  $S_r(t)$  and  $L_r(t)$  can be understood as short-time and long-time memory terms respectively.

For simplicity we denote  $P(X_t = 1 | S_r(t) = i, L_r(t) = j) = p_{ij}$ . It is straightforward to see that the number of parameters for a SLMC $_k(r)$  is  $(k+1-r)(r+1)$ . Moreover the number of parameters can also be reduced by considering logistic linear relations between the  $p_{ij}$ 's. For instance, for the SLMC $_k(k)$ , a model without long-time memory term, we could consider  $P(X_t = 1 | S_k(t) = j) = p_j$ , for  $j = 0, 1, \dots, k$ , and  $\log(p_j/(1-p_j)) = \alpha + \beta \cdot j$ . This is a model with only two parameters. These models will be indicated with an asterisk. All these models can be fitted by using logistic regression procedures.

### 2.1 Example 1: epileptic data

This data set of MacDonald and Zucchini (1997) is analyzed in Berchtold and Raftery (2002) by using fully parameterized and MTD models. Here we have a binary time series describing for each day whether a specific patient had at least one epileptic seizure (1) or not (0):

```
0100001101111001111110110111111101011111000111010101000000001
0000010001000100010010011011000110111111101111111011101100111
```

01001010001001000000011000001011100000101110000001100000000000  
 00000000000000000000

Table 1 displays the Bayes information criteria (BIC) for several fitted models with different orders. Thus we have fitted five full parameterized Markov chain models, MC1 to MC5, where the number indicates the order of the model. Notice that MC4 is the full parameterized model that has presented the smallest BIC. Table 1 also includes the results corresponding to the  $SLMC_k(k)$  models ( $k = 1, 2, \dots, 5$ ) in two different approaches: using  $S_k(t)$  as an ordinal factor (these has been indicated as SLMC) and using  $S_k(t)$  as a continuous covariate in a logistic regression model (indicated as SLMC\*). The best model according to the BIC statistic has been, in both cases, the 4-order model.

TABLE 1. Higher-order Markov Chain Models for the epileptic data. The model with the best BIC is shown in bold.

Model	n. parameters	LL	BIC
MC1	2	-122.569	255.632
MC2	4	-119.331	259.650
MC3	8	-117.161	276.298
MC4	16	-111.489	306.930
MC5	32	-101.907	371.719
$SLMC_2(2)$	3	-119.394	254.529
$SLMC_3(3)$	4	-117.354	255.695
$SLMC_4(4)$	5	-113.283	252.802
$SLMC_5(5)$	6	-115.607	262.697
$SLMC_2^*(2)$	2	-119.564	249.622
$SLMC_3^*(3)$	2	-117.815	246.123
<b><math>SLMC_4^*(4)</math></b>	<b>2</b>	<b>-113.814</b>	<b>238.123</b>
$SLMC_5^*(5)$	2	-116.652	243.799

It is interesting to remark that, for this data set, the best MTD model obtained by Berchtold and Raftery (2002) was of order 8 with 9 parameters and BIC=251.9. Figure 1 shows the predicted probabilities using our best model ( $SLMC_4^*(4)$ ), and the corresponding observed probabilities obtained by cross tabulation. We can observe a quite good performance. Consequently, it seems that in a specific moment of the series, the probability to have an epileptic seizure depends of the total number of epileptic seizures observed in the last four days.

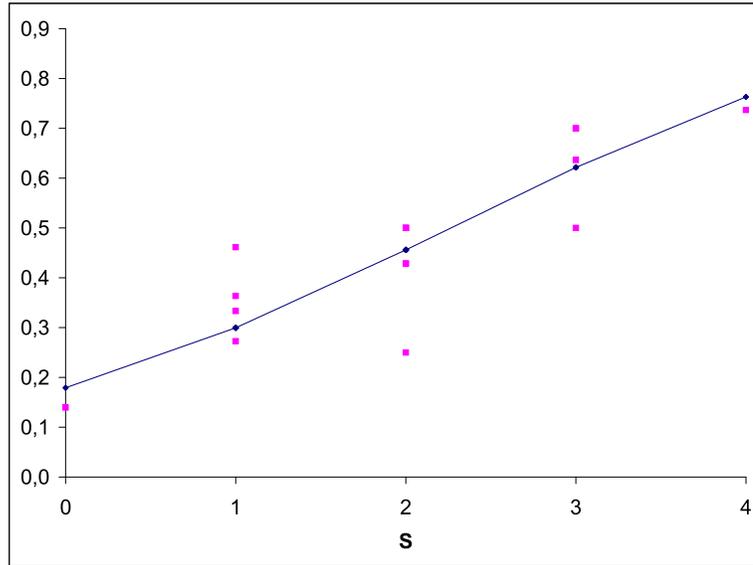


FIGURE 1. Predicted and observed probabilities for  $X_t = 1$  as a function of  $S = X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4}$ .

## 2.2 Example 2: randomness perception

Nineteen students of Political Sciences were asked to simulate a perfect coin. Each student was asked to produce 50 responses of head and tails. It is known that people have the mistaken belief that a coin alternates from heads to tails and back again more often than it really does. Memory for previous responses is difficult to eradicate and it produces a dependent series of head and tails. Our goal is to study how the memory mechanism of the individuals acts in order to produce these series.

Budescu (1987) proposed that human randomizing behavior could be modeled as a Markov chain. However, using again the BIC statistic to select a model, we obtain that the  $SLMC_5^*(1)$  models have a smaller BIC values than the corresponding full parameterized models up to order 5 for most individuals (13 of 19). For these models, the long-time memory has been considered a continuous covariate and consequently we have 3 parameters. Table 2 shows the BIC values for these individuals. Specifically, we finally establish for each individual a model where,

$$P(X_t = 1 | X_{t-1} = i, X_{t-2} + X_{t-3} + X_{t-4} + X_{t-5} = j) = p_{ij}$$

$$\log(p_{ij}/(1 - p_{ij})) = \alpha_i + \beta j, \quad i = 0, 1, \quad j = 0, 1, 2, 3, 4$$

TABLE 2. BIC statistics for the  $SLMC_5^*(1)$  models and for the best full parameterized MC models

	IND3	IND4	IND5	IND6	IND8	IND9	IND11
BIC	65.13	68.00	70.28	70.57	71.43	37.28	69.99
best BIC	69.43	73.15	72.21	72.11	72.93	39.90	73.19
best MC	MC1	MC0	MC1	MC1	MC1	MC1	MC1
	IND12	IND13	IND14	IND15	IND17	IND19	
BIC	68.56	66.53	61.73	62.03	70.20	69.34	
best BIC	73.15	73.15	64.42	72.21	72.50	70.32	
best MC	MC0	MC0	MC1	MC1	MC0	MC0	

Consequently, it seems that in a specific moment of the series, the head or tail choice depends on the last choice (short-time memory) and the memory of the last five choices (long-time). More or less, it corresponds to the following idea: "People remember what just doing and have a vague idea of which they did previously".

**Acknowledgments:** The authors would like to thank Isabel Serra from the Department of Mathematics, Universitat Autònoma de Barcelona, for collect the data analyzed in Example 2. This research was partially supported by grant MTM2006-01477 from the Ministry of Education of Spain.

## References

- Berchtold, A. and Raftery, A.E. (2002). The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series. *Statistical Science*, **17**, 328-356.
- Budescu, D.V. (1987). A Markov Model for Generation of Random Binary Sequences. *Journal of Experimental Psychology*, **13**, 25-39.
- MacDonald, I.L., and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.
- Raftery, A.E. (1985). A Model for High-Order Markov Chains. *Journal of the Royal Statistical Society. Series B*, **47**, 528-539.
- Zeger, S.L. and Qaqish, B. (1988). Markov Regression Models for Time Series: A Quasi-Likelihood Approach. *Biometrics*, **44**, 1019-1031.

# Nonparametric Principal Components Analysis for Ecological Data

C. Ferguson<sup>1</sup> and A. Bowman<sup>1</sup>

<sup>1</sup> Department of Statistics, 15 University Gardens, University of Glasgow, G12 8QW, Email: [claire@stats.gla.ac.uk](mailto:claire@stats.gla.ac.uk)

**Abstract:** Smooth principal components analysis will be used to assess whether the covariance structure in an ecological system is changing throughout the year. These techniques will be applied to the system at Loch Leven, Scotland.

**Keywords:** Loch Leven; Nonparametric; Principal Components Analysis.

## 1 Introduction

Principal components analysis (PCA) is a widely used dimension reduction technique for multivariate data (see, for example, Jolliffe 2002). In particular, for environmental and ecological data it is often used to reduce dimensionality of species but also in its dynamic form to consider how time series evolve over time. For example, it can be used to identify common trends (Zuur *et al.* 2003a & 2003b).

Ecological systems comprise physical, chemical and biological variables which are generally interrelated and consequently many of the variables are potentially both responses and covariates within the system. It is known that relationships between these variables can change throughout time but it is also of interest to investigate if the covariance structure between them also changes throughout time. Such assessments are particularly important for legislation such as the European Community (EC) Water Framework Directive (WFD) 2000, which states that good ecological status should be achieved in all lakes by 2015. Whether such targets are achievable depend on a thorough understanding of all processes and relationships at work within such a system.

It is therefore of interest to consider approaches to assess if the covariance structure in an ecological system is changing throughout time. This paper considers a nonparametric time dependent principal components analysis (smooth PCA) and explores its application to the ecological system at Loch Leven, Scotland.

## 2 Smooth Principal Components Analysis

To assess if the covariance structure for related ecological responses is changing throughout time PCA could be performed at each time point  $t$ . However, if time  $t$  does not coincide with a data point or the data are not replicated at each time point then it would not be possible to construct the covariance matrix. Nonparametric time dependent principal components analysis (Prvan & Bowman, 2003) consists of performing PCA at each time  $t$  considered using a smoothed covariance matrix, where the amount of neighbouring data that contributes is controlled by the choice of the smoothing parameter. A local covariance matrix can be constructed as:

$$A_w(t) = \sum_{i=1}^n w_i(t)(\mathbf{y}_i - \bar{\mathbf{y}}_w(t))(\mathbf{y}_i - \bar{\mathbf{y}}_w(t))^T$$

for each time point  $t$  (over a grid of points). In this paper, it is of interest to assess if the covariance structure is changing throughout the year and hence the time points of interest are months of the year. The weights  $w_i(t)$ , are therefore generated using a circular weight function defined as:

$$w_i(t_i - t, h) = e^{\frac{1}{h} \cos(2\pi \frac{(t_i - t)}{r})} \quad (1)$$

where  $r = 12$  (corresponding to months in a year),  $t$  is a particular time point and the smoothing parameter  $h$  is determined by specifying the degrees of freedom (df) required, where  $\text{df} = \text{tr}(S)$ , and  $S$  is a smoothing matrix based on local constant smoothing with the above weights. The estimate at a particular time point is constructed as  $\bar{\mathbf{y}}_w(t) = S\mathbf{y}$  and is therefore a smooth function of month here. However, the mean function could be generalised to incorporate both trend and seasonality or other covariates as required. With environmental data, it is common for variables to be on a different scale and hence data may also have to be standardised before implementing PCA.

## 3 Inference

It is of interest to assess formally if the covariance matrix remains constant over time. Therefore, the reference model of interest is

$$H_0 : \Sigma(t) = \Sigma$$

Investigation of the sampling properties of the coefficients and variances of the sample principal components (PCs) is equivalent to looking at the sampling properties of the eigenvalues and eigenvectors (Jolliffe, 2002). Under the null hypothesis the covariance structure between response variables does not change over time and Schott (1991) considers one way to assess

this by providing test statistics for common principal component subspaces across several groups based on eigenprojections, where each group would correspond to a particular month in this context. Alternatively, Anderson (1963) and Kollo & Neudecker (1993) state asymptotic results for eigenvalues and normalised eigenvectors of sample variance and correlation matrices. However, it is difficult to see how such results can be modified to incorporate the general weight matrix required here for different response variables.

Therefore, to investigate if there is a relationship with time, reference bands can be constructed to illustrate the direction that would be expected for a particular principal component if the null hypothesis is true. In this paper the main focus will be on principal component one and hence the direction of maximum variability.

Reference bands can be constructed by permuting the variable for time and performing smooth PCA for each new sample. However, environmental and ecological time series often possess autocorrelation, and this approach does not account for this.

Alternative approaches for inference include creating the reference band based on block bootstrapping of time series, to retain autocorrelation, and simulating data from a fitted model with known covariance structure to provide a reference distribution. Bootstrapping can also be used to produce approximate distributions for eigenvalues and eigenvectors, to investigate specific changes in the variance or coefficients of PCs.

## 4 Loch Leven Application

Loch Leven (Scotland) is the largest shallow, eutrophic lake in the UK and it has been monitored by the Centre for Ecology & Hydrology (CEH) in Edinburgh since 1967. The variables measured at the loch cover the chemistry, biology and meteorology with key variables of chlorophyll<sub>a</sub> (phytoplankton biomass), phosphorus, and zooplankton (*Daphnia*, water fleas which graze on phytoplankton).

Monthly mean data from Loch Leven for the period January 1988 to December 2002 will be considered here as a result of changing regulations and substantive missing periods. The relationship between chlorophyll<sub>a</sub> (chl<sub>a</sub>) and Total Phosphorus (TP) throughout the year will be presented initially. However, Soluble Reactive Phosphorus (SRP) and *Daphnia* (Daph) are also related responses, see Figure 1, and techniques can be extended to incorporate these or other related variables. Figure 2 displays the seasonal pattern for each variable with a smoothed mean calculated using local constant smoothing with a circular weight function, equation (1). It highlights a strong seasonal pattern for log TP with a weaker signal evident for log chlorophyll<sub>a</sub> and it is of interest to assess if the relationship between these variables changes throughout the year after accounting for this mean

structure. A natural log transform has been applied to each variable before standardising to reduce variability and subsequently a smooth PCA with 4 degrees of freedom was performed on this data to assess if the covariance structure changes throughout the year.

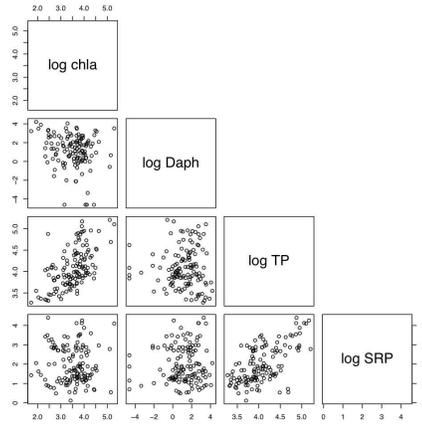


FIGURE 1. Matrix plot of log transformed variables for the Loch Leven data.

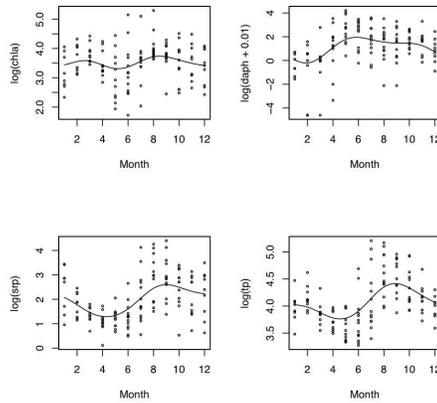


FIGURE 2. Time series plots, over months of the year, for variables measured at Loch Leven with a smoothed mean.

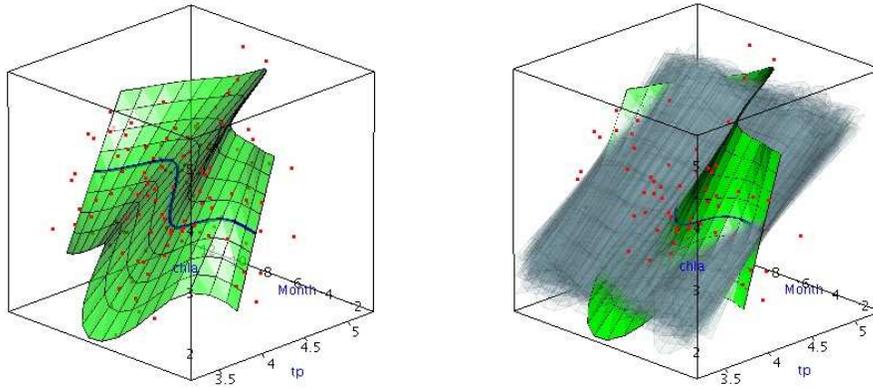


FIGURE 3. (Left) 3d plot of log transformed variables for chlorophyll<sub>a</sub> (vertical axis) and TP against month with surface to display the direction of principal component one in each month. (Right) 3d plot of log transformed variables for chlorophyll<sub>a</sub> (vertical axis) and TP against month with surface to display the direction of principal component one in each month with additional reference band to assess changing covariance.

## 5 Results, Conclusions and Future Work

Figure 3 (left) displays a 3d plot of the log data for chlorophyll<sub>a</sub> and TP over the months of the year with a surface displaying the direction of the first principal component produced for each month, which indicates the direction of maximum variability. The plot highlights a sharply positive relationship in January, which flattens out throughout the year before returning to sharply positive in December.

Since TP is used in the production of chlorophyll<sub>a</sub> this highlights the strong relationship in winter months. However, moving through the year both of these variables become affected by and affect other variables, such as *Daphnia*, changing this relationship and hence the direction of variability. Therefore, it appears that the covariance structure between these two variables is changing throughout the year. Figure 3 (right) is a repeat of the plot on the left with a reference band (currently ignoring temporal correlation) included to illustrate the direction of maximal variability if the covariance structure does not change throughout time. This was created by producing samples of data based on permuting the months of the year and performing smooth PCA for each sample. This highlights that the covari-

ance structure is changing throughout the year since the direction of the PCs lie outwith the reference band for months early and late in the year. This paper considers methods to assess whether the covariance structure changes throughout the year between two ecological variables. Such methods can be extended to consider relationships between  $p$  variables throughout the year and more generally over time. The local covariance matrices produced at each time point, the autocorrelation in individual response variables and the asymptotic nature of existing distributional results mean that bootstrapping techniques provide the most promising route for inferential methods.

**Acknowledgments:** The authors gratefully acknowledge Laurence Carvalho and his colleagues at the Centre for Ecology & Hydrology for assistance with collaborations involving these data, Loch Leven Estates for providing CEH with access to the loch and assistance with fieldwork over the years and Loch Leven Estates Data providers for *Daphnia* data.

## References

- Anderson, T.W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*. **34**, 1, 122-148.
- Jolliffe, I.T. (2002). *Principal Component Analysis*. New York: Springer-Verlag.
- Kollo, T. & Neudecker, H. (1993) Asymptotics of Eigenvalues and Unit-Length Eigenvectors of Sample Variance and Correlation Matrices. *Journal of Multivariate Analysis*. **47**, 283-300.
- Prvan, T. & Bowman, A. (2003). Nonparametric time dependent principal components analysis. *Anziam J.* **44** (E), 627-643.
- Schott, J.R. (1991). Some Tests for Common Principal Component Subspaces in Several Groups. *Biometrika*, **78**, 4, 771-777.
- Zuur, A.F., Fryer, R.J., Jolliffe, I.T., Dekker, R. and Beukema, J.J. (2003a). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics*. **14**, 665-685.
- Zuur, A.F., Tuck, I.D. and Bailey, N. (2003b). Dynamic factor analysis to estimate common trends in fisheries time series. *Can. J. Fish. Aquat. Sci.* **60**, 542-552.

# Modelling Individual Animal Growth in Random Environments

Patrícia A. Filipe<sup>1</sup> and Carlos A. Braumann<sup>1</sup>

<sup>1</sup> Universidade de Évora, Centro de Investigação em Matemática e Aplicações  
Rua Romão Ramalho, 59, 7000-671 Évora, Portugal  
pasf@uevora.pt, braumann@uevora.pt

**Abstract:** We have considered, as general models for the evolution of animal size in a random environment, stochastic differential equations of the form  $dY(t) = b(A - Y(t))dt + \sigma dW(t)$ , where  $Y(t) = g(X(t))$ ,  $X(t)$  is the size of an animal at time  $t$ ,  $g$  is a strictly increasing function,  $A = g(a)$  where  $a$  is the asymptotic size,  $\sigma$  measures the effect of random environmental fluctuations on growth, and  $W_t$  is the Wiener process. We have considered the stochastic Bertalanffy-Richards model ( $g(x) = x^c$  with  $c > 0$ ) and the stochastic Gompertz model ( $g(x) = \ln x$ ). We have studied the problems of parameter estimation for one path and also considered the extension to several paths. We also used bootstrap methods. Results and methods are illustrated using bovine growth data.

**Keywords:** growth models; stochastic differential equations; estimation; cattle weight.

## 1 Introduction

The most common models used to describe the growth of an individual animal in terms of its size  $X(t)$  at time  $t$  have assumed the form of a differential equation  $dY(t) = b(A - Y(t))dt$ ,  $Y(t_0) = y_0$ , where we made a change of variable  $Y(t) = g(X(t))$  with  $g$  a strictly increasing function (which we assume known). We have  $y_0 = g(x_0)$  and  $A = g(a)$ , where  $x_0$  is the size at birth and  $a$  is the asymptotic size or size at maturity of the animal. The parameter  $b > 0$  is a rate of approach to maturity.

The *Bertalanffy-Richards model* (Bertalanffy (1957) and Richards (1959)) corresponds to the choice  $g(x) = x^c$  for  $c > 0$  (typical choices are  $c = 1$  and  $c = 1/3$ ) and the *Gompertz model* corresponds to  $g(x) = \ln x$  (can be considered the limiting case of Bertalanffy-Richards model when  $c \rightarrow 0$ ).

If the animal is growing in a randomly fluctuating environment, we can model growth through a *stochastic differential equation* (SDE) of the form

$$dY(t) = b(A - Y(t))dt + \sigma dW(t), \quad (1)$$

where  $\sigma > 0$  measures the strength of environmental fluctuations and  $W(t)$  is a standard Wiener process. Garcia (1983) applied these type of models to tree growth.

The solution of (1) is a homogeneous diffusion process with drift and diffusion coefficient, respectively,  $b(A - y)$  and  $\sigma^2$ . The solution of this SDE is  $Y(t) = A + e^{-bt}(y_0 - A) + \sigma e^{-bt} \int_0^t e^{bs} dW(s)$  (see, for instance, Braumann (2005)). The distribution of  $Y(t)$  is Gaussian with mean  $A + e^{-bt}(y_0 - A)$  and variance  $\frac{\sigma^2}{2b}(1 - e^{-2bt})$  and converges, as  $t \rightarrow +\infty$ , to a Gaussian distribution with mean  $A$  and variance  $\frac{\sigma^2}{2b}$ . The data used for illustration is the weight of "mertolengo" cattle of the "rosillo" strand and was provided by Carlos Roquete (ICAM-UE).

## 2 Parameter estimation

In Filipe et al. (2007), we have considered, for a single path, the statistical problems of parameter estimation and of prediction of future sizes of an animal for model (1). Subsection 2.1 gives a brief summary of the estimation part. Subsection 2.2 presents the extension of the estimation methods to the case of several paths, assumed to be independent. We have also studied bootstrap estimation methods, shown on subsection 2.3.

### 2.1 Parameter estimation for a single path

Let us assume we observe the evolution of the weight of one animal at times  $0 = t_0 < t_1 < \dots < t_n$ , and represent the weight of the animal at time  $t_k$  ( $k = 1, 2, \dots, n$ ) by  $X_k = X(t_k)$ . Let  $Y_k = g(X_k)$  and  $\mathbf{Y} = (Y_0, Y_1, \dots, Y_n)$ . We want to estimate  $\mathbf{p} = (A, b, \sigma)$ . Since we know the transition distributions of  $Y(t)$ , using the fact that it is a Markov process and given  $Y_0 = y_0$ , assumed known, we can obtain the log-likelihood function

$$L_{\mathbf{Y}}(\mathbf{p}) = -\frac{n}{2} \ln \left( \frac{2\pi\sigma^2}{2b} \right) - \frac{1}{2} \sum_{k=1}^n \ln (1 - E_k^2) - \frac{b}{\sigma^2} \sum_{k=1}^n \frac{(y_k - A - (y_{k-1} - A) E_k)^2}{1 - E_k^2},$$

with  $E_k = e^{-b(t_k - t_{k-1})}$ . In terms of  $X$  the log-likelihood function is  $L_{\mathbf{X}}(\mathbf{p}) = L_{\mathbf{Y}}(\mathbf{p}) + \sum_{k=1}^n \ln \left( \frac{dY}{dX} \Big|_{x=x_k} \right)$ . The *maximum likelihood estimator* (MLE),  $\hat{\mathbf{p}}$ , is obtained by maximization of  $L_{\mathbf{Y}}$  (equivalent to maximization of  $L_{\mathbf{X}}$ ). Using the properties of MLE and  $Y(t)$  we can obtain the Fisher information matrix and construct approximate confidence intervals for the parameters. In Filipe et al. (2007), we used data of the weight in kg of a single animal for which we had 79 observations. We have applied model (1) for the particular cases  $g(x) = x^c$  ( $c > 0$ ) and  $g(x) = \ln x$  ( $c = 0$ ) (Table 1). Some choices of  $c$  were considered and the models which turned out to be the best choices were correspondent to  $c = 0$  and  $c = 1/3$ .

TABLE 1. Maximum likelihood estimates, log-likelihood value and approximate 95% confidence intervals (data from one animal).

	$a$	$b$	$\sigma$	$L_X$
$c = 0(\text{Gompertz})$	$407.1 \pm 60.5$	$1.472 \pm 0.354$	$0.226 \pm 0.036$	$-338.12$
$c = 1/3$	$422.4 \pm 81.6$	$1.096 \pm 0.525$	$0.525 \pm 0.083$	$-337.88$

## 2.2 Parameter estimation for several paths

Assume we have data on  $m$  animals. The weight of animal number  $j$  ( $j = 1, 2, \dots, m$ ) is observed at times  $0 = t_{j,0} < t_{j,1} < \dots < t_{j,n_j}$ , and is, respectively,  $X_{j,0} = X(t_{j,0})$ ,  $X_{j,1} = X(t_{j,1})$ ,  $\dots$ ,  $X_{j,n_j} = X(t_{j,n_j})$ . Let  $Y_{j,k} = Y(t_{j,k}) = g(X_{j,k})$  ( $j = 1, 2, \dots, m$ ;  $k = 1, 2, \dots, n_j$ ) and  $\mathbf{Y}_j = (Y_{j,0}, Y_{j,1}, \dots, Y_{j,n_j})$ .

For animal number  $j$  we can obtain the log-likelihood  $L_{\mathbf{Y}_j}$  by proceeding as in the case of a single path. From independence, the overall log-likelihood for the  $m$  animals is  $L_{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_m}(\mathbf{p}) = \sum_{j=1}^m L_{\mathbf{Y}_j}(\mathbf{p})$ . The MLE  $\hat{\mathbf{p}}$  is obtained, now, by maximization of  $L_{\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_m}$ .

In Filipe and Braumann (2007), we have applied the procedure for the stochastic Bertalanffy-Richards model, for the cases  $c = 0$  and  $c = 1/3$ , to the data of  $m = 5$  animals of the same strand raised under similar conditions. In Figure 1 we can see the observed weights for these 5 animals. For one animal we have 79 observations and the other four have 38 observations each. Table 2 shows the results obtained.

TABLE 2. Maximum likelihood estimates, log-likelihood value and approximate 95% confidence intervals (data from 5 animals).

	$a$	$b$	$\sigma$	$L_{X_1, \dots, X_5}$
$c = 0$	$352.4 \pm 28.3$	$1.708 \pm 0.193$	$0.253 \pm 0.023$	$-958.84$
$c = 1/3$	$384.1 \pm 46.2$	$1.147 \pm 0.211$	$0.506 \pm 0.047$	$-941.85$

## 2.3 Bootstrap methods

The asymptotic confidence intervals obtained from the Fisher information matrix may be quite unreliable for small sample sizes. In such case, bootstrap methods are recommended.

In Efron and Tibshirani (1993) we can find two types of bootstrap procedure, respectively, *parametric bootstrap* (PB) and *nonparametric bootstrap* (NPB). We have applied these two bootstrap methods for the cases  $c = 0$  and  $c = 1/3$ .

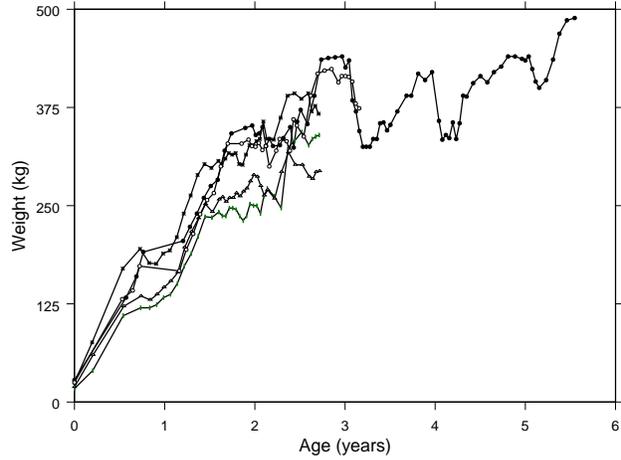


FIGURE 1. Observed growth curves for the 5 animals.

For PB, we have considered the Gaussian distribution of  $Y_k$ , mentioned on section 1, and, using the MLE  $\hat{\mathbf{p}}$  to approximate  $\mathbf{p}$ , generated 1000 independent "samples",  $\mathbf{y}^{*i} = (y_0^{*i}, y_1^{*i}, \dots, y_n^{*i})$  ( $i = 1, \dots, 1000$ ). For each one of these "samples" we have computed the estimates  $\hat{\mathbf{p}}^{*i}$  ( $i = 1, \dots, 1000$ ) (following the procedure described in subsection 2.1), and consequently, by calculating the mean, obtained the bootstrap estimate  $\hat{\mathbf{p}}^*$ .

Extending this procedure to  $m$  animals, we have generated  $\mathbf{y}_j^{*i} = (y_{j,0}^{*i}, y_{j,1}^{*i}, \dots, y_{j,n_j}^{*i})$  ( $i = 1, \dots, 1000; j = 1, \dots, m$ ) using as an approximation of  $\mathbf{p}$  the overall MLE,  $\hat{\mathbf{p}}$ , presented in subsection 2.2. We have obtained, for each  $i = 1, \dots, 1000$ , the maximum likelihood estimates  $\hat{\mathbf{p}}^{*i}$  as in subsection 2.2. From the 1000 replicates  $\hat{\mathbf{p}}^{*i}$  ( $i = 1, \dots, 1000$ ), to obtain  $\hat{\mathbf{p}}^*$  the procedure is similar to the one presented for a single animal.

For the NPB method we can find in Efron and Tibshirani (1993) how to approach the problem of dependency between observations, wich must be considered in our case. We can see that

$$e_k = (e^{t_k} Y_k - e^{t_{k-1}} Y_{k-1} - A(e^{t_k} - e^{t_{k-1}})) / \sqrt{\frac{\sigma^2(e^{2t_k} - e^{2t_{k-1}})}{2b}}, \quad (2)$$

for  $k = 1, \dots, n$ , are i.i.d with standard Gaussian distribution. We have obtained 1000 independent replicates,  $\mathbf{e}^{*i} = (e_0^{*i}, e_1^{*i}, \dots, e_n^{*i})$  ( $i = 1, \dots, 1000$ ) where the  $e_k^{*i}$  ( $k = 1, \dots, n; i = 1, \dots, 1000$ ) are obtained by sampling with

replacement the empirical distribution of the observed values of  $e_1, \dots, e_n$ . For each  $i=1, \dots, 1000$ , we have used  $\mathbf{e}^{*i}$  to reconstruct, using the inverted expression of (2), a vector of  $n$  observations  $\mathbf{y}^{*i}$ . We can, then, obtain the bootstrap estimates of the parameters in the same way as in PB.

In case we have  $m$  animals, in a similar way we must consider  $e_{jk}$ , i.i.d with standard Gaussian distribution. For each path we proceed as described above for a single animal.

For both PB and NPB, the standard bootstrap confidence intervals are obtained using normality and the sample standard deviation of the 1000 replicates of the estimates. We can also obtain bootstrap confidence intervals using the empirical quantiles, which, in our example, gives very similar results.

Although our data has a reasonably large sample size, for illustration purposes we still obtained the bootstrap estimates and 95% confidence intervals for both PB and NPB (see Table 3).

TABLE 3. Bootstrap estimates and 95% confidence intervals

		$a$	$b$	$\sigma$
1 animal (PB)	$c = 0$	$405.3 \pm 59.5$	$1.517 \pm 0.376$	$0.222 \pm 0.036$
	$c = 1/3$	$418.3 \pm 84.1$	$1.180 \pm 0.466$	$0.516 \pm 0.085$
1 animal (NPB)	$c = 0$	$404.8 \pm 59.5$	$1.519 \pm 0.371$	$0.223 \pm 0.043$
	$c = 1/3$	$419.1 \pm 80.3$	$1.179 \pm 0.453$	$0.518 \pm 0.097$
5 animals (PB)	$c = 0$	$352.4 \pm 31.4$	$1.714 \pm 0.184$	$0.252 \pm 0.024$
	$c = 1/3$	$384.9 \pm 46.7$	$1.159 \pm 0.210$	$0.504 \pm 0.047$
5 animals (NPB)	$c = 0$	$362.2 \pm 31.4$	$1.630 \pm 0.189$	$0.250 \pm 0.031$
	$c = 1/3$	$392.5 \pm 50.8$	$1.094 \pm 0.201$	$0.501 \pm 0.057$

### 3 Conclusions

Stochastic differential equations models for the growth of individual animals where considered and parameter estimation were developed for the case of several animals. In progress is the study of nonparametric estimation of the drift and diffusion coefficients, with the goal of finding a more general growth model. We have also considered the more realistic case in which we have different asymptotic expected size for different animals (to appear).

**Acknowledgments:** Both authors are members of CIMA, research center financed by FCT (Fundação para a Ciência e a Tecnologia) within its 'Programa de Financiamento Plurianual'. This work was financed by FCT within the research project PTDC/MAT/64297/2006.

**References**

- Bertalanffy, L. von. (1957). Quantitative laws in metabolism and growth. *The Quarterly Review of Biology*, **34**, 786-795.
- Braumann, C. A. (2005). *Introdução às Equações Diferenciais Estocásticas*. Edições SPE.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.
- Garcia, O. (1983). A stochastic differential equation model for the height of forest stands *Biometrics*, **39**, 1059-1072.
- Filipe, P. A. and Braumann C. A. (2007). Animal growth in random environments: estimation with several paths. *Bulletin of the International Statistical Institute*, **vol. LXII** (in press).
- Filipe, P. A., Braumann C. A. and Roquete, C. J. (2007). Modelos de crescimento de animais em ambiente aleatório. In: *Estatística Ciência Interdisciplinar, Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística*. 401-410, Edições SPE.
- Richards, F. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, **10**, 290-300.

# Variation in Mortality: Estimation via a Meta-Analytic Approach

Jutta Gampe<sup>1</sup>, Paul H.C. Eilers<sup>2</sup>

<sup>1</sup> Max Planck Institute for Demographic Research,  
Konrad-Zuse-Str. 1, Rostock, Germany. [gampe@demogr.mpg.de](mailto:gampe@demogr.mpg.de)

<sup>2</sup> Methodology and Statistics, Faculty of Social and Behavioural Sciences,  
Utrecht University, and Data Theory Group, Leiden University,  
The Netherlands. [P.H.C.Eilers@uu.nl](mailto:P.H.C.Eilers@uu.nl)

**Abstract:** If variation in mortality is to be studied, direct comparison of empirical death rates is a simple solution. However, reliability is a serious issue because population sizes differ strongly between countries. We propose to estimate a latent mortality distribution, properly accounting for population size. This introduces a meta-analytic framework into mortality studies, offering additional opportunities to study various aspects of mortality development.

**Keywords:** Latent density; mixture of GLMs; smooth EM-algorithm.

## 1 Introduction

Mortality conditions vary across countries, or regions within countries, even if we focus on areas with rather homogeneous living conditions. If we want to assess the variation in mortality, disaggregated by age and time, one approach would be to immediately compare death rates across the respective countries or regions. However, this naive approach can be hampered by the fact that countries or regions vary considerably in size, and this variation has to be taken into account when judging individual estimates.

To estimate variation in mortality, we obtained data from the Human Mortality Database (HMD, 2008). In this paper we consider European countries only. The HMD currently contains data on most countries in Northern, Western and Southern Europe, but also on several of the former communist states of Central and Eastern Europe, which underwent dramatic changes in mortality over the last decades. (The 26 countries included here were: Austria, Belgium, Bulgaria, Czech Republic, Denmark, England and Wales, Estonia, Finland, France, Germany –East and West separately–, Hungary, Iceland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Ukraine.) Besides trends in mortality levels it is of interest how the heterogeneity in mortality has changed in Europe since the fall of the Iron Curtain.

For each age and each year we have information on the number of deaths and the corresponding exposures. We will only consider data for females.

## 2 Latent mortality distribution

Mortality at age  $a$  in year  $t$  is denoted by  $\mu(a, t)$  and we assume that  $\mu(a, t)$  varies across units, i.e. countries or regions, according to a distribution with density  $f_{a,t}(m)$ . That is, mortality  $\mu(a, t)$  is itself considered to be a random variable, having a latent distribution, which can only be inferred indirectly. Mortality  $\mu_j(a, t), j = 1, \dots, J$ , in any of the  $J$  units considered is a realization from this density  $f_{(a,t)}(m)$ .

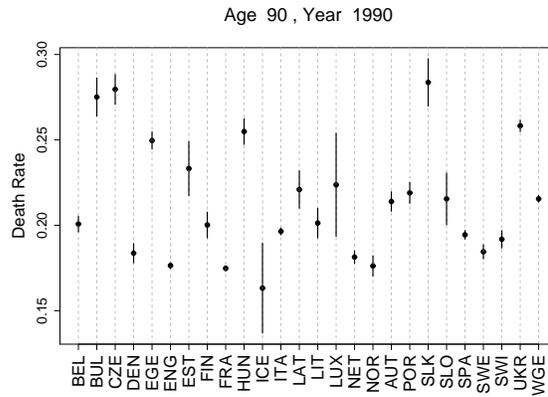


FIGURE 1. Death rates for females at age 90 in year 1990 for the  $J = 26$  countries included in this study. Shown are estimates and  $\pm 2$  standard errors.

To make inference on the distribution of  $\mu(a, t)$  an immediate solution would be to consider the empirical rates

$$\hat{\mu}_j(a, t) = \frac{y_j(a, t)}{n_j(a, t)},$$

where  $y_j(a, t)$  is the number of deaths in unit  $j$  at age  $a$  in year  $t$  and  $n_j(a, t)$  the corresponding exposures. Therefrom we could derive mean, variance or other sample statistics of interest. One drawback of this strategy is that the accuracy of the rates in the different units is not taken into account. As an example Figure 1 shows female mortality estimates at age 90 in 1990. Not only vary death rates widely, but so does their precision. Small countries, like Iceland or Luxembourg, show large standard errors. Besides the level information given by the  $\hat{\mu}_j(a, t)$  we also want to take this information into account when making inference on  $\mu(a, t)$ .

### 3 A smooth EM-algorithm

Therefore we suggest a different approach. For ease of presentation we focus on one age and one year and drop the dependence on  $a$  and  $t$  in the notation in the following. We consider a discrete distribution for  $\mu$  with a dense grid of mass-points  $m_k, k = 1, \dots, K$ , and probability masses  $p_k = P(\mu = m_k)$ . Naturally, the  $p_k$  sum to one. The grid can be equally spaced on the scale of the  $m_k$ , but usually equidistant values on the log-scale  $\eta_k = \ln m_k$  will be more appropriate, especially for small values of  $\mu$ .

For a given value of mortality  $m_k$  the number of deaths  $y_j$  in unit  $j$  is a Poisson variable with mean  $\nu_j = n_j m_k$ . That is

$$w_{jk} = P(y_j | m_k) = \frac{\exp\{-n_j m_k\} (n_j m_k)^{y_j}}{y_j!} = \alpha_j \exp\{-n_j m_k\} m_k^{y_j} \quad (1)$$

with  $\alpha_j = n_j^{y_j} / y_j!$  independent of  $k$ . The marginal distribution of the  $y_j$  is therefore

$$P(y_j) = \sum_{k=1}^K w_{jk} p_k.$$

To estimate the mixing distribution  $p_k, k = 1, \dots, K$ , the EM-algorithm is a natural choice. The E-step results from the

$$P(m_k | y_j) = \frac{P(y_j | m_k) p_k}{P(y_j)} = \frac{w_{jk} p_k}{\sum_l w_{jl} p_l}. \quad (2)$$

In the M-step we obtain  $p_k^{(s+1)}$  from the current values  $p_k^{(s)}$  as

$$p_k^{(s+1)} = \frac{1}{J} \sum_{j=1}^J \frac{w_{jk} p_k^{(s)}}{\sum_{l=1}^K w_{jl} p_l^{(s)}}. \quad (3)$$

This procedure is not limited to observations from a Poisson distribution but can be used more generally in mixtures of generalized linear models (Aitkin, 1999). Consequently, it is also possible to study the mixture of Binomial variables if probabilities of death, instead of death rates  $\mu(a, t)$ , are to be modeled.

Without any further restrictions on the  $p_k$  the EM-algorithm will converge to the nonparametric maximum likelihood estimate of the mixing distribution of  $\mu(a, t)$ . Usually only a few mass-points carry positive probabilities, leading to a rather spiky and far from smooth mixing distribution. Furthermore, convergence of the EM-algorithm typically is very slow.

To get round both drawbacks Eilers (2007) introduced the following strategy: In each iteration a smoothing step is introduced. That is, starting from the current values of the  $p_k^{(s)}$  steps (2) and (3) are performed as before. However, before the results of the current M-step,  $\tilde{p}^{(s+1)}$ , are introduced into the next EM-iteration they are smoothed by an additional smoothing step:

$$p^{(s+1)} = S_\lambda(\tilde{p}^{(s+1)})$$

with  $p^{(s+1)} = (p_1^{(s+1)}, \dots, p_K^{(s+1)})'$  and  $\tilde{p}^{(s+1)}$  accordingly. The smoothing function  $S_\lambda(\cdot)$  depends on an additional parameter  $\lambda$  that controls the amount of smoothness introduced in this step. Naturally, the smoothing step should preserve the property  $\sum_k p_k^{(s+1)} = 1$  of the mixing distribution.

### 3.1 The smoothing sub-step

The smoothing is performed by applying a discrete Whittaker smoother (Eilers, 2003). The function  $S_\lambda(\cdot)$  solves, for a given value of the smoothing parameter  $\lambda$ , the following penalized least-squares problem (for simplicity we drop the iteration index  $s + 1$  here):

$$S_\lambda(p) = \arg \min_p \{ (p - \tilde{p})'(p - \tilde{p}) + \lambda^2 p' D_2' D_2 p + 2\lambda p' D_1' D_1 p \},$$

where the matrices  $D_1$  and  $D_2$  calculate first and second order differences of the elements of  $p$ , respectively.

While the sum of squares  $(p - \tilde{p})'(p - \tilde{p})$  forces  $p$  to stay close to the outcome of the most recent M-step, the variation between neighboring elements of  $p$  is restricted by the penalty terms  $\lambda^2 p' D_2' D_2 p + 2\lambda p' D_1' D_1 p$ . The larger the value of  $\lambda$ , the stronger the effect of the penalization and the smoother the resulting vector  $p$  will be. A value of  $\lambda = 0$  implies no smoothing and the unmodified EM-algorithm would be performed.

The combination of first and second order penalty terms is due to the fact that we want to estimate a mixing distribution whose individual elements  $p_k$  have to be non-negative. This is properly taken care of by the second penalty term  $2\lambda p' D_1' D_1 p$ , as had been demonstrated by Eilers & Goeman (2004). Also, the definition of the penalty guarantees that  $\sum_k p_k$  will equal one whenever based on a vector  $\tilde{p}$  whose elements sum to one. Solving this penalized least-squares problem leads to a system of linear equations for  $p$

$$(I + \lambda^2 D_2' D_2 + 2\lambda D_1' D_1) p = \tilde{p}, \tag{4}$$

which can be easily solved for given values of  $\lambda$ . This algorithm is implemented in the following steps:

- A dense uniform grid of mass-points is chosen on the log-scale  $\eta_k = \ln m_k$ . The starting values for the mixing distribution are commonly chosen as a uniform distribution  $p_k^{(0)} = P(m_k) = 1/K, k = 1, \dots, K$ .
- The values of the  $w_{jk}$ , see (1), are determined, usually neglecting the values of  $\alpha_j$ . The  $w_{jk}$  only have to be calculated once during the whole procedure.
- E-step (2) and M-step (3) are performed, followed by the smoothing step, i.e. the system (4) is solved. For this a value of  $\lambda$  is required and currently has to be chosen by the user. Practically this is done

by inspecting results for a grid of  $\lambda$ -values. Generally the optimal values of the smoothing parameter depends on the mass-point grid, the number of observations included, and on the variance of the latent distribution to be estimated. In our applications commonly values of  $\lambda$  between 10 and 50 showed good results. The resulting  $p$  is inserted into the next EM-iteration.

- The iterations are continued until the maximum difference between elements of  $p^{(s)}$  and  $p^{(s+1)}$  is below a threshold  $\varepsilon$ . We commonly chose  $\varepsilon = 10^{-6}$ . In our application the number of necessary iterations never exceeded five, demonstrating the efficiency introduced by the additional smoothing step.

The resulting final estimate  $\hat{p} = (\hat{p}_1, \dots, \hat{p}_K)'$  is a discrete but smooth approximation to the latent density  $f_{a,t}(m)$  of  $\mu(a, t)$ , from which further parameters of interest, such as the mean, variance or certain quantiles, can be derived.

#### 4 Results and discussion

Figure 2, left panel, illustrates the procedure for females in one year, 1990, at one age, 90, (see also Figure 1). The larger the number of exposures, the sharper the scaled likelihoods. However, small countries have a broad

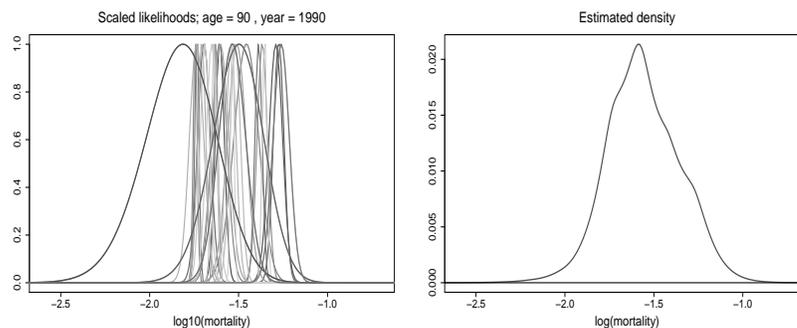


FIGURE 2. Scaled likelihoods (left) for 26 countries, and estimated latent mortality distribution (right).

likelihood, the case of Iceland being most clearly visible. The right panel shows the resulting latent mortality density (on log-scale). The density is asymmetric and reflects the non-uniform living conditions across the continent. High mortality in Eastern European countries contributes to the skew to the right.

Figure 3 illustrates the results for a range of ages, namely 60 to 100, in 1990 (dashed lines) and in year 2000 (solid lines).

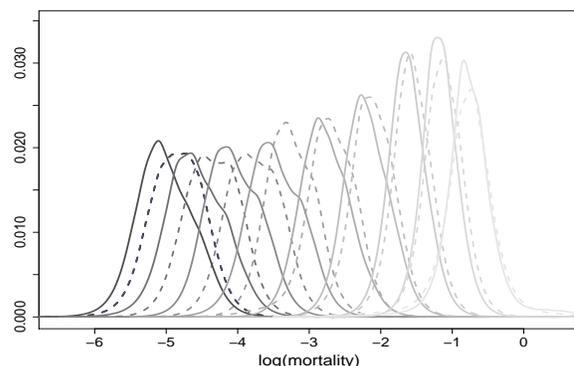


FIGURE 3. Estimated latent mortality distributions for ages 60 to 100 (multiples of five) in year 1990 (dashed lines) and year 2000 (solid lines).

In general mortality fell, i.e. the densities shifted to the left. But also some clear mixture pattern shows, particularly in lower old ages 60–75. This ability to uncover such features is a clear advantage of the nonparametric density estimate. In more homogenous situations the results may suggest a more parsimonious summary, e.g. by Normal densities. In this case fitting a parametric distribution can replace the smoothing step.

Currently the latent distributions for different ages are fitted independently, however, it is reasonable to assume that mortality in the same year does vary smoothly with age. Therefore moving to a bivariate smoothing procedure is an obvious extension of the current approach. This, together with the development of automatic selection of the smoothing parameter, is work in progress.

## References

- Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics*, **55**: 117–128.
- Eilers, P.H.C. (2003). A Perfect Smoother. *Analytical Chemistry*, **75**(14): 3631–3636.
- Eilers, P.H.C. (2007). Data exploration in meta-analysis with smooth latent distributions. *Statistics in Medicine*, **26**: 3358–3368.
- Eilers, P.H.C. and Goeman, J. (2004). Enhancing scatterplots with smooth densities. *Bioinformatics*, **20** (5): 623–628.
- Human Mortality Database (2008). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at [www.mortality.org](http://www.mortality.org).

# An Application of GAMLSS: An Insurance Type Model for the Health Cost of Cold Housing

Robert Gilchrist<sup>1</sup>, Alim Kamara<sup>1</sup> and Janet Rudge<sup>1</sup>

<sup>1</sup> STORM, London Metropolitan University, UK, Holloway Road, London, N7 8DB, U.K., [r.gilchrist@londonmet.ac.uk](mailto:r.gilchrist@londonmet.ac.uk)

**Abstract:** In this paper we consider the relationship between fuel poverty and excess winter morbidity amongst older people, and show how the GAMLSS suite of programs ([www.gamlss.com](http://www.gamlss.com)) can be used to provide a very flexible method of modelling both the number of hospital admissions and the corresponding lengths of stay in hospital. The approach is closely related to the models that have been used to model the number of insurance claims, and their cost. We fit the Beta Binomial distribution to the number of episodes, and we fit a variety of continuous distributions to the lengths of stay.

**Keywords:** Beta Binomial; GAMLSS; insurance; morbidity; costs.

## 1 Introduction

This paper shows how new facilities in GAMLSS (Rigby and Stasinopoulos (2005)) can be used to understand an important health problem, namely the relationship between fuel poverty and excess winter morbidity amongst older people. Section 2 describes the data and the statistical methodology, with section 3 giving a more formal definition of the model. We initially describe the substantive problem.

Fuel poverty is defined as the inability to afford adequate warmth in the home and is related to poor energy efficiency of homes as well as householders' incomes. Older households are the group most vulnerable to fuel poverty, and are also particularly susceptible to cold-related health effects. The significant numbers recognised as fuel poor have as yet unrecognised implications for costs to public services. These costs are not being identified to inform planning processes for health and social care.

Conventionally, research has referred to effects of cold homes in terms of excess winter *deaths* (e.g. Wilkinson et al. (2001)). These deaths are known to be associated with outdoor winter temperatures, but direct evidence of links to low indoor temperatures is limited. Mortality statistics disguise the full extent of potentially long-term chronic conditions exacerbated by cold. Hence we have concentrated on measuring excess winter *morbidity*

(illness), rather than mortality, because of the consequent implications for winter pressures on health services.

We have previously demonstrated links between fuel poverty risk and excess winter hospital episodes among older people in Newham, using this excess as a measure of associated morbidity (e.g. Rudge and Gilchrist (2005)). In this paper we refer to that work and describe the means by which this measure could be developed as a costing element for health impact assessment.

## 2 Data and statistical methodology

The main source of data here considered is our existing database for Newham hospital admissions over 1993-96. These data are anonymised with respect to individuals, having been provided at enumeration district (ED) level, showing age but not date of birth.

Our previous work examined the excess morbidity for different ages and genders. We extend this work by analysing length of episodes and investigating the associated costs for such episodes. Our proposed methodology is based upon the modelling of the propensity for an individual to be an emergency respiratory hospital admission, together with the duration of stay in hospital for such admissions. This approach is similar to that used for insurance claims (see e.g. Heller, et al.(2007)) in which the probability of a claim and the size of a claim are both modelled. Having modelled the probability of being a hospital respiratory admission and the length of the consequential *episodes* in hospital, we will use data on the cost of such hospital admissions, to give a model for the cost of the Newham admissions. Our methodology utilises the R-based GAMLSS package; see Rigby and Stasinopoulos (2005) and [www.gamlss.com](http://www.gamlss.com). We consider that the probability of being an emergency hospital admission follows a Beta Binomial distribution, this being a more flexible extension of the more traditional Binomial distribution. The 'default' approach is to use the traditional logistic regression approach which considers the log odds of being admitted as being linearly related to a combination of explanatory variables and nominal factors, and their interactions. The corresponding length of stay is modelled from continuous distributions such as the Gamma, Inverse Gaussian, Normal or Tweedie distribution. The parameters of the distributions (e.g. mean, variance, skewness, kurtosis) are modelled in terms of the risk factors. This enables us not only to assess the expected average stay (and cost) but also the expected variability of the stay (and cost). Our analysis makes use of a fuel poverty index FPR as defined by Rudge and Gilchrist (2007). We utilise nominal factors ED, gender and age to allow differing parameters to be fitted for the differing numbers  $n_i$  of 'at risk' males and females, of differing ages, in each enumeration district. Potential confounding factors are considered, using 1991 UK Census data. Daily weather data were obtained for 1993-1997. We have concluded that an assumption of in-

dependence of the observed admissions and of the observed lengths of stay is not too unreasonable.

### 3 A more formal definition of the statistical model

We explain the observed illness counts in each enumeration district (ED), in each month, in terms of the potential explanatory variables, and notably FPR. We consider data for each of 48 months. Our particular interest is in the difference between the counts observed in summer and winter, and whether we can explain this difference in terms of the explanatory variables. Our initial assumption is of a logistic model, so the probability  $p_{ijkl}$  of an individual of gender  $i$ , in age group  $j$ , in ED  $k$ , being ill in month  $l$ , is given by  $p_{ijkl} = 1/(1 + \exp(-\eta_{ijkl}))$ ,  $i = 1, 2$ ;  $j = 1, \dots, 3$ ;  $k = 1, \dots, 450$ ,  $l = 1, \dots, 48$ , where  $\eta_{ijkl}$  is a linear predictor based upon the explanatory variables. The observed counts  $Y_{ijkl}$  are assumed to follow a Beta Binomial distribution.

The duration  $d_{ijkl}$  of observed stays of patients for cell  $i, j, k, l$  are modelled with a Gamma, Inverse Gaussian, Normal or Tweedie distribution, with a random effect to allow for any lack of independence. Our default link is the log link.

### 4 A model for the number of people admitted to hospital

We fitted the Beta Binomial distribution to the observed numbers of people aged 65+ who were admitted to hospital as emergency respiratory patients. Our 'Binomial denominator' consisted of the number of people aged 65+ from each of 450 enumeration districts in Newham, UK. We considered a wide range of potential explanatory factors, mostly only available at 'enumeration district' level. We could distinguish the age and sex of the people concerned, and we had monthly weather data. A subset of the variables considered is shown in the **Table 1**.

#### 4.1 The Beta Binomial distribution (BB)

The probability function of a random variable,  $Y$  which follows the Beta Binomial (BB) distribution denoted here as  $\mathbf{BB}(n, p, \sigma)$ , is given by

$$p_Y(y|p, \sigma) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(\frac{1}{\sigma})\Gamma(y+\frac{p}{\sigma})\Gamma[n+\frac{(1-p)}{\sigma}-y]}{\Gamma(n+\frac{1}{\sigma})\Gamma(\frac{p}{\sigma})\Gamma(\frac{1-p}{\sigma})} \quad (1)$$

for  $y = 0, 1, 2, \dots, n$ , where  $0 < p < 1$  and  $\sigma > 0$  (and  $n$  is a known positive integer). Note that  $E(Y) = np$  and  $Var(Y) = np(1-p) \left[1 + \frac{\sigma}{1+\sigma}(n-1)\right]$ .

Variable)	Description
hh1 #	% households with one or more pensioner(s)
hh2 *	% small households (one or two persons households)
undoc #	% households under-occupied (1 person with 4 rooms; 2 person with 5 rooms)
lowsap #*	% dwellings with poor energy efficiency (below SAP35**)
ctb #*	% households in receipt of Council Tax Benefit (indicator of low income)
tow	Townsend deprivation score
ch *	% households with no central heating
pens*	% lone pensioner households with no central heating
pre*	% dwellings built before 1945
pensm	total male pensioners as % of total population
pensf	total female pensioners as % of total population
penswh	% of white pensioners in the ED.
FPR	Fuel Poverty Risk Index = $(hh1*undoc*lowsap*ctb)*10^{-3}$
pwh	White pensioners (% total pensioners)
mmaxt	Monthly maximum air temperature, ° C
pop	% population 65 years old or more
age*	(1) 65-74, (2) 75-84, (3) 85+
nage	Age with 2 levels only: (1) 65-84 (2) 85+
sex*	(1) Male, (2) female
q*	Season factor with 3 levels: (1) Summer, (2) November, January, February, (3) December
nq*	Season factor with 2 levels: (1) Not December (2) December
z	Factor with 48 levels denoting month
E	factor with 450 levels specifying enumeration district (ED)

TABLE 1. Explanatory variables and factors. \*\* *SAP35* is energy rating, or measure of energy efficiency, on a scale of 0 - 100, where 0 is poorest. # denotes component of FPR. A \* indicates a variate in the final model for the mean number of hospital admissions.

The Binomial  $\mathbf{BI}(n,p)$  distribution is the limiting distribution of  $\mathbf{BB}(n,p,\sigma)$  as  $\sigma \rightarrow 0$ . For  $p = 0.5$  and  $\sigma = 0.5$ ,  $\mathbf{BB}(n,p,\sigma)$  is a uniform distribution. For our modelling we have a random variable  $Y_{ijkl}$  where we model both  $p_{ijkl}$  and  $\sigma_{ijkl}$  in terms of our explanatory variables and factors.

## 5 A model for the length of episode of people admitted to hospital

The basic model for the duration  $d_{ijkl}$  of observed episodes of patients for cell  $i, j, k, l$  uses a Gamma distribution, with log link. GAMLSS allows us

to explain not only the mean  $\psi_{ijkl}$  of the Gamma distribution but also the 'scale' parameter  $\lambda_{ijkl}$  where  $Var(d_{ijkl}) = \lambda_{ijkl}^2 \psi_{ijkl}^2$ .

### 5.1 The Gamma distribution (GA)

The gamma distribution is appropriate for positively skew data. The parameterization of the gamma distribution,  $\mathbf{GA}(\psi, \lambda)$ , is given by

$$f_Y(y) = f_Y(y|\psi, \lambda) = \frac{1}{(\lambda^2\psi)^{1/\lambda^2}} \frac{y^{\frac{1}{\lambda^2}-1} e^{-y/(\lambda^2\psi)}}{\Gamma(1/\lambda^2)} \quad (2)$$

for  $y > 0$ , where  $\psi > 0$  and  $\lambda > 0$ , and  $E(Y) = \psi$  and  $Var(Y) = \lambda^2\psi^2$ .

### 5.2 Model selection strategy

With a large number of potential explanatory variables and factors, we adopted the GAMLSS AIC criterion for model selection. For a distribution with two parameters, we first fitted the 'mean' parameter for a constant 'scale' parameter, then fitted a model for the scale parameter using the 'current' mean model, then refitting the mean parameter for the new 'current' scale, and so on. We then removed parameters for which the change in scaled deviance was not significant on a  $\chi^2$  scale. We obtained a more parsimonious model by combining appropriate factor levels that were not significantly different.

### 5.3 Results of the model fitting

The variates in the final **BB** model, for the probability of an individual being an emergency hospital respiratory admission are starred in **Table 1**. The corresponding scale coefficient,  $\sigma$ , is always negative; it is larger for the over 84 year olds than for the other over 64 year olds, and is larger for men than women. One of our main interests is in the effect of the variable FPR, this being a measure of fuel poverty. The linear predictor for  $p_{ijkl}$  has an interaction between 'season' and FPR, showing that morbidity counts rise with increasing fuel poverty risk index in 'winter', with a notably large effect in December. This is over and above the underlying effect of winter itself, irrespective of FPR. Effects are evident for age, with higher counts for older people, and sex, with lower counts for women. There was a strong month effect. To understand this further, we considered monthly weather-related factors. Of all these, including looking for a lag effect, maximum temperature was most significant, with a higher maximum leading to lower morbidity counts. Having allowed for the maximum temperature effect, other weather related variables were not significant.

For the length of stay in hospital, we found that the mean level of episode depended mainly upon age and sex, with older people staying in hospital longer, and older women more so than men.

Combining data on the cost of hospital admissions with our model for the probability of being admitted and length of stay in hospital, allows us to develop a simple model for the cost to the UK National Health Service of fuel poverty. The oral presentation will discuss the results of the model fitting in more detail.

## References

- Bardsley M. (2000). Healthier homes: the role of health authorities. In: Rudge J and Nicol F (Eds.): *Cutting the Cost of Cold. Affordable Warmth for Healthier Homes*. London: E&FN Spon Ltd.
- DEFRA. (2006). The UK Fuel Poverty Strategy 4th Annual Progress Report 2006. At: [www.dti.gov.uk/files/file29688.pdf](http://www.dti.gov.uk/files/file29688.pdf). (Accessed 7/9/06).
- Department of Health. (2001). National Service Framework for Older People. At: [www.doh.gov.uk/nsf/olderpeople/pdfs/nsfolderpeople.pdf](http://www.doh.gov.uk/nsf/olderpeople/pdfs/nsfolderpeople.pdf). (Accessed 12/3/03).
- Heller, G., Stasinopoulos, D.M., Rigby, R.A. and De Jong, P. (2007). Mean and dispersion modelling for policy claim costs. *Scandinavian Actuarial Journal*, 1-12.
- Rigby, R. and Stasinopoulos D.M. (2005). Generalized additive models for location, scale and shape (with Discussion). *Applied Statistics* **54**, 507-554.
- Rudge, J. and Gilchrist, R. (2005). Excess winter morbidity among older people at risk of cold homes. *Journal of Public Health* **27** 4, 353-358.
- Rudge, J. and Gilchrist, R. (2007). Measuring the health impact of temperatures in dwellings: investigating excess winter morbidity and cold homes in the London Borough of Newham. *Energy and Buildings* **39**, 847-858.
- Wilkinson, P., Landon, M., Armstrong, B. et al (2001). *Cold Comfort: the social and environmental determinants of excess winter death in England, 1986-96*. Bristol: The Policy Press.
- Wilkinson, P., Pattenden, S., Armstrong, B. et al (2004). Vulnerability to winter mortality in elderly people in Britain: population based study. *British Medical Journal* **29** **7467**, 647.

# Cold and Heat waves: Modelling the mortality in Évora — Portugal

Dulce Gomes<sup>1</sup>, Carla Nunes<sup>2</sup> and Luísa Canto e Castro<sup>3</sup>

<sup>1</sup> University of Évora, CIMA-UE. Dep. of Maths., R. Romão Ramalho 59, 7000-671 Évora. Portugal

<sup>2</sup> CIESP, National School of Public Health. University Nova of Lisbon. Av. Padre Cruz 1600-560 Lisboa. Portugal

<sup>3</sup> Faculty of Sciences of the University of Lisbon, Dep. of Stat. and Oper. Research, C6, Campo Grande, 1749-016, Lisbon. Portugal

**Abstract:** In this work, models for discrete non-negative integer-valued time series, using as covariates daily maximum and minimum temperatures, are presented in order to model the number of deaths occurred in the district of Évora (Portugal). Two methodologies were applied to define heat and cold waves (one based on classical definition and other using temporal clustering processes) to identify critical time periods and to validate the model with a particular emphasis on these specific periods.

**Keywords:** Integer-valued time series; cold and heat waves; temporal clustering.

## 1 Introduction

In Portugal, a heat wave in 2003 Summer caused a large number of fires, costing about 925 million euros, destroyed over 270,000 ha of forest and 25,000 ha of agricultural land (according to the European Commission). 14 people died due to sudden and unexpected high temperatures (Min. Saúde, 2004). These facts support and justify many studies concerning the impact of weather conditions in several fields, with a special emphasis on Public Health. We present a model for mortality occurrences based on integer time series. A brief summary of this model will be presented. Our goal isn't only to model general climatic conditions, but to evaluate if its performance is robust and useful in critical situations.

We also propose a new approach using temporal clustering to determine a more consistent and robust (with statistical support) heat and cold waves definition. In this study we present the results obtained by temporal clustering through spatial scan statistics (Kulldorff, 1997). Finally, we compare results obtained by classical definitions and temporal clustering processes.

### 1.1 Motivation

In order to evaluate the impact of weather conditions on Public Health, it was proceeded to model the number of deaths in Évora between 1980 and 1997 through models for non-negative integer-valued time series, using as covariates daily maximum temperatures and minimum temperatures. This option dues to the fact that the observed values are not large enough, so the use of methods based on Gaussian noises is inadequate. In what respects, in particular, the district of Évora, daily deaths present values that oscillate between 0 and 19, which justifies the option. We will also identify cold and heat waves based on temporal clustering processes and compare the results. The correct identification of these critical time periods is essential to predict and organize useful Public Health interventions. Impacts differ if critical temperatures arrive after a slowly increasing or abruptly.

### 1.2 Cold and Heat Waves

There's no universal definition of a heat or cold wave. But many countries have their own definitions of heat waves. The same does not happen with cold waves. According to the definition used by the Portuguese Meteorological Institute, a heat wave occurs when the daily maximum temperature of six or more consecutive days exceeds the 90% quantile value. As in many other countries, in Portugal there isn't an official definition of cold waves. So, we say that a cold wave occurs when the daily minimum temperature of six or more consecutive days is lower than 2°C (35.6°F), and in more than two of these days the daily minimum temperature is lower than 0°C (32°F). Table 1 presents identified cold and heat waves based on the previous definitions, here called classical definitions. Using temporal clustering methodologies, it will be analyzed if these definitions and results are statistically supported.

TABLE 1. Cold and heat waves identification based on classical definitions.

	Time Frame	Days
Heat Waves	1981/6/11 — 1981/6/17	7
	1982/8/07 — 1982/8/12	6
	1986/7/15 — 1986/7/21	7
	1988/9/04 — 1988/9/09	6
	1991/7/12 — 1991/7/17	6
Cold Waves	1983/2/09 — 1983/2/17	8
	1985/1/08 — 1985/1/17	10

## 2 Temporal Clustering

Another approach for classifying a sequence of daily high (low) temperatures has a heat (cold) wave will be based on the space-time scan statistic

TABLE 2. Characterization of identified waves using temporal clustering.

	Time Frame	Days	Mean Inside	Mean Outside	Test Statistic
Heat Waves	1981/6/11 — 1981/6/17	7	38.56	29.27	11.26
	1982/8/07 — 1982/8/12	6	34.35	27.38	7.62
	1986/7/15 — 1986/7/21	7	36.17	28.10	14.04
	1988/9/04 — 1988/9/09	6	37.62	27.61	10.70
Cold Waves	1991/7/12 — 1991/7/17	6	38.63	29.35	9.61
	1983/2/09 — 1983/2/17	8	-0.29	7.43	24.56
	1985/1/08 — 1985/1/17	10	-0.30	7.81	46.47

TABLE 3. Cold and heat waves identification based on temporal clustering.

	Time Frame	Days	Mean Inside	Mean Outside	Test Statistic
Heat Waves	1981/6/11 — 1981/6/18	8	38.28	29.21	12.26
	1982/8/03 — 1982/8/25	23	32.07	26.72	15.73
	1986/7/13 — 1986/7/21	9	35.51	28.01	15.50
	1988/9/04 — 1988/9/12	9	36.33	27.44	12.51
	1991/7/09 — 1991/7/18	10	37.80	29.34	12.78
Cold Waves	1983/2/07 — 1983/2/18	12	0.86	7.55	27.44
	1985/1/07 — 1985/1/17	11	-0.07	7.85	49.24

(Kulldorff, 1997), which identifies the most significant cluster of a particular shape in space and/or time. This method identifies the zone showing the strongest evidence of representing a discordant cluster. In particular, the study of time series temperatures clustering assumes that temperatures occur uniformly in time; *i.e.*, the observed temperatures are about the same in one interval as in previous or succeeding intervals. To implement this approach Scan Statistics from the software SaTScan ([www.satscan.org](http://www.satscan.org)) has been used.

To apply Kulldorff space-time scan statistic the maximum length of the cluster must be previously given. Table 2 show the results obtained when considering the waves lengths presented in table 1 (based on classical definitions) as a maximum possible length wave. In table 3, waves length were not limited (it was used a sufficient high value of 25 days, that can be seen as a non restrict criteria, considering that natural seasonality requires time series under consideration were not too wide).

In order to compare both cases, indicators — mean inside, mean outside and test statistic — were included in tables 2 and 3. Both tables present test statistic to characterize each wave, instead of common p-values, due to the fact that all identified waves have p-values  $< 0.001$ .

Comparing the results of tables 2 and 3, it can be observed that critical periods, for Winter and Summer, are approximately the same, *i.e.*, using classical or temporal approach. Also, wave lengths are always bigger in the second approach (table 3) and more significant (bigger test statistic values). Means inside and outside are always lower in summer periods, but higher in winter periods (table 3) when comparing with the values in table 2. Although these clustering processes do not use the common Portuguese

reference threshold temperature for Summer season, 35°C (95°F), (Min. Saúde, 2007), it is important to see that almost all heat clusters present mean inside temperatures above this value.

As the results were not very different and considering that classical definition identified short waves with high means in summers (and respectively small means in winters), we will use the cold and heat wave definition presented in table 1 (classical definitions) to validate the model in critical climacteric situations.

### 3 Model for Non-Negative Integer-valued Time Series

#### 3.1 The Model

We use a Generalized Doubly Stochastic INteger Autoregressive model (GDSINAR(1)) (Gomes *et al.* (2005)) to model the daily mortality in Évora. This model is based on a generalized thinning operator, represented by  $\circ^G$ , where  $G$  is a discrete type distribution associated to the operation, and generalizes the thinning or binomial thinning operation,  $*$ , (proposed by Steutel and Van Harn (1979)) defined by

$$\{U_i\}_{i \in \mathbf{N}} * Y = \begin{cases} \sum_{i=1}^Y U_i, & Y > 0 \\ 0, & Y = 0, \end{cases}$$

where  $\{U_i\}_{i \in \mathbf{N}}$  is a sequence of i.i.d. Bernoulli random variables with mean  $\alpha$  and independent of  $Y$ . The generalized thinning operation  $\circ^G$  between two random variables  $\alpha$  and  $Y$  will be defined as follows:

**Definition** *Let  $Y$  be a non-negative integer-valued random variable and  $\alpha$  a random variable with support on  $\mathbf{R}_+$ . Then  $\alpha \circ^G Y$  is a random variable that verifies*

$$(\alpha \circ^G Y | \alpha, Y) \sim G(\mu, \sigma^2),$$

where  $G(\mu, \sigma^2)$  is a given discrete type distribution associated to the generalized thinning operation, with mean  $\mu = \alpha Y$  and finite variance  $\sigma^2 [\sigma^2 \equiv \sigma^2(\alpha, Y)]$ . In the following results, we take  $\sigma^2 = \delta Y$  with  $\delta$  possibly dependent of  $\alpha$  and  $Y$ .

Our approach incorporates in the model two explanatory time series, the minimum (Tmin) and maximum (Tmax) daily temperatures. The more satisfying modulation (in terms of residual analysis) was obtained by the following model:

$$Y_t = \alpha_t \circ^P Y_{t-1} + Z_t,$$

where  $Z_t \sim \text{Po}(\lambda)$ ,  $Y_t | Y_{t-1}, \text{Tmax}_{t-1}, \text{Tmin}_{t-8}^* \sim \text{Po}(\alpha_t Y_{t-1} + \lambda)$  and  $\alpha_t = \exp(\omega_1 \text{Tmin}_{t-8}^* + \omega_2 \text{Tmax}_{t-1})$ , with  $\text{Tmin}^*$ :

$$\text{Tmin}_t^* = \begin{cases} \text{Tmin}_t - \overline{\text{Tmin}} & , 0 < \text{Tmin} < 5^\circ\text{C} \\ \text{Tmin}_t & \text{c. c.}, \end{cases}$$

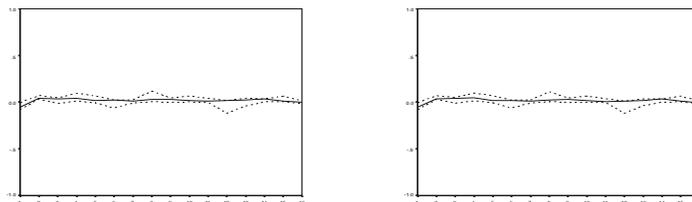


FIGURE 1. Residuals ACF (left) and PACF (right) of adjusted GDSINAR(1). Horizontal lines represent 95% bootstrap confidence bands.

$\overline{T_{\min}}$  is the sample mean of  $T_{\min}$  when  $T_{\min}$  is over  $5^{\circ}\text{C}$  ( $41^{\circ}\text{F}$ ).  $\{Z_t\}_{t \in \mathbf{Z}}$  is a sequence of independent and identically distributed (i.i.d.) random variables of mean  $\mu_Z$  and finite variance  $\sigma_Z^2$ .

### 3.2 The Results

In table 4, conditional maximum likelihood estimates of the model parameters are presented. These estimates are obtained using algorithms developed in Fortran language.

TABLE 4. Conditional maximum likelihood estimates of model parameters  $(\omega_1, \omega_2, \lambda)$ .

$\omega_1$	$\omega_2$	$\lambda$
-0,162194	-3,874511E-02	5,055070

Since the values observed during the heat waves doesn't deviate significantly from the values usually observed, we only present the study of the impacts of cold waves. Estimates of extreme values occurring during cold waves between 1980 and 1997 are presented in table 5. In figure 1 residuals ACF and PACF of adjusted GDSINAR(1) model are presented.

## 4 Conclusions

Cold and heat waves definitions must be clarified and well defined. Temporal clustering methods seem to be promising specially if they would be able to account for the impact of low (high) temperatures. In our opinion the definition of a cold or heat wave must depend on the effect it has on the variable under study. For instance, when dealing with mortality these definitions should depend on discrepant high values for the number of deaths and can vary from region to region.

To Public Health promotion and prevention it is important to detect a wave in its beginning. Classical definitions are used to alert a critical situation:

TABLE 5. Extremes values caused by cold waves and theirs estimates.

Date	Num. of mortality	Pontual estimates of mortality
26/Fev/83	19	20.89864
28/Jan/85	17	17.92792

as an example, five consecutive days with daily temperatures above 35°C constitute a real-time alert of a heat wave (Min. Saúde, 2007). The question here is if temporal clustering methods can be adapted to statistically support these real-time alerts. Further studies must be done in order to answer this question.

In what respects daily mortality modelling using maximum and minimum temperature as explanatory variables the results, as expected, are not totally satisfactory. However, the used model seems to be able to capture the data variability and to give good estimates of extreme values in mortality due to cold waves. Other explanatory variables will be introduced in order to improve the model.

## References

- Gomes, D., Canto e Castro, L. (2005). Preocupar-se-ão os Eborenses com o Frio? In: *Estatística Jubilar. Actas do XII Congresso Anual da SPE. Edições SPE, Lisboa*, pp.343-354.
- Huynen, M., Martens, P., *et al.* (2001). The impact of heat waves and cold spells on mortality rates in the Dutch population. *Environmental Health Perspectives*, **109**, **5**, 15-51.
- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: theory and methods*, **26(6)**, 1481-1496.
- Ministério da Saúde (2004). Ondas de calor. <http://www.portugal.gov.pt>.
- Ministério da Saúde (2007). Plano de contingência para ondas de calor. <http://www.min-saude.pt/NR/rdonlyres/233DB43D-7781-4F70-B428-585C74ACF1E8/0/PlanoOndasCalor2007.pdf>.
- Steutel, F. e Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. probab.*, **7**, 893-899.

# Modelling Changes in Irish Forest Carbon Stocks

M.J.Hawkins,<sup>1</sup> K.Black<sup>1,2</sup> and J.Connolly<sup>1</sup>

<sup>1</sup> UCD Dublin, School of Mathematical Sciences, Environmental and Ecological Modelling Group, Rm. L552, Library Bldg., Belfield, D.4, Ireland

<sup>2</sup> FERS Ltd, 117 East Courtyard, Tullyvale, Cabinteely, D. 18, Ireland

**Abstract:** We are interested in estimating changes in forest carbon sequestration levels in Ireland. Our overall approach is to calibrate a distance-independent individual tree growth model using empirical data gathered from silvicultural experiments that have been running in Ireland for more than fifty years. Actual increment estimates can then be derived, using data from a recent National Forest Inventory to drive the prediction model. We discuss our initial modelling which involves interpolating missing values in the calibration dataset which have not been recorded due to cost constraints. This is a common feature of such datasets.

**Keywords:** tree-growth modelling; nonlinear regression; mixed-effects

## 1 Introduction

The context of this work is the estimation of the amount of carbon that is sequestered in Irish forests and the degree to which that may change in the future. Our overall approach to this problem can be summarised by the following schema:

$$Biomass(Growth) \propto Carbon \quad (1)$$

e.g. biomass estimates are derived as functions of growth estimates, and the resulting output is assumed proportional to the carbon level. In this poster we describe our proposed derivation of the individual tree growth estimates and present some preliminary growth modelling results.

## 2 Materials and Methods

We utilise two datasets, a permanent plot (PP) database developed by Coillte Teoranta and the 2006 National Forest Inventory (NFI). The former contains periodic records (with approximate intervals of five years) of a variety of individual-tree measurements from silvicultural thinning and spacing trials that have been conducted from 1953 to the present day. The NFI dataset was assembled through a survey conducted over the period 2004-2006 to quantify the volume of Ireland's forest stock. Our approach

is to calibrate a distance-independent growth model, based on the Prognosis modelling framework (Stage, 1973) using the PP data. This model will then be used to estimate growth increments from the starting point defined by the NFI dataset. These estimated increments in turn drive the biomass function estimates of carbon storage increments (Equation 1).

The Prognosis model framework was selected because it does not make explicit use of the spatial coordinates of individual trees, which were not recorded in the PP dataset. By contrast, spatially explicit approaches typically estimate an optimal growth and then modify this estimate as a function of the number and proximity of the subject tree's neighbours. The Prognosis framework rather incorporates proxies of these competition and neighbour effects directly into the model equation of the estimated individual-tree growth increment in the form of tree-level or plot-level covariates, such as plot density, plot basal area, or basal area in larger trees. The equation for the individual-tree growth increment may also include site-level covariates describing climatic and topographical characteristics (Equation 2).

$$\text{Increment} = \alpha + \beta(\text{tree}) + \gamma(\text{comp.}) + \delta(\text{site}) + \epsilon \quad (2)$$

where  $\beta(\text{tree})$ ,  $\gamma(\text{comp.})$  and  $\delta(\text{site})$  are estimators of tree-level, competition, and site-level effects, respectively. For example, some researchers use the following model for the effect of current diameter on diameter growth increments :  $\beta(\text{tree}) = \beta_0 + \beta_1 \ln(\text{Diameter}) + \beta_2 \text{Diameter}^2$ .

### 3 Preliminary results : Missing values in the calibration dataset

When deriving and calibrating growth models for Sitka Spruce and other tree-species using permanent plot data that has been recorded over several decades, a common problem encountered is that of missing tree-level data which has not been recorded because of cost constraints. For example, it is not uncommon for permanent sample plot datasets to contain measurements of the tree breast height diameter (DBH) for every tree and on every measurement occasion, but that tree height (H) measurements would only be recorded on much fewer occasions. That is the case with the PP growth model calibration dataset, where the ratio of tree height measurements to diameter measurements is approximately 1:19.

Temesgen and von Gadow (2004) showed that commonly-used nonlinear models of the DBH-H relationship could be improved in permanent sample plots by incorporating concomitant site-level stand-density information. Our initial data exploration shows that the functional form suggested by Yang et al. (1978) and subsequently verified by Temesgen and von Gadow (2004) is a reasonable model of the DBH-H relationship (Equation 3).

$$H = 1.3 + a(1 - \exp(b.DBH^c)) \quad (3)$$

where  $a$ ,  $b$  and  $c$  are parameters and 1.3 is a lower bound for model predictions. Following the approach of Temesgen and von Gadow (2004) we have attempted to extend this model to incorporate the estimated effects of physically relevant covariates on the model parameters. As a further extension of the model, we have attempted to model certain model effects as fixed and random. By estimating separate parameters for each plot in the dataset we have seen that the asymptote is the most variable parameter between forest plots. We have successfully estimated the effects of more covariates on this parameter than on the other two parameters. Our working model derived through this exercise describes the data well and has low prediction errors.

#### 4 Discussion

Our adopted approach is to derive a model of individual tree growth increments on a calibration dataset which we will then use to estimate increments using inventory data. In the calibration dataset we have observed a high level of missing data which we will attempt to impute using models that incorporate the effects of physically relevant covariates at the level of tree, plot and site. Preliminary results suggest that including these covariates significantly improves the baseline nonlinear model. In particular, the asymptote of the nonlinear imputation model was observed to be the most variable across plots and we tested for a variety of covariate effects on this parameter. Our adopted approach to modelling forest carbon sequestration represents an improvement on existing stand-based methods because the individual-tree equations we derive can be applied to mixed-species and uneven-aged stands.

#### References

- Stage, A.R. (1973). *Prognosis model for stand development*. Research Paper INT-137, Ogden, Utah, USDA Forest Service.
- Temesgen, H. and von Gadow, K. (2004). *Generalized height-diameter models – an application for major tree species in complex stands of interior British Columbia*. European Journal of Forest Research. **123**, 45 - 51.
- Yang, R.C., Kozak, A., Smith, J.H.G. (1978). *The potential of Weibull-type functions as flexible growth curves*. Canadian Journal of Forest Research **8**, 424 - 431.

# The Behaviour of the Self-Protective Parameter Estimation in Models for Randomized-Response Data.

Tamara A. M. Hendrick <sup>1</sup>, Maarten J. L. F. Cruyff <sup>2</sup> and Peter G. M. van der Heijden <sup>3</sup>

<sup>1</sup> Department of Methodology and Statistics, Utrecht University, 3508 TC Utrecht, the Netherlands, T.A.M.Hendrick@students.uu.nl and Peter G. M. van der Heijden, Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht

<sup>2</sup> Maarten J. L. F. Cruyff, Department of Methodology and Statistics, University Utrecht, Heidelberglaan 1, 3584 CS Utrecht, the Netherlands, M.Cruyff@uu.nl

**Abstract:** Randomized Response is an interview technique that provides confidentiality when answering sensitive questions. Despite this protection some of the respondents choose to answer in a evasive way, known as the Self-Protective response pattern. In this abstract four models used to handle Randomized Response data, Profile Likelihood, Log Linear model, Item Randomized Response and the Zero-Inflated Poisson Model, for Randomized Response data are used to discuss the behaviour of an additional parameter for the Self-Protective response pattern.

**Keywords:** Randomized Response; SP parameter.

## 1 Introduction

An interview method to eliminate the evasive response bias, normally found under questioning in a direct way, is the randomized response method. The responses obtained by this method partly depend on the respondent's true status and partly on a randomizing device operated by the respondents. The design used within this abstract is the Forced Choice Randomized-Response (FCRR) design, for other designs see Fox and Tracy (1986). In the FCRR design used, the respondent rolls two virtual dice each time a question is asked and keeps the outcome of the dice hidden from the researcher. In advance it is predetermined which outcomes of the dice evoke either a forced or truthful answer. This protects the answer of the respondent since the researcher does not know whether it is a forced or a truthful response. The misclassification of the response partly depends on the true response of the respondent and partly on the probability distribution of the two dice.

Despite the confidentiality provided, a number of studies have shown that not all the respondents follow the randomized response design. One of the examples is an experimental randomized response study with in advanced fixed randomized response outcomes performed by Edgell, Himmelfarb and Duncan (1982). This study suggested that about 25% of the respondents answer *no* when asking if they had homosexual experiences, while according to the randomizing device they should have said *yes*. This evasive response pattern of consistently responding *no* without regarding the outcome of the randomizing device is denoted by Böckenholt and Van der Heijden (2007) as the self-protective response pattern. This self-protective answering (SP) implies that the respondent does not wish to reveal any information, neither about his/her true response nor the response determined by the randomizing device by consistently answering *no*. Given this definition, it is evident that SP leads to an overrepresented no-response. To correct, an additional parameter is a necessity when estimating the sensitive behaviour in question, to avoid underestimating the behaviour.

Currently there are four models estimating fairly unbiased the prevalence of the sensitive behaviour measured with a randomized response design, while correcting for SP. Cruyff, Van den Hout, Van der Heijden and Böckenholt (2007) used the profile likelihood which provides an interval estimate of the SP under the best fitting model, and the log-linear randomized response model which associated the sensitive characteristics and the estimated prevalence's correct for SP. Böckenholt and Van der Heijden (2007) used an item response model, to model the sensitive characteristics corrected for SP. The last model introduced by Cruyff, Böckenholt, Van den Hout, and Van der Heijden (2008a) is a zero-inflated poisson model, which assess the individual number of sensitive characteristics based on the observed sum score corrected for SP. The question that is nevertheless unanswered, is: do all of the four model provide approximately the same estimates of SP. The main objective of this abstract is to compare the robustness of the estimated SP parameters under the four models mentioned, by creating a correlation matrix and checking the variance-covariance structure present. In the remainder of this abstract, the general randomized response model is discussed, the empirical data used for the seed sampling bootstrapping, concluding with a discussion of the results on the comparison of the SP parameter.

### 1.1 The Randomized Response model

The randomized response method obtains estimates of the prevalence of the sensitive behaviour, by modelling the observed response to the true status of the respondent, taken into account the misclassification of the responses. The four models briefly described in the introduction all use this general randomized response method, for more information see references given. Given a multivariate FCRR design with  $K$  dichotomous sensitive

questions, the general randomized response model is,

$$\begin{aligned}\boldsymbol{\pi}^* &= (1 - \theta)\mathbf{P}_K\boldsymbol{\pi} + \mathbf{I} \\ &= \mathbf{Q}_K\boldsymbol{\pi}\end{aligned}\quad (1)$$

where  $\theta$  is the probability of Self-Protective No response,  $\boldsymbol{\pi}^*$  is the vector containing the observed response profile probabilities,  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_D)^t$  is a vector containing the true-response profile probabilities,  $\mathbf{I}$  is the indicator variable and  $\mathbf{Q}_k$  is the transition matrix with the elements,

$$q_{ij} = \begin{cases} (1 - \theta)p_{ij} & \text{for } i \neq D, j \in \{1, \dots, D\} \\ (1 - \theta)p_{ij} + \theta & \text{for } i = D, j \in \{1, \dots, D\} \end{cases} \quad (2)$$

Note that multivariate data is used, therefore the Kroneckers product of the univariate transition matrices is needed and  $p_{ij}$  is a probability found under the distribution of the two dice.

## 2 The dataset

The empirical dataset consist of the nationwide survey to measure compliance with the rules of the Unemployment Insurance Act conducted by the Department of Social Affairs in the Netherlands. This large dataset is divided into three smaller datasets using only the questions asked about health rules violation under respondents who receive a Disability benefit income over the years 2002, 2004 and 2006. See for more information Van der Heiden, Van Gils and Laudy (2005).

## 3 Results and Conclusions

The main objective of this abstract is to show robust SP parameters estimated by the profile likelihood (prof), log-linear randomized response (LLRR) model, item randomized response (IRR) model and the zero-inflated poisson (ZIP) model. In table 1 the mean, variance and standard deviation of the estimates SP per model and per year are shown.

TABLE 1. The descriptive statistics of the estimated SP parameter over the years 2002, 2004 and 2006

statistic	Health WAO 2002				Health WAO 2004				Health WAO 2006			
	prof	LLRR	IRR	ZIP	prof	LLRR	IRR	ZIP	prof	LLRR	IRR	ZIP
mean	0.101	0.114	0.141	0.220	0.106	0.145	0.138	0.219	0.242	0.263	0.255	0.292
variance	0.0005	0.0007	0.0007	0.0005	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.001
st. deviation	0.022	0.026	0.027	0.022	0.033	0.041	0.043	0.034	0.027	0.031	0.032	0.036

To test for robustness, a test for covariance is used and to see if the estimated SP deviates in the same way for each model. The analysis of the

variance-covariance structure for the four models shows that an autoregressive structure is present within each year. Given this structure, a Pearson correlation matrix is created showing high and significant correlations ( $p < .001$ ) between each model for each year. (lowest correlation between prof and IRR of .713 for 2002, between prof and LLRR of .735 in 2004 and between ZIP and LLRR of .660 in the year 2006)

Concluding, with reference to the significant and high correlation and the variance-covariance structure analysis over all three years there is a significant association between the SP parameters. Meaning that the estimation process over the three years of the SP parameters are robust, given that each year gives back the same correlation and variance-covariance structure results. Although the variance-covariance structure analysis, shows that the zero-inflated poisson model and then the item randomized response model covary in lesser extent than the profile likelihood and the log-linear randomized response model. It should be mentioned that there remains a high correlation between the estimation of the SP over the four models. Therefore it is safe to reason that the behaviour of the SP parameter over the four models are different but this difference is not relevant.

## References

- Böckenholt, U. & van der Heijden, P.G.M. (2007) Item Randomized-Response Models For Measuring Noncompliance: Risk-return perceptions, Social Influences, and Self-Protective Responses. *Psychometrika* **72**, 245–262.
- Cruyff, M.J.F., Böckenholt, U., Hout van den, A. and Heijden van der, P.G.M (2008a) Accounting for Self-protective Response in Randomized-Response Data From A Social Security Survey Using The Zero-Inflated Poisson Model. *The Annals of Applied Statistics* **2**, 316-331.
- Cruyff, M.J.F., Hout van den, A., Heijden van der, P.G.M and Böckenholt, U. (2007) Log-Linear Randomized-Response Models Taking Self-protective Response Behaviour Into Account. *Sociological Methods and Research* **36**, 266–282.
- Edgell, S. E., Himmelfarb, S. and Duncan, K. L. (1982) Validity of Forced Response in a Randomized Response Model. *Sociological & Methods and Research* **11**, 89-110.
- Fox, J.A. and Tracy, P.E. (1986) *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills CA: Sage.
- Van der Heijden, P.G.M., van Gils, G. and Laudy, O. (2005) *Regelovertreding in de WAO, WW en ABW/WWB in 2004. (vergeleken met de jaren 2000 en 2002)* Ministerie van Sociale Zaken en Werkgelegenheid. (in Dutch)

# Constant Latent Odds-Ratios Models and Marginal Maximum Likelihood Estimation

David J. Hessen<sup>1</sup>

<sup>1</sup> Utrecht University

**Abstract:** In this paper, the focus is on parameter estimation under a general parametric CLORs model for dichotomously scored items. For estimation of the structural parameters a marginal maximum likelihood procedure is proposed. Special attention is given to three special cases of the general parametric CLORs model. For assessing the appropriateness of CLORs models likelihood ratio tests are presented.

**Keywords:** Item response theory; Rasch model; MML estimation; Likelihood ratio test.

## 1 Introduction

A relatively new class of dichotomous item response models is the class of constant latent odds-ratios (CLORs) models (Hessen, 2004, 2005). A favorable property of the class of CLORs models is that the total score (the unweighted sum of the item scores) is a minimal sufficient statistic. As a consequence, computational procedures under CLORs models are simpler than under models without this property. For the estimation of the parameters of the CLORs models a marginal maximum likelihood (MML) procedure is proposed. In addition, likelihood ratio tests for assessing the goodness-of-fit of these models are presented. For assessing the overall appropriateness of a specific CLORs model, the usual likelihood ratio test against the saturated multinomial model is proposed. For assessing the relative appropriateness of two CLORs models a chi-square difference test is considered. In the presentation of this paper, all procedures are demonstrated by real data examples.

## 2 CLORs Models

Fundamental to dichotomous item response models is the item response function (IRF). The IRF gives the probability of a correct response to an item as a function of the latent trait. In the present paper, the focus is on CLORs models in which the IRF of item  $i$  is given by

$$P_i(\theta) = \frac{\exp(\alpha_i)g(\theta)}{1 + \exp(\alpha_i)g(\theta)}, \quad (1)$$

where  $\alpha_i$  is the location parameter of item  $i$ , and can be interpreted as the simplicity or easiness of item  $i$ . The function  $g$  is a positive parametric function of  $\theta$  independent of the item. The parameter vector of this function is denoted by  $\varepsilon$ , and its number of entries is  $c$ . Models that satisfy (1) are called CLORs models because under each of these models the odds-ratio function (ORF) for any two items  $i$  and  $j$  is a constant function of  $\theta$ , i.e.,

$$\omega_{ij}(\theta) = \frac{P_i(\theta) \{1 - P_j(\theta)\}}{\{1 - P_i(\theta)\} P_j(\theta)} = \exp(\alpha_i - \alpha_j). \quad (2)$$

An important property that holds for any CLORs model is that the total score is a sufficient statistic. Let the number of items be  $k$ . Then, it can be shown that the probability of a score pattern given  $\theta$  is

$$P(\mathbf{x}|\theta) = \frac{\exp\left(\sum_{i=1}^k \alpha_i x_i\right) g(\theta)^t}{\prod_{i=1}^k \{1 + \exp(\alpha_i) g(\theta)\}}, \quad (3)$$

where  $t = \sum_{i=1}^k x_i$  is the total score, and sufficiency of the total score for  $g(\theta)$  follows by the Neyman factorization theorem. A very general CLORs model is the model with

$$g(\theta) = \frac{\gamma + \exp(\beta_1 \theta + \beta_2 \theta^2 + \dots + \beta_r \theta^r)}{1 + \delta \exp(\beta_1 \theta + \beta_2 \theta^2 + \dots + \beta_r \theta^r)}. \quad (4)$$

The parameters of this model are however difficult to interpret. Moreover, it is computationally difficult to estimate the parameters  $\beta_1, \beta_2, \dots, \beta_r$  simultaneously. Therefore, the focus is on simpler CLORs models with

$$g(\theta) = \frac{\gamma + \exp(\beta \theta)}{1 + \delta \exp(\beta \theta)}, \quad (5)$$

where  $\beta$  is a scaling parameter and can be interpreted as a common discrimination parameter. When  $\beta > 0$ , the common parameter  $\gamma$  determines the values of the lower asymptotes of the IRFs, and the common parameter  $\delta$  determines the values of the upper asymptotes of the IRFs. So in a restricted way the model can account for both guessing and the phenomenon that examinees with high ability give incorrect responses to the items because they have information beyond that assumed by the test item writer. Special cases of the model can be considered by setting  $\delta$  and/or  $\gamma$  to zero. Setting both  $\delta$  and  $\gamma$  to zero and  $\beta$  to one, yields the Rasch model (Rasch, 1960).

### 3 MML Estimation

For a randomly sampled examinee the marginal probability of score pattern  $\mathbf{x}$  is

$$P(\mathbf{x}) = \exp\left(\sum_{i=1}^k \alpha_i x_i\right) \int_{-\infty}^{\infty} g(\theta)^t \prod_{i=1}^k Q_i(\theta) \phi(\theta) d\theta, \quad (6)$$

where  $\phi(\theta)$  is the density function of the random latent variable  $\Theta$  in the population of examinees, and  $Q_i(\theta) = 1 - P_i(\theta) = \{1 + \exp(\alpha_i)g(\theta)\}^{-1}$ . Taking all possible response patterns and total score groups in the sample into account, it follows that the marginal log-likelihood equals

$$l = \sum_{i=1}^k \alpha_i s_i + \sum_{t=0}^k n_t \ln \left\{ \int_{-\infty}^{\infty} g(\theta)^t \prod_{i=1}^k Q_i(\theta) \phi(\theta) d\theta \right\}, \quad (7)$$

where  $s_i$  is the total number of examinees in the sample with a correct score on item  $i$ , and  $n_t$  is the number of examinees in the sample with total score  $t$ . There are  $k$  easiness parameters and  $c$  parameters in  $g(\theta)$ . In estimating these parameters with the data from a random sample of examinees, it is customary to assume that the population density  $\phi(\theta)$  is normal, and in order to eliminate indeterminacy, the standard normal density can be selected. Therefore, the number of structural parameters to be estimated is  $k + c$ . Let  $\varepsilon_u$  be the  $u$ th structural parameter in  $g(\theta)$ . The maximum likelihood estimates of the structural parameters that jointly maximize the marginal log-likelihood function can be found by solving the likelihood equations  $\frac{\partial l^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})}{\partial \alpha_i} = 0$ , for all  $i$ , and  $\frac{\partial l^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})}{\partial \varepsilon_u} = 0$ , for all  $u$ . An approximate solution to these equations can be obtained by means of the Newton-Raphson procedure. In the Newton-Raphson procedure the first and second derivatives of the marginal log-likelihood function with respect to the structural parameters are used. The first derivative of the marginal log-likelihood function with respect to  $\alpha_i$  is

$$\frac{\partial l^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})}{\partial \alpha_i} = s_i - \sum_{t=0}^k n_t \int_{-\infty}^{\infty} P_i(\theta) \pi(\theta|t) d\theta, \quad (8)$$

where

$$\pi(\theta|t) = \frac{g(\theta)^t \prod_{i=1}^k Q_i(\theta) \phi(\theta)}{\int_{-\infty}^{\infty} g(\theta)^t \prod_{i=1}^k Q_i(\theta) \phi(\theta) d\theta} \quad (9)$$

is the conditional density of  $\theta$  given total score  $t$ . The first derivative of the marginal log-likelihood function with respect to  $\varepsilon_u$  is

$$\frac{\partial l^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})}{\partial \varepsilon_u} = \sum_{t=0}^k n_t \int_{-\infty}^{\infty} \frac{\partial}{\partial \varepsilon_u} \{ \ln g(\theta) \} \left\{ t - \sum_{i=1}^k P_i(\theta) \right\} \pi(\theta|t) d\theta. \quad (10)$$

In the case of  $\boldsymbol{\varepsilon} = (\beta, \gamma, \delta)$ , the derivatives  $\frac{\partial}{\partial \varepsilon_u} \{ \ln g(\theta) \}$ , for  $u = 1, 2, 3$ , are

$$\begin{aligned} \frac{\partial}{\partial \beta} \{ \ln g(\theta) \} &= \theta \left\{ e^{\beta\theta} / (e^{\beta\theta} + \gamma) - \delta e^{\beta\theta} / (1 + \delta e^{\beta\theta}) \right\}, \\ \frac{\partial}{\partial \gamma} \{ \ln g(\theta) \} &= (e^{\beta\theta} + \gamma)^{-1}, \\ \frac{\partial}{\partial \delta} \{ \ln g(\theta) \} &= -e^{\beta\theta} / (1 + \delta e^{\beta\theta}). \end{aligned}$$

Since the derivatives of the marginal log-likelihood function involve integrals that are difficult to evaluate analytically, the numerical approximation technique called Gauss-Hermite quadrature can be employed to compute the values of the derivatives to any desired degree of accuracy. Any integral of the form

$$\int_{-\infty}^{\infty} h(\theta)\phi(\theta) d\theta \quad (11)$$

can be approximated by

$$\sum_{q=1}^m h(y_q)w(y_q), \quad (12)$$

where  $y_q$  is the  $q$ th tabled quadrature point or node and  $w(y_q)$  is the corresponding weight. The number of nodes is denoted by  $m$ , and the larger this number, the more accurate the approximation will be. See Stroud and Secrest (1966) for tables of nodes and corresponding weights, or Press et al. (1992) for computer programs. The nodes and weights can also be obtained by means of the freely available software packages R and Octave which have built-in functions for quadrature integration. So by using Gauss-Hermite quadrature the first derivative of the marginal log-likelihood function with respect to  $\alpha_i$  can be approximated by

$$\frac{\partial l^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \alpha_i} = s_i - \sum_{t=0}^k n_t \sum_{q=1}^m P_i(y_q) \pi(y_q | t), \quad (13)$$

where

$$\pi(y_q | t) = \frac{g(y_q)^t \prod_{i=1}^k Q_i(y_q)w(y_q)}{\sum_{q=1}^m g(y_q)^t \prod_{i=1}^k Q_i(y_q)w(y_q)}. \quad (14)$$

The first derivative of the marginal log-likelihood function with respect to  $\varepsilon_u$  can be approximated by

$$\frac{\partial l^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \varepsilon_u} = \sum_{t=0}^k n_t \sum_{q=1}^m \frac{\partial}{\partial \varepsilon_u} \{ \ln g(y_q) \} \left\{ t - \sum_{i=1}^k P_i(y_q) \right\} \pi(y_q | t). \quad (15)$$

The MML estimates of the structural parameters can be obtained using an EM algorithm (Dempster, Laird, & Rubin, 1977). In the E-step, for provisional parameter values the expected number of examinees at the  $q$ th node with total score  $t$  is computed for all  $q$  and  $t$ . This expected number of examinees is given by  $r_{qt} = n_t \pi(y_q | t)$ , for all  $q$  and  $t$ , where  $\pi(y_q | t)$  is  $\pi(y_q | t)$  evaluated for the provisional parameter values at the  $q$ th node and total score  $t$ . Next, the  $r_{11}, \dots, r_{mk}$  are treated as the data at the M-step in which improved parameter estimates are obtained. The improved parameter estimates are the solutions to the likelihood equations

$$\frac{\partial \bar{l}^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \alpha_i} = s_i - \sum_{q=1}^m f_q P_i(y_q) = 0, \quad (16)$$

where  $f_q = \sum_{t=0}^k r_{qt}$ , and

$$\frac{\partial \bar{l}^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \varepsilon_u} = \sum_{t=0}^k \sum_{q=1}^m r_{qt} \frac{\partial}{\partial \varepsilon_u} \{ \ln g(y_q) \} \left\{ t - \sum_{i=1}^k P_i(y_q) \right\} = 0. \quad (17)$$

Applying at each M-step the Newton-Raphson procedure requires second derivatives. Differentiation of the derivative in (16) with respect to  $\alpha_i$  gives the second derivative

$$\frac{\partial^2 \bar{l}^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \alpha_i^2} = - \sum_{q=1}^m f_q P_i(y_q) Q_i(y_q), \quad (18)$$

and differentiation with respect to  $\varepsilon_u$  yields the second derivative

$$\frac{\partial^2 \bar{l}^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \alpha_i \partial \varepsilon_u} = - \sum_{q=1}^m f_q P_i(y_q) Q_i(y_q) \frac{\partial}{\partial \varepsilon_u} \{ \ln g(y_q) \}. \quad (19)$$

Differentiation of (17) with respect to  $\varepsilon_v$  finally yields the second derivative

$$\begin{aligned} \frac{\partial^2 \bar{l}^M(\boldsymbol{\alpha}, \boldsymbol{\varepsilon})^*}{\partial \varepsilon_u \partial \varepsilon_v} &= \sum_{t=0}^k \sum_{q=1}^m r_{qt} \left[ \frac{\partial^2}{\partial \varepsilon_u \partial \varepsilon_v} \{ \ln g(y_q) \} \left\{ t - \sum_{i=1}^k P_i(y_q) \right\} \right. \\ &\quad \left. - \frac{\partial}{\partial \varepsilon_u} \{ \ln g(y_q) \} \frac{\partial}{\partial \varepsilon_v} \{ \ln g(y_q) \} \sum_{i=1}^k P_i(y_q) Q_i(y_q) \right]. \quad (20) \end{aligned}$$

In case of  $\boldsymbol{\varepsilon} = (\beta, \gamma, \delta)$ , the derivatives  $\frac{\partial^2}{\partial \varepsilon_u \partial \varepsilon_v} \{ \ln g(y_q) \}$ , for  $u = 1, 2, 3$ , are

$$\begin{aligned} \frac{\partial^2}{\partial \beta^2} \{ \ln g(y_q) \} &= y_q^2 \{ e^{\beta y_q} \gamma / (e^{\beta y_q} + \gamma)^2 - \delta e^{\beta y_q} / (1 + \delta e^{\beta y_q})^2 \}, \\ \frac{\partial^2}{\partial \beta \partial \gamma} \{ \ln g(y_q) \} &= -y_q e^{\beta y_q} / (e^{\beta y_q} + \gamma)^2, \\ \frac{\partial^2}{\partial \beta \partial \delta} \{ \ln g(y_q) \} &= -y_q e^{\beta y_q} / (1 + \delta e^{\beta y_q})^2, \\ \frac{\partial^2}{\partial \gamma^2} \{ \ln g(y_q) \} &= -(e^{\beta y_q} + \gamma)^{-2}, \\ \frac{\partial^2}{\partial \gamma \partial \delta} \{ \ln g(y_q) \} &= 0, \\ \frac{\partial^2}{\partial \delta^2} \{ \ln g(y_q) \} &= \{ e^{\beta y_q} / (1 + \delta e^{\beta y_q}) \}^2. \end{aligned}$$

The EM cycles can be repeated until the improved estimates have converged satisfactorily. The final estimates are taken as the MML estimates.

#### 4 LR Tests

Let  $n_{\mathbf{x}}$  be the number of examinees with response pattern  $\mathbf{x}$  and let  $n$  be the total sample size. If the vectors  $\hat{\boldsymbol{\alpha}}$  and  $\hat{\boldsymbol{\varepsilon}}$  contain the MML estimates

under a parametric CLOrS model with  $k + c$  parameters and  $P(\mathbf{x}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\varepsilon}})$  is the corresponding estimated probability, then the likelihood ratio statistic

$$G^2 = 2 \sum_{\mathbf{x}} n_{\mathbf{x}} \ln \left\{ \frac{n_{\mathbf{x}}/n}{P(\mathbf{x}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\varepsilon}})} \right\} \quad (21)$$

can be used to test the parametric CLOrS model against the general multinomial model. When the parametric CLOrS model is true, the test statistic is asymptotically chi-square distributed with  $2^k - 1 - (k + c)$  degrees of freedom (Bishop, Fienberg, & Holland, 1975). Note that  $P(\mathbf{x}_a; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\varepsilon}})$  is approximated by

$$P^*(\mathbf{x}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\varepsilon}}) = \exp \left( \sum_{i=1}^k \hat{\alpha}_i x_i \right) \sum_{q=1}^m \hat{g}(y_q)^t \prod_{i=1}^k \hat{Q}_i(y_q) w(y_q), \quad (22)$$

where  $\hat{g}(\cdot)$  and  $\hat{Q}_i(\cdot)$  are  $g(\cdot)$  and  $Q_i(\cdot)$  evaluated for the MML estimates. Other likelihood ratio tests are presented for testing specific CLOrS models against more general unsaturated alternatives.

## References

- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM-algorithm. *Journal of Royal Statistical Society, Series B*, **39**, 1-22.
- Hessen, D.J. (2004). A new class of parametric IRT models for dichotomous item scores. *Journal of Applied Measurement*, **5**, 385-397.
- Hessen, D.J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika*, **70**, 497-516.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge, UK: University Press.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas*. Englewood Cliffs, NJ: Prentice-Hall.

# Methods for aggregation and linkage analysis of human longevity in selected families.

Jeanine J Houwing-Duistermaat<sup>1</sup>, Andrea Callegaro<sup>1</sup>, Marian Beekman<sup>1</sup>, Eline P Slagboom<sup>1</sup> and Hans C van Houwelingen<sup>1</sup>

<sup>1</sup> Leiden University Medical Centre, Department of Medical Statistics and Bioinformatics, S-5-P, P.O.Box 9600, 2300 RC Leiden, The Netherlands, j.j.houwing@lumc.nl

**Abstract:** Typically long-lived sibling pairs have been collected for linkage analysis of human longevity and information on life span of first degree relatives is available to assess familial aggregation of life span within the families. We derive methods for aggregation and linkage analysis of human longevity. To illustrate the methods we analyze data from the Leiden Longevity Study.

**Keywords:** demographic data, family history score, frailty models.

## 1 Introduction

To reveal the genetic basis of human longevity, families displaying exceptional longevity have been studied. In a simulation study it was shown that stringent selection criteria of at least one sibling above the age of 98 years should be used (Tan et al 2004). In Dutch and European studies on longevity (Francheschi et al 2007) however, sibling pairs of age above 90 years will be genotyped for linkage. To improve efficiency, one may want to use a weighted score statistic.

For genetic linkage analysis of survival data observed in randomly selected families, Commenges (1994) proposed a frailty model where the linkage effect is decomposed into the sum of two random effects representing the paternal and maternal alleles at a locus. To test for linkage, Commenges derived the score statistic from the likelihood of the phenotypes given the marker data. For genetic analysis of human longevity however, the subjects need to be alive and for all siblings the outcome of interest (age at death) is censored. To apply survival methods, population based information on mortality has to be used to standardize the age at entry of the siblings. Further a score statistic should be based on the retrospective likelihood of the marker data given the phenotypes.

Before linkage analysis is performed, aggregation of the trait within families may be assessed. Clustering of an outcome within families can be studied by testing for the presence of a relationship between the outcome of an

individual and a family history score based on the outcomes of the relatives. For binary data, Houwing-Duistermaat et al (1998) derived a family history measure. This approach will be adapted for longevity studies.

In this paper we consider long lived sibling pairs and their first degree relatives (parents, siblings and offspring). For all family members information on current age or age at death is available. For the long lived sibling pairs genetic markers are typed to perform linkage analysis. Because ascertainment of the families depends not only on the selected sibling pairs themselves but also on the sizes of the sibships and the distribution of life span within the whole sibship, we study the relationship between the excess survival in the whole sibship and the excess survival of the parental and offspring generations.

## 2 Theory

*Aggregation analysis* Let  $D_{ij}$  be 1 if subject  $j$  of family  $i$  is deceased and 0 otherwise and let  $H_{ij}$  be the sex and birth cohort specific cumulative hazard for subject  $j$  of family  $i$ . Ignoring the correlation between family members, the likelihood function for parents or offspring of sibships is equal to

$$L(\lambda|D_{ik}, H_{ik}) = \prod_i \prod_k (H_{ik} \exp(\lambda))^{D_{ik}} \exp(-H_{ik} \exp(\lambda)). \quad (1)$$

The parameter  $\lambda$  models deviation of the survival of the parents or the offspring from the corresponding Dutch birth cohort. The numerator of the score statistic  $U$  to test the null hypothesis  $\lambda = 1$  versus the alternative  $\lambda < 1$  is given by

$$U = \sum_i \sum_k^{n_i} (D_{ik} - H_{ik}) = \sum_i \text{sumMR}_i, \quad (2)$$

with  $n_i$  the number of relatives (parents or offspring) in family  $i$ . The variance of this statistic can be empirically estimated.

A family history measure for relative (offspring or parent)  $k$  of family  $i$  is the sum of the kinship coefficient  $\Gamma_{jk}^i$  between relative  $k$  and the sibling  $j$  times the sibling's martingale residual  $D_{ij} - H_{ij}$ :

$$x_{ik} = \sum_j^{m_i} \Gamma_{jk}^i (D_{ij} - H_{ij}), \quad (3)$$

with  $m_i$  the number of siblings. For parents  $\Gamma_{jk}^i$  will be  $\frac{1}{4}$  for all  $j$ . For offspring,  $\Gamma_{jk}^i$  will be  $\frac{1}{4}$  if  $j$  is the parent and  $\frac{1}{8}$  if  $j$  is the aunt or uncle of  $k$ . Now to test for a relationship between the excess survival in the sibships and the survival in the offspring and parents, replace  $\lambda$  in likelihood (1) by

TABLE 1. Results of testing for excessive survival in Leiden Longevity Study.

type of relative	number	number deceased	family specific sumMR*	Unweighted statistic <sup>#</sup>	Weighted statistic <sup>%</sup>
siblings	1249	686	-9.5 (3.9)	-	-
parents	330	330	-0.8 (1.9)	-5.63	-5.43
offspring	1317	138	-0.4 (1.0)	-4.70	-5.12

\*mean (sd), see formula (4)

<sup>#</sup>Statistic 2 divided by the square root of its variance

<sup>%</sup>Statistic 4 divided by the square root of its variance

$\lambda_{ik} = \theta x_{ik}$ . The statistic  $U(x_{ik})$  corresponding to this parametrization is given by

$$U(x_{ik}) = \sum_i \sum_k^{n_i} x_{ik}(D_{ik} - H_{ik}). \tag{4}$$

*Linkage analysis* Let for sibship  $i$ ,  $\hat{\pi}_i$  be a vector with as elements the proportions alleles shared identical by descent (IBD) by the long lived sibling pairs at a certain position at the genome. Then for additive effects, the score statistic  $Z$  to test for linkage is given by:

$$\hat{Z} = \frac{\sum_i w'_i(\hat{\pi}_i - \frac{1}{2})}{\sqrt{\sum_i w'_i \text{var}_0(\hat{\pi}) w_i}}, \tag{5}$$

with  $w_i$  a vector of known weights of the same lengths as  $\hat{\pi}_i$  (Kruglyak et al 1996). The variance  $\text{var}_0(\hat{\pi})$  may be computed using multipoint simulations. Information available in the ages of the siblings at entry of the study can be incorporated via the weights. To assess genetic linkage of a region, statistic (5) is computed for a grid of genomic positions. The final statistic is the maximum of these statistics.

Based on the frailty model of Commenges (1994), we derived the score statistic for genetic linkage from the retrospective likelihood of the marker data given the phenotypes. The numerator is given by

$$\hat{U} = \sum_i \sum_{lk} H_{il} H_{ik} (\hat{\pi}_{i,lk} - \frac{1}{2}). \tag{6}$$

This statistic can be interpreted as regression of the excess IBD sharing on the product of the cumulative hazards (Lebec et al. 2004).

### 3 The Leiden Longevity Study

Families participating in the Leiden Longevity Study have at least two nonagenarian siblings. For 166 families we had data on current age or age

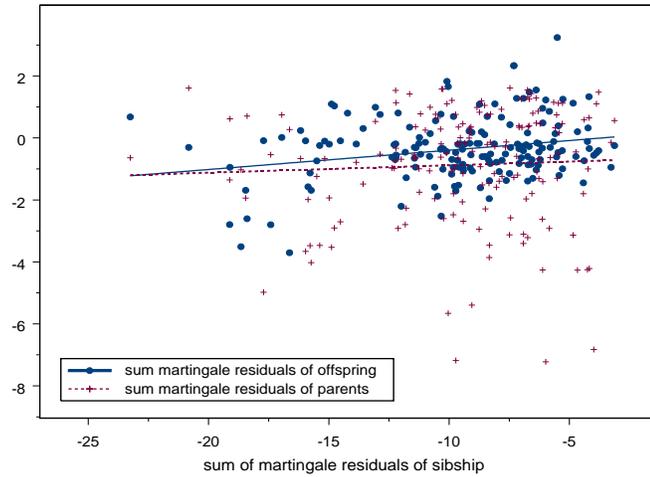


FIGURE 1. Relationship between excess survival of sibship generation and survival of relatives.

at death for parents, siblings and offsprings. Descriptives per generation (sibships, parents, offsprings) are given in table 1. The sizes of the sibships varied from 2 to 17 with a mean of 7.5 siblings.

The relationships between the sumMRs of the siblings and the parental and offspring generations are depicted in figure 1. There was no relationship between excess survival of the sibship generation and the parental generations ( $\text{cor}=0.02$ ). Between the sibship and the offspring generation some correlation was present ( $\text{cor}=0.25$ ). The values of the standardized statistics to test for excess survival are also given in table 1. For the offspring generation weighting reduced the standardized statistic.

For linkage analysis, genotypes of six micro satellite markers were available for 160 sibling pairs (see Beekman et al 2006 for description of the genetic data). The lod-score curves ( $\text{sign}(Z) \frac{1}{2 \log_{10}} Z^2$  with  $Z$  given by (5) computed at a grid of positions) for the standard unweighted statistic, the statistic applied to the oldest 80 sibships and the weighted statistic are given in figure 2. Although weighting increased the lod score, linkage for this region was not statistically significant ( $P=0.12$  at position 95 cM).

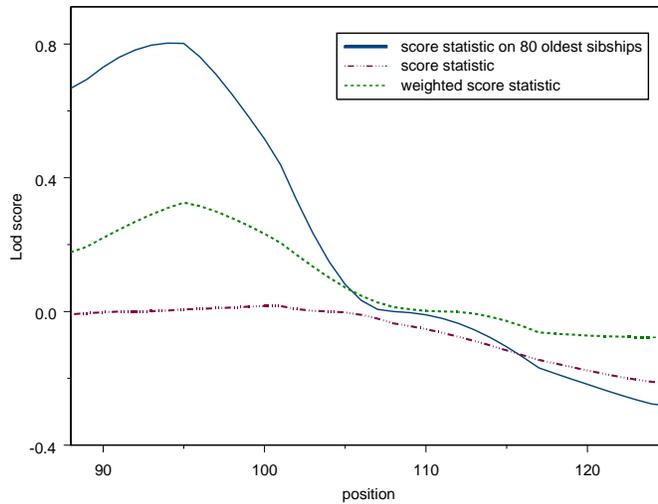


FIGURE 2. Lod Scores.

#### 4 Discussion

Early mortality is under represented in the parental generation, because parents with large offspring sizes are more likely to be in the sample. In contrast early mortality is present in the siblings and the offsprings of the nonagenarians. The lack of correlation between the parental generation and the sibship generation may be caused by the lack of early mortality within the parental generation as well as underestimation of the correlation due to ascertainment of the families. Further since the offspring generation cannot express the longevity trait yet, the found correlation between the offspring and sibling generation must be based on mean survival at middle age. For genetic linkage analysis, the maximum lod score corresponding to the weighted statistic was higher than the lod score of the unweighted statistic, but still far from significant.

The model of Commenges (1994) contains only a frailty effect for linkage and does not allow for residual correlation. For the longevity trait this model is not realistic. Residual correlation will be present due to shared environment and genetic factors not linked to the marker locus. Li and Zhong (2002) extended the model of Commenges with an additional frailty which models other shared effects. The weights of the statistic derived from the retrospective likelihood based on the Li and Zhong (2002) model

depends on the marginal correlation of the frailties of the siblings. For the Dutch population this correlation is unknown. Therefore we based our statistic on the model of Commenges. Note that although the model for the phenotype may not be correct, the statistic is still valid.

To conclude we identified aggregation of life span within the Leiden Longevity families. Hence efficiency of future analyses will be improved if the information on the age distributions is used. In addition to enhancing power, weighting provides a tool to take into account various life expectancies for various European countries when combining linkage results of various countries.

## Acknowledgements

This study was supported by grants from the Netherlands Organization for Scientific Research (NWO 917.66.344) and IOP Genomics/Senter (IGE0101014).

## References

- Beekman, M. et al. (2006) Chromosome 4q25, microsomal transfer protein gene, and human longevity: novel data and a meta-analysis of association studies. *J Gerontol A Biol Sci Med Sci*, **61**, 355–362.
- Commenges, D. (1994) Robust genetic linkage analysis based on a score test of homogeneity: The weighted pairwise correlation statistic. *Genet Epidemiol* **11**, 189–200.
- Franceschi, C. et al. (2007) Genetics of healthy aging in Europe: the EU-integrated project GEHA (GENetics of Healthy Aging). *Ann N Y Acad Sci* **1100**, 21–45.
- Houwing-Duistermaat, J.J. and van Houwelingen, J.C. (1998). Incorporation of family history in logistic regression models *Stat Med* **17**, 2865–2882.
- Kruglyak, L. et al. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, **58**, 1347–1363.
- Lebrec, J. et al. (2004) Score test for detecting linkage to complex traits in selected samples. *Genet Epidemiol*, **27**, 97–108.
- Li, H. and Zhong, X. (2002) Multivariate survival models induced by genetic frailties, with application to linkage analysis *Biostatistics* **3**, 57–75.
- Tan, Q. et al (2004) Power of non-parametric linkage analysis in mapping genes contributing to human longevity in long-lived sib-pairs. *Genet Epidemiol* **26**, 245–253.

# Explaining Self-Protective "No"-Saying in Randomized Response

Marianne Hubregtse <sup>1</sup> and Maarten Cruyff <sup>2</sup>

<sup>1</sup> Student at Utrecht University, The Netherlands, m.hubregtse@students.uu.nl

<sup>2</sup> Utrecht University, The Netherlands, m.cruyff@uu.nl

**Abstract:** In randomized response (RR), some respondents do not follow the rules of the design. This may lead to a response pattern of only ('no') responses; self-protective "no"-saying (SPN). Although SPN can be accounted for by modeling it in an IRT model, it would be better to prevent SPN altogether. This research tries to cast forward some theories regarding SPN and to identify some important aspects of SPN. Trust, understanding, law abidance and the social environment are predictors for SPN. Extensive recommendations for further research are given in the discussion.

**Keywords:** Randomized Response; cheating parameter; evasive response.

## 1 Introduction

### 1.1 Randomized Response

Respondents are reluctant to answer questions on sensitive topics, such as sexuality or fraud. This typically leads to response bias and hence underreporting of sensitive behavior in surveys. A meta-analysis proved that randomized response (RR) is a technique that reduces this response bias, yielding more valid estimates (Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005b). In RR the answers are partly determined by a randomizing device, such as a pair of dice (Forced Response) or two decks of cards (Kuk's design) (for specifics, refer to Fox and Tracy (1986)). Since only the respondent knows the value of the randomizing device, the origin of a certain answer is only known to the respondent. This increases confidentiality of the responses and thus decreases response bias. This research focuses on fraud as a sensitive topic.

### 1.2 Self-Protective "No"-Saying

RR does not completely solve response bias, however. Some respondents appear to disregard the instructions of the response design, leading to response patterns with all non-incriminating answers. Disregarding the RR instructions is called self-protective "no"-saying (SPN) (for a more elaborate discussion, see Böckenholt and Van der Heijden (2007)).

TABLE 1. Overview of Covariates Used

	name	description
1	ethnicity	the country the participant is born in
2	education	respondent's level of highest followed education
3	language	respondent's understanding of the Dutch language
4	letter of law	respondent thinks the DSA clings to the letter of the law
5	secrecy	respondent feels that RR yields more secret answers
6	privacy	respondent feels that privacy is guaranteed by RR
7	trust	respondent's trust in protection RR offers
8	honesty	respondent feels that the RR method yields more honest answers
9	clarity	respondent feels that the RR instructions were clear
10	costs	respondent feels that following rules does not cost much effort
11	benefits	respondent feels it is beneficial to commit fraud
12	acceptance	respondent's acceptance of rules of the benefit act
13	abide	respondent always abides by the rules of the benefit act
14	social control	amount of social control perceived by the respondent
15	social environment	respondent feels his social environment may turn respondent in
16	official control	respondent feels that governments strictly control rule abiding
17	discovery	respondent's perceived chance of discovery of fraud

### 1.3 Theory on SPN

Three theories regarding reasons for SPN were laid out. First, a rational choice based theory dictates that behavior is governed by goals, limitations and available information. Important for this theory are variables that indicate these goals, identify possible limitations and reveal how well respondents absorb the available information. Relevant for this theory, respondents that absorb the information regarding the protection the RR design offers display less SPN. Respondents that understand when to say 'yes' or 'no' to certain questions however, seem to exhibit more SPN.

Second, a norm-based theory was expounded, where decisions are mostly influenced by respondents' and their peers' attitudes towards law abidance, their trust in and understanding of the RR design and the costs and benefits they perceive to be associated with law abidance. When respondents' tend to abide by the law or have peers that do not condone law breaking behavior, they do not display as much SPN, confirming that respondents base their behavior on imbedded norms. Understanding of the RR instructions seems to rather facilitate than inhibit SPN, which was not expected. The costs and benefits agree with the law abidance results, where high costs and low benefits of law abidance increase SPN.

From a methodological viewpoint, respondents may be mostly influenced by response burden resulting in satisficing behavior; SPN. The above mentioned increase in SPN from a decrease in trust in the protection of the RR design, confirms that trust is a prerequisite for successfully following the RR instructions. Since understanding of the RR instructions increases SPN, there may be a problem in trying to prevent SPN. While these results tempt the researcher to keep the RR instructions vague, this may lead to more mistakes, which is just as unwanted as SPN.

## 2 Data

Because of the exploratory nature of this research, no experimentation was done and already available data was used. In the Netherlands, employees

are insured under various Social Benefit Acts. In case of loss of income due to unemployment or disability, a (previously) employed person is eligible for financial benefits, provided that certain conditions are met. Two benefit acts are the Unemployment Insurance Act (UIA) and the Disability Benefit Act (DBA). Should some of the rules that have to be followed in order to receive benefits be transgressed, this is considered fraud. The RR questions regarding fraud in the UIA are described in set A. This set is called 'application', because all RR questions ask after neglecting to apply for a fitting job. Set B entails questions that ask after fraud in the DBA, where benefit recipients are obliged to report their health status to SS. This set is called 'health'. In 2000, 2002, 2004 and 2006 the Department of Social Affairs conducted nationwide surveys to track the amount of fraud with the two acts. In this paper, the data of 2000, 2002 and 2004 for the two benefit acts are analyzed. The exact phrasings to the RR questions analyzed as well as some particulars of the design can be found in the Appendix. The focus of the analysis lies not with the prevalence estimates of different types of fraud, but rather with identifying which covariates contribute to the model by predicting the SPN parameter.

### 3 Model

For this research, the model from Böckenholt and Van der Heijden (2007) is used. The model used is based on the regular RR model as shown in equation 1 and builds on a two-parameter logistic model. In the model, the probability of a 'yes' response for person  $i$ , given person parameter  $\theta$ , forced 'yes' response chance  $c$  and chance on giving real answer  $d$ , is equal to

$$P(\theta) = c + d \frac{\exp[a(\theta - \mathbf{b})]}{1 + \exp[a(\theta - \mathbf{b})]} \quad (1)$$

where  $\mathbf{b}$  and  $a$  are parameters characterizing the item-response function. The person parameter  $\theta$  is specified to follow a normal distribution  $\phi(\theta)$ . The log-likelihood of the model then becomes

$$\log(L_{\text{RR}}) = \sum_{i=1}^n \int_{-\infty}^{\infty} \prod_{j=1}^J P(\theta_i)^{y_{ij}} (1 - P(\theta_i))^{1-y_{ij}} \phi(\theta_i) d\theta \quad (2)$$

As mentioned before, this model can be extended to contain an SPN parameter  $\pi_i$  as follows

$$\log(L_{\text{SPN}}) = \log \{ (1 - \pi_i) (L_{\text{RR}}) + \pi_i I(y_{ij} = 0 \forall j) \} \quad (3)$$

TABLE 2. Significant Covariates Per Data Set

covariate (x)		SPN estimate					
name	mean(sd)	$p(SP\bar{N})$ -1 sd of x	$p(SP\bar{N})$ mean x	$p(SP\bar{N})$ +1 sd of x	-2 log( $\Lambda$ )		
set A - application							
2	education	5.10 (1.67)	.1098	.1560	.2168	17.105	***
3	language	4.28 (.14)	.1792	.1790	.1789	4.852	*
10	costs	2.52 (1.14)	.1597	.1726	.1863	10.493	**
11	benefits	2.83 (1.21)	.1106	.1567	.2174	10.724	**
13	abide	2.67 (.91)	.1466	.1745	.2064	4.655	*
17	discovery	3.72 (.91)	.1412	.1717	.2072	6.964	**
set B - health							
2	education	4.74 (1.71)	.0605	.1051	.1765	12.380	***
9	clarity	1.67 (.83)	.1160	.1083	.1011	3.922	*
11	benefits	2.51 (1.13)	.0715	.1086	.1616	10.176	**
13	abide	2.54 (.93)	.0786	.1054	.1400	4.060	*
14	social control	2.70 (1.06)	.0858	.1107	.1416	6.142	*
15	social environment	2.63 (1.00)	.0916	.1082	.1274	7.606	**
17	discovery	3.60 (.96)	.0770	.1053	.1424	5.299	*

NOTE 1: \*  $p \leq .05$ , \*\*  $p \leq .01$ , \*\*\*  $p \leq .001$

NOTE 2: set A: -2LL = 8601.68, n = 2,951,  $p(SP\bar{N}) = .179$ ;

set B: -2LL = 8137.66, n = 3,898,  $p(SP\bar{N}) = .106$

NOTE 3: In the first 2 columns, the name and number of the covariates are given. Column 3 and 4 contain the means and standard deviations of the involved covariate. Columns 5, 6 and 7 contain probabilities of SPN. The last column contains a likelihood ratio test statistic and the associated significance.

where  $I(y_{ij} = 0 \forall j)$  is an indicator function that is 1 for all respondents with only 'no' answers and 0 otherwise. In order to investigate the relation between the covariates and  $\pi_i$ , the SPN parameter  $\pi_i$  is modeled to be a function of the covariates matrix  $X$  as follows

$$\pi = \frac{\exp(X\beta)}{1 + \exp(X\beta)} \tag{4}$$

where  $\beta$  is a vector of the covariates' regression coefficients. The loglikelihood function as in equation 3 can be maximized whilst modeling  $\pi_i$  as in equation 3. The data are divided in two data sets of dependent variables. For each set a model without covariates for  $\pi_i$  was computed and for each covariate out of  $X$  every set of dependent variables was run. A comparison on the basis of difference in the -2 loglikelihood (-2LL) of the model without covariates and the model with a specific covariate is then made, by performing a  $\chi^2$  test with 1 degree of freedom. Covariates used to predict the SPN parameter  $\pi_i$  are listed in table 1.

## 4 Results

For all of the theories some corroborating results were found, as shown in table 2. These three theories need not be mutually exclusive. There is something to say to see the three theories as three different sides of the response processes going on for each respondent. Furthermore, even should only one of the theories be ultimately correct, there is so much overlap with the other two, it may be wise to sometimes view the overlapping aspects from the point of another theory. Besides, each respondent is an individual and all respondents need not have the exact same response processes.

## 5 Discussion

Researchers that want to use RR response, may do well to explain very clearly and elaborately how the RR design protects the respondent's privacy. Also, it should be ascertained that respondents know that they cannot get in trouble for the answers that are given in the survey. This may put some respondents more at ease. Furthermore, increasing the seriousness of the survey may help respondents take the questionnaire seriously. Should the RR questionnaire be administered face-to-face, the interviewer may try to act as a peer by establishing rapport and subsequently propagate law abiding norms. Incorporating an SPN parameter in the model may help increase the validity of the prevalence estimates. Costs of complying with the RR instructions should be minimized, that means that the RR method should be easy to use, although not necessarily easy to understand. The use of computer-rolled dice may be easier, especially if with every roll a message appears that tells the respondent what to do. Furthermore, computer assisted interviews can restrict the dice to be rolled only once and not until the respondent likes the outcome.

In addition to the covariates tested in this research, there are many other variables that may help gain insight in processes regarding SPN. A start would be to gather more information about the respondent, especially regarding reading skills, cognitive learning ability and memory. These covariates make the limitations of respondents more clear. Attitudes of the respondent regarding law or rule abidance in less important situations than fraud could be measured better. Furthermore, the perceived response burden could be charted by asking respondents about their time and effort investments. Likewise, respondents could be asked for experiences with earlier questionnaires. The amount of questionnaires filled out, as well as their general experiences with those questionnaires may influence SPN. Also, respondents could be asked about how important they think their answer is, that is, how much it counts in the total. If they feel their answer does not matter so much, because there are many other people that fill out the questionnaire, they may display less SPN, since there is small chance of repercussions. Alternatively, if respondents feel their answer is not important at all, they may not take the survey seriously. Last, some qualitative questions could be asked to gather more information. It may be very interesting to let respondents rephrase the instructions in their own words, or to ask respondents about what could be better or made easier in the survey.

Adding some covariates to the model to predict SPN is not enough. Some solid research should be done to clarify important issues. A major issue that is in need of clarification is whether SPN occurs as much or even at all in research with RR questions on non-sensitive topics. In addition, there are several questions where some aspects of the RR design can be varied, which may have an influence on SPN. First, the amount of information

about the RR instructions that is available may be varied. In computer assisted interviews, the RR questions may be placed at the beginning of the questionnaire, where survey tiring should not be applicable. Furthermore, varying the actual probability on a forced response and in other research the perceived probability, for instance through visual manipulation, may influence SPN by changing the perception of the protection the RR design offers. To reveal the influence of respondents' familiarity with RR on SPN, the amount of practice questions may be varied. Furthermore, it may be interesting to ask respondents that participate in RR research to fill out a RR questionnaire at a later time, to track learning and familiarity with the RR design. A final issue that may be interesting to look at is the respondent's familiarity with probabilities.

## Bibliography

- Binmore, K. (1994). *Game Theory and the Social Contract II: Just Playing*. The MIT Press.
- Böckenholt, U., & Van der Heijden, P. G. M. (2007). Item Randomized-Response Models for Measuring Noncompliance: Risk-Return Perceptions, Social Influences, and Self-Protective Responses. *Psychometrika*, *72*(2), 245–262.
- Cruyff, M. J. L. F., Van den Hout, A., Van der Heijden, P. G. M., & Böckenholt, U. (2007b). Log-Linear Randomized-Response Models Taking Self-Protective Response Behavior into Account. In Press.
- Fox, J., & Tracy, P. (1986). *Randomized Response: A Method for Sensitive Surveys*. Sage Publications.
- Huang, K.-C., Lan, C.-H., & Kuo, M.-P. (2005). Detecting Untruthful Answering in Randomized Response Sampling. *Quality and Quantity*, *39*(5), 659–669.
- Landsheer, J. A., Van der Heijden, P. G. M., & Van Gils, G. (1999). Trust and Understanding, Two Psychological Aspects of Randomized Response. *Quality & Quantity*, *33*, 1–12.
- Lensvelt-Mulders, G. J., Hox, J. J., van der Heijden, P. G., & Maas, C. J. (2005b). Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation. *Sociological Methods & Research*, *33*(3), 319–348.
- Sharp, L. M., & Frankel, J. (1983). Respondent Burden: A Test of Some Common Assumptions. *The Public Opinion Quarterly*, *47*(1), 36–53.
- Soeken, K. L., & Macready, G. B. (1982). Respondents' Perceived Protection When Using Randomized Response. *Psychological Bulletin*, *92*(2), 487–489.

# Climate impacts on Sudden Infant Death Syndrome: a GAMLSS approach

Irene Hudson<sup>1</sup>, Alethea Rea<sup>2</sup> and Michelle Dalrymple<sup>3</sup>

<sup>1</sup> School of Math and Stats, University of South Australia, Adelaide, Australia (irene.hudson@unisa.edu.au)

<sup>2</sup> The Bioinformatics Centre, Dept. of Math, University of Auckland, NZ

<sup>3</sup> C/O Dept. of Math and Stats, University of Canterbury, Christchurch, NZ

**Abstract:** GAMLSS models identify both lower and upper thresholds of sudden infant death syndrome (SID) risk as they pertain singly and jointly to climate variables (above seasonality) for part of a unique total ascertainment study of SIDS in Canterbury, NZ (1968-2000). Significant non-linear relationships between climate and SIDS are established by GAMLSS and lead to a better understanding, and prediction of SIDS. SIDS are shown to be associated with a complex, multivariate array of climatic predictors, namely impacts of changing wind direction, rainfall, temperature (variants), dew point and relative humidity. New thresholds for climatic predictors are established which allow identification of (seasonal specific) climate profiles that lead to in/decreased SIDS.

**Keywords:** sudden infant death (SID); Climatic thresholds of risk; GAMLSS.

## 1 Introduction

The health of populations in Australasia will be affected by global climate change. It is known that marked short-term fluctuations in weather cause acute adverse health effects, leading to a greater number of hospital admissions/deaths (McMichael et al., 2006). Wind has long been associated with ill health in NZ. SIDS is still the most predominant cause of death in infants under one year of age, and currently accounts for between 10-20% of all infant deaths in developed countries (Malloy & Freeman, 2000). Health research has, for some decades, looked at specific short/long-term climate variables as a contributor to health *per se* e.g. SIDS. Recently Dalrymple, 2004, in a complete ascertainment study of all SIDS deaths in Christchurch NZ, 1968-2000, showed the first consistent pattern of climatic risk on SIDS. This work identified new SIDS risk factors; dewpoint, wind direction, wind speed and humidity; apart from temperature (seasonality) and formulated a new mixture approach for discrete count time series (Dalrymple 2004). Earlier findings on climatic impacts on SIDS had not been consistent. No study to that date has included dewpoint, wind direction nor wind chill as possible SIDS risk factors. The major aim of this present study is to inves-

tigate the value of GAMLSS (Rigby & Stasinopoulos (2005)) in modelling climate with SIDS using the 1973-1989 SIDS data of Dalrymple (2004). The SIDS counts derive from a unique retrospective, complete ascertainment study of SIDS incidence in Canterbury, NZ (1968 to 2000). The climate data analysed was obtained from NIWA [<http://www.niwa.co.nz/>]. This data is unusual in the field of SIDS-climate research, as SIDS deaths were localised around the site of the meteorological data collection. All climate variables are day of SIDS-death specific. This study investigates the most comprehensive set of climatic predictors for SIDS, to date, sourced from Dalrymple (2004). All climate variables are scaled between 0-1. Monthly SIDS counts were analysed (as in Hudson et al., 2005, 2007). The logarithm of the number of infants at risk of SIDS, namely the number of infants in the postnatal age group per month (NAR), was used to account for underlying population changes.

## 2 Methods & Results

A stepwise GAMLLS modelling using cubic smoothing splines (cs) of the log mean ( $\log \mu$ ) of monthly SIDS counts was performed, assuming a Poisson link (Stasinopoulos pers. comm., 2005). The stepwise procedure chooses the best model according to the AIC. An autoregressive lag 1 term (SID-Slag1) was used to account for the correlated nature of the data; an AR(1) approach was also adopted using mixture time series by Hudson et al., 2007.

### 2.1 Best fit model and Upper and Lower Climatic Thresholds

The best fit model, where cs denotes a cubic smoothing spline effect, is:  $SIDS \sim 7.04 - 1.58NAR - 5.33cs(\text{Temperature}) + 0.47cs(\text{Rain}) - 0.89cs(\text{WindDir}) + 0.87cs(\text{Radiation}) + 1.18cs(\text{DewPoint}) - 2.48\text{Humidity} - 0.12\text{SIDSlag1}$ .

The AIC value for this model is 710.94. Variables deleted by the stepwise routine are air pressure, sunshine hours, wind chill and wind speed. Note that the wind variant, wind direction, was retained. In agreement with Hudson et al., 2007, a low SIDS number the previous month is likely to result in an increase in SIDS this month. Figure 1 shows that wind direction, rainfall, temperature, radiation and dewpoint have cubic spline effects. Figure 1 clearly shows increased SIDS at south-westerly wind directions ( $\text{WindDir} > 0.6$ ) and at low temperature (and variants). GAMLSS modelling demonstrates that temperature acts as a seasonal proxy, which agrees with the work of Hudson et al., 2007. Lower thresholds of increased SIDS can be identified via GAMLSS as follows: Temperature less than  $11.5^{\circ}\text{C}$  ( $< 0.6$  for the scaled variable) leads to higher SIDS particularly in winter (3.62 SIDS/month) and autumn (2.97 SIDS/mth). Winter SIDS excess occurs

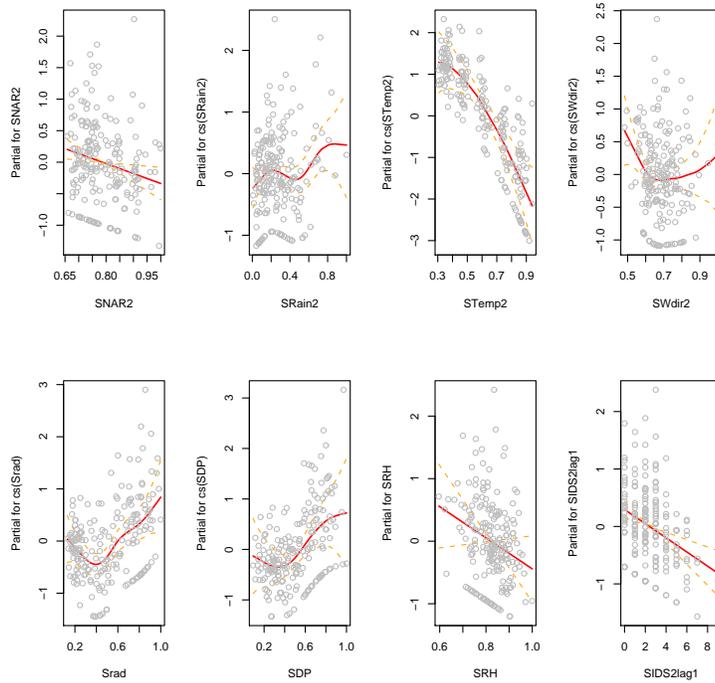


FIGURE 1. SIDS counts data and cubic smoothing splines vs predictors.

for temperature variants below the following cutpoints; radiation  $< 15.2$  Mj/hr and dew point  $< 8$  °C (Fig. 1).

GAMLSS models also allowed us to establish upper thresholds of increased radiation and dew point for increased SIDS; and with increased incidence of SW breezes. Specifically, dew point temperatures above 8 °C ( $> 0.6$  scaled) give rise to higher SIDS in Autumn (1.83) and Spring (1.70 SIDS/month). Radiation levels  $> 15.2$  Mj/hr ( $> 0.6$  scaled) give rise to increased SIDS particularly in spring (2.35 SIDS/month). Changing wind direction from E to SW leads to significantly increased SIDS with predicted winter, spring and autumn mean monthly SIDS of 4.35, 3.01 and 2.64 when the predominant direction is SW. Seasonality acts at both the lower and upper thresholds, whether for temperature variants or wind direction. Months where wind direction is mainly SW have higher levels of predicted SIDS deaths, than when winds are easterly. This is especially so in winter (4.35 vs 3.54) and spring (3.01 vs 2.31 SIDS/month), SW vs E contrast.

### 3 Conclusions

The lower thresholds for increased SIDS with cold weather are new, but possibly to be expected, given the winter peaks followed by autumn peaks of SIDS (Hudson et al., 2007, 2005). However, the existence of the upper thresholds of the temperature variants, above seasonality, are unexpected, and show an increased risk of SIDS in spring or autumn with increased heat. This supports a previous hypothesis of Dalrymple, 2004, namely that increased temperature adds to heightened SIDS risk due to possible over-wrapping and overheating of babies, in what was a significantly warmer ENSO period (1973-1989). The spring excess of SIDS may also be allergen related. Wind variations, more than air pressure changes, are easily perceived and may change parental care, i.e. over-wrapping. The GAMLSS model alludes to possible causal pathways, where wind, rain and humidity act on changing parental care, which may have an impact on SIDS. GAMLSS results are highly comparable to the findings from a discrete counts time series mixture approach (see Hudson et al., 2007). Indeed the GAMLSS AIC fit is very close to this mixtures approach, begging the question as to whether our new mixture approach could be easily implemented using the GAMLSS platform. This is the topic of future work.

### References

- Dalrymple, M. (2004). *Poisson Mixture Methods and Change Point Analyses to Study the Relationship Between Temporal Profiles of Sudden Infant Death Syndrome and Climate*. PhD thesis, University of Canterbury, Christchurch, New Zealand.
- Hudson, I., Fukuda, K., & Dalrymple, M. L. (2005). Climate-pollution impacts on sudden infant death. In *16th congress of the Modelling and Simulation Society of Australia and NZ*, Dec., Melbourne, (pp. 1-7).
- Hudson, I. L. Dalrymple M. L. & M. J. Faddy (2007) New Mixture Models for Discrete Counts Time Series: with an Application to Modelling Mortality and Climate in NZ. In *17th congress of the Modelling and Simulation Society of Australia and NZ*, 10-13 Dec., Christchurch, NZ (pp. 3024-3030).
- Malloy, M. & Freeman, D. (2000). Birth weight- and gestational age-specific sudden infant death syndrome mortality: United States, 1991 versus 1995. *Pediatrics*, 105, 1227-1231.
- McMichael, A.J., R.E. Woodruff and S. Hales (2006) Climate change and human health: present and future risks. *Lancet*, 367, 859-869.
- Rigby, R. A. and Stasinopoulos D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics, JRSS Series C*, 54(3), 507-554.

# Correlated gamma frailty model for linkage analysis on twin data with application to interval censored migraine data

M.A. Jonker<sup>1</sup>, S. Bhulai<sup>1</sup>, D.I. Boomsma<sup>2</sup>, R.S.L. Ligthart<sup>2</sup>, D. Posthuma<sup>2</sup> and A.W. van der Vaart<sup>1</sup>

<sup>1</sup> Department of Mathematics, Faculty of Sciences, VU University, De Boelelaan 1081a, 1081 HV Amsterdam, Nederland.

<sup>2</sup> Faculty of Psychology and Education, VU University Amsterdam.

**Abstract:** Several studies showed that the age at onset of some diseases might be genetically influenced. For these diseases ages at onset are often collected to map disease genes. Because ages are often censored, many standard techniques can not be used for mapping. In this paper we describe a correlated gamma frailty model for disease onset data of twins, to test whether the onset ages are genetically influenced and to find regions on the chromosomes where disease genes are located. The model is applied to interval censored age at onset of migraine.

**Keywords:** correlated gamma frailty model; family data; likelihood ratio test; statistical genetics; interval censoring.

## 1 Introduction

For many diseases (like breast cancer and Alzheimer disease) ages at onset of family members are highly correlated. For these diseases not only the occurrence but also the age at onset might be genetically influenced. Therefore, disease onset data of family members are often collected to find positions on the chromosomes where the disease gene(s) and the gene(s) that have influence on the age at onset of the disease are located. For many individuals the age at onset of the disease is censored and/or truncated, what makes many standard techniques for gene mapping less useful. Instead, statistical methods that combine techniques from survival analysis and from quantitative genetics can be used.

Our first aim is to test whether the age at which people experience their first migraine attack is genetically influenced (whether heritability is positive). Our second aim is to find regions on the chromosomes where genes might be located that influence the age at the first migraine attack. This means that we try to find regions on the chromosomes for where genotypic similarity between dizygotic twins are highly correlated with similarity of their phenotypes (ages at which the twin experience their first migraine

attack). Genotypic similarity is defined in terms of IBD numbers. Two alleles are IBD (identical by descent) if they are both copies of a common ancestral allele. The number of alleles a twin-pair has IBD at a marker (location on the chromosome with at least two possible alleles) is a measure of genotypic similarity between the two twins at the marker.

The data we use are from a longitudinal study of Dutch twins and their family members. The participants are members of the Netherlands Twin Registry. Every two years, between 1991 and 2002, all participants were asked to fill in a questionnaire on health, lifestyle, and personality. Based on the individual's answers, it was concluded whether the individual had ever experienced migraine before. So, the ages at migraine onset are interval censored; no exact ages are available, just age intervals in which the age at onset falls. For 3975 twin pairs migraine data is available. Moreover, for only 258 dizygotic twin pairs, migraine data and IBD information for 63 to 284 markers on the autosomes is available.

## 2 The Model

Define  $(T_1, T_2)$  as the ages at migraine onset for a twin pair. We want to build a regression model for  $(T_1, T_2)$  on the number of alleles IBD at a marker, so that

1. marginally  $T_1$  and  $T_2$  are independent of the IBD number (the IBD number does not contain any information on survival times),
2. marginally  $T_1$  and  $T_2$  are equal in distribution,
3. if the marker and the gene that influences the age at onset are in proximity on the chromosome, the association between  $T_1$  and  $T_2$  should increase with the number of alleles the twin has IBD at the marker; a dizygotic twin pair with two alleles IBD should have more similar survival times than a pair with zero alleles IBD at the marker.

We model the survival times of a twin,  $(T_1, T_2)$ , and the number of alleles the twin has IBD at a particular marker, with a correlated gamma frailty model in which the random effects ("frailties") account for the dependence between the survival times of the twins. Let  $(Z_1, Z_2)$  be the pair of frailties. Then, given  $(Z_1, Z_2)$  the survival times  $T_1$  and  $T_2$  are independent and have hazard functions  $t \rightarrow Z_1\lambda(t)$  and  $t \rightarrow Z_2\lambda(t)$  for  $\lambda$  an unknown baseline hazard function.

The frailties are decomposed as

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} U_{G,1} + U_C + U_{E,1} \\ U_{G,2} + U_C + U_{E,2} \end{pmatrix}.$$

The first term in the decomposition represents the genetic contribution to the age at onset of migraine due to part of the genome at the marker.

So, this term should be modeled dependently on the number of alleles IBD at the marker. Because the marginal distributions of  $T_1$  and  $T_2$  are independent of the IBD number, the correlation between  $U_{G,1}$  and  $U_{G,2}$  and not their marginal distributions depends on the IBD number. The second term in the decomposition represents the common environment and sharing alleles at markers that are unlinked to the disease gene and the third term represents specific environmental effects and non-sharing alleles. We assume that conditional  $IBD = k$ ,  $(U_{G,1}, U_{G,2}), U_C, U_{E,1}$  and  $U_{E,2}$  are independent and gamma distributed with

$$\rho_k := \text{cor}(Z_1, Z_2 | IBD = k) = \alpha + \beta k$$

for unknown parameters  $\alpha$  and  $\beta$ . If the marker is in proximity with a gene that influences the age at which people experience their first migraine attack,  $\beta$  should be positive. Then, an increasing number of alleles IBD at the marker coincides with an increasing correlation between the frailty variables, and thus, between the survival times.

The pair of frailty variables, conditional the IBD number, follow a bivariate gamma distribution. In that specific situation, an explicit expression of the bivariate survival function (conditional the IBD number) in terms of the marginal survival function  $S$  exists:

$$\begin{aligned} S_k(t_1, t_2) &= P(T_1 > t_1, T_2 > t_2 | IBD = k) \\ &= S(t_1)^{1-\rho_k} S(t_2)^{1-\rho_k} \left( S(t_1)^{-\sigma^2} + S(t_2)^{-\sigma^2} - 1 \right)^{-\rho_k/\sigma^2}, \end{aligned} \tag{1}$$

with  $\rho_k$  as defined before and  $\sigma^2 = \text{var}Z_1 = \text{var}Z_2$ . We consider two different models. In the first model  $S$  is known up to a finite-dimensional parameter (a parametric model) and in the second model  $S$  is completely unknown (a semi-parametric model).

For testing heritability for the age at onset of migraine (that is whether this age is genetically influenced) we use the same model, but with a different decomposition of the frailty variables. We decompose

$$Z_i = A_i + C + E_i \quad i = 1, 2$$

where  $A_i$  represents the additive genetics for the  $i$ th twin (half),  $C$  the common environment between the twins, and  $E_i$  the non-shared, specific environmental effects for the  $i$ th twin (half) (see also Yashin et al (1999)). We assume that  $(A_1, A_2), C$ , and  $E_1$  and  $E_2$  are independent and gamma distributed with the correlation between  $A_1$  and  $A_2$  equal to 1 for monozygotic twins and equal to 1/2 for dizygotic twins (monozygotic twins share all alleles IBD and dizygotic twins on average half of the alleles). Usually heritability is defined as the proportion of variance of the quantitative trait associated with genetic effects. Because the variance of the age at onset of migraine is difficult to compute, we define the trait in this context as the

frailty, and hence define heritability as the proportion of variance of the frailty associated with genetic effects:  $h^2 = \text{var}A_i/\text{var}Z_i$ . Basing heritability on a latent trait in this way is not unusual, and is closely linked to our method to test linkage (that is to test whether a marker is in proximity with a gene that influence the age at onset of migraine). Gamma frailty models for testing linkage or heritability have been proposed before (see, e.g., Li and Zhong (2002), Zhong and Li (2002), Yashin et al (1999)). The model we propose can be used for testing both.

### 3 Estimation and Testing

For testing heritability and linkage we use the likelihood ratio test. In case of linkage, we test  $H_0 : \beta = 0$  versus  $H_1 : \beta > 0$ ; so whether the IBD number for the marker influences the correlation between the survival times. The likelihood ratio statistic for  $n$  twin pairs is given by

$$\frac{\sup_{\alpha, \beta, \sigma^2, S} \prod_{i=1}^n L(\alpha, \beta, \sigma^2, S)[\text{twin pair } i]}{\sup_{\alpha, \sigma^2, S} \prod_{i=1}^n L(\alpha, 0, \sigma^2, S)[\text{twin pair } i]}$$

with  $L(\alpha, \beta, \sigma^2, S)[\text{twin pair } i]$  the likelihood for the  $i$ th twin in  $(\alpha, \beta, \sigma^2, S)$ . For the parametric model, and under the null hypothesis, the limit distribution of the likelihood ratio test statistic is a 1/2-1/2 mixture of a point mass at zero and a chi-square distribution with one degree of freedom. For the semi-parametric model the limit distribution of the test-statistic is unknown. Maximizing the likelihood with respect to  $S$  (and the finite-dimensional parameters) for all markers separately is very time consuming. As an alternative we estimate  $S$  in the numerator and the denominator of the likelihood ratio statistic by the NPMLE estimator based on survival data of independent individuals, and next maximize the likelihoods with respect to the remaining (finite-dimensional) parameters. We now have to estimate  $S$  only once. Generally, inserting an estimator for  $S$  into the likelihood ratio statistic destroys the asymptotic distribution of the likelihood ratio statistic. If we estimate  $S$  separately, we can also use data of individuals of whom only phenotypic data, and no genotypic data, is available. The number of observations for estimating  $S$  might then be considerably larger than is used to construct the test-statistic. In this case, we expect that the asymptotic distribution of the likelihood ratio statistic will not deviate much from the asymptotic distribution of the likelihood ratio statistic for the parametric model. This expectation was supported by the results of a simulation study and we also verified it by (heuristic) theoretical arguments. For some (quite) common diseases, like breast cancer, estimates of  $S$  are based on huge numbers of data. Then, the mixture distribution as mentioned above can be used for testing.

## 4 Application to interval censored migraine data

We applied the correlated frailty model to the interval censored migraine data. For all twin pairs the ages at onset,  $T_1$  and  $T_2$ , are never observed. Instead, it is observed that the ages fall into the intervals  $[U_1, V_1]$  and  $[U_2, V_2]$ . We assume that the censoring times  $(U_1, V_1)$  and  $(U_2, V_2)$  are independent of the survival times  $T_1$  and  $T_2$  and that the distribution of the observation times are uninformative. The likelihood function for the data of the  $i$ th dizygotic twin for the frailty model for linkage is proportional to

$$L(\alpha, \beta, \sigma^2, S)[twinpair] = \sum_{k=0}^2 Pr(IBD = k|MD) \\ \times (S_k(U_1, U_2) - S_k(U_1, V_2) - S_k(V_1, U_2) + S_k(V_1, V_2)),$$

with  $S_k$  the conditional bivariate survival function as in expression (1) and  $MD$  the observed marker data. A similar expression of the likelihood holds for the model for heritability. We estimated the unknown parameters and tested heritability and linkage for the parametric and the semi-parametric model as described before. The results for both models were similar.

More women than men suffer from migraine. We estimated the marginal survival function for the age at which people experience migraine for the first time separately for men and women. For ease of notation this is not shown in the expression of the likelihood. We estimated the NPMLE of the age at the first migraine attack based on 4,791 males and 6,796 females. These estimates were inserted into the likelihood before estimating the other parameters and testing heritability and linkage. When testing we assumed that the asymptotic distribution of the likelihood ratio test statistic equals the 1/2-1/2 mixture of zero and a chi-square distribution with 1 degree of freedom. In the parametric model we assumed that the marginal survival functions for males and females equal shifted exponential distributions with unknown shift and intensity parameter. These choices were based on the shape of the NPMLE's.

Heritability was estimated and tested based on data of almost 4,000 twin pairs. In the parametric model heritability was estimated as 0.42 (95% c.i.: [0.374; 0.461]) and as 0.37 (95% c.i.: [0.323; 0.415]) in the semi-parametric model. Both values were significant unequal to zero. So, there is a genetic contribution to the variability of age at onset of migraine.

Linkage analysis was based on 258 genotyped dizygotic twin pairs. The highest lod-score was 1.86 (parametric model) and 1.36 (semi-parametric model) both at the end of chromosome 19 (see Figure 1). (The lod-score is defined as the common logarithm with base 10 (instead of the natural logarithm with base  $e$ ) of the likelihood ratio.) In practice the value 3 is often taken as a threshold for significant lod-scores. In that case none of the lod-scores is significant. So, for none of the markers we may conclude that they are in proximity with genes that influence the age of onset of

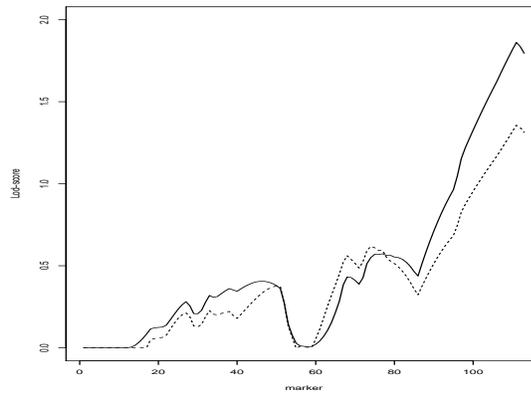


FIGURE 1. Lod-scores for testing linkage for the markers at chromosome 19 for the parametric model (solid curve) and the semi-parametric model (dashed curve).

migraine. More twins will be genotyped in the near future, so that it gets easier to detect interesting locations on the chromosomes.

**Acknowledgments:** This project was financially supported by NDNS+ and NWO. I acknowledge Piet Groeneboom for using his computer program for estimating the NPMLE for  $S$  based on interval censored data.

### References

- Jonker, M.A., Bhulai, S., Boomsma, D.I., Ligthart, R.S.L., Posthuma, D. van der Vaart, A.W. (2008). Gamma frailty model for linkage analysis with application to interval censored migraine data. *To appear in Biostatistics*.
- Li, H. and Zhong, X. (2002). Multivariate survival models induced by genetic frailties, with application to linkage analysis. *Biostatistics*, **3**(1).
- Yashin, A.I., Begun, A.Z., and Iachine, I.A. (1999). Genetic factors in susceptibility to death: a comparative analysis of bivariate survival models. *Journal of Epidemiology and Biostatistics*, **4**(1).
- Zhong, X. and Li, H. (2002). An additive genetic gamma frailty model for two-locus linkage analysis using sibship age of onset data. *Statistical Applications in Genetics and Molecular Biology*, **11**(1), Article 2.

# Modelling and synchronization of four Eucalyptus species via MTD and EKF

S. W. Kim<sup>1</sup>, I. L. Hudson<sup>1</sup>, M. Agrawal<sup>1</sup> and M. R. Keatley<sup>2</sup>

<sup>1</sup> School of Maths & Stats, UniSA, GPO Box 2471, SA 5001, Australia  
(Susan.Kim@postgrads.unisa.edu.au)

<sup>2</sup> School of Forest & Ecosystem Science, Melbourne University, Australia

**Abstract:** The extended Kalman filter (EKF) is a method to estimate the past, present and future status of non-linear time series data by minimising the mean square error using a set of pre-defined mathematical equations. Mixture transition distribution (MTD) is a method to estimate high order Markov chains with a reduced number of parameters. We combined these two approaches by using the functional and parameterization of a new extended MTD with interactions for the EKF to model flowering data of four eucalypts species. By adapting Moran's synchronization method to MTD and EKF residuals, three species, *E. leucoxylo*, *E. polyanthemos*, and *E. tricarpa* are shown to be synchronizing, while *E. microcarpa* is asynchronizing with *E. leucoxylo*. EKF estimates the Kalman gain and covariance matrix at each time point and so better detects asynchronous species pairs.

**Keywords:** Mixture Transition Distribution (MTD); Extended Kalman Filter (EKF); Phenology; Synchronization; Climate.

## 1 Introduction

Separation or lack of overlap of flowering time in eucalypts has been suggested as a mechanism for maintaining overall 'generic identity' of a species. If, however, flowering times and pollinators overlap in sympatric species, hybridization can occur between closely related eucalypts species (Keatley et al. 2004; Hudson et al. 2006). Examination of long-term synchrony establishes a baseline of flowering behaviour which may assist in detecting recent or future changes. Few studies have quantified eucalypts flowering overlap, within or between species, because of the rarity of such data in Australia (Keatley et al. 2004 and Hudson et al. 2006). This paper examines flowering synchrony over a 30 year period, 1940 to 1970, at the population level among four eucalypts species. The aim of this study is to test the combination of MTD (Berchtold 2004; Kim et al. 2005) and the extended Kalman filter (EKF) models (van der Merwe 2004), using the functional and parameterizations from the mixture transition distribution (MTD) analysis for the EKF model. Synchronization is tested via an adaptation of Moran's

classical synchrony statistic (Moran 1953) which uses residuals from our resultant model.

## 2 Data

Flowering data are from Havelock Forest Block, Maryborough, Victoria, Australia. These records are sourced from the former Forest Commission of Victoria covering the flowering of *E. leucoxyton* (Leu), *E. microcarpa* (Mic), *E. polyanthemos* (Pol) and *E. tricarpa* (Tri) in observation plots, between 1940 and 1970. Flowering intensity was calculated by using a rank score (from 0 to 5) based on the quantity and distribution of flowering (Keatley et al. 2004). We used discrete state on/off (1/0) flowering (Figure 1) as in Kim et al. (2005). We also used mean monthly diurnal temperature ( $meanT$ ) and monthly rainfall ( $rain$ ) as covariates and their interaction effects in the MTD model.

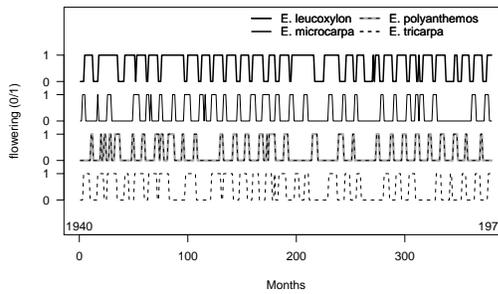


FIGURE 1. Flowering of four eucalypts species.

## 3 Methods and Models

The extended Kalman filter (EKF) is a method to estimate past, present and future status of non-linear time series data which minimizes the mean square error using a set of mathematical equations. EKF was implemented via a in-house modification of ReBel (van der Merwe 2004) on Matlab 7. We also adapted Moran's (Moran 1953) synchrony statistic to test the synchrony between species pairs via a novel combination of MTD and EKF.

### 3.1 Mixture Transition Distribution

An earlier study by Kim et al. (2005) used MARCH (Berchtold 2004) for MTD modelling. For this current study, an extended model for MTD analysis which accommodates interactions was developed using the AD Model

Builder<sup>TM</sup> (Fournier 2000). This work extends both MARCH and the work of Kim et al. (2005). The MTD model with covariates was defined by Kim et al. (2005) and also by Berchtold (2004). The probabilities associated with an order  $f$  MTD model with an interaction between two covariates,  $C_h$  for  $h = 1, 2$  ( $C_1 = meanT$  and  $C_2 = rain$ ), can be modeled by

$$P(Y_t = i_0 | Y_{t-f} = i_f, \dots, Y_{t-1} = i_1, C_1 = c_1, C_2 = c_2) \\ = \sum_{g=1}^f \lambda_g q(i_g, i_0) + \sum_{h=1}^2 \lambda_{f+h} d^h(c_h, i_0) + \lambda_{f+3} s(c_1, c_2, i_0),$$

where the  $q(i_g, i_0)$  are probabilities of the transition matrix from the previous flowering state,  $Y_{t-1}$ , to the current flowering state,  $Y_t$ , where  $Y_t \in \{0, 1\}$ ;  $d^h(c_h, i_0)$  are the probabilities of the transition matrices between covariate  $C_h \in \{0, 1\}$  and  $Y_t$ ;  $s(c_1, c_2, i_0)$  are the probabilities of the transition matrix between interactions and  $Y_t$ ; and  $\lambda_g$  are weights on probabilities from each order, covariates, and interactions with  $\sum_{g=1}^{f+3} \lambda_g = 1$ . A preliminary study of this model was tested for a NZ trawling study (Kim et al. 2006). This new model is different from MARCH in terms of its minimization process because ADMB<sup>TM</sup> uses auto-differentiation as a minimization tool, it is also shown to be computationally less intensive than MARCH. ADMB<sup>TM</sup> is widely used in fisheries and uses basic C++ language for coding because the likelihood can be defined in any term and function can be easily written using C++ language.

The major advantage of our new model is that its run-time is more than 10 times shorter (less than 1 minute vs 2 days) and can be run from a batch file in DOS. Hence multiple models can be tested one after the other in remote mode. The outputs can also be appended into one file to be easily accessed by any graphical software.

## 4 Results

Parameters and the goodness of fit (likelihood) of the MTD model for four species are shown in Table 1 for the given significant lags ( $O_j$ , where  $j=1, \dots, 12$  months), covariates ( $meanT$  and  $rain$ ) and interactions ( $meanT * rain$ ). Estimated parameters for the MTD model generally shows a 1 month lag effect, and 9 to 12 months lag effects. Mean diurnal temperature ( $meanT$ ) has a significant effect on flowering for all species;  $rain$  impacts *E. tricarpa* (Tri) only and an interaction effect between  $rain$  and  $meanT$  exists only for *E. polyanthemos* (Pol). Both *E. leucoxydon* (Leu) and Tri flower when  $meanT$  was low, while the other two species when  $meanT$  is high. Pol flowers at low  $meanT$  with below average rainfall and also flowers at high  $meanT$  with above average rainfall.

Synchronization among species was tested using Moran's correlation method on cross-residuals. The significant correlations from both the MTD and

TABLE 1. Likelihood and model parameters of the MTD model.

Species	$L$	model parameters					
Leu	123.81	$O_1$ (0.63), $O_{11}$ (0.15), $meanT$ (0.18)					
Mic	115.98	$O_1$ (0.53), $O_{12}$ (0.27), $meanT$ (0.14)					
Pol	139.68	$O_1$ (0.53), $O_9$ (0.06), $O_{11}$ (0.16), $O_{12}$ (0.1), $meanT$ (0.08), $meanT * rain$ (0.07)					
Tri	129.47	$O_1$ (0.61), $O_9$ (0.05), $O_{11}$ (0.09), $meanT$ (0.17), $rain$ (0.06)					

Species	Previous flowering		$meanT$		$rain$	
	off	on	low	high	less	more
Leu	0.09	1.00	1.00	0.00	0.99	0.73
Mic	0.00	1.00	0.00	1.00	0.39	0.28
Pol	0.01	1.00	0.00	0.33	0.33	0.41
Tri	0.00	1.00	1.00	0.00	0.00	1.00

Species	Interaction			
	low $meanT$		high $meanT$	
	less $rain$	more $rain$	less $rain$	more $rain$
Pol	0.73	0.18	0.44	0.59

EKF models showed that results from the two models generally agreed but that models with EKF showed stronger synchronization between Leu and Tri (0.33 and 0.26 vs 0.11, Table 2). Residuals between Leu and Pol showed significant synchronization between these two species via both the MTD and EKF models. The EKF model, however, showed significant asynchrony between Pol and *E. microcarpa* (Mic) in agreement with Hudson et

TABLE 2. Significant Moran correlations. (p<0.05.)

MTD				EKF			
leu	mic	pol	tri	leu	mic	pol	tri
pol	tri	leu	mic	tri	tri	leu	leu
(0.16)	(0.14)	(0.14)	(0.15)	(0.33)	(0.12)	(0.19)	(0.26)
tri				pol	<b>leu</b>	<b>mic</b>	
(0.11)				(0.18)	<b>(-0.17)</b>	<b>(-0.1)</b>	
<b>mic</b>							
<b>(-0.14)<sup>+</sup></b>							

<sup>+</sup>: bolded denotes asynchronous species pairs.

TABLE 3. Significant Synchronous and asynchronous groups via MTD/EKF models.

MTD			EKF		
Species	Synch	Asynch	Species	Synch	Asynch
leu	pol, tri	<b>mic</b>	leu	tri, pol	
mic	tri		mic	tri	<b>leu</b>
pol	leu		pol	leu	<b>mic</b>
tri	<b>mic</b>		tri	<b>leu</b>	

al. (2006) while the MTD model showed no such significant relation. Overall, Leu synchronizes with Pol and Tri, but asynchronizes with Mic. Also, Mic synchronizes with Tri (Table 3). Interestingly, the EKF model shows that Pol asynchronizes with Mic. This relationship is also shown schematically in Figure 2 for the EKF model. EKF estimates the Kalman gain and covariance matrix at each time point and so better detects asynchronous species pairs.

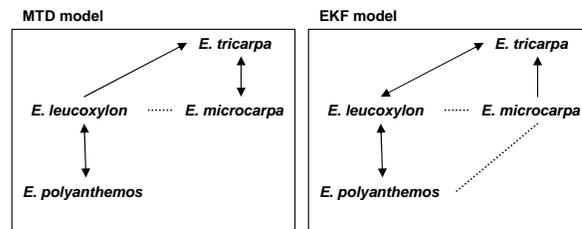


FIGURE 2. Relationship among four eucalypts species. Solid line: synchronous pairs; dashed line: asynchronous pairs of species.

## 5 Discussion

Our results agree with Keatley et al. (2004), who established that a cross between Leu and Mic is impossible; and also with a recent study by Hudson et al (2006) using Moran's ARIMA method without covariates (nor MTD) on this flowering data, which however showed that Pol and Tri are asynchronous. This present study shows however no such significant relationship. Our synchrony values assist in quantifying the likelihood of hybridization between species, and provide a baseline which may assist in detecting recent or future changes. The highest degree of synchrony occurs

between Leu and Tri (EKF) which indicates the potential for intense competition for potential pollinators, and therefore the prospect for a high level of hybridization. MTD model showed Leu and Tri have decreased flowering with increasing temperature opposite to Mic and Pol which have increased flowering with decreasing temperature. Tri has increased flowering with increasing rain. High temperature and high rain increase flowering for Pol.

## References

- Berchtold, A. (2004). March v.2.01. Markovian models Computation and Analysis Users guide. <http://www.andreberchtold.com/march.html>.
- Fournier, D.A. (2000). AD Model Builder, Version 5.0.1. Otter Research Ltd, Canada.
- Fournier, D.A. (1996). AUTODIFF, A C++ array language extension with automatic differentiation for use in nonlinear modeling and statistics, Otter Res. Ltd., Canada.
- Hudson, I.L., Keatley, M.R., Kim, S.W., Kang, I. (2006). Synchronicity in Phenology: from PAP Moran to now. In: *Australian Statistical Conference/New Zealand Statistical Association (ASC/NZSA) conference*, 3-6 July, 2006, Auckland, New Zealand.
- Keatley, M.R. (1999). The flowering phenology of box-ironbark eucalypts in the Maryborough Region, Victoria. PhD Thesis, University of Melbourne.
- Keatley, M.R., Hudson, I.L. and Fletcher, T.D. (2004). Long-term flowering synchrony of Box-Ironbark Eucalypts. *Australian Journal of Botany*, **52**(1), 47-54.
- Kim, S.W. and Hudson, I.L. (2006). Extending Mixture Transition Distribution (MTD) methods to incorporate interactions: Links to species synchrony and phenology. In: *Australian Statistical Conference/New Zealand Statistical Association (ASC/NZSA) conference*, 3-6 July, 2006, Auckland, New Zealand.
- Kim, S.W., Hudson, I.L., Keatley, M.R. (2005). MTD analysis of flowering and climatic states. In: *Proceedings of International Workshop in Statistical Modelling (IWSM)*, 10-15 July, 2005, Sydney, Australia. 305-312.
- Moran, P.A.P. (1953). The statistical analysis of the Canadian lynx cycle II. *Australian Journal of Zoology*, **1**, 291-298.
- Van der Merwe, R. (2004). Quick-start guide for ReBel toolkit, Oregon Health and Science University.

# Dose-response modeling with bivariate binary data under model uncertainty

Bernhard Klingenberg<sup>1</sup>

<sup>1</sup> Department of Mathematics and Statistics, Williams College, Williamstown, MA, 01267 and Institute of Statistics, Graz University of Technology, Steyrergasse 17/IV, 8010 Graz, Austria. *E-mail: bklingen@williams.edu*

**Abstract:** When modeling a dose-response for a drug based on bivariate binary data such as two co-primary efficacy endpoints in early stages of development, there is usually uncertainty about the form of the true underlying dose-response shape. Often, investigators fit several different models that are deemed plausible, but later fail to acknowledge this uncertainty in inference that is based on a single model selected via e.g. the minimum AIC criterion. This leads to an inflation in the error of the proof of activity decision and may also result in poor estimation of a target dose that is used in future trials. In this article we acknowledge model uncertainty by fitting several candidate models for a bivariate binary response and develop a principled approach to establish proof of activity and a target dose.

**Keywords:** Dose estimation; Multiplicity adjustment; Proof of Concept.

## 1 Introduction

Dose-response studies are important tools for investigating the existence, nature and extent of a dose effect on efficacy or safety outcomes in drug development, toxicology and related areas. The following four questions, usually in this order, are of prime interest: i.) Is there any evidence of a dose effect (i.e., *Proof of Activity*), ii.) Which doses exhibit a response different from the control response, iii.) What is the nature of the dose-response relationship and iv.) What dose should be selected for further studies/marketing (i.e., *target dose estimation*)? Here we suggest to answer questions i) to iv) in a unified framework using statistical modelling. To meet the criticism that results based on modelling depend too much on assumptions on the underlying dose-response shape, we incorporate model uncertainty into our methods of establishing Proof of Activity (PoA) and estimating a target dose. We do this by considering several candidate models for the true dose-response which we simultaneously use for inference. Recently, Klingenberg (2008) has shown via simulation that for univariate binary dose-response data from simple parallel group designs, common methods of establishing PoA with a model based approach (i.e., basing the PoA decision on the P-value of a test using the model with the smallest AIC

among a set of candidate models) lead to inflated type I errors. This means that the probability of carrying forward an ineffective drug into Phase III is not well controlled. Motivated by the example below, in this article we consider *bivariate* binary response data  $(Y_{1j}, Y_{2j})$ , where  $Y_{ij}$  is the binary indicator of efficacy (or safety) for endpoint  $i$  ( $i=1,2$ ) for subjects randomized to one out of  $k$  dose levels  $d_j$ ,  $j = 1$  (Placebo),  $\dots, k$ . We develop a principled approach to both, establishing PoA and estimating a target dose (such as the minimum effective dose, MED) under model uncertainty, controlling the type I error of an incorrect PoA decision.

## 2 Proof of Activity under model uncertainty

We assume independent multinomial distributions  $\text{Mult}\{n_j; \pi_{00}(d_j), \pi_{01}(d_j), \pi_{10}(d_j), \pi_{11}(d_j)\}$  for the counts in each of the  $j = 1, \dots, k$   $2 \times 2$  tables that crossclassify the bivariate response at dose level  $d_j$ . Here,  $n_j$  is the number of subjects randomized to dose level  $d_j$  and  $\pi_{ab}(d_j) = P_{d_j}(Y_{1j} = a, Y_{2j} = b)$ ,  $a, b \in 0, 1$  is the probability of response  $(a, b)$  at dose level  $d_j$ . From now on, we assume that  $Y_{1j}$  and  $Y_{2j}$  are both measuring efficacy, for instance when there are two co-primary endpoints in a clinical trial (see example below). One simple approach for establishing PoA and estimating the MED would be to collapse the two efficacy endpoints into a single one, recording an overall success if both endpoint show efficacy, and failure otherwise. However, one drawback of this method is that it declares outcomes of type  $(0, 1)$  and  $(1, 0)$  as failures, which might lead to a loss of power in establishing PoA and too high an estimate for the MED. In a multivariate approach, we model separately the marginal success probabilities  $\pi_{1+}(d_j)$  and  $\pi_{+1}(d_j)$  for each endpoint in terms of dose, taking into account the dependence between them by modeling the log-odds (Palmgren, 1989).

Let  $M_s$  denote a specific constellation of two marginal models and one model for the association in  $(Y_{1j}, Y_{2j})$ . To accommodate model uncertainty in the PoA decision and subsequent target dose estimation, we start by considering several plausible dose-response models  $M_s$ ,  $s = 1, \dots, m$  that differ in the way they model the two margins and/or the association. Let  $\mathcal{M} = \{M_1, \dots, M_m\}$  be a candidate set spanned by  $m$  such models. Since target doses such as the MED depend on the assumed shapes for each margin, considering several shapes a priori makes the procedure more robust to model misspecification. To decide which of the candidate models, if any, significantly pick up the dose-response signal, we compare each one to the no-effect model via a signed and penalized likelihood ratio statistic

$$T_s = \pm \{-2[l_0 - l_s]\} - 2df_s,$$

where  $l_s$  is the maximized multinomial log-likelihood under candidate model  $M_s$ , and  $M_0$  and  $l_0$ , respectively correspond to the no dose effect model

$\pi_{1+}(d_j) = \alpha_1, \pi_{+1}(d_j) = \alpha_2, OR_j = \frac{\pi_{00}(d_j)\pi_{11}(d_j)}{\pi_{01}(d_j)\pi_{10}(d_j)} = \alpha_3$  that assumes constant margins and odds ratios across all dose levels. Naturally, we are only interested in models that show a positive dose effect. Although straightforward for monotone marginal profiles, in general we define a dose effect as positive if  $\hat{\pi}(d_{\max}) > \hat{\pi}(d_1)$ , where  $d_{\max} = \operatorname{argmax}_d |\hat{\pi}(d) - \hat{\pi}(d_1)|$  is the dose at which the maximum absolute effect relative to placebo occurs. This condition must be met for both, the first ( $\hat{\pi} \equiv \hat{\pi}_{1+}$ ) and second ( $\hat{\pi} \equiv \hat{\pi}_{+1}$ ) margin, otherwise we declare the dose effect as negative. The purpose of the  $\pm$  sign in  $T_s$  is then to give models with an estimated (but potentially significant) *negative* dose effect a small value of the test statistic, moving it to the lower tail. Finally, we penalize fitting more complex models by subtracting two times the difference in the number of parameters between  $M_s$  and  $M_0$ . Up to the  $\pm$  sign,  $T_s$  is equivalent to the differences in the AICs of  $M_0$  and  $M_s$ , a statistic favored for model selection by Burnham and Anderson (2002).

Under the null hypothesis of no dose effect, bivariate responses  $(Y_{1j}, Y_{2j})$  are exchangeable among the  $k$  dose levels. To evaluate the significance of  $T_s$  under simultaneous inference with all  $m$  candidate models, we build the permutation distribution of its maximum,  $\max_s T_s$ , by fitting each  $M_s$  to a random sample of all possible assignments of the  $(Y_{1j}, Y_{2j})$  responses to dose levels. (Although the asymptotic distribution of  $T_s$  is proportional to a Chi-square, the distribution of the maximum is not straightforward to derive due to correlation between the test statistics.) The permutation distribution yields raw and, using the full closed-testing methodology or the step-down approach in Westfall and Young (1993), multiplicity adjusted P-values for the PoA test with each candidate model. These adjusted P-values now appropriately account for the uncertainty in the dose-response shapes and control the familywise error rate of a wrong PoA decision under the family of candidate dose-response models. This control is in the strong sense, that is, under any combination of true and false null hypotheses, where the  $s$ -th null hypothesis concerns the testing of no dose effect under model  $M_s$ .

### 3 MED estimation

After establishing PoA with at least one candidate model (smallest multiplicity adjusted P-value less than the chosen overall type I error rate), the procedure moves to estimating the MED. For a given model, the MED is the smallest dose that shows a clinically relevant and statistical significant improvement over placebo, i.e.,

$$\widehat{\text{MED}} = \operatorname{argmin}_{d \in (d_1, d_k)} \{ \hat{\pi}(d) > \hat{\pi}(d_1) + \Delta, \hat{\pi}^L(d) > \hat{\pi}(d_1) \}, \quad (1)$$

where  $\Delta$  is the clinically relevant effect (may differ by margins) and  $\hat{\pi}^L(d)$  is the lower limit of a  $100(1 - \gamma/2)$  confidence interval for  $\pi(d)$ . We say

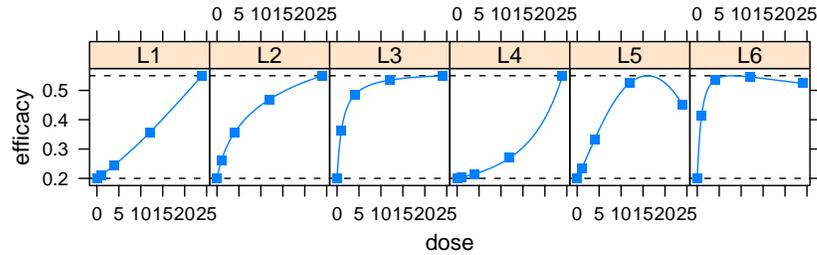


FIGURE 1. Some plausible dose-response models for margins of the IBS compound

that the MED does not exist if there is no  $d \in (d_1, d_k]$  for which the conditions are satisfied. There are several possibilities of defining the MED with bivariate data. Here, we estimate the MED from each of the two margins (i.e., with  $\hat{\pi} \equiv \hat{\pi}_{1+}$  and  $\hat{\pi} \equiv \hat{\pi}_{+1}$ ) and take the maximum over the two MED estimates as the MED estimate for the model, so that clinical relevance and statistical significance is guaranteed for both margins. A different approach would derive the MED by setting  $\hat{\pi} \equiv \hat{\pi}_{11}$ . We can take the MED that corresponds to the model with the smallest adjusted P-value as the overall estimate for the target dose. However, a MED estimator that incorporates model uncertainty is based on a weighted average of the MEDs from each candidate model (so they exist),  $w\widehat{\text{MED}} = \sum_s w_s \widehat{\text{MED}}_s / \sum_s w_s$ , with weights  $w_s = \exp(T_s/2)$ . In this way, the ratio of weights  $w_s$  and  $w_{s'}$  attached to the MED from two candidate models  $M_s$  and  $M_{s'}$  reflects their (penalized) likelihood distance.

#### 4 Example: PoA and MED for a diarrhea compound

Guidelines from the European Agency for the Evaluation of Medicinal Products for proving efficacy for compounds against Irritable Bowl Syndrome (IBS) demand that two endpoints, relief of abdominal pain and relief from overall GI-tract symptoms be considered jointly. Figure 1 and Table 1 show several shapes for the marginal probability of each of this endpoints (plotted using information such as the expected placebo and maximum dose effect from prior studies) that are deemed plausible by the clinical team developing a compound against IBS at dose levels  $d = (0, 1, 4, 12, 24)mg$ . The different shapes of these models are generated with linear predictors of fractional polynomial form (Roystone and Altman, 1994) that allow for a broad dose-response space. For instance, the clinical team was uncertain about the rate of increase in both margins at low doses and about the mono-

TABLE 1. Various shapes for the marginal efficacy of the IBS compound

Shape	Linear Predictor	Shape	Linear Predictor
$L_0$ :	$\alpha$	$L_4$ :	$\alpha + \beta \exp(\exp(d_j / \max(d_j)))$
$L_1$ :	$\alpha + \beta d_j$	$L_5$ :	$\alpha + \beta d_j + \gamma d_j^2$
$L_2$ :	$\alpha + \beta \log(d_j + 1)$	$L_6$ :	$\alpha + \beta \log(d_j + 1) + \gamma / (d_j + 1)$
$L_3$ :	$\alpha + \beta / (d_j + 1)$		

TABLE 2. Candidate dose-response models for the efficacy of the IBS compound. The triplets in the first column refer to which linear predictor is used to model the two margins  $\text{logit}[\pi_{1+}(d_j)]$  and  $\text{logit}[\pi_{+1}(d_j)]$  and the log-odds ratio  $\log(OR_j)$ .

candidate model	AIC	$T$	perm. P-value	adj. P-value	MED (mg)	weight (%)
$M_0$ : ( $L_0, L_0, L_0$ )	900.1	—	—	—	—	—
$M_1$ : ( $L_1, L_2, L_0$ )	898.2	1.92	0.4196	0.5875	15.9	4.3
$M_2$ : ( $L_2, L_2, L_0$ )	894.3	5.80	0.0031	0.0101	6.0	29.8
$M_3$ : ( $L_3, L_2, L_0$ )	895.6	4.55	0.0060	0.0218	9.6	15.9
$M_4$ : ( $L_4, L_2, L_0$ )	900.7	-0.55	0.6408	0.7593	NA	1.2
$M_5$ : ( $L_1, L_6, L_0$ )	900.0	0.08	0.6212	0.7593	15.4	1.7
$M_6$ : ( $L_2, L_6, L_0$ )	896.1	4.03	0.2682	0.4780	4.0	12.3
$M_7$ : ( $L_3, L_6, L_0$ )	894.0	6.08	0.0030	0.0101	1.0	34.3
$M_8$ : ( $L_4, L_6, L_0$ )	902.6	-2.45	0.7631	0.7631	NA	0.5

tonicity of the dose-response curve for the second margin. Hence, the candidate set should include models that incorporate various scenarios for the slope or non-monotonicity. A potential candidate set for the IBS compound is shown in Table 2. For instance, consider model  $M_6$  which postulates  $\text{logit}[\pi_{1+}(d_j)] = \alpha_1 + \beta_1 \log(d_j + 1)$ ,  $\text{logit}[\pi_{+1}(d_j)] = \alpha_2 + \beta_2 \log(d_j + 1) + \gamma_2 / (d_j + 1)$ ,  $\log(OR_j) = \log(\pi_{00}(d_j)\pi_{11}(d_j) / \pi_{01}(d_j)\pi_{10}(d_j)) = \alpha_3$ . When fitting  $M_6$  to (slightly modified) clinical trials data on the efficacy of the IBS compound, a test of no dose effect compares this model to the no-effect model with  $\beta_1 = \beta_2 = \gamma_2 = 0$ . Its penalized likelihood ratio statistic  $T_6 = + \{-2[-447.06 - (-442.04)]\} - 2 * 3 = 4.03$ . However, the maximum value of 6.08 for the test statistic over all candidate models occurs under model  $M_7$  that allows for a steeper rate of increase in the marginal odd of abdominal pain. This model would also be considered for inference (testing PoA and estimating the MED) when the decision is based on the minimum AIC criterion. The likelihood ratio PoA test with this model has an asymptotic P-value of 0.0035, very similar to the raw permutation P-value displayed in Table 2 that shows the proportion of all 10,000 permutations that yielded a  $T_7$  value as large or larger than the observe one. However, this

statement of significance is conditional on the selected model and ignores the uncertainty the clinical team had at the start of the trial, as expressed in the candidate set. Under simultaneous inference with all candidate models, i.e., testing PoA under each model, the multiplicity adjusted P-value for the PoA test in model  $M_7$  equals 0.0101, which is about three times larger but still significant at a conventional overall type I error rate of 2.5%, say. This multiplicity adjusted P-value is derived from the closed testing framework using the permutation distribution of  $\max_s T_s$ . The MED derived from  $M_7$  (with a clinically relevant effect of  $\Delta = 10\%$  for both margins and  $\gamma = 5\%$ ) equals 1.0mg. Incorporating model uncertainty, the weighted estimate  $\widehat{\text{wMED}} = 5.2\text{mg}$  uses the MED estimates from all candidate models with weights displayed in Table 2 and may be more appropriate.

## 5 Discussion

The use of multiplicity adjusted P-values guarantees that the type I error rate for one or more incorrect PoC decision (based on various candidate models) when the drug is actually ineffective is controlled at the desired level (e.g., 2.5%). This is in contrast to the common habit of “model fishing” or gambling on an a priori specified shape in the study protocol. A disadvantage of our method is that non-linear (on the link scale) dose response models are harder to incorporate because they would not provide convergent fits for many permutations. Here we treated the case where both variables describe efficacy. Equally interesting is the case where  $Y_{1j}$  is a binary primary efficacy variable and  $Y_{2j}$  a binary variable describing safety at dose level  $j$ .

### References

- Burnham, K., and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer.
- Klingenberg, B. (2008). Proof of Concept and dose estimation with binary responses. Under second review by Statistics in Medicine.
- Palmgren, J. (1989). Regression models for bivariate binary responses. Tech. Rep. 101, Department of Biostatistics, University of Washington, Seattle.
- Roystone, P., and Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, **43**, 429–467.
- Westfall, P., and Young, S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley & Sons.

# The Robustness of the Parameter and Standard Error Estimates in Trials with Partially Nested Data. A Simulation Study

Elly J.H. Korendijk<sup>1</sup>, Cora J.M. Maas<sup>1</sup>, Joop J. Hox<sup>1</sup>, Mirjam Moerbeek<sup>1</sup>

<sup>1</sup> Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University. P.O. Box 80140, 3508 TC Utrecht, the Netherlands

**Abstract:** In experimental research trials in which the experimental condition consists of subjects nested within clusters while the control condition consists of individuals who are not nested within clusters, are not uncommon. Though the specific structure of such data can be modelled in special programs for multilevel data like MLwiN, it can not be done in general statistics programs like SPSS and SAS. In practice most researchers use approximate models when analyzing such data. The current study uses simulation to assess the consequences on the parameter and standard error estimates of this practice.

**Keywords:** multilevel analysis; clustering versus non-clustering; cluster randomized trials; effective sample size

## 1 Introduction

Multilevel analysis has become general practice for data with a hierarchical structure. In such data, observations on the lowest level are nested within a higher level, and hence these observations are generally not independent, for example pupils in classes, employees in corporations and patients in general practices.

A specific example of nested data is the so called cluster randomized trial. The name refers to an experimental trial in which groups or clusters are at random assigned to the experimental or control condition. In most cluster randomized trials, the experimental as well as the control condition consist of clusters. However, trials in which the experimental treatment is delivered to clusters, while the subjects in the control condition receive no treatment or individual treatment are not uncommon. In these trials, data in the experimental condition is nested while the data in the control condition is not. From here on we will refer to such designs as trials with partially nested data.

The specific structure of partially nested data can be modelled in the multilevel program MLwiN, but it is not possible in standard statistical software

like SPSS and SAS. Therefore, in practice researchers often use approximate models. In general this is either an “ordinary” multilevel analysis in which the non-nesting in the control condition is ignored, or an analysis of variance in which the nesting in the experimental condition is ignored. Until now it is not systematically investigated what the consequences are on the parameter and standard error estimates when the partially nested data structure is not correctly modelled. In the current study, simulation is used to assess these consequences.

## 2 Estimation models

In this section the three above named models are presented, to start with the appropriate model for partially nested data (see also Roberts & Roberts, 2005), followed by the ordinary multilevel model and the model for an analysis of variance.

Since in a trial with partially nested data there is nesting in the experimental condition while in the control condition there is not, the outcome in each condition is described by different models. In a model without covariates the outcome for a subject in the control condition is described by

$$Y_i = \gamma_0 + r_i, \quad (1)$$

in which  $Y_i$  is the continuous outcome for subject  $i$ ,  $\gamma_0$  is the mean outcome for the control condition and  $r_i$  is the residual for subject  $i$ . The residuals  $r_i$  are assumed to be normally distributed with zero mean and variance  $\sigma_r^2$ . The outcome for a subject in the experimental condition is described by

$$Y_{ij} = \gamma_0 + \gamma_1 + u_j + e_{ij} \quad (2)$$

in which  $Y_{ij}$  is the continuous outcome for subject  $i$  in cluster  $j$ ,  $\gamma_0$  the mean outcome in the control condition and  $\gamma_1$  the mean treatment effect. Note that, since all subjects in the experimental condition receive treatment, a treatment indicator is redundant here. The nesting in this condition is modelled by the two separate residual terms:  $u_j$  is the cluster level residual for cluster  $j$ , and  $e_{ij}$  the subject level residual for subject  $i$  in cluster  $j$ . Both residuals are assumed to be identically and independently normally distributed, with zero mean and variance  $\sigma_u^2$  and  $\sigma_e^2$  respectively.

The combination of equation (1) and (2) gives the model for partially nested data

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + (u_j + e_{ij}) T_j + r_i (1 - T_j), \quad (3)$$

in which  $T_j$  is the treatment indicator for cluster  $j$ . It should be noted that in the control condition there are no clusters and in this condition the  $j$  for all subjects is arbitrary. The treatment indicator  $T_j$  is coded 0 for the control condition and 1 for the experimental condition, so that the intercept  $\gamma_0$  still can be interpreted as the mean outcome in the control condition and

the slope  $\gamma_1$  as the mean treatment effect. In the experimental condition the residual is  $(u_j + e_{ij})$  and in the non nested control condition it is  $r_i$ . Equation (3) can easily be extended with covariates on the cluster as well as on the subject level. In the current study we use a model with one covariate on the subject level

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_{ij} + (u_j + e_{ij})T_j + r_i(1 - T_j), \quad (4)$$

in which  $\gamma_2$  is the slope for the subject level covariate  $X$ .

The ordinary multilevel model, with one covariate at the subject level is given by

$$Y_{ij} = \gamma_0 + \gamma_1 T_j + \gamma_2 X_{ij} + u_j + e_{ij}, \quad (5)$$

in which  $Y_{ij}$  is the continuous outcome for subject  $i$  in cluster  $j$ .  $T_j$  is the treatment indicator which coded  $T = 1$  in the experimental condition and  $T = 0$  in the control condition, the intercept  $\gamma_0$  again can be interpreted as the mean outcome in the control condition and the slope  $\gamma_1$  as the mean treatment effect. The differences with model (4) are twofold. Firstly, since the subjects in the control condition in model (5) are supposed to be nested within one cluster, the  $j$  is the same for all subjects in this condition. Secondly, in the control condition model (5) has not just a residual  $e_{ij}$  at the subject level, but a residual  $u_j$  at the cluster level as well.

The second approximate model that is used in practice, is the analysis of variance model which is given by

$$Y_i = \gamma_0 + \gamma_1 T_i + \gamma_2 X_i + e_i, \quad (6)$$

in which  $Y_i$  is a continuous outcome variable for subject  $i$ . In this model the subjects are assumed not to be nested within clusters and so in this model there is only a subscript for the subject. The main difference with model (5) is the absence of the cluster level residual  $u_j$ .

Summing up, the model for partially nested data has residual terms at both levels in the experimental condition and just a subject level residual in the control condition, the ordinary multilevel regression has residual terms at both levels in both conditions, and the analysis of variance model has just a residual at the subject level in both conditions.

### 3 Intraclass correlation coefficient and effective sample size

The dependency between individual outcomes within a cluster is expressed by the intraclass correlation coefficient (ICC), which is equal to the proportion variance of the outcome variable at the cluster level:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}. \quad (7)$$

In which  $\rho$  is the ICC,  $\sigma_u^2$  the variance at the cluster level and  $\sigma_e^2$  the variance at the subject level. The ICC affects the design effect, which in its turn is an important feature for the determination of the effective sample size. Sample sizes are said to be effective when the standard errors of the fixed parameter estimates are comparable. Since clustering affects the standard error estimates, clustering should be taken into account by means of the design effect, when the sample size is determined.

The design effect is equal to  $1 + (\text{cluster.size} - 1)\rho$ . In a trial with partially nested data it is desirable that the standard errors of the means of both conditions are comparable. Hence, when the number of clusters and the cluster size in the experimental condition are known, the sample size in the control condition is determined by

$$n_{effective} = \frac{k * m}{design.effect}. \quad (8)$$

In which  $n_{effective}$  is the number of subjects in the control condition,  $k$  the number of clusters and  $m$  the cluster size in the experimental condition.

#### 4 Method

In each condition 3,000 data sets based on model (4) are simulated, and estimated with model (4), (5) and (6) in sequence. Both, simulation and estimation are done with the MLwiN (2.1) software (Rasbash, J., Steele, F., Browne, W. & Prosser, B., 2004)

Based on practice and earlier studies on robustness of parameters with nesting in both conditions (e.g. Maas & Hox, 2005), two variables are varied: the number of clusters in the experimental condition (10, 30 and 50) and the ICC (.05, .1 and .3).

The following variables are fixed: the cluster size ( $m = 5$ ) the treatment effect ( $\gamma_1 = 0.3$ ), the effect of the subject level covariate  $X$  ( $\gamma_2 = 0.3$ ) and the first level variance in the experimental condition ( $\sigma_e^2 = 1$ ). The second level variance  $\sigma_u^2$  follows from the first level variance and the value of the ICC. Homoscedasticity is assumed, so it follows that the variance in the control arm is equal to the sum of both variance components in the experimental condition  $\sigma_r^2 = \sigma_u^2 + \sigma_e^2$ . Equal effective sample sizes are used in all conditions.

#### 5 Preliminary results and remarks

When the data is analyzed with the appropriate model for partially nested data (i.e. using model (4)), all parameters are estimated without bias. The standard errors of the fixed parameters are also well estimated. The standard errors of the random parameters are mostly underestimated (see Table

TABLE 1. The coverage for each estimated parameter per model per condition

Model / Condition	$\sigma_u^2$	$\sigma_e^2$	$\sigma_r^2$	Intercept	Treatment effect	Effect of $X$
Model(4)						
ICC = .05	.9204*	.9346*	.9343*	.9406	.9378	.9381
ICC = .10	.9134*	.9390	.9357*	.9458	.9489	.9439
ICC = .20	.9052*	.9367	.9287*	.9456	.9470	.9396
k = 10	.8804*	.9236*	.9152*	.9376	.9404	.9301*
k = 30	.9228*	.9401	.9404	.9460	.9449	.9417
k = 50	.9359*	.9466	.9430	.9474	.9483	.9498
Model (5)						
ICC = .05	.6773**	.9447		.9081	.9578	.9472
ICC = .10	.7737**	.9330**		.9526	.9701*	.9431
ICC = .20	.8426**	.8870**		.9833*	.9864*	.9469
k = 10	.6930**	.9527		.9033*	.9453	.9430
k = 30	.7788**	.9244**		.9611	.9802*	.9487
k = 50	.8218**	.8876**		.9796*	.9888*	.9456
Model (6)						
ICC = .05		.9340***		.9444	.9402	.9469
ICC = .10		.8364***		.9486	.9282*	.9457
ICC = .20		.4903***		.9437	.9076*	.9498
k = 10		.9172***		.9432	.9223*	.9471
k = 30		.7327***		.9456	.9249*	.9474
k = 50		.6109***		.9480	.9288*	.9478

\* significant outside the interval  $.9365 < CI < .9635$ ; \*\* significant outside the interval  $.9369 < CI < .9631$ ; \*\*\* significant outside the interval  $.9377 < CI < .9623$ .

1), which confirms earlier findings that the Wald test is not suitable to test whether the variance equals zero (Berkhof & Snijders, 2001). Moreover, less than 100 clusters leads to underestimation of the standard error of the estimated variances (Maas & Hox, 2005).

Analyzing the data with an ordinary multilevel regression model (i.e. model (5)), results again in unbiased estimates of the fixed parameters. Since the variance is modelled incorrectly, it is not surprising that the estimated random parameters are all biased. As Table 1 shows, in most conditions the standard errors of the random parameters are underestimated. Especially the standard errors of the second level variance are severely underestimated. The underestimation is not only due to the reasons named above, but also because the estimated parameters themselves are biased. In some situations the standard errors of the fixed parameters are less well estimated as well. Since the intercept is a nuisance parameter, incorrect standard error of this

parameter is ignorable. Non ignorable is the overestimation of the treatment effect, which is seen in four of the six conditions. The overestimation is due to the fact that the subjects in the control condition are treated as being in one cluster, while they are in fact independent of each other. Since overestimation of the standard errors of the treatment effect results in a deflation of type I error, one risks a higher probability of not detecting an existing treatment effect when an ordinary multilevel model is used for the analysis of partially nested data.

When an analysis of variance model (model (6)) is used for analyzing partially nested data, again all fixed parameters are estimated without bias. The random parameter estimate is biased, like expected, since two of the three variance components are ignored in this analysis model. Hence, the standard error estimates of this only estimated random parameter, are expected to be underestimated. As they are. The lowest coverage, 49% in the condition  $ICC = .20$ , is even below the lowest coverage of the former analysis. The standard errors of the intercept and the first level covariate  $X$  are well estimated. However, in all but one condition ( $ICC = .05$ ), the standard errors of the treatment effect are underestimated, resulting in an inflation of type I error. This is like expected as well, since the data is analyzed assuming independency of the observations, which assumption is violated in the experimental condition. Analyzing partially nested data with an ANOVA model may lead to declaring a non existing treatment effect to be significant.

Thus, analyzing partially nested data with an approximate model may lead to either a deflation or an inflation of type I error, and hence to erroneous conclusions. So whenever data is partially nested, we advise to model this correctly.

## References

- Berkhof, J. & Snijders, T.A.B. (2005). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, 26, 133-152.
- Goldstein, H. (2003). *Multilevel Statistical Models.*, London: Edward Arnold.
- Maas, C.J.M. & Hox, J.J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology*, 1, 86-92.
- Rasbash, J., Steele, F., Browne, W. & Prosser, B. (2004) *A user's guide to MLwiN version 2.1.* London: Multilevel Models Project, University of London.
- Roberts, C., & Roberts, S.A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2, 152-162.

# Variable Selection in Joint Modelling of Mean and Covariance Structures for Longitudinal Data

Chaofeng Kou<sup>1</sup>, Jianxin Pan<sup>1</sup>

<sup>1</sup> School of Mathematics, The University of Manchester, Manchester, M13 9PL, U.K. Email: ckou@maths.man.ac.uk; jpan@maths.man.ac.uk.

**Abstract:** Joint modelling of mean-covariance structures is important in longitudinal studies. Correct modelling of covariance structures improves the efficacy of statistical inferences. Like the mean structure, covariances may depend on various explanatory variables of interest. We thus propose a new approach, based on penalty methods including LASSO, HARD thresholding and SCAD techniques, to select the most important explanatory variables that affect the modelling of mean and covariances structures for longitudinal data.

**Keywords:** Variable selection; Cholesky decomposition; Joint modeling; Penalized maximum likelihood.

## 1 Introduction

In longitudinal studies, the main objective may be to find out how the average value of the response varies over time and how this average response profile is affected by different treatments or various explanatory variables of interest. In many circumstances the within-subject covariance matrices are treated as nuisance parameters or assumed to have a simple structure. However, misspecification of covariance structures may lead to inefficient estimates of the mean parameters.

On the other hand, the within-subject covariance structure itself may be of interest, for example, in prediction problem arising in econometrics and/or finances. Furthermore, like the mean the covariances may be dependent on various explanatory variables. However, many of such variables may not be informative to statistical inferences and so should be removed from the mean and covariance models in order to increase modelling accuracy and avoid overfitting problem. This work aims to develop an efficient method to select important explanatory variables that really make significant contributions to the joint modeling of mean and covariance structures for high dimensional longitudinal/clustered data.

## 2 Variable Selection in Joint Mean-Covariance Models

### 2.1 Joint Mean-Covariance Models

Let  $y_{ij}$  be the  $j$ th of  $m_i$  measurements on the  $i$ th of  $n$  subjects. We assume that  $y_i = (y_{i1}, y_{i2}, \dots, y_{im_i})' \sim \mathcal{N}(\mu_i, \Sigma_i)$  where  $\mu_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{im_i})'$  and  $\Sigma_i$  are the  $(m_i \times 1)$  mean vector and  $(m_i \times m_i)$  covariance matrix of the responses  $y_i$ . According to the modified Cholesky decomposition of Pourahmadi (1999), the subject-specific covariance matrix  $\Sigma_i$  can be decomposed into  $T_i \Sigma_i T_i' = D_i$  where  $T_i$  is a lower triangular matrix with 1's as diagonal entries and  $-\phi_{ijk}$  as others, and  $D_i$  is a diagonal matrix with positive diagonal entries  $\sigma_{ij}^2$ . The reparameterized parameters  $\phi_{ijk}$  and  $\sigma_{ij}^2$  have clear statistical interpretations, that is,  $\phi_{ijk}$  are the autoregressive coefficients of  $\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik})$ , the linear least squares predictor of  $y_{ij}$  based on its predecessors  $y_{i1}, \dots, y_{i,j-1}$ , and  $\sigma_{ij}^2$  are the innovation variances  $\sigma_{ij}^2 = \text{var}(y_{ij} - \hat{y}_{ij})$ .

Based on the modified Cholesky decomposition, the unconstrained parameters  $\mu_{ij}$ ,  $\phi_{ijk}$  and  $\log \sigma_{ij}^2$  may be modelled in terms of linear regression models:

$$\mu_{ij} = x'_{ij}\beta \quad \phi_{ijk} = z'_{ijk}\gamma \quad \log \sigma_{ij}^2 = h'_{ij}\lambda \quad (1)$$

where  $\beta$ ,  $\gamma$  and  $\lambda$  are  $p$ -,  $q$ - and  $d$ -dimensional parameter vectors associated with the explanatory variables  $x_{ij}$ ,  $z_{ijk}$  and  $h_{ij}$ , respectively. When modelling stationary growth curve data using polynomials in time, for example, these explanatory variables may be chosen as  $x_{ij} = (1, t_{ij}, t_{ij}^2, \dots, t_{ij}^{p-1})'$ ,  $z_{ijk} = (1, (t_{ij} - t_{ik}), (t_{ij} - t_{ik})^2, \dots, (t_{ij} - t_{ik})^{q-1})'$  and  $h_{ij} = (1, t_{ij}, t_{ij}^2, \dots, t_{ij}^{d-1})'$ . An advantage of the model (1) is that the resulting estimators of the covariance matrices  $\Sigma_i$  are guaranteed to be positive definite as long as the parameter estimators of  $\gamma$  and  $\lambda$  are obtained.

### 2.2 Penalized Maximum Likelihood

In the initial step we assume all explanatory variables, and their interactions as well if any, are included into the model (1) through  $x_{ij}$ ,  $z_{ijk}$  and  $h_{ij}$ . We aim to remove unnecessary explanatory variables from the joint mean-covariance models. Traditional methods such as AIC and BIC model selection criteria become extremely computationally intensive (Pan and MacKenzie, 2003).

Let  $\ell(\theta)$  denote the log-likelihood function. We consider the following penalized likelihood for the joint models:

$$Q(\theta) = \ell(\theta) - n \sum_{i=1}^p p_{\tau(1)}(|\beta_i|) - n \sum_{j=1}^q p_{\tau(2)}(|\gamma_j|) - n \sum_{k=1}^d p_{\tau(3)}(|\lambda_k|) \quad (3)$$

where  $p_{\tau^{(l)}}(\cdot)$  is a known penalty function with the tuning parameter  $\tau^{(l)}$  ( $l = 1, 2, 3$ ) and  $\theta = (\theta_1, \dots, \theta_s)' = (\beta_1, \dots, \beta_p; \gamma_1, \dots, \gamma_q; \lambda_1, \dots, \lambda_d)'$  where  $s = p + q + d$ . The optimal value  $\hat{\theta}$  that maximizes  $Q(\theta)$  in (3) is called penalized maximum likelihood estimator (PMLE) of  $\theta$ . Commonly used penalty functions include LASSO (Tibshirani, 1996), HARD thresholding (Antoniadis, 1997), SCAD (Fan and Li, 2001), etc. The PMLE  $\hat{\theta}$  can be calculated using Newton-Raphson algorithm. For the two scenarios: a) the number of explanatory variables  $s$  is fixed, and b) the number of explanatory variables  $s = s_n \rightarrow +\infty$  as  $n \rightarrow +\infty$ , we showed that the PMLE  $\hat{\theta}_n$  is consistent and asymptotically Normally distributed under certain mild assumptions. The resulting PMLE  $\hat{\theta}$  may contain many zero components, implying that the associated explanatory variables are not necessarily entered in the models.

The tuning parameters  $\tau = (\tau^{(1)}, \tau^{(2)}, \tau^{(3)})'$  control the amount of penalties, which are usually unknown. We then propose a method based on  $K$ -fold cross-validation to choose the most appropriate  $\tau$ . We may randomly split the data set  $\mathcal{D}$  into  $K$  subsets,  $\mathcal{D}^v$  ( $v = 1, 2, \dots, K$ ) say. We then use the data in  $\mathcal{D} - \mathcal{D}^v$  to estimate the parameters and use  $\mathcal{D}^v$  to validate the inferences. Specifically, the optimal value of  $\tau$  can be chosen by minimizing the  $K$ -fold cross-validated log-likelihood criterion

$$\text{CV}(\tau) = \frac{1}{K} \sum_{v=1}^K \left\{ \sum_{i \in I_v} \log(|\hat{\Sigma}_i^{-v}|) + \sum_{i \in I_v} (y_i - X_i \hat{\beta}^{-v})' (\hat{\Sigma}_i^{-v})^{-1} (y_i - X_i \hat{\beta}^{-v}) \right\} \quad (4)$$

where  $I_v$  is the index set of the data in  $\mathcal{D}^v$ , and  $\hat{\beta}^{-v}$  and  $\hat{\Sigma}_i^{-v}$  are the estimators of  $\beta$  and  $\Sigma_i$  using the training data in  $\mathcal{D} - \mathcal{D}^v$ . Usually, the number  $K$  may take  $K = 5$  or  $K = 10$ .

### 3 Real Data Analysis

CD4+ cell data (Diggle *et al.*, 1994) comprise CD4+ cell counts for 369 HIV-infected men. Altogether there are 2376 values of CD4+ cell counts, with repeated measurements being made for each individual at different points in time (about 8.5 years). The number of measurements for each individual,  $m_i$ , varies from 1 to 12 and their measuring time is unequally spaced. The longitudinal data are highly unbalanced. Figure 1 gives the scatter plot of the CD4+ cell counts against time with some randomly selected individuals profiles superimposed. The explanatory variables in the data include:

- $X_0 =$  intercept,
- $X_1 =$  time,
- $X_2 =$  age,
- $X_3 =$  smoking habit (number of packs of cigarettes smoked per day),
- $X_4 =$  recreational drug use (1, yes; 0, no),
- $X_5 =$  number of sexual partners, and
- $X_6 =$  score on center for epidemiological studies of depression scale.

We aim to select the explanatory variables that have significant impacts on the mean and covariance models for the CD4+ Cell counts. Polynomials in time with 6 degrees for the mean  $\mu_{ij}$ 's and cubic ones for the generalized autoregressive parameters (GARP)  $\phi_{ijk}$ 's and the log innovation variances (IV)  $\log \sigma_{ij}^2$ 's, together with  $X_1, \dots, X_6$ , are all included in the modelling for initial considerations (Ye and Pan, 2006). The tuning parameter  $\tau = (\tau^{(1)}, \tau^{(2)}, \tau^{(3)})'$  is chosen by minimizing (4) with the use of 5-fold cross-validation. Table 1 presents the estimators of parameters in the mean, GARP and log-IV's.

#### 4 A Simulation Study

In this section we conduct simulation study to access the small sample performance of proposed procedures. We simulated 100 subjects, each of which has 5 observations from the model

$$\mu_{ij} = x_{ij}^T \beta, \quad \phi_{ijk} = z_{ijk}^T \gamma, \quad \log \sigma_{ij}^2 = h_{ij}^T \lambda,$$

where

$$\begin{aligned} \beta &= (3, 0, 0, -2, 1, 0, 0, 0, 0, -4)^T, \\ \gamma &= (-4, 0, 0, 2, 0, 0, 0)^T, \\ \lambda &= (0, 1, 0, 0, 0, -2, 0)^T. \end{aligned}$$

$x_{ij} = (x_{ijt})_{t=1}^{10}$  follows multivariate normal distribution with mean 0 and variance AR(1) with  $\rho = 0.5$ . Then we choose  $z_{ijk} = (x_{ijt} - x_{ikt})_{t=1}^7$ ,  $h_{ij} = (x_{ijt})_{t=1}^7$ . There are 1000 repetitions of sample are generated. The unknown parameters  $\tau^{(l)}$ ,  $l = 1, 2, 3$  of penalty functions are chosen by 5-fold cross-validation.  $\tau^{(1)} = 0.3072, 0.0264, 0.6812$ ,  $\tau^{(2)} = 0.1853, 0.0019, 0.5382$  and  $\tau^{(3)} = 0.6325, 0.1035, 0.9011$  for SCAD, LASSO and Hard-thresholding respectively. The average of 0 coefficients over 1000 simulated datasets is reported in Table 2. Notes that 'True' in Table 2 means the average number is restricted to true 0 coefficients and 'Wrong' depicts the average number of coefficients erroneously set to 0. The estimated coefficients and standard errors are also summarized in Table 3. As we can see from the simulation results, the SCAD outperforms the LASSO and the hard thresholding and it significantly reduces both model error and model complexity.

TABLE 1. Simulation results for  $\beta$ ,  $\gamma$  and  $\lambda$ .

coefficient	MLE	SCAD	LASSO	Hard-thresholding
$\beta_1(X_0)$	776.60(20.96)	776.68(20.31)	775.35(20.96)	776.60(20.96)
$\beta_2(X_1)$	-209.05(14.24)	-209.10(9.40)	-209.04(14.25)	-209.05(14.24)
$\beta_3(X_1^2)$	-14.47(8.36)	-14.49(8.04)	-14.51(8.37)	-14.47(8.36)
$\beta_4(X_1^3)$	32.68(5.93)	32.74(2.17)	32.72(5.93)	32.68(5.93)
$\beta_5(X_1^4)$	-1.97(1.05)	-1.97(1.02)	-1.96(1.05)	-1.97(1.05)
$\beta_6(X_1^5)$	-1.84(0.57)	-1.84(0.21)	-1.85(0.55)	-1.84(0.57)
$\beta_7(X_1^6)$	0.25(0.08)	0.26(0.02)	0.26(0.08)	0.25(0.08)
$\beta_8(X_2)$	0.88(1.34)	0.88(0.007)	0.88(1.35)	0.88(1.34)
$\beta_9(X_3)$	61.27(5.36)	61.32(6.35)	61.04(6.30)	61.27(6.36)
$\beta_{10}(X_4)$	45.70(18.84)	45.71(18.71)	45.61(18.84)	45.70(18.84)
$\beta_{11}(X_5)$	-3.61(2.09)	-3.60(2.09)	-3.64(2.09)	-3.61(2.09)
$\beta_{12}(X_6)$	-2.24(0.80)	-2.30(0.82)	0(-)	-2.24(0.80)
$\gamma_1(X_0)$	0.29(0.06)	0.29(0.02)	0.29(0.06)	0.29(0.06)
$\gamma_2(X_1)$	-0.33(0.09)	-0.33(0.02)	-0.33(0.09)	-0.33(0.09)
$\gamma_3(X_1^2)$	0.20(0.04)	0.20(0.01)	0.20(0.04)	0.20(0.04)
$\gamma_4(X_1^3)$	-0.03(0.004)	-0.03(0.002)	-0.03(0.003)	-0.03(0.004)
$\gamma_5(X_2)$	-0.001(0.0008)	0(-)	0(-)	0(-)
$\gamma_6(X_{31})$	-0.01(0.008)	-0.01(0.005)	-0.01(0.007)	-0.01(0.007)
$\gamma_7(X_{32})$	0.007(0.008)	0(-)	0(-)	0(-)
$\gamma_8(X_{41})$	-0.01(0.02)	0.01(0.06)	0.01(0.01)	0.01(0.02)
$\gamma_9(X_{42})$	0.02(0.02)	0.02(0.07)	0.02(0.01)	0.02(0.02)
$\gamma_{10}(X_{51})$	0.001(0.002)	0(-)	0(-)	0(-)
$\gamma_{11}(X_{52})$	-0.005(0.003)	0(-)	0(-)	0(-)
$\gamma_{12}(X_{61})$	0.004(0.0009)	0(-)	0(-)	0(-)
$\gamma_{13}(X_{62})$	0.006(0.001)	0(-)	0(-)	0(-)
$\lambda_1(X_0)$	11.64(0.07)	11.63(0.04)	11.63(0.08)	11.64(0.07)
$\lambda_2(X_1)$	-0.22(0.03)	-0.22(0.01)	-0.22(0.03)	-0.22(0.03)
$\lambda_3(X_1^2)$	-0.03(0.01)	-0.03(0.04)	-0.03(0.01)	-0.03(0.01)
$\lambda_4(X_1^3)$	-0.02(0.003)	-0.02(0.001)	-0.02(0.004)	-0.02(0.003)
$\lambda_5(X_2)$	-0.005(0.004)	0(-)	0(-)	0(-)
$\lambda_6(X_3)$	0.21(0.02)	0.21(0.01)	0.21(0.02)	0.21(0.02)
$\lambda_7(X_4)$	-0.12(0.07)	-0.12(0.005)	-0.12(0.06)	-0.12(0.07)
$\lambda_8(X_5)$	-0.02(0.008)	-0.02(0.004)	-0.02(0.008)	-0.02(0.009)
$\lambda_9(X_6)$	-0.006(0.003)	0(-)	0(-)	0(-)

## 5 Conclusions

It is concluded that the proposed penalty-based method produces good estimation results and is able to identify zero regression coefficients in joint models of mean-covariance structures. Furthermore, the SCAD approach

TABLE 2. Simulation results: Average number of 0 coefficients

Parameter	SCAD		LASSO		Hard-thresholding	
	True	Wrong	True	Wrong	True	Wrong
$\beta$	5.42	0.00	4.76	0.00	4.92	0.00
$\gamma$	4.18	0.06	3.28	0.08	3.55	0.21
$\lambda$	4.53	0.00	3.70	0.00	4.06	0.00

TABLE 3. Simulation results for non-zero coefficients

coefficient	True value	SCAD	LASSO	Hard-thresholding
$\beta_1$	3	3.08(0.95)	3.08(0.95)	3.09(0.93)
$\beta_4$	-2	-1.94(0.68)	-1.93(0.63)	-1.95(0.65)
$\beta_5$	1	0.95(0.32)	0.96(0.39)	0.97(0.39)
$\beta_{10}$	-4	-4.12(1.65)	-4.13(1.74)	-4.14(1.75)
$\gamma_1$	-4	-4.13(1.88)	-4.07(2.14)	-4.10(2.14)
$\gamma_4$	2	1.77(0.79)	1.71(0.85)	1.75(0.85)
$\lambda_2$	1	1.05(0.05)	1.03(0.06)	1.03(0.06)
$\lambda_6$	-2	-2.20(0.83)	-2.11(0.81)	-2.11(0.82)

outperforms the LASSO and the HARD thresholding methods in the joint model framework.

## References

- Antoniadis, A. (1997). Wavelets in Statistics: A Review (with discussion). *Journal of the Italian Statistical Association*, **6**, 97-144.
- Fan, J., Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of American Statistical Association*, **96**, 1348-60.
- Pan, J.X., Mackenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239-44.
- Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677-90.
- Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of Royal Statistical Society B*, **58**, 267-88.
- Ye, H.J., Pan, J.X. (2006). Modelling of Covariance Structures in Generalized Estimating Equations for Longitudinal Data. *Biometrika*, **93**, 927-941.

# *P*-spline ANOVA-type interaction models for spatio-temporal smoothing

Dae-Jin Lee<sup>1</sup> and María Durbán<sup>1</sup>

<sup>1</sup> Department of Statistics, Universidad Carlos III de Madrid, SPAIN.  
e-mail: [dae-jin.lee@uc3m.es](mailto:dae-jin.lee@uc3m.es) and [mdurban@est-econ.uc3m.es](mailto:mdurban@est-econ.uc3m.es)

**Abstract:** In recent years, spatial and spatio-temporal modelling have become an important area of research in many fields (epidemiology, environmental studies, disease mapping, ...). However, most of the models developed are constrained by the large amounts of data available. We propose the use of Penalized splines (*P*-splines) in a mixed model framework for smoothing spatio-temporal data. Our approach allows the consideration of interaction terms which can be decomposed as a sum of smooth functions similarly as an ANOVA decomposition. The properties of the basis used for regression allow the use of algorithms that can handle large amount of data. We show that imposing the same constraints as in a factorial design it is possible to avoid identifiability problems. We illustrate the methodology for Europe ozone levels in the period 1999-2005.

**Keywords:** *P*-splines; Mixed Models; Spatio-temporal data; space-time interactions; Smooth-ANOVA models.

## 1 Spatio-temporal smoothing with *P*-splines

Suppose we have normal spatio-temporal data which are located in geographical locations,  $s = (\mathbf{x}_1, \mathbf{x}_2)$ , and measured over time periods,  $\mathbf{x}_t$ . The response  $\mathbf{y}_{ijt}$  is then indexed in their spatial locations and over time. A smooth model for the data would be given by:

$$\mathbf{y} = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (1)$$

where  $\boldsymbol{\theta}$  is the vector of coefficients, and  $\mathbf{B}$  is a regression basis constructed from *B*-spline basis products. Currie et al. (2006) developed an approach based on Kronecker products, known as generalized linear array methods (GLAM) for data in a grid. When data are scattered (as is the case of spatial data), Eilers et al. (2006) proposed the use of the “row-wise” Kronecker or box-product of individual basis (denoted as  $\square$ ).

Most of the common approaches in spatio-temporal smoothing assume an additive function for the temporal dimension, ignoring the interaction between space and time (MacNab, 2001; Kneib, 2006). This formulation implies a spatio-temporal covariance structure given by separable terms for a

spatial and temporal components respectively. This could be too simplistic in some situations. As an alternative, we propose non-separable models of the form:

$$\hat{\mathbf{y}} = f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_t), \quad (2)$$

which explicitly considers the interaction between space and time. The regression basis for a  $3d$  interaction model (2) is:

$$\mathbf{B} = (\mathbf{B}_1 \square \mathbf{B}_2)_s \otimes \mathbf{B}_t = \mathbf{B}_s \otimes \mathbf{B}_t, \quad nt \times c_1 c_2 c_3, \quad (3)$$

where  $\mathbf{B}_1$ ,  $\mathbf{B}_2$  and  $\mathbf{B}_t$  are the marginal *B*-spline basis of dimensions  $n \times c_1$ ,  $n \times c_2$  and  $t \times c_3$  respectively.

Model (2) and basis given by (3) can easily be set into GLAM framework. We can express the data in a compact notation replacing  $\mathbf{y}$  of length  $nt \times 1$  by the matrix  $\mathbf{Y}$  of dimension  $n \times t$  and the coefficient vector  $\boldsymbol{\theta}$  of length  $c_1 c_2 c_3 \times 1$  by an array of coefficients  $\boldsymbol{\Theta}$ , of dimension  $c_1 \times c_2 \times c_3$ .

In matrix notation, the model can be written as

$$\mathbb{E}[\mathbf{Y}] = \mathbf{B}_t \boldsymbol{\Theta} \mathbf{B}'_s \quad (4)$$

Smoothness is imposed via the penalty matrix  $\mathbf{P}$  based on second order difference matrices  $\mathbf{D}_1$ ,  $\mathbf{D}_2$  and  $\mathbf{D}_t$ . The penalty term in 3-dimensions is:

$$\mathbf{P} = \lambda_1 \mathbf{D}'_1 \mathbf{D}_1 \otimes \mathbf{I}_{c_2} \otimes \mathbf{I}_{c_3} + \lambda_2 \mathbf{I}_{c_1} \otimes \mathbf{D}'_2 \mathbf{D}_2 \otimes \mathbf{I}_{c_3} + \lambda_t \mathbf{I}_{c_1} \otimes \mathbf{I}_{c_2} \otimes \mathbf{D}'_t \mathbf{D}_t, \quad (5)$$

which implies placing penalties over each dimension of the array  $\boldsymbol{\Theta}$ . For the spatio-temporal case, the penalty (5) allows spatial *anisotropy* considering a different amount of smoothing for longitude and latitude ( $\lambda_1 \neq \lambda_2$ ) and for the temporal component ( $\lambda_t$ ).

The mixed model representation of *P*-splines consists in setting a new basis which allows the reparameterization of (1) and its associated penalty into a mixed model of the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{G}), \quad \text{and } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (6)$$

where  $\mathbf{G}$  is a diagonal matrix which depends on the smoothing parameters  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_t$ . Following a similar approach to Currie et al. (2006), and using the properties of the Kronecker and ‘‘row-wise’’ Kronecker products it can be shown that using the singular value decomposition (SVD) of (5) the penalty becomes block-diagonal and basis and coefficients are reparameterized into:  $\mathbf{B} \equiv [\mathbf{X} : \mathbf{Z}]$  and  $\boldsymbol{\theta}' \equiv (\boldsymbol{\beta}' : \boldsymbol{\alpha}')$ .

## 2 Smooth-ANOVA decomposition models

Sometimes the interest lies in fitting complex models with functional form given by

$$\hat{\mathbf{y}} = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_3(\mathbf{x}_3) + f_{1,2}(\mathbf{x}_1, \mathbf{x}_2) + f_{1,2,3}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3), \quad (7)$$

where  $f_1$ ,  $f_2$  and  $f_3$  are smooth functions for the main effects ( $\mathbf{x}_1$ ,  $\mathbf{x}_2$  and  $\mathbf{x}_3$ ),  $f_{1,2}$  the 2d-interaction effects for ( $\mathbf{x}_1$ ,  $\mathbf{x}_2$ ) and  $f_{1,2,3}$  the 3d-interaction effects. Chen (1993) proposed Smoothing Spline Analysis-of-Variance (SS-ANOVA) decompositions to model main effects and interactions which can be interpreted as in classical ANOVA. In contrast, the approach presented in this paper allow a more computationally efficient methodology based on low-rank Penalized splines. Wood (2006) also considers smooth-ANOVA decompositions with  $P$ -splines, and notes the need of imposing constraints to maintain the model identifiability. However, the way how these constraints are imposed and how the basis for each component of the decomposition are constructed are not clear. In this paper we use the properties of the SVD of the penalty (5) and show how to fit each component of the model and establish an intuitive connection with the usual ANOVA.

In the case of spatio-temporal data this interpretation may be very useful, since we can model not only main effects of latitude and longitude, (or other covariates effects) but also the spatial effects (2-way interactions) and specially the interaction between space and time (3-way interactions).

The basis  $\mathbf{X}$  and  $\mathbf{Z}$  of the mixed model representation can be expanded to allow the representation of the 3d model as the sum of smooth main and interaction terms as in (7). However, this representation does not account for independent and separate penalties since we have 3 smoothing parameters  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_t$  for each of the dimensions of the model, with penalty matrix given in (5), but we do not allow separate parameters for interaction terms. Alternatively, ANOVA-type models which explicitly consider different amount of smoothing for each smooth function in (7) can be considered. The corresponding new  $B$ -splines regression matrix would not be of full rank, given the linearly dependent columns, and the model would not be identifiable.

The identifiability problem can be avoided by removing the columns of the basis of  $f(\mathbf{x}_1)$  which are repeated in those for  $f(\mathbf{x}_1, \mathbf{x}_2)$  and  $f(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$  and so on. Therefore, we need to impose constraints so that the model (7) is identifiable. We demonstrate that these constraints are applied on the  $P$ -spline regression coefficients  $\boldsymbol{\theta}$ , and are exactly equivalent to those applied in a 3-way factorial design, i.e.

$$\text{Main Effects: } \sum_i \boldsymbol{\theta}_i^{(1)} = \sum_j \boldsymbol{\theta}_j^{(2)} = \sum_t \boldsymbol{\theta}_t^{(3)} = 0 \quad (8)$$

$$\text{2-Way Interactions: } \sum_{i,j} \boldsymbol{\theta}_{ij}^{(12)} = \sum_{i,t} \boldsymbol{\theta}_{ik}^{(23)} = \sum_{j,t} \boldsymbol{\theta}_{jt}^{(13)} = 0 \quad (9)$$

$$\text{3-Way Interactions: } \sum_{i,j,t} \boldsymbol{\theta}_{ijt}^{(123)} = 0 \quad (10)$$

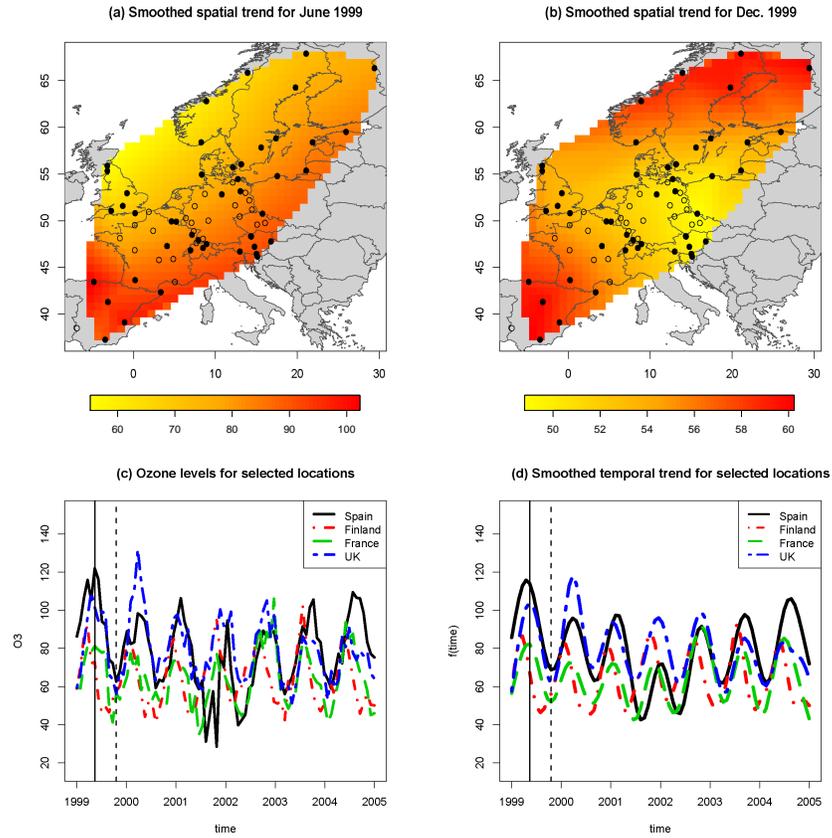


FIGURE 1.  $3d$  *P*-spline model: (a) spatial trend for June 1999, (b) spatial trend for december 1999. The symbol  $\bullet$  denotes the stations where monthly average measurements are available for period 1999-2005, and  $\circ$  the stations with missing data. (c) Time series plot of a sample of stations of four countries which reflects the seasonality and temporal patterns in the data. (d) Smoothed temporal trends for the four stations selected. The vertical solid line corresponds to June 1999 and the dashed line to December 1999.

### 3 Application to ozone levels in Europe

We apply the methodology proposed to the analysis of air pollution by ozone levels (in  $ug/m^3$  units) over Europe from 1999 to 2005. The data set are collected by the EMEP monitoring network which includes 126 stations in 28 countries. The ozone data are reported hourly in each monitoring station. We consider monthly averages in a regular temporal pat-

tern, but due to limited number of sites available, we selected a sample of 70 monitoring stations covering 15 countries. Data can be obtained at [www.emep.int](http://www.emep.int) and further information and annual reports about air pollution trends are available in the European Environmental Agency (EEA) web site ([www.eea.europa.eu](http://www.eea.europa.eu)).

We fitted a  $2d$   $P$ -spline model for the spatial component with an additive smooth function for time which does not consider space-time interaction, i.e.  $f(\mathbf{x}_1, \mathbf{x}_2) + f(\mathbf{x}_t)$ , and  $3d$   $P$ -spline interaction model (2). In addition,  $P$ -spline ANOVA models were fitted considering the appropriate constraints proposed in the previous section depending on the interaction terms included in the model. The model selection criteria was the Akaike Information Criteria (AIC). In general, better AIC results were obtained for interaction models.

Figure 1 shows the results for the  $3d$  space-time interaction model: (a) and (b) are the fitted spatial trends for two periods (June and December of 1999). It can be noticed the different spatial trend pattern and also the different overall level in each period, reflecting a seasonal variation which is very common in environmental data. Figure 1(c) shows this cyclic pattern in the data for selected monitoring stations in Spain, Finland, France and the UK. As reported by the EEA for ozone levels, summer periods show the highest values in contrast to winter months. Finally, Figure 1(d) shows the smooth function for time covariate, i.e.  $f(\mathbf{x}_t)$ , for the four selected stations.

## Concluding remarks

We presented a computationally efficient methodology for multidimensional smoothing. The ANOVA-Type models present an attractive alternative due to their interpretability in terms of decompositions of smooth functions and basis which are identifiable. From our  $P$ -spline approach, the mixed model representation and the decomposition of the basis used, allow more flexibility in contrast to existing SS-ANOVA models. The analysis of the ozone level data showed that a model where the time dimension is additive could ignore important features in the data.

## References

- Chen, Z. (1993). Fitting Multivariate Regression Functions by Interactions Spline Models. *J. R. Statist. Soc. B*, **55**, 473-491.
- Currie, I. D., Durbán, M. and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. 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 & Data Analysis*, **50**(1), 61–76.
- Gu, C. (2002). Smoothing Spline ANOVA Models. *Springer*, New York.
- Kneib, T. and Fahrmeir, L. (2006). Structured Additive Regression for Categorical Space-Time Data: A Mixed Model Approach. *Biometrics*, **62**, 109–118.
- MacNab, Y. C. and Dean, C.B. (2001). Autoregressive Spatial Smoothing and Temporal Spline Smoothing for Mapping Rates. *Biometrics*, **57**, 949–956.
- Wood, S. N. (2006). Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, **62**, 1025–1036.

# ROC.Regression: A new R software for ROC Regression Analysis

I. López-de-Ullibarri<sup>1</sup>, M. X. Rodríguez-Álvarez<sup>2</sup> and C. Cadarso-Suárez<sup>2</sup>

<sup>1</sup> Dept. of Mathematics, University of A Coruña, 15405 Ferrol, Spain, [ilu@udc.es](mailto:ilu@udc.es)

<sup>2</sup> Dept. of Statistics and OR, University of Santiago de Compostela, Spain

**Abstract:** The aim of this work is to present an user-friendly R (R Development Core Team, 2007) software, called ROC.Regression, developed to analyze the possible effect of covariates over the Receiver Operating Characteristic (ROC) curve. In this software, different approaches to regression analysis of ROC curves have been implemented, allowing the user to choose between the 'Normal Method' (Faraggi, 2003), the 'Parametric ROC-GLM Method' (Alonzo and Pepe, 2002) and the 'Semiparametric ROC-GLM Method' (Cai, 2004). This software can be used to fit ROC regression models including a set of continuous and/or categorical covariates, and their possible interactions. To illustrate usage of the program we analyse data from a computer-aided diagnostic (CAD) system dedicated to early detection of clustered microcalcifications, a primary sign of breast cancer.

**Keywords:** Receiver operating characteristic curve; Regression Model; Accuracy measures; Thresholds values; R software

## 1 Introduction

Continuous biomarkers are often used to discriminate between diseased and healthy populations. The 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: the **induced methodology** (Pepe, 1998; Faraggi, 2003) and the **direct methodology** (Alonzo and Pepe, 2002; Cai, 2004). In spite of the advantages of using these methodologies in practice, sufficient standard software is not quite available at the present time. The scarcity of implemented ROC regression software is probably responsible for the lack of popularity of these models in the medical community. To overcome some of these difficulties, we have developed a user-friendly R program, called ROC.Regression, providing numerical and graphical outputs of different methods for ROC regression. Our software allows the user to fit all regression approaches with just one input command. With this software, users can easily obtain numerical and graphical output for all models, and make decisions accordingly.

## 2 Implemented methods

The **induced ROC methodology** 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  $\varepsilon$  is a random variable with zero mean, unit variance and survival function  $S$ . From expression (1) the covariate-specific ROC curve can be obtained as follows:

$$ROC_X(t) = S\left(\frac{\mu(0, X) - \mu(1, X)}{\sigma(1)} + \frac{\sigma(0)}{\sigma(1)}S^{-1}(t)\right).$$

In ROC.Regression we have implemented 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**'.

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,  $\eta$  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**').

## 3 Numerical and Graphical Outputs

From the estimated ROC curve obtained by the several methods described in Section 2, summary measures of the accuracy, such as the area under the curve (AUC) and the generalized Youden index (YI) can be obtained, and also thresholds values based on the YI criterion. Numerical output provides coefficient estimates, standard errors and p-values. Direct methodology's inference is based on bootstrap procedures. Graphical output offers ROC curves, and AUC, YI and threshold values when requested. Also, the current version allows the user to incorporate flexibility on the continuous covariate effects by using regression splines.

## 4 ROC.Regression usage example

We have applied the ROC.Regression software to the output of a CAD system dedicated to the automated detection of clustered microcalcifications on digital mammograms. The database (called 'radio') consists of

numerical information of mammograms selected from a screening program that is currently underway, from 1992, at the Galicia Community (Spain), among women aged 50-64 years. The mammograms were classified as fatty and dense, according to their breast tissue. A total of 71 true clusters and 740 false detections were yielded by the CAD system. The marker (Y) considered is the ratio of the cluster size to mean distance between the microcalcifications of each cluster detected on digital mammograms. The covariate vector consists of a continuous covariate, X1, the ratio of the cluster average grey level to that of the image, and a categorical one, X2, the tissue type. As an illustration, we present the results obtained by using the Parametric ROC-GLM Method (Alonzo and Pepe, 2002), including the interaction between X1 and X2. We also included in the model the interaction of X1 and X2 with False Positive Rates (FPR).

The fitted model can be obtained by using ROC.Regression with input command

```
>fit.ROC <- ROC.Regression(method = "PROCGLM", model = c("~1", "~x1*x2"),
marker="x", group = "group", tag.healthy = 0, data = radio, nboot=500,
control = controlROCreg(ROC.model="binormal", card.T=50, FPFint = TRUE))
```

The numerical results presented in Figure 1 are provided using the input command:

```
> summary(fit.ROC)
```

```
Call:
ROC.Regression(method = "PROCGLM", model = c("~1", "~x1*x2"),
tag.healthy = 0, marker = "x", data = radio,
control = controlROCreg(ROC.model = "binormal", card.T = 50,
FPFint = TRUE))

*****
Parametric ROC-GLM Method
*****

ROC Coefficients:
-----
              Estimate      Std. Error  95% Conf. In:
(Intercept)    1.3517         0.4867   ( 0.3977, 2.
x1             -1.1287         4.4174   (-9.7868, 7.
h              0.7483         0.3170   ( 0.1271, 1.
x2Fat         -1.1114         1.2318   (-3.5258, 1.
x1:h           0.9275         3.1460   (-5.2387, 7.
h:x2Fat        1.2932         0.6537   ( 0.0119, 2.
x1:x2Fat       13.2094        19.5199  (-25.0497, 51.
x1:h:x2Fat    -16.9465         9.5084  (-35.5831, 1.
```

FIGURE 1. Numerical Results for the illustration data.

Finally, the graphical output in Figures 2 and 3 can be obtained through the following input command:

```
> plot(fit.ROC, AUC=TRUE, theta=25)
```

**Acknowledgments:** The authors would like to thank the support of Spanish MEC Grants MTM2005-00818 and MTM2005-00429.

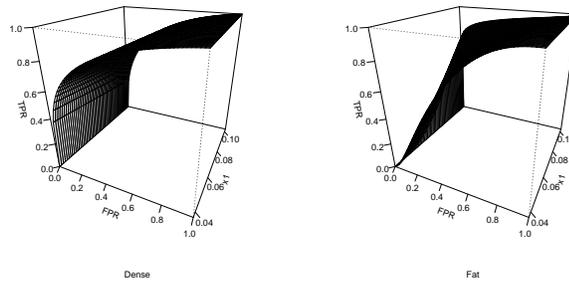


FIGURE 2. Graphical output for the ROC Curve and the AUC. z-axis: Continuous covariate.

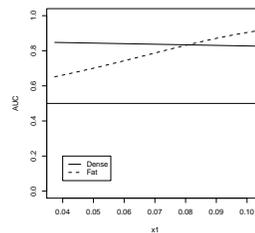


FIGURE 3. Graphical output for the AUC. x-axis: Continuous covariate.

## 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.
- R Development Core Team (2007). *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>.

# Piece-wise exponential PH models

Joseph Lynch<sup>1</sup> and Gilbert MacKenzie<sup>1</sup>

<sup>1</sup> Centre of Biostatistics, University of Limerick, Ireland.  
Emails: joseph.lynch@ul.ie; gilbert.mackenzie@ul.ie

**Abstract:** In the Cox proportional hazard (PH) regression model, the baseline hazard function,  $\lambda_0(t)$ , is unspecified and therefore of potentially high-dimension. Here we compare an alternative semi-parametric estimator of  $\Lambda_0(t)$  based on a piece-wise exponential approach, which we anticipate will prove useful in developing  $h$ -likelihood versions of a focussed AIC for selecting frailty structures in PH frailty models. Working with a large dataset of women diagnosed with breast cancer in the West Midlands, U.K., we compare and contrast the performance of piece-wise and Breslow estimators for  $\hat{\Lambda}_0(t)$

**Keywords:** Piece-wise exponential; PH model; CIs, Model Selection.

## 1 Introduction

In recent work on model selection in Proportional Hazard frailty models, Ha, Lee & MacKenzie, (2007) dealt with the potentially large dimensional baseline hazard function  $\lambda_0(t)$  by an Extended Restricted Likelihood (ERL) method of nuisance parameter elimination within the framework of  $h$ -likelihood. Whilst the focussed AIC derived by this method performed better than a less-focussed AIC, a simulation study suggested that there was clearly room for improvement.

In PH frailty modelling inference a complicating factor is that in some cases the dimension of  $\lambda_0(t)$  increases with  $n$ , the sample size, being exactly  $n$ -dimensional when all the subjects fail, thus violating the usual assumptions for regular estimation. Accordingly, if one is to remove this difficulty we need to replace the non-parametric estimator of  $\Lambda_0(t)$  with one of fixed dimension, say  $k \ll n$ . For this reason, as a first step, we turn to investigate the usefulness of piece-wise exponential estimators in simple cases.

## 2 Cox Model

In the Cox proportional hazard (PH) regression model, the hazard function is given by

$$\lambda(t; x) = \lambda_0(t) \exp(x' \beta) \tag{1}$$

where  $T \geq 0$  and  $\beta$  is a  $p \times 1$  vector of regression parameters associated with fixed covariates  $x' = (x_1, \dots, x_p)$ . The unspecified form of the baseline hazard function,  $\lambda_0(t)$ , means that the data essentially determine the shape of the function. The censored log likelihood for the Cox model is given by

$$l(\beta, \lambda_0(t)) = \sum_{i=1}^n \left[ \delta_i \log_e(\lambda_0(t_i) \cdot \exp(x'_i \beta)) - \int_0^{t_i} \lambda_0(u) du \exp(x'_i \beta) \right] \quad (2)$$

### 3 Piece-wise Exponential Model

The Piece-wise Exponential (PWE) model avails of the same form of log likelihood as the Cox, but the form of the baseline hazard is structured differently. In this model, the time-frame of the study is divided into an arbitrary number of intervals,  $m = k + 1$ , the last interval being a reference interval about which no information is contributed by the data. Within the  $j$ th. interval, the hazard function is assumed to be  $\lambda_{0j} = \exp(\alpha_j)$ , in accordance with the piece-wise exponential model.

The division of the time-axis can be based on having, for example, intervals of approximately equal length or intervals of equal numbers. The contribution made by an individual to the cumulative hazard is found as  $\sum_j d(t)_j \times \lambda_{0j}$ , where  $d(t)_j$  is the time spent in the  $j$ th. interval. Thus, the likelihood becomes

$$l(\beta, \lambda_0(t)) = \sum_{i=1}^n [\delta_i \cdot \log_e(\exp(a'_i \alpha) \cdot \exp(x'_i \beta)) - w_i^* \epsilon \exp(x'_i \beta)] \quad (3)$$

where:  $a'_i$  is a row vector from a matrix of indicator variables,  $a$ , with  $a_{ij} = 1$  if  $t_i \in$  interval  $I_j$ ;  $w_i^*$  is the  $i$ th. row of the  $(n \times k)$  matrix,  $w^*$ , representing the subjects' times on study;  $\epsilon = (e^{\alpha_1}, \dots, e^{\alpha_k})^T$  and  $\int_0^{t_i} \lambda_0(u) du = w_i^* \epsilon$ , represents the baseline cumulative hazard for the  $i$ th. subject.

### 4 Estimation and Inference

The partial derivatives with respect to the parameters are:

$$\frac{\partial l(\lambda_0(t), \beta)}{\partial \lambda_{0r}} = \sum_{i=1}^n [\delta_i(a_{ir}) - \exp(x'_i \beta) w_{ir}^* e^{\alpha_r}]$$

for  $r = 1, \dots, k$ , yielding the closed form estimator

$$\hat{\lambda}_{0r} = \frac{\sum_{i=1}^n \delta_i(a_{ir})}{\sum_{i=1}^n \exp(x'_i \beta) w_{ir}^*} \quad (4)$$

where  $a_{ir}$  and  $w_{ir}^*$  are elements of the  $a$  and  $w^*$  matrices respectively.

The  $\beta$  derivatives are given by

$$\frac{\partial l(\lambda_0(t), \beta)}{\partial \beta_u} = \sum_{i=1}^n [\delta_i(x_{iu}) - x_{iu} \exp(x'_{iu}) w_i^* \lambda_0] \quad (5)$$

where  $x_{iu}$  is an element of the  $x$  matrix and  $w_i^* \lambda_0$  is a scalar product. The  $\hat{\beta}$ s are obtained by using the `nlm` function in R.

The mixed partial derivatives are given by:

$$\frac{\partial^2 l(\lambda_0(t), \beta)}{\partial \lambda_{0r} \partial \beta_u} = \sum_{i=1}^n [-x_{iu} \exp(x'_i \beta_u) w_{ir}^*] \quad (6)$$

If we let  $H$  be the Hessian matrix, then the observed information matrix is,  $I_0 = -H$ , viz:

$$I_0 = \begin{bmatrix} I_\alpha & I_{\alpha\beta} \\ I_{\alpha\beta} & I_\beta \end{bmatrix} \quad (7)$$

where  $I_\alpha$  is a diagonal matrix. The covariance matrix is  $\Sigma = I_0^{-1}$  and is given by

$$\Sigma = \begin{bmatrix} I_\alpha^{-1} + I_\alpha^{-1} I_{\alpha\beta} Z^{-1} I'_{\alpha\beta} I_\alpha^{-1} & -I_\alpha^{-1} I_{\alpha\beta} Z^{-1} \\ -Z^{-1} I'_{\alpha\beta} I_\alpha^{-1} & Z^{-1} \end{bmatrix} \quad (8)$$

where  $Z = I_\beta - I'_{\alpha\beta} I_\alpha^{-1} I_{\alpha\beta}$  and in the absence of  $\beta$ s,  $\Sigma = I_\alpha^{-1}$ .

## 5 Correlation between $\alpha$ & $\beta$

The basic dataset considered in this paper comprised survival information on c16,000 incident cases of carcinoma of the female breast (West Midlands Survival Study, UK, 2004).

We sampled this dataset to produce a variety of scenarios in terms of sample size ( $n$ ) and number of intervals ( $k$ ). In each scenario, we analysed two covariates, one categorical (screen) and the other continuous (age) in order to explore the correlation between  $\alpha$  and  $\beta$ .

As the  $\hat{\alpha}$ s and  $\hat{\beta}$ s are estimated simultaneously, we report on the range of correlation values between the two sets of parameters in selected scenarios (Table 1). We note that the correlation between the parameter estimates can be quite significant.

## 6 Estimation of $\Lambda_0(t)$

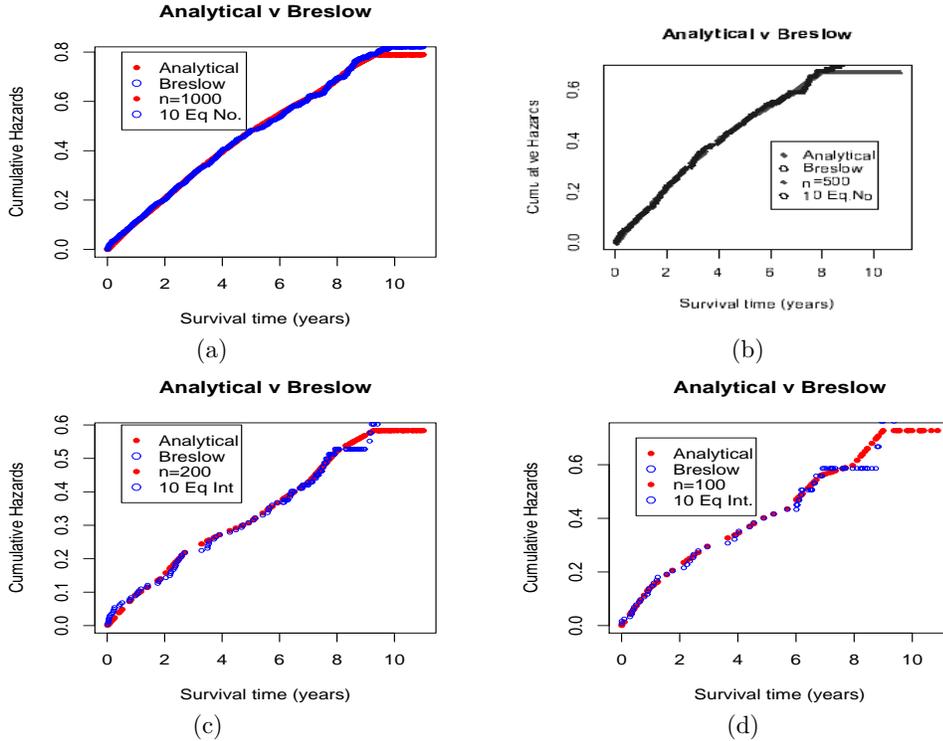
Next we compared the PWE estimates for the cumulative baseline hazard with the usual Breslow estimates in a different range of scenarios. It can be

TABLE 1. Range of Correlations between  $\hat{\alpha}$ s and  $\hat{\beta}$ s

$n = 1000, k = 20$		$n = 1000, k = 10$	
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
(-0.1 : -0.03)	(-0.21 : -0.009)	(-0.13 : -0.008)	(-0.26 : -0.025)
$n = 100, k = 20$		$n = 100, k = 10$	
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
(-0.1 : -0.04)	(-0.3 : 0.06)	(-0.13 : -0.05)	(0.047 : 0.38)

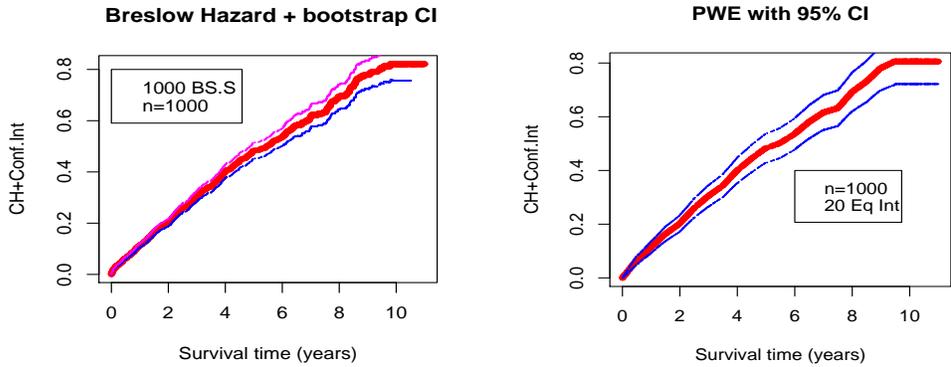
$\hat{\beta}_1$  and  $\hat{\beta}_2$  are categorical and continuous covariates respectively.

seen from Figures (a) – (d) that the estimated cumulative hazard functions obtained via both models are similar in the scenarios considered. The PWE model performed satisfactorily when the  $n \geq 100$ .



### 7 CI for Cumulative Baseline Hazard

We also compared the 95% confidence intervals for the cumulative baseline hazard using the estimates obtained via the PWE method and 1000 bootstrap samples when  $n = 1000$ . In the calculation of the standard errors, the classical delta method is applied once in the case of the PWE estimates.  $\hat{\Lambda}_0(t)$ .



From the diagram, the confidence intervals are tighter all the way for the boot-strap estimator than for the PWE estimator. The boot-strap estimates involves only an estimate of the  $\beta$  parameters, while the PWE method involves the joint estimation of the  $\alpha$ s and  $\beta$ s.

### 8 Wald Test

From Section 4, it is easy to formulate a Wald test for an underlying exponential distribution,  $\alpha_r = \alpha^*$ , when  $\beta = 0$ .

$$\begin{aligned}
 H_0 : \alpha_r &= \alpha^* \quad \forall r \in 1 \dots k \\
 H_1 : \alpha_r &\neq \alpha_s \quad \forall r, s \in 1 \dots k
 \end{aligned}
 \tag{9}$$

When  $H_0$  is true,

$$l(\lambda_0(t)) = \sum_{i=1}^n \left[ \delta_i \left( \alpha^* \sum_{j=1}^k a_{ij} \right) - e^{\alpha^*} \sum_{j=1}^k w_{ij}^* \right]
 \tag{10}$$

$$\frac{\partial l(\lambda_0(t), \beta)}{\partial \alpha^*} = \sum_{i=1}^n \left[ \delta_i \left( \sum_{j=1}^k a_{ij} \right) - e^{\alpha} \sum_{j=1}^k w_{ij}^* \right]
 \tag{11}$$

leading to

$$\hat{\alpha}^* = \log_e \left( \frac{\sum_{i=1}^n \delta_i (\sum_{j=1}^n a_{ij})}{\sum_{i=1}^n (\sum_{j=1}^k w_{ij}^*)} \right) \quad (12)$$

whence the Wald statistic is

$$\chi_{k-1}^2 = (\hat{\alpha} - \hat{\alpha}^*)' \Sigma(\alpha)_{|\alpha^*}^{-1} (\hat{\alpha} - \hat{\alpha}^*) \quad (13)$$

where  $(\hat{\alpha} - \hat{\alpha}^*)'$  is a  $(1 \times k)$  vector and  $\Sigma(\alpha)^{-1} = \text{diag}[\sigma_{11}, \dots, \sigma_{kk}]$ , and  $\sigma_{rr} = w_{ir} e^{\alpha^*}$ .

Using a sample size of  $n = 1000$  with 20 intervals of approximately equal sizes, the value of the  $\chi^2$  statistic is 72.56727, which is greater than the critical value of 28.869 on  $(k-1)$  degrees of freedom at the 5% significance level. Thus, the Null Hypothesis is rejected.

## 9 Discussion

The estimation of  $\beta$  is relatively invariant to the choice of parametrization for  $\lambda_0(t)$  in the PWE. Despite this, the magnitude of correlations between the  $\hat{\alpha}$ s and  $\hat{\beta}$ s are sometimes high. Later, we intend to extend both models by accommodating frailty and simplifying the focussed AIC based on the ERL in Ha et al (2007).

## References

- Cox DR (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187-220.
- Gillon N, MacKenzie G, Li Xuefang (2004) The North Staffordshire Breast Cancer Study. Centre for Medical Statistics: Technical Report Series No 2, 1-80.
- Sy JP, Taylor JMG. (2001) Standard Errors for the Cox Proportional Hazards Cure Model. *Mathematical and Computer Modelling*. **33**, 1237-1251.
- Ha ID, Lee Y, & MacKenzie G. (2007) Model Selection for Frailty Structures. *Statistics in Medicine*. **46**, 4790-4807.

# Robustness of the Regression Parameter in PH and Non-PH Frailty Survival Models

Gilbert MacKenzie<sup>1</sup> and Il Do Ha<sup>2</sup>

<sup>1</sup> Centre of Biostatistics, Department of Mathematics & Statistics, University of Limerick, Ireland email: gilbert.mackenzie@ul.ie

<sup>2</sup> Department of Asset Management, Daegu Haany University, Gyeongsan, 712-715, South Korea email: idha@dhu.ac.kr

**Abstract:** Correlated survival times can be modelled by introducing a random effect, or frailty component, into the hazard function. For multivariate survival data we extend a non-PH model, the generalized time-dependent logistic survival model, to include random effects. The hierarchical-likelihood procedure, which obviates the need for marginalization over the random effect distribution, is derived for this extended model and its properties discussed. The extended model leads to a robust estimation result for the regression parameters against the mis-specification of model assumptions about the basic hazard function or frailty distribution, which is comparable to Cox-PH frailty model. The proposed method is illustrated with two practical examples and simulation studies

**Keywords:** Frailty; GTDL; h-likelihood; Robustness; Survival Modelling.

## 1 Introduction

As an alternative to Cox's well-known PH model, MacKenzie (1996) introduced the generalized time-dependent logistic (GTDL) non-PH model for univariate survival data. This distribution is a wholly parametric competitor for the PH model which generalizes the relative risk (RR) in Cox's semiparametric PH model to time-dependent form. It is thus able to accommodate a wider class of univariate survival data including PH survival data. Accordingly, in this paper we introduce a flexible non-PH frailty model for multivariate survival data based on the GTDL model. In general, frailty models require the specification of two main terms, the basic hazard function, which is usually multiplied by the frailty term, and the assumed frailty distribution. Here, the basic hazard depends on time and fixed covariates. It has been shown that parametric inference on the regression parameters can be sensitive to the choice of the hazard function. In parametric PH models, this amounts to sensitivity to different choices of the baseline hazard function (Ha and Lee, 2003). Accordingly, here, we study the effect of the mis-specification of the basic hazard function and the frailty distribution on the regression parameter estimates in various scenarios. We also

compare the GTDL frailty model and Cox's PH frailty model, since both models are non-PH. The simulation studies show that the proposed GTDL frailty model is comparable to Cox's PH frailty model and inference on the regression parameter is robust against mis-specification of the basic hazard function and/or the frailty distribution. For inference we use hierarchical likelihood (h-likelihood, Lee and Nelder, 1996, 2001) which obviates the need for marginalization over the frailty distribution. The h-likelihood approach provides a unified inferential framework and a numerically efficient fitting algorithm for various random-effect models including frailty models (Ha et al., 2001; Lee, Nelder and Pawitan, 2006).

## 2 GTDL Frailty Models

First we define the multivariate data structures as follows. Let  $T_{ij}$  ( $i = 1, \dots, n$ ,  $j = 1, \dots, n_i$ ,  $N = \sum_i n_i$ ) be the survival time for  $j$ 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. Denote by  $U_i$  the random variable denoting the unobserved frailty (or random effect) for the  $i$ th subject. We extend the model (1) to the multivariate survival data setting by inclusion of a frailty term acting multiplicatively on the individual hazard rate of (1). The GTDL non-PH frailty model is then defined as follows. Given  $U_i = u_i$ , the conditional hazard function of  $T_{ij}$  takes the form

$$\lambda_{ij}(t|u_i; x_{ij}) = \lambda_{ij}(t; x_{ij})u_i, \quad (4)$$

where  $\lambda_{ij}(t; x_{ij})$  is a basic hazard function not depending on  $u_i$  and from (1) it is given by

$$\lambda_{ij}(t; x_{ij}) = \lambda_0[\exp(t_{ij}\alpha + x_{ij}^T\beta)/\{1 + \exp(t_{ij}\alpha + x_{ij}^T\beta)\}]$$

with  $x_{ij} = (1, x_{ij1}, \dots, x_{ijp})^T$ . We have found by numerical studies that the inclusion of frailty term into the model (1) is not necessary to allow the parameter  $\lambda_0$ . Thus, hereafter we take  $\lambda_0 = 1$  in (4). The frailties  $U_i$  are assumed to be independent and identically distributed random variables with a density function depending on the frailty parameter  $\theta$ .

Recall that if  $U_i$  is log-normal or gamma, then  $V_i = \log U_i$ , the log-frailty, becomes Normal or log-Gamma, respectively. The corresponding model is usually called log-normal or gamma frailty model, based on frailty  $U_i$ . For convenience, we also shall refer to it as normal or log-gamma frailty model, based on log-frailty  $V_i$ .

If a basic hazard function in (4) is of the form

$$\lambda_{ij}(t; x_{ij}) = \lambda_0(t) \exp(x_{ij}^T\beta), \quad (5)$$

TABLE 1. Results of fitting the two frailty models to the litter-matched rat data

Variable	GFM		CFM	
	Est.	SE	Est.	SE
Intercept	-9.271	0.645	—	—
Drug	0.927	0.329	0.903	0.322
Time ( $\alpha$ )	0.048	0.008	—	—
Frailty ( $\sigma$ )	0.679	—	0.636	—

GFM, GTDL non-PH lognormal frailty model;

CFM, Cox's PH lognormal frailty model;

Est., estimate; SE, standard error;

$\alpha$ , time effect of GTDL;  $\sigma$ , a square root of log-frailty variance in  $N(0, \sigma^2)$

where  $\lambda_0(t)$  is a baseline hazard function, we have a PH frailty model. Here, the term  $x_{ij}^T \beta$  in (5) does not include an intercept term for identifiability reasons. Note that  $\lambda_0(t)$  can be parametric (e.g. Weibull) or non-parametric. In particular, the latter gives a semi-parametric Cox-PH frailty model, an extension of Cox's PH model (1972): see for example McGilchrist and Aisbett (1991) and Ha et al. (2001).

We note that the marginal model,  $\lambda^M(t; x)$ , will not be PH unless  $U$  is positive stable. In particular, when  $U$  is log-normal or gamma frailty the non-PH model (4) and the PH model (5) are marginally non-PH. Accordingly, this affords an interesting opportunity to compare the behaviour of these two models.

### 3 Applications

Mantel et al. (1977) presented a data set on a tumorigenesis study of 50 litters of female rats. For each litter, one rat was selected to receive the drug and the other two rats were placebo-treated controls. The survival time is the time to development of tumor, measured in weeks. Death before occurrence of tumor yields a right-censored observation; forty rats developed a tumor, leading to 73% censoring. The survival times for rats in a given litter may be correlated due to a random effect representing shared genetic or environmental effects. The two frailty models were fitted with a single fixed covariate ( $x_{ij}$ ) having  $x_{ij}^T \beta = \beta_0 + \beta_1 x_{ij}$  ( $i = 1, \dots, 50$ ;  $j = 1, 2, 3$ ). Here  $x_{ij} = 1$  if the  $j$ th rat in the  $i$ th litter received the drug and 0 otherwise.

The results of fitting the two models (GFM & CFM) are summarized in Table 1. In GFM the estimated time coefficient  $\hat{\alpha} = 0.048$  (with SE=0.008) suggests that the effect of time is significantly different from zero, indicating an increasing time-trend in hazard. The intercept term is also well-defined. In addition, both models give similar estimates of the treatment effect and frailty parameter,  $(\beta_1, \sigma)$  where  $\sigma = \sqrt{\theta}$  is the square root of the log-frailty variance in  $N(0, \theta)$ . In particular, the Wald test of  $H_0 : \beta_1 = 0$ , yields similar  $\chi^2$ -values and ( $p$ -values) with 1 d.f., namely: 7.94 (0.0048) for GFM and 7.86 (0.0050) for CFM, respectively.

## 4 Discussion

In the main paper we show the simulation results in some detail comparing the GTDL, Weibull and Cox PH frailty models with various mis-specified frailty distributions including the log-Normal (N) and extreme value (EV) for a range of frailty variance values. The overall results confirm that the parametric GTDL frailty model performs just as well as Cox's PH frailty model in these scenarios and it has the advantage of modelling PH or non-PH data in the basic hazard.

**Acknowledgments:** We thank Professor Youngjo Lee for helpful discussions.

## References

- Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187-220.
- Ha, I. D., Lee, Y. and Song, J.-K. (2001) Hierarchical likelihood approach for frailty models. *Biometrika*, **88**, 233-243.
- Ha, I. D., Lee, Y. and Song, J.-K. (2002) Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis*, **8**, 163-176.
- Lee, Y. and Nelder, J. A. (1996) Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc. 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.
- Lee, Y. and Nelder, J. A. (2006) Double hierarchical generalised linear models (with discussion). *J. R. Statist. Soc. C*, **55**, 139-185.

# New Estimators for Mortality Ratios through Loglinear Modelling with Linear Constraints

Nirian Martín<sup>1</sup>

<sup>1</sup> Department of Statistics and O.R. III, Complutense University of Madrid, School of Statistics, Avda. Puerta de Hierro s/n, 28040 Madrid, Spain, Email: nirian@estad.ucm.es

**Abstract:** Since loglinear models were proved to be useful for analyzing categorical data, commonly considered sampling schemes have been the multinomial, product-multinomial and Poisson sampling plans. This work deals also with marginal models widely applied for comparing rates. Both models jointly form the so-called loglinear models with linear constraints (LMLC). In the same way as in Haberman (1974) was shown the equality of the maximum likelihood estimator for the three kinds of sampling plans, this equality is also holds for loglinear models with linear constraints. This result is proved in the framework of a family of minimum  $\phi$ -divergence estimators, power divergence estimators, which are presented for LMLC as extension of the definition given before in Martín and Pardo (2008a, 2008b). An application of LMLC for a useful model in biomedical sciences is shown: estimation of mortality ratios in categorical data.

**Keywords:** Loglinear models; Poisson sampling; Maximum likelihood estimators; Minimum power divergence estimators; Mortality ratios.

## 1 Loglinear Modelling with Linear Constraints

We consider a table of  $k$  cells which contains counts,  $\mathbf{n} \equiv (n_1, \dots, n_k)^T$  obtained from a Poisson sampling plan. Define  $\mathbf{m}(\boldsymbol{\theta}) = (m_1(\boldsymbol{\theta}), \dots, m_k(\boldsymbol{\theta}))^T \equiv E[\mathbf{n}]$  and denote  $N \equiv E[\sum_{i=1}^k n_i]$ , the expected total. The loglinear model is defined according to a coordinate-free notation (see Zelterman (1999), p. 138), by  $\log \mathbf{m}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta}$ , where  $\mathbf{X} = (\mathbf{J}_k, \mathbf{W})$  is a known  $k \times t$  full rank matrix such that  $t < k$ ,  $\mathbf{J}_k \equiv (1, \dots, 1)^T$ , and  $\boldsymbol{\theta} \in \mathbb{R}^t$  is an unknown vector of parameters.

If one conditions on the random variable of totals  $\sum_{i=1}^k n_i$ , the vector  $\mathbf{n} = (n_1, \dots, n_k)^T$  becomes multinomial, i.e.  $\mathbf{n}$  is multinomial with parameters  $(\sum_{i=1}^k n_i; \pi_1(\boldsymbol{\theta}), \dots, \pi_k(\boldsymbol{\theta}))$ , where  $\pi_i(\boldsymbol{\theta}) = m_i(\boldsymbol{\theta})/N$ ,  $i = 1, \dots, k$ . If the observations come from a product-multinomial sampling scheme, being  $c$  the number of independent multinomial groups and  $k_i$  the number of cells in the  $i$ -th subtables ( $k = \sum_{i=1}^c k_i$ ),  $c$  margins  $\sum_{j=1}^{k_i} n_j(i)$  have been fixed for every subtable by conditioning in the same way as previously has been done with all the cells in the table. In both cases, we can

consider with any loss of generality that the data come from a Poisson sampling scheme and we should introduce linear constraints on the expected frequencies to carry out any inferential procedure equivalent to the original one. This is mathematically distinguished through the parameter space  $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^t : \mathbf{X}_0^T \mathbf{m}(\boldsymbol{\theta}) = \mathbf{X}_0^T \mathbf{n}\}$ , where  $\mathbf{X}_0 = \bigoplus_{i=1}^c \mathbf{J}_{k_i}$  for the multinomial ( $c = 1$ ) or product-multinomial sampling ( $c \geq 1$ ) and  $\Theta = \mathbb{R}^t$  for the Poisson sampling. Note that for the Poisson sampling we can consider  $c = 0$ , referred to the number of linear constraints. Inferential results were given in Martín and Pardo (2008b) dealing simultaneously with the three sampling schemes and focussing on minimum  $\phi$ -divergence estimators, which cover the maximum likelihood estimator as a special case.

Now in addition to the previous model we shall assume  $r \leq t - c$  linear constraints,  $\mathbf{C}^T \mathbf{m}(\boldsymbol{\theta}) = \mathbf{d}^*$ , therefore jointly the first  $c$  linear constraints the new parameter space is given by

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^t : \mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) = \mathbf{d}\}, \quad (1)$$

being  $\mathbf{L} = (\mathbf{X}_0, \mathbf{C})$ ,  $\mathbf{d}^T = (\mathbf{n}^T \mathbf{X}_0, (\mathbf{d}^*)^T)$ , for  $c \geq 1$ , and  $\mathbf{L} = \mathbf{C}$ ,  $\mathbf{d} = \mathbf{d}^*$ , for  $c = 0$ . It is also assumed to hold  $k \geq t - c - r$ , and  $\text{rank}(\mathbf{L}) = \text{rank}(\mathbf{L}, \mathbf{d}) = c + r$ . In most cases  $\mathbf{d}^* = \mathbf{0}_r$ , being  $\mathbf{0}_r \equiv (0, \dots, 0)^T$ .

The theory of loglinear models with linear constraints (LMLC), was at the beginning considered only for the (product) multinomial sampling (see Haber and Brown (1986)). Its extension to Poisson sampling is interesting not only for applications such as one shown in Section 3, but also for allowing a flexible notation and for providing unified estimation procedures for different sampling plans in LMLC.

## 2 Minimum power-divergence estimators for LMLC

Let  $\Phi$  be the class of all convex and differentiable functions  $\phi : [0, \infty) \rightarrow \mathbb{R} \cup \{\infty\}$ , such that at  $x = 1$ ,  $\phi(1) = \phi'(1) = 0$ ,  $\phi''(1) > 0$ . For defining any statistical tool based on  $\phi$ -divergence estimators is usual to take probabilistic arguments (see Pardo (2006)), however in the Poisson sampling setting it is not sensible to use such measures because of the meaningless of probabilistic parameters, for this reason in Martín and Pardo (2008b)  $\phi$ -divergence measures with non-negative  $k$ -dimensional vectors  $\mathbf{a}$  and  $\mathbf{b}$  as arguments were proposed

$$D_\phi(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^k b_i \phi\left(\frac{a_i}{b_i}\right), \quad \phi \in \Phi, \quad (2)$$

where  $0\phi(0/0) \equiv 0$  and  $0\phi(p/0) \equiv p \lim_{u \rightarrow \infty} \phi(u)/u$  conventions are assumed. In particular by taking

$$\phi_\lambda(x) = \frac{x^{\lambda+1} - x - \lambda(x-1)}{\lambda(\lambda+1)}, \quad \lambda(\lambda+1) \neq 0, \quad (3)$$

and  $\phi_{\lambda^*}(x) = \lim_{\lambda \rightarrow \lambda^*} \phi_\lambda(x)$ , if  $\lambda^*(\lambda^* + 1) = 0$ , power divergence measures are obtained (see Read and Cressie (1988)). A minimum power divergence estimator for LMLC is defined as follows

$$\hat{\boldsymbol{\theta}}_{\phi_\lambda} = \arg \min_{\boldsymbol{\theta} \in \Theta} D_{\phi_\lambda}(\mathbf{n}, \mathbf{m}(\boldsymbol{\theta})), \tag{4}$$

with  $\Theta$  defined by (1). By taking  $\phi_0(x) = x \log x - x + 1$  in (4), the maximum likelihood estimators (minimum Kullback divergence estimators) for LMLC are obtained.

In the same way as in Haberman (1974) was shown the equality of the maximum likelihood estimator for the three kinds of sampling plans, in the following Theorem is established the same equality for LMLC and it is also extended “partially” for minimum power divergence estimators.

**Theorem 1** (Haberman’s Theorem for LMLC). *Let  $\log \mathbf{m}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta} = \mathbf{J}_k u + \mathbf{W}\bar{\boldsymbol{\theta}}$  be a loglinear with  $\mathbf{L}^T \mathbf{m}(\boldsymbol{\theta}) = \mathbf{d}$ , being  $\mathbf{d}^* = \mathbf{0}_r$ . Then, the minimum power divergence of  $\bar{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_{\phi_\lambda}$ , does coincide for the three types of sampling plans and the estimated expected total mean  $\hat{N}_{\phi_\lambda}$  is functionally obtained,*

$$\hat{N}_{\phi_\lambda} = \begin{cases} \left( \sum_{i=1}^k \frac{n_i^{\lambda+1}}{(\bar{m}_i(\bar{\boldsymbol{\theta}}_{\phi_\lambda}))^\lambda} \right)^{\frac{1}{\lambda+1}}, & c = 0, \\ n, & c \geq 1, \end{cases} \tag{5}$$

where  $\bar{\mathbf{m}}(\bar{\boldsymbol{\theta}}) = (\bar{m}_1(\bar{\boldsymbol{\theta}}), \dots, \bar{m}_k(\bar{\boldsymbol{\theta}}))^T = \exp\{\mathbf{W}\bar{\boldsymbol{\theta}}\} / (\mathbf{J}_k^T \exp\{\mathbf{W}\bar{\boldsymbol{\theta}}\})$ .

Note that  $\hat{N}_{\phi_\lambda}$  does not coincide for the three types of sampling plans unless  $\lambda = 0$ , however this result remains being important because the algorithm needed for calculating  $\bar{\boldsymbol{\theta}}$  is the same for any sampling scheme.

### 3 Multiplicative models for mortality ratios

Loglinear models are applied in a broad class of research works, specially in social, behavioral, and biomedical sciences. A clear interest of considering a unified estimation algorithms for the three kinds of sampling plans in loglinear modelling arises specially dealing with data from medical or epidemiological investigations where the mortality rates are the main characteristic to be estimated. Let us introduce some basic ideas, concepts and notation related to a study of mortality ratios. Suppose  $n_{1ij}$  is the number of deaths in stratum  $i$  of population  $j$ , and let  $n_{2ij}$  be the number of survivors, i.e. being  $n_{ij} = n_{1ij} + n_{2ij}$  the number of individuals at risk and  $p_{ij}$  the corresponding probability of death,  $(n_{1ij}, n_{2ij})$  is distributed according to a multinomial distribution with parameters  $(n_{ij}; p_{ij}, 1 - p_{ij})$ . By considering independence of number of deaths among stratum and populations, we deal with data with product-binomial sampling, being  $c = IJ$

and  $k_h = 2$ ,  $h = 1, \dots, c$  ( $k = 2c = 2IJ$ ). When  $n_{ij}$  is large enough compared to  $n_{1ij}$ , it holds  $n_{1ij} \sim \text{Bin}(n_{ij}, p_{ij}) \rightarrow \mathcal{P}(m_{ij})$ , being  $m_{ij} = n_{ij}p_{ij}$ , and therefore we can consider  $k = IJ$  individuals independently obtained through Poisson sampling ( $c = 0$ ).

Following the multiplicative model in Gail (1978), mortality ratios for stratum on one hand and for populations on the other hand are considered, and taking into account that now we deal with weighted data  $(\frac{1}{n_{11}}n_{111}, \dots, \frac{1}{n_{IJ}}n_{1IK})^T$  rather than the data  $(n_1, \dots, n_k)^T$  given in Section 1 (see Agresti (1990), Section 6.7.2), we carry out a simulation study for analyzing some power divergence estimators of mortality ratios under the assumption that a subset of populations have equal mortality ratios.

**Acknowledgments:** This work was supported by Grants MTM2006-06872 and CAM-UCM2007-910707. I would like to thank Professor Leandro Pardo who was the advisor of my Ph.D. dissertation on which is this work based.

## References

- Agresti, A. (1990). *Categorical Data Analysis*. John Wiley & Sons.
- Gail, M. (1978). The Analysis of Heterogeneity for Indirect Standardized Mortality Ratios. *Journal of the Royal Statistical Society. Series A*, **141**, 224–234.
- Haber, M., and Brown, M.B. (1986). Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. *Journal of the Royal Statistical Society, Series B*, **81**, 477–482.
- Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press.
- Martín, N., and Pardo, L. (2008a). Minimum phi-divergence estimators for loglinear models with linear constraints and multinomial sampling. *Statistical Papers*, **49**, 15–36. DOI: 10.1007/s00362-006-0370-3
- Martín, N., and Pardo, L. (2008b). New families of estimators and test statistics in loglinear models. *Journal of Multivariate Analysis*, In Print. DOI: 10.1016/j.jmva.2008.01.002
- Pardo, L. (2006). *Statistical inference based on divergence measures*. Chapman & Hall/CRC.
- Read, T. R. C., and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer.
- Zelterman, D. (1999). *Models for discrete data*. Oxford University Press.

# Bilinear Varying-Coefficient Models for Seasonal Time Series and Tables

Brian D. Marx<sup>1</sup>, Paul H.C. Eilers<sup>2</sup>, Jutta Gampe<sup>3</sup> and Roland Rau<sup>4</sup>

<sup>1</sup> Dept of Experimental Statistics, Louisiana State University, Baton Rouge, USA

<sup>2</sup> Dept of Methodology and Statistics, Utrecht University, The Netherlands & Data Theory Group, Leiden University, The Netherlands

<sup>3</sup> Max Planck Institute for Demographic Research, Rostock, Germany

<sup>4</sup> Terry Stanford Institute of Public Policy, Duke University, Durham, USA

**Abstract:** We propose generalized linear models for time series or age-time tables of seasonal counts. The linear predictor contains a smooth component for the trend and the product of a smooth varying-coefficient (the modulation) and a periodic time series (the carrier wave) of arbitrary shape. The model is applied to female deaths in the US due to respiratory diseases.

**Keywords:** Array regression, Life table, P-splines, Tensor product

## 1 Introduction

Judging from observed monthly counts of deaths (by year and age), many diseases show seasonal patterns. It is of interest to model these patterns, their overall strength and the relative strengths in different months. In our earlier work (Eilers et al., 2008), we proposed a modulation model: the pattern within years is described by a (co)sine and its amplitude is described by smooth varying-coefficient surfaces, over year and age. To this a trend is added.

Such a modulation model can describe seasonal patterns quite well, but from studying the residuals it became clear that the (co)sine function is too simple: it cannot handle relatively sharp peaks in winter and relatively flat troughs in summer. One possible solution is to introduce modulated harmonics of double or triple frequency, in the spirit of Fourier analysis. Instead we opt for a different approach, one that offers better opportunities for generalization. In a time series (of disease counts), we assume that there exists a “carrier wave” (a term we borrow from radio technology), which is modulated over time. The period of carrier wave is 12, and it is parameterized by the vector  $\gamma$ . This leads to a bilinear modulation model. We propose an iterative algorithm, cycling between estimation of the carrier wave and the modulation. With proper normalization of  $\gamma$  this leads to an

identifiable model that can easily be estimated by iterative weighted least squares.

## 2 One-dimensional Model Details

To start simply, we describe the models in one dimension (over time), suppressing age, and move into two dimensions in the following application section. Thus the observed data are a time series of counts,  $y_t$ ,  $t = 1, 2, \dots, T$ . The linear predictor is  $\eta_t$ , which models the logarithm of the expected value, i.e.  $\log(\mu_t) = \eta_t$ .

**The modulation model.** Eilers et al. (2008) described the model

$$\eta_t = v_t + f_t \cos(\omega t) + g_t \sin(\omega t), \quad (1)$$

where  $\omega = 2\pi/12$  and the series of coefficients  $v$ ,  $f$ , and  $g$  are smooth, and are modeled using P-splines. This is a modulation model: the amplitude of the seasonal waveforms is varied in strength by  $f$  and  $g$ . Technically this is a varying-coefficient model (VCM); it is estimated without backfitting (Eilers and Marx, 2002). As mentioned, this model can be extended with sine and cosine waves of double and triple frequency, each having their own modulation.

**A bilinear model.** Imagine a more general and unrestricted carrier wave, characterized by the 12-vector  $\gamma$  (one entry for each month). The model is

$$\eta_t = v_t + h_t \gamma_{[t]}, \quad (2)$$

where we use the notation  $[t]$  for  $((t - 1) \bmod 12) + 1$ . The model is bilinear, because both  $h$  and  $\gamma$  are unknown. We propose the following iterative algorithm: (i) For given  $\gamma$ , we have a VCM again to estimate  $h$  and  $v$ . (ii) For given  $h$  and  $v$ , estimation of  $\gamma$  can be done using (generalized) linear regression, which can be viewed as an “estimated” basis. We cycle back and forth until convergence. Identifiability conditions are required on  $h$  or  $\gamma$ ; we choose  $\sum_{k=1}^{12} \gamma_k = 0$  and  $\sum_{k=1}^{12} \gamma_k^2 = 12$ .

Both the modulation and the bilinear models can be viewed to have their own complications and simplifications. Certainly (1) and (2) can also be combined in one model: the (co)sine then models the main periodic pattern, while  $\gamma$  and  $h$  would explain additional shocks.

## 3 The Two-dimensional Bilinear Model

Eilers et al. (2008) also formulated the modulation model in two dimensions; here we extend the bilinear model as such. The counts are now arranged in a two-dimensional table  $Y = [y_{ta}]$ , e.g., indexed by both time,  $t = 1, \dots, T$ , and by age,  $a = 1, \dots, A$ . The bilinear model is expressed as,

$$\log(\mu_{ta}) = \eta_{ta} = v_{ta} + h_{ta} \gamma_{[t]}. \quad (3)$$

Notice that (3) has a double (time and age) subscript for the varying trend and modulation coefficients, producing varying coefficient surfaces. The carrier has the same periodic structure as before. The varying coefficients  $v$  and  $h$  are assumed to be smooth along time and age, and thus we model them using tensor product B-spline bases, allowing general surfaces. In the spirit of a P-spline approach, we avoid knot selection by: (i) using a sufficiently rich  $K \times L$  gridded tensor product basis, and (ii) imposing penalties on the coefficients associated with the rows and columns of the tensor product basis, where each penalty is regularized by its own positive tuning parameter,  $\lambda$ , hence allowing anisotropic smoothing.

Let  $B = [b_{tk}]$  ( $\check{B} = [\check{b}_{al}]$ ) be the  $T \times K$  ( $A \times L$ ) B-spline basis on the time (age) domain. Denote  $\mathcal{A}$  and  $\mathcal{B}$  as the  $K \times L$  matrices of the tensor product coefficients for  $V = [v_{ta}]$  and  $H = [h_{ta}]$ , respectively. We can rewrite (3) as

$$\log(M) = V + \Gamma H = B A \check{B}' + \Gamma B \check{B}' \check{B}', \quad (4)$$

where  $M = [\mu_{ta}]$  and  $\Gamma = \text{diag}(\gamma_{[t]})$ , for  $t = 1, \dots, T$ . Again, for fixed  $\Gamma$ , we have a 2D VCM; for fixed  $V$  and  $H$ , we have a generalized linear regression. Penalties are now applied to both rows and columns of  $\mathcal{A}$  and  $\mathcal{B}$ . Denote the (second order) difference penalty matrices  $D$  and  $\check{D}$  with dimensions  $(K-2) \times K$  and  $(L-2) \times L$ , respectively. The penalty is defined as

$$P = P_{\mathcal{A}} + P_{\mathcal{B}} = (\lambda_1 \|D\mathcal{A}\|_F + \check{\lambda}_1 \|\mathcal{A}\check{D}'\|_F) + (\lambda_2 \|D\mathcal{B}\|_F + \check{\lambda}_2 \|\mathcal{B}\check{D}'\|_F), \quad (5)$$

where  $\|\cdot\|_F$  indicates the Frobenius norm, or the sum of the squares of all elements. The penalty is composed of two parts, e.g. the first is equivalently  $P_{\mathcal{A}} = \text{vec}(\mathcal{A})' [\lambda_1 (I_L \otimes D'D) + \check{\lambda}_1 (\check{D}'\check{D} \otimes I_K)] \text{vec}(\mathcal{A})$ , where  $I$  is the identity matrix. The tensor product coefficients,  $\mathcal{A}$  and  $\mathcal{B}$ , are found by maximizing the penalized Poisson log-likelihood function

$$l^*(\mathcal{A}, \mathcal{B}) = l(\mathcal{A}, \mathcal{B}) - P. \quad (6)$$

Formulating the bilinear model in such a way has the advantage that it plays into the hands of the generalized linear array modelling (GLAM) (Currie et al., 2006). GLAMs are extensions of the generalized linear model to tensor product structures for data in multi-dimensional arrays. They offer very efficient computation with increases in fitting speed (of far more than 10-fold in most cases) when compared to the following unfolded representation

$$\begin{aligned} \text{vec}(M) &= (\check{B} \otimes B) \text{vec}(\mathcal{A}) + (I_A \otimes \Gamma) (\check{B} \otimes B) \text{vec}(\mathcal{B}) \\ &= X\beta, \end{aligned} \quad (7)$$

where  $X = [(\check{B} \otimes B) | (I_A \otimes \Gamma) (\check{B} \otimes B)]$  and  $\beta = (\text{vec}(\mathcal{A})', \text{vec}(\mathcal{B})')'$ . The gain in memory and speed using the GLAM formulation in (3), when compared to the unfolded representation above in (7), is due to the fact that

the former completely avoids the use of Kronecker products when computing the elements in the  $KL \times KL$  information matrix,  $X'WX$  with  $W = \text{diag}(\mu)$ . Thus GLAM specifically provides efficiency in the iterative scoring algorithm

$$(X'\tilde{W}X + P)\beta = X'\tilde{W}\tilde{z},$$

$\tilde{z} = \tilde{\eta} + \tilde{W}^{-1}(\text{vec}(Y) - \text{vec}(\tilde{M}))$  is the Poisson “working vector” at the current iterate.

#### 4 Optimization of the Penalty

In the bilinear model, we have four tuning parameters: a penalty on the rows and columns for each of the varying trend and the varying modulation (carrier wave) coefficient surfaces. The penalty in (5) is constructed in such a way that each row (or column) of the tensor product coefficients has the same amount of penalization, but with a breakage in linkage from one row (or column) to the next. Large  $\lambda$  enforces smoothness, whereas small values encourages roughness in either the row or column orientation. By virtue of the fact that the general carrier wave can adapt to spikes (troughs) in the winter (summer) months, overdispersion was not as problematic as was encountered with the smoother (co)sine modulation model. Nonetheless, during optimization only, we put zero weights on observations with months December through March; the remaining months have enough information to optimize smoothness. Such a scheme ensures that winter epidemics do not unduly influence optimal smoothing, and we have the added benefit of measuring any specific annual epidemic effect.

We choose to optimize the tuning parameters by monitoring and minimizing  $\text{AIC} = \text{deviance}(Y; \mathcal{A}, \mathcal{B}) + 2 \times \text{ED}$ , where ED is the approximate effective dimension computed from the trace of the corresponding “hat” matrix. We implement the search in a greedy way: each of the four  $\lambda$ s is changed in turn, by one step up and one step down on the grid. The step that gives the largest decrease gives the largest improvement in AIC and is kept. If there is no improvement in AIC, then the current value is kept.

#### 5 Two-Dimensional Bilinear Application

We present an application in two dimensions. We model counts of females of ages 51–100 in the United States who died of respiratory diseases during 1960–1998. The data can be viewed as a large contingency table of monthly death counts with 23,400 cells (50 ages  $\times$  [39 years  $\times$  12 months]).

Both the trend and the modulation surfaces were constructed on the year and age grid using a basis with  $K \times L = 13 \times 13$  tensor products of (cubic) B-splines. A second order difference penalty was used on both the rows and on the columns of the tensor product coefficients, each having their

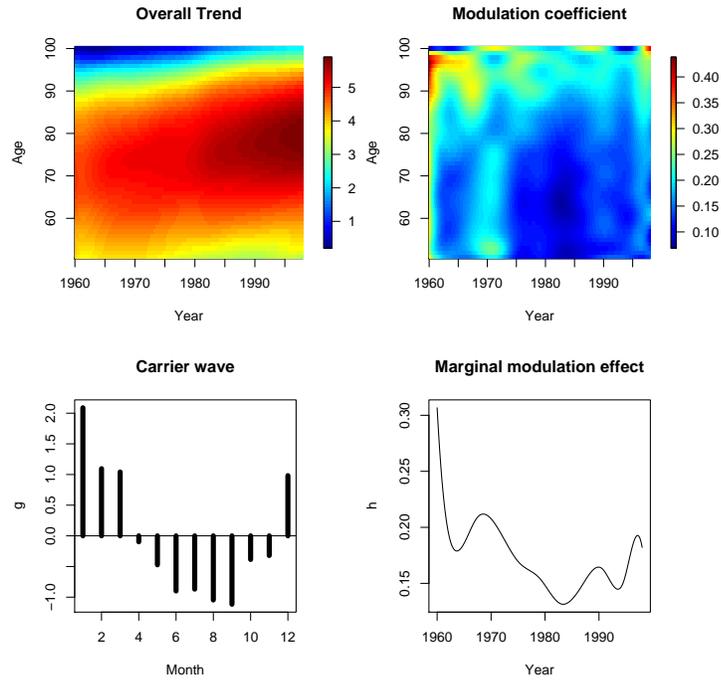


FIGURE 1. Image plot of the varying intercept term or overall trend (upper, left); 2D modulation effect for the carrier wave, (upper, right); the estimated carrier wave (lower, left); Marginal modulation effect (lower, right).

own tuning parameter to allow anisotropic smoothing. The  $\lambda$ s could take values on a grid with width 0.5 (on a base 10 log scale), and were changed in turn to test for possible downhill turns in (weighted) AIC. The results were:  $\lambda_1 = 0.8$ ,  $\check{\lambda}_1 = 2.1$ ,  $\lambda_2 = 0.04$ ,  $\check{\lambda}_2 = 0.78$ . The combined ED=46.8, associated with both estimated varying coefficient surfaces. We see that  $ED < 2 \times 13^2$ , indicating that we chose a sufficiently rich set of bases.

## 6 Extensions

The bilinear model assumes that the carrier wave has the same pattern for each year and age, only the amplitude changes. This assumption can be relaxed along the lines of the BAYSEA model (Akaike, 1980). Instead of a 12-vector series  $\gamma$ , we introduce a series with as many elements as there are

months in the data. To reduce its freedom, penalties are introduced, between identical months in adjacent years. This forces the parameter value for, e.g., May 1990 not to deviate too much from the average of the parameter values for May 1989 and May 1991. This scheme is applied to all years and all months.

In one dimension the model is  $\eta_t = v_t + s_t$ , where  $v$  is the trend,  $s$  is the seasonal component. A second order difference penalty is put on  $v$ . The penalty on  $s$  is seasonal, having components  $(s_1 - 2s_{13} - s_{25})^2 + (s_2 - 2s_{14} - s_{26})^2 + \dots$ . Thus the elements of  $s$  corresponding to, e.g., January readings get a second order penalty, as do other months. Essentially this model contains one long series, the trend  $v$ , and 12 shorter smooth series, one for each month. To gain efficiency in computation, we propose to use P-spline smoothing for  $v$  as well as for each of the twelve smooth series corresponding to the months. This model is more flexible than the periodic carrier wave modulation model, but is more difficult to interpret.

In one dimension, the BAYSEA model has no modulation component. In two dimensions we can imagine a carrier wave in BAYSEA style, which is modulated by a surface along age and time.

## 7 References

- Akaike, H. (1980). Seasonal Adjustment by a Bayesian Modeling. *Journal of Time Series Analysis*, **1**, 114.
- 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**, 259-280.
- Eilers, P.H.C., Gampe J., Marx, B.D., and Rau R. (2008) Modulation models for seasonal time series and incidence tables. *Statistics in Medicine*. In press.
- 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. (1993). Varying coefficient models. *Journal of the Royal Statistical Society*, **55**(4), 757-796.

# Model building for series of block designed experiments

Iwona Mejza<sup>1</sup>, Stanislaw Mejza<sup>1</sup>

<sup>1</sup> Department of Mathematical and Statistical Methods,  
University of Life Sciences in Poznań, Wojska Polskiego 28, 60-637 Poznań,  
Poland, e-mail: imejza@up.poznan.pl; smejza@up.poznan.pl

**Abstract:** The paper deals with the problem of a model building for data obtained from a series of experiments carried out in block designs. In the modeling of the series, the structure of the experimental material and an appropriate scheme of randomization play the main role. Two cases of modeling are considered, i.e. when environments are randomized and when they are not. In the resulting models, treatment effects are assumed to be fixed. Some information about an application of these models is given.

**Keywords:** master plan, pure effect, scheme of randomization, series of block designs, technical error, zero yield

## 1 Introduction

The mixed models presented here and used for the analysis of the series of experiments are randomization models which are based on the ideas of Nelder (1954, 1965a, 1965b), (cf. Bailey, 1981, Hinkelmann and Kempthorne, 1994, Caliński and Kageyama, 2000, 2003). According to this, the method for deriving the linear model utilizes an experimental situation and a design of the experiment (in each environment). The starting points of the method are a notion of experimental units, a concept of zero yield (conceptual yield, true response) and the use of randomization in the design (in each environment). The process of randomization plays a central role in the paper. The considered schemes of the randomization are strictly connected with the two-stage structure of plots in each single experiment of the series. Three-step randomization is applied to the units, i.e. to environments, to the blocks and to the units within each block, (independently in each environment). Moreover, we consider two experimental situations connected with the randomization of environments, i.e. (a) environments are not randomized, (b) environments are randomized (three-step randomization). The basic assumption in the building of the model refers to the meaning of an observed yield in the experiment. In the paper we assume that 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 and "*technical error*" connected with measurements (additivity assumed). Let

us note that every unit possesses some kind of fertility which gives some yield both in the case when treatments do not occur on a unit and in the case where no treatments have an effect on yield. This yield is called "zero yield". The increase (or decrease) in the zero yield due to the treatment (combination) used on experimental unit is called "pure effect". Many times the sum of "zero yield" and "pure effect" is called "pure yield" and it constitutes the basis for further statistical analysis. But the "pure yield" is changed by the "technical error", i.e. by the inaccuracy of performing the experiment or by the measurements. This fact is included in our approach to modeling of observation (cf. Mejza, 1994).

Furthermore, we assume that in all environments we observe  $v$  treatments. The treatments will not be randomized. Hence, the treatment effect will affect the fixed part of the linear model for observed yield.

Let a population of units in each environment,  $P_g$  ( $g = 1, \dots, S$ ) be divided into  $B_g$  blocks, and let each block be additionally divided into  $K_g (\leq v)$  units. The approach presented in the paper is applicable to the series of incomplete block designs or complete block designs (as a particular case). Let  $D = \{D_1, D_2, \dots, D_S\}$  be the theoretical design of the series. The subdesigns  $D_g$  may be the same or different with respect to the material structure and to the treatment structure. In the paper the blocks and units within blocks will always be randomized.

## 2 Environments are not randomized

There are experimental situations where the environments are not randomized, for example when the experiments are performed in some chosen experimental stations. Then the environments represented by the stations are treated as fixed. This means that the environmental effects affect the fixed part of the linear model for the observed data.

Let the master design  $D_g$  utilizes  $b_g (\leq B_g)$  blocks of the size  $k_g (\leq K_g)$  all. The scheme of randomization assigns (randomly) only the theoretical blocks to the experimental ones and, independently within blocks, to all units.

Then the observed yield obtained in the  $g$ -th environment ( $g = 1, 2, \dots, S$ ), the  $h$ -th block ( $h = 1, 2, \dots, b_g$ ) concerning the  $j$ -th treatment ( $j = 1, 2, \dots, v$ ) has a form

$$E(y_{ghj}) = \mu + \pi_g + \gamma_j + (\pi\gamma)_{gj} \quad (1)$$

where  $\mu$  denotes the general parameter,  $\pi_g$  stands for the effect of the  $g$ -th environment,  $\gamma_j$  denotes the effect of the  $j$ -th treatment, and  $(\pi\alpha)_{gj}$ , stand for interaction effects. All the treatment effects and the environment effects are considered to be fixed.

The dispersion structure of the observed yield has the form

$$Cov(y_{ghj}, y_{g'h'j'}) =$$

$$\begin{cases} \sigma_{gh}^2 + \sigma_{ghj}^2 + \sigma_\varepsilon^2, & g = g', h = h', j = j', \\ \sigma_{gh}^2 - \frac{1}{K_g - 1} \sigma_{ghj}^2, & g = g', h = h', j \neq j', \\ -\frac{1}{B_g - 1} \sigma_{gh}^2, & g = g', h \neq h', \\ 0, & g \neq g', \end{cases} \quad (2)$$

where  $\sigma_{gh}^2, \sigma_{ghj}^2$  denotes the variance of blocks within environments, variance of units and variance of technical errors. Two practical situations are of great interest. The first one we have when we draw units from infinite population of units,  $B_g \rightarrow \infty, K_g \rightarrow \infty$  and  $\sigma_{gh}^2 = \sigma_\pi^2$ . Then the observed yields from different units can be treated as uncorrelated. In the second practical case we utilize whole population of units ( $b_g = B_g, k_g = K_g$ ) and additionally we assume that  $\sigma_{gh}^2 = \sigma_\pi^2, \sigma_{ghj}^2 = \sigma_\eta^2$ .

### 3 Environments are randomized

It may happen that experiments are conducted in places not necessarily connected with experimental stations. Then the places can be randomized. Usually in this case the whole population of potential places (environments) takes part in the experiment. In this case the environments (represented by places or/and years) affect the dispersion structure of the model.

Let us assume that an experiment utilizes  $s \leq S$  environments (e.g. places) and the master design  $D_g$  utilizes  $b_g \leq B_g$  blocks of the size  $k_g \leq K_g$ . The scheme of randomization assigns (randomly) the  $s$  master plans to the environments and, independently within environment, theoretical blocks to the experimental ones and, independently within blocks, to all units.

This kind of randomization leads to a linear model of observed yield of the form:  $E(y_{ghj}) = \mu + \gamma_j$ , while the dispersion structure is as follows:

$$\begin{aligned} & Cov(y_{ghj}, y_{g'h'j'}) = \\ & \begin{cases} \sigma_\pi^2 + \sigma_\beta^2 + \sigma_\eta^2 + \sigma_\varepsilon^2, & g = g', h = h', j = j', \\ \sigma_\pi^2 + \sigma_\beta^2 - \frac{1}{K_g - 1} \sigma_\eta^2, & g = g', h = h', j \neq j', \\ \sigma_\pi^2 - \frac{1}{B_g - 1} \sigma_\beta^2, & g = g', h \neq h', \\ -\frac{1}{S - 1} \sigma_\pi^2, & g \neq g', \end{cases} \quad (3) \end{aligned}$$

where  $\sigma_\pi^2, \sigma_\beta^2$  and  $\sigma_\varepsilon^2$ , denotes the variance of environments, variance of blocks within environments, variance of units and variance of technical errors, respectively. As in the previous case two practical situations are of great interest. The first one is when we take units from an infinite population of units,  $S \rightarrow \infty, B_g \rightarrow \infty, K_g \rightarrow \infty$ . Then the observed yields from different units can be treated as uncorrelated. In the second practical case we utilize whole population of units ( $b_g = B_g, k_g = K_g$ ).

## 4 Application

The particular cases of the models presented above are used in the analysis of series of agricultural experiments. In the case when experiments are repeated over some places the randomization is easily to perform. In case when the experiments are performed over years then the model (1) with (2) seems to be appropriate. In both cases, treatment effects are assumed to be fixed. This is a typical situation in the modeling of data from agricultural field experiments, especially in plant breeding and variety testing experiments. Problems with analyzing series of multi-environment variety trials using randomization-derived mixed linear model (for different design than considered here) are presented in Caliński, et al. (2005). Before using one of the proposed linear models it is necessary to check if the proper assumptions are satisfied.

### References

- Bailey R.A. (1981). A unified approach to design of experiments. *J. Roy. Statist. Soc. Ser. A*, **144**, 214-223.
- Caliński, T., Czajka, S., Kaczmarek, Z., Krajewski, P., Pilarczyk, W., (2005). Analyzing multi-environment variety trials using randomization derived mixed linear. *Biometrics*, **61**, 448-455.
- Caliński, T., Kageyama, S. (2000). *Block Designs: A Randomization Approach, Volume I: Analysis*. Springer, New York.
- Caliński, T., Kageyama, S. (2003). *Block Designs: A Randomization Approach, Volume II: Design*. Springer, New York.
- Hinkelmann, K., Kempthorne, O. (1994). *Design and Analysis of Experiments*. Wiley, New York.
- Nelder, J.A. (1954). The interpretation of negative components of variance. *Biometrika*, **41**, 544-548.
- Nelder, J.A. (1965a). The analysis of randomized experiment with orthogonal block structure. 1. Block structure and the null analysis of variance. In: *Proc. Roy. Soc., A*, **283**, 147-162.
- Nelder, J.A. (1965b). The analysis of randomized experiment with orthogonal block structure. 2. Treatment structure and general analysis of variance. In: *Proc. Roy. Soc., A*, **283**, 163-178.
- Mejza, S. (1994). On modelling of experiments in natural sciences. *Listy Biom. - Biom. Letters*, **31**, 79-100.

# A note on a modelling environmental indexes

Stanisław Mejza<sup>1</sup>, João T. Mexia<sup>2</sup>, Dulce Pereira<sup>3</sup>

<sup>1</sup> Department of Mathematical and Statistical Methods, University of Life Sciences in Poznań, Wojska Polskiego 28, 60-637 Poznań, Poland, [smejza@up.poznan.pl](mailto:smejza@up.poznan.pl)

<sup>2</sup> Departamento de Matemática, Universidade Nova de Lisboa, Quinta da Torre, 2825 Monte da Caparica, Portugal

<sup>3</sup> 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 of the Genotype x Environmental Interaction in an analysis of series of experiments. Regression analysis is one of the most often applied statistical techniques for this purpose. In regression analysis we should have two sets of variables, one characterizing genotypes while the second characterizing environments. The so-called adjusted means (or some other genotype characteristics) for genotypes usually constitute observations of the dependent variable. The problem is how to model the environmental indexes, these being the observation of independent variable. In the paper we examine three approaches to modelling the environmental indexes; two are based on so-called adjusted means for environments, while the third method uses iterative ("zig-zag") algorithm for estimation of the considered 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 performed without any reference to applications. But here we will refer the data to a 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 to select genotypes that are consistently high-yielding over the range of observed or potential environments. The main problem of inference from series of experiments is connected with modelling (structuring) the GEI (Genotype x Environment Interaction) effects. Usually, the GEI is non-orthogonal. Hence, for its analysis it is necessary to use very advanced statistical tools (cf. for example Aastveit and Mejza, 1992). In this paper our interest in GEI analysis is limited to two cases. The first one involves a modification of the analysis of a two-way table as we do in incomplete block designed experiments. The second approach is based on an application of joint regression analysis.

## 2 Adjusted environmental effects

In this approach the GEI will be expressed by a 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 in matrix form as:

$$Y = 1\mu + \Delta'\tau + D'\beta + e$$

where  $1$  denotes the vector of ones,  $\Delta'$  and  $D'$  are design matrices for environments and genotypes,  $\mu$  is the general mean, while  $\tau$  and  $\beta$  are the vectors of environment and genotype effects and  $e$  denotes the vector of errors. In this approach the results known from the theory of block designs are used. Then so-called adjusted means for environments can be calculated as:  $\tilde{\beta} = \tilde{\mu} + G^{-}Q$ , where  $\tilde{\mu} = n^{-1}Y'1$  - estimates 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,  $k = N1$ ,  $r = N'1$ ,  $n = r'1$ ,  $k^{\delta} = \text{diag}(k_1, k_2, \dots, k_b)$ ,  $r^{-\delta} = \text{diag}(1/r_1, 1/r_2, \dots, 1/r_j)$ ,  $Q = T - Nr^{-\delta}B$ ,  $B = DY$ ,  $T = \Delta Y$ , and  $G^{-}$  - denotes the generalized inverse of the matrix  $G$ . Let us note that generalized inverse of the matrix is not unique. Many different methods of calculating the general inverse lead to different values of adjusted means. We would like to check if there exist any relationships between normally used algorithms of calculating adjusted means. In the paper we consider two ways of obtaining a generalized inverse of matrix  $G$ . The first method of calculation  $G^{-}$  comes from Tocher (1952). This method can be used when  $r(G) = b - 1$ . Then we have  $G^{-} = (k^{\delta} - Nr^{-\delta}N' + kk'/n)^{-1}$ . In the second case we use a general algorithm leading to the unique Moore-Penrose generalized inverse.

## 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  $(\alpha_j, \beta_j)$ ,  $j = 1, \dots, J$  are the regression coefficients and the  $x_i$ ,  $i = 1, \dots, b$ , are the environmental indexes.

One can observe that the lately proposed so called zig-zag algorithm is very efficient in finding the estimates of  $(\alpha_j, \beta_j)$  and the  $x_i$  (cf. Mexia *et al.*, 1999, Pereira and Mexia, 2002, 2003).

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 an environment can be a good initial value (cf. Gusmao, 1985). Moreover, in designs the initial values may be the average for the superblock. Then the goal function is minimized firstly for the regression coefficients and then for environmental index. At the end of each iteration the environmental indexes are rescaled so that the range of environmental indexes is kept unchanged. The process always converges, but in a time which depends on the initial values. Hence, the iteration procedure is called zig-zag algorithm.

The aim of this paper is to compare three approaches, described briefly above, to estimating the 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 (cf. Mejza *et al.*, 2007). In the paper we will compare three approaches on the basis of yield/plot observed genotypes. The data are represented in a matrix of 32 environments (rows) and 20 genotypes (columns). The 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 non-orthogonal. For these three data sets we calculated environmental indexes using three approaches. Let us call these 1: Tocher's inverse matrix, 2: the generalized inverse matrix, and finally, 3: the zig-zag algorithm. Then the coefficient of correlations, calculated for 1/5, 1/2 and 3/5 removed observations, were as follows:  $r(1,2)=0.9999$ ,  $r(1,3)=0.9999$ ,  $r(2,3)=0.9998$ ;  $r(1,2)=0.9978$ ,  $r(1,3)=0.9999$ ,  $r(2,3)=0.9977$ ;  $r(1,2)=1$ ,  $r(1,3)=0.9986$ ,  $r(2,3)=0.9986$ ; respectively. In this example 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 any of the environments exactly 4 genotypes were observed. Together we had 80 observations. The same genotypes were observed in the same environments but in two years. The coefficient of correlations calculated for the first year are:  $r(1,2)=0.9450$ ,  $r(1,3)=0.8278$ ,  $r(2,3)=0.8274$ ; and those for the second year are:  $r(1,2)=0.9997$ ,  $r(1,3)=0.4276$ ,  $r(2,3)=0.4179$ ; respectively. 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. This results from the low variability of the environments in the second year.

## 5 Discussion

In the paper the example with rye was considered only. Rye belongs to a quiet stable variety over different environments. Hence, there is probably very good correspondence between environmental indexes obtained by the considered methods. Our observations, from other examples, suggest that zig-zag algorithm is more suitable in the case when the environments are non-homogenous. In the second example the variability of environments was small and environmental indexes obtained by zig-zag algorithm were less correlated with the others. Another comparison relates to calculations. It seems to us that adjusting the environmental indexes (Section 3) is much easier and numerically more efficient. The calculation of the inverse or generalized inverse, in the case of large number of environments, is numerically difficult and biased by numerical errors. This is a very weak point of the methods based on the matrix inverses.

## References

- Aastveit, A., and Mejza, 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.
- Mejza S., Mexia, J.T., Pereira, D.G. (2007). On a modelling environmental indexes. *Proc. of the 22nd International Workshop on Statistical Modelling*, Barcelona, J. del Castillo, A. Espinal, P. Puig (Eds), 445-448.
- 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. (2003). Reproducibility of Joint Regression Analysis. *Colloquium Biometryczne*, **33**, 279-299.
- Tocher, K. D. (1952). The design and analysis of block experiments (with discussion). *J. Roy. Statist. Soc. Ser. B***14**, 45-100.

# Integrated statistical analysis to identify associations between DNA copy number and gene expression in microarray data

Renée X. de Menezes<sup>1,2,4</sup>, Marten Boetzer<sup>1</sup>, Melle Sieswerda<sup>1</sup>,  
Claudia Gaspar<sup>3</sup>, Riccardo Fodde<sup>3</sup>, Gert-Jan van Ommen<sup>1</sup> and  
Judith M. Boer<sup>1</sup>

<sup>1</sup> Center for Human and Clinical Genetics, Leiden University Medical Center,  
PO Box 9600, 2300 RC Leiden, The Netherlands

<sup>2</sup> Pediatric Oncology Laboratory, Erasmus Medical Center, The Netherlands

<sup>3</sup> Josephine Nefkens Institute, Erasmus Medical Center, The Netherlands

<sup>4</sup> Author to which correspondence should be addressed.

**Abstract:** One of the mechanisms of gene expression regulation is copy number change, which plays an important role in tumorigenesis. We propose a model that allows simultaneous analysis of copy number and expression microarray data. Gene sets, rather than individual genes, are used to construct a robust and sensitive model. This is a general and flexible tool, providing a powerful approach to prioritize putative targets for functional validation. We illustrate the method's performance in two breast and one colon cancer examples. The method can be also applied to other types of microarray data.

**Keywords:** high-dimensional data; global test; gene networks.

## 1 Introduction

Several approaches have been described to identify genes whose expression levels are associated with copy number changes of the corresponding genomic region. Most of those involve arbitrariness, however, for example via discretization of copy number or selection of features to be studied. Moreover, focus has remained on finding associations with high-level copy number changes, partly because methods were powerless to handle low-level changes. The more prevalent low-level gains and losses were shown to have a significant influence on expression levels of genes in the regions affected, but these effects were more subtle on a gene-by-gene basis, even though its impact on the dysregulation of gene expression patterns in cancer may be equally important if not more important than that of high-level amplifications (Hyman *et al*, 2002). Therefore, the search for DNA regions that might be involved in the initiation and progression of cancer must

be powerful enough to detect subtle gene-specific effects that are possibly consistent across many genes.

We developed a model to jointly analyse copy number and gene expression array data. Our model is able to detect subtle effects of mild copy number alterations by taking into account multiple genes in the altered region. We will present the method and illustrate its applicability in examples in the following sections.

## 2 A model for integrated analysis of microarrays

We search for copy number changes affecting gene expression within the same chromosomal region. So we consider copy number as the dependent variable, while expression is the independent variable. Since the expression of many genes may be affected by each copy number change, we propose using the model, for each  $i$ ,

$$E(Y_{ni}) = \alpha + \sum_{j=1}^J \beta_j X_{nj}, \quad n = 1, \dots, N. \quad (1)$$

where  $Y_{ni}$  represents the copy number measured for sample  $n$  and array-CGH copy number probe  $i$  ( $i = 1, \dots, I$ ) and  $X_{ni}$  represents the expression level for sample  $n$  and expression probe  $i$ . We assume that  $\beta \equiv (\beta_1, \dots, \beta_J)^t$  is a vector of independent random variables, each with a certain distribution and, under the null hypothesis of no association, has mean 0 and variance  $\tau^2 \equiv 0$ . This makes (1) a random-effects model, and a natural distribution to assign to  $\beta$  is the multivariate normal. This guarantees that model (1) is identifiable, since  $J \gg N$  in this context.

Under the alternative hypothesis of association, the mean of each  $\beta_j$  may still be zero, but their variance should be strictly positive ( $\tau^2 > 0$ ), suggesting that a non-empty subset of the  $\{X_{nj}, j = 1, \dots, J\}$  is associated with copy number measurements for probe  $i$ . Therefore, we shall focus on testing  $H_0 : \tau^2 = 0$  against  $H_a : \tau^2 > 0$ . A test to compare such null and alternative hypotheses was proposed by Goeman *et al* (2004) for testing association between expression levels of many genes with a clinical outcome. The regression framework means that confounders can be included in the analysis. Obtained p-values are corrected for multiple testing allowing for dependence (Benjamini & Yekutieli, 2001). Individual contributions of gene expressions can be calculated allowing for finding genes whose expression levels are most associated with the copy number changes.

The method has been implemented as an R package and is available from BioConductor.

### 3 Examples

#### 3.1 Breast cancer examples

Here we analyse two independent, publicly available datasets relating to breast cancer samples. The first was produced and first analysed by Pollack *et al* (2002), consisting of copy number and expression array data generated using the same Stanford cDNA arrays for 37 breast tumors and 4 breast-tumor cell lines. The second, analysed and made publicly available by Chin *et al* (2006), comprised 89 samples profiled using CGH (2.5K BAC) and expression (Affymetrix U133A) arrays. Note that genomic coverage differ considerably between the two sets of arrays used. Nevertheless, association patterns found were strikingly similar. We shall focus on chromosome arms 8p and 17q.

On 8p an area of association between copy number and expression near the centromere (8p11-12) is found on both datasets. Note that copy number change in this region is mostly of low level, and at most 25% of the samples display any change. Some of the genes we found have been shown to be breast cancer oncogenes that work in combination to influence the transformed phenotype in human mammary epithelial cells. Amplifications of 8p11-12, occurring in 15% of breast cancers, were found significantly associated with disease-specific survival and distant recurrence (Chin *et al*, 2006). Finally, two of the genes we identified were confirmed as drugable targets associated with gene dosage of 8p11-12 in a larger breast cancer study (Chin *et al*, 2006).

The long arm of chromosome 17 is very interesting as is known to harbour oncogenes *ERBB2*, *TRAF4* and *BRCA1*. Copy number of Pollack's data shows 3-4 regions of high amplification, and although the same amplitude of change is displayed in Chin's data the regions are not confirmed, due to the much lower genomic resolution yielded by BACs compared with cDNAs. Nevertheless, strikingly similar patterns of association are found in the datasets, mapping exactly to the regions identified with Pollack's copy number data. One of these regions includes *ERBB2*, another one *TRAF4*. For the samples studied by Chin some clinical variables were known, such as estrogen-receptor status. We have repeated the analysis using this variable as a confounder in the model. Using for example 0.01 as threshold for the FDR-corrected p-values, results were similar to those without considering the confounder for most chromosomes, except for five of them: for 1p, 5q, 6p and 12q, no features were selected with ER-status adjustment whilst the unadjusted model selected between 20% and 60% of the probes, and for 19q, no features were found with the unadjusted model, but about 40% of the BACs were selected with ER-status adjustment. More importantly, in each of these chromosome arms a handful of BACs was assigned an FDR-corrected p-value in one analysis below 0.01, whilst in the other the p-value was larger than 0.20. These results suggest that copy number-based mech-

anisms of gene expression regulation differ according to estrogen-receptor status in breast cancer.

### 3.2 Colon cancer

We also applied the model to a dataset relating to 68 colonic cancers, for which genomic copy number and expression were measured using 3K BAC arrays and Affymetrix U133 2.0 Plus arrays, respectively. Patients were followed up for at least five years after diagnosis and sample collection, so prognosis information was available, namely that 40 patients were disease-free after 5 years, whilst 28 had at least local recurrence in the same period. Joint analysis of the copy number and expression data suggests that, from among the associations found, some are specific to one prognostic type. For example, copy number changes on 1p, 14q, 18p and 18q only display association with expression amongst good prognosis samples, whilst for 5p and 16p the reverse is observed. The knowledge of where differential expression-regulating mechanisms are activated according to the prognosis suggests for example new targets for therapy.

## 4 Conclusions

We have proposed a new approach to identify association between high-throughput genomic copy number and gene expression profiling data, which can be used to identify putative candidate genes involved in tumorigenesis. By considering the expression levels of many genes simultaneously in the model, our approach takes advantage of the typically larger signal-to-noise ratio in copy number compared with expression data. It is able to control for confounder effects, and avoids the arbitrariness of discretizing the data. It requires neither matching between copy number and expression probes on the genome, as it rightly focuses on finding relevant associations regardless of their genomic location, nor categorization of copy number, both of which are possible sources of bias. Finally, it can also be applied to other types of microarray, such as methylation and SNP arrays.

**Acknowledgments:** We thank P.A.C. 't Hoen, J.T. den Dunnen and J. J. Goeman for fruitful discussions. Samples analysed in the colon cancer example are part of a collaboration with Dr. G.A. Patijn and Prof. D. Gotley, University of Queensland, Australia. This work was conducted within the Centre for Medical Systems Biology (CMSB), established by the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research (NGI/NWO). This work is part of the BioRange bioinformatics research program supported by the Netherlands Bioinformatics Centre (NBIC).

## References

- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.
- Chin, K., DeVries, S., Fridlyand, J., Spellman, P.T., Roydasgupta, R. *et al* (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiology. *Cancer Cell*, **10**, 529–541.
- Eilers, P.H.C. and de Menezes, R.X. (2005). Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.
- Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–109.
- Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S. *et al* (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Research*, **62**, 6240–6245.
- Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S. *et al* (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 12963–12968.

# An Estimator of the Variance of Measurement Error

Indrajit Mitra<sup>1</sup>, Anirban Majumdar<sup>2</sup>

<sup>1</sup> Thomson Financial, 425 Market Street, 6th Floor, San Francisco, CA 94105.

<sup>2</sup> Graduate School of Business, Gleacher Center, University of Chicago, Chicago, IL 60611

**Abstract:** In this letter we propose a new estimator for the strength (variance) of additive, unbiased, measurement errors in a linear simultaneous equation model with a single predictor variable. This estimator, which relies on the in-sample correlation between the  $T$  measured values of the independent variable and a principal component estimate of the latter, is unbiased up to quadratic order in intrinsic noise in the system. We also outline a second estimator for the variance of measurement error which can be used if the sensitivities of the dependent variables on the independent variable, (i.e.  $\vec{\beta}$ ) are uncorrelated with the intrinsic noise for every observation. This latter estimator serves as a benchmark against which we compare the former estimator, and we find that the estimator which uses PCA is superior in a mean squared error sense.

**Keywords:** Measurement Error; Principal Component Analysis; Time Series .

## 1 Introduction

The simplest linear model of dependent variables  $\vec{y}_i$  supposes that these are proportional to a single underlying variable  $\vec{x}$  plus homoskedastic, mean zero shocks  $\vec{e}_i$  which are uncorrelated with the independent variable  $\vec{x}$ :

$$y_{it} = \beta_i x_t + e_{it}, \quad \text{E}[e_{it}] = 0, \quad \text{E}[\vec{e}_i \vec{e}_j] = \sigma_e^2 \delta_{ij} \delta_{tt'}, \quad \text{E}\left[\sum_t x_t e_{it}\right] = 0, \quad \forall i, \quad (1)$$

where we have explicitly included an index  $t$  to label each of the  $T$  observations of  $\vec{y}_i$  and  $\vec{x}$ . We shall often refer to  $x_t$  as the predictor variable. In this letter, we look at the model with no-intercept, and additionally assume  $\text{E}[x_t] = 0, \forall t$ . (The  $x_t$  are not constants, so in standard parlance, this is a Structural Model).

Given the  $N \times T$  observations  $y_{it}$  and measured values of the predictor  $x_t$  (which we henceforth denote by  $X_t$ ), the OLS estimates of the sensitivities  $\beta_i$  are unbiased and efficient if  $X_t$  have no error. This is however extremely rare, and a more realistic scenario is when  $x_t$  is measured with

some measurement error. In this paper, we consider unbiased, additive, homoskedastic, measurement errors, i.e. the supplied values of the independent variable  $X_t$  have been measured with measurement error  $\eta_t$  which has mean zero and variance  $\sigma_\eta^2$ . We further assume that the errors are uncorrelated with the “true”  $x_t$ :

$$X_t = x_t + \eta_t; \quad \text{s.t.} \quad \text{E}[\eta_t] = 0, \quad \text{Var}[\eta_t] = \sigma_\eta^2, \quad \text{E} \left[ \sum_t (\eta_t x_t) \right] = 0. \tag{2}$$

It is well known [Carroll] that in this case, the OLS estimate of  $\beta_i$  is biased towards zero:  $\text{E} \left[ \hat{\beta}_i^{\text{OLS}} \right] = \frac{\sigma_\varepsilon^2}{\sigma_x^2 + \sigma_\eta^2} \beta_i$ . In cases where  $\sigma_\eta^2$  is not known, it is important to estimate it to obtain improved, unbiased estimates of  $\beta_i$  (using the SIMEX method, for instance). Moreover, if an estimate of  $\sigma_\eta^2$  indicates that it is very large, it is better not to use the measured  $X_t$ , but instead estimate past realizations of the predictor variable and  $\beta_i$  by principal component analysis.

In Section 2 of this letter, we propose a new method to estimate  $\sigma_\eta^2$ . This technique is different from some existing ones in the literature in that it uses internal data and therefore is immune from data transportability issues. It also does not require us to use sub-samples. The technique uses the fact, that measurement error in the supplied values of the independent variable  $X_t$  causes it to lose correlation with the principal component estimate of past realizations of the predictor [Jolliffe]. We first derive an analytic expression (correct up to order  $\sigma_\varepsilon^2$ ) of how this correlation depends on  $\sigma_\eta^2$ . Inverting this result gives us an estimator of  $\sigma_\eta^2$  which, by construction, is unbiased to leading order in  $\sigma_\varepsilon^2$ . We end the section with another estimator for  $\sigma_\eta^2$  which is inferior to the previous one because it has a bias of order  $\sigma_\varepsilon^2$ . The idea for the second estimator is quite straightforward – errors in  $x_t$  of the kind mentioned in equation 2 cause the realized residuals to be correlated with the estimated sensitivities  $\hat{\beta}_i$  by an amount proportional to measurement errors  $\eta_t$ . In models where it is known that the shocks  $e_{it}$  and the true  $\beta_i$  are uncorrelated for all times, the amount of measurement error ( $\eta_t$ ) may be extracted by regressing the realized residuals against the estimated  $\hat{\beta}_i$ :  $\hat{\sigma}_\eta^2$  is the variance of  $\hat{\eta}_t$ . We include this second estimator in this letter not only because of its simplicity, but because it serves as a good benchmark for the former technique which relies on PCA. Results of simulations are included.

## 2 Estimating the variance of measurement error

### 2.1 An estimator of $\sigma_\eta^2$ using principal component analysis

The presence of measurement error in the supplied values of the independent variable  $X_t$  causes it to lose correlation with the “true”  $x_t$ . Since the

latter cannot be observed, it is interesting to estimate the loss in correlation of  $X_t$  with a proxy for past realizations of  $x_t$ . The one we choose is the PCA estimate which we shall call  $\hat{s}_t$ . Recall that  $\hat{s}_t$  is the principal eigenvector of the  $T \times T$  matrix:  $\Omega_{tt'} = \sum_i y_{it} y_{it'}$ .

Let us denote the normalized magnitude of the dot product between the measured values  $X_t$  and the PCA estimate  $\hat{s}_t$  by  $\hat{\Theta}$ :

$$\hat{\Theta} = \left| \frac{X}{\sqrt{(X \cdot X)}} \cdot \frac{\hat{s}}{\sqrt{(\hat{s} \cdot \hat{s})}} \right|. \quad (3)$$

The quantity  $\hat{\Theta}$  is a random number and we are interested in its expectation value as a function of  $\sigma_\eta^2$ . By assumption, the measurement errors  $\eta_t$  are uncorrelated with the true  $x_t$  (and hence also with the amount by which the PCA estimate is off – we denote this difference  $\hat{s}_t - x_t \equiv \theta_t$ ). Up to first order in  $\frac{\sigma_\eta^2}{\sigma_x^2}$  and  $\frac{\sigma_\varepsilon^2}{\sigma_x^2}$  the expected value of  $\hat{\Theta}$  is:

$$\mathbb{E}[\hat{\Theta}] \approx 1 - \frac{1}{2} \frac{\sigma_\eta^2}{\sigma_x^2} - \frac{1}{2T\sigma_x^2} \mathbb{E} \left[ \sum_t \theta_t^2 \right] = 1 - \frac{1}{2} \frac{\sigma_\eta^2}{\sigma_x^2} - \frac{T\sigma_\varepsilon^2}{2\lambda}, \quad (4)$$

where we have used Equation 2, and  $\lambda$  is the largest eigenvalue of  $\Omega_{tt'} = \sum_i y_{it} y_{it'}$ . In the last equality we have used the expectation value of the norm of  $\theta_t$  (This proof appears in one of the author's papers [Mencheró].) Our estimator for  $\sigma_\eta^2$  is obtained by inverting the above equation:

$$\frac{\hat{\sigma}_\eta^2}{\sigma_x^2} \approx 2 \left( 1 - \hat{\Theta} - \frac{T\hat{\sigma}_\varepsilon^2}{2\lambda} \right). \quad (5)$$

This is the main result of this paper. Note that by construction, this estimator is unbiased up to order  $\frac{\sigma_\varepsilon^2}{\sigma_x^2}$ . The linear model Equation 1 allows us to choose a normalization for the signal: we choose to work with  $\sigma_x^2 = 1$ . In order to use equation 5, one estimates in the usual way:

$$\hat{\sigma}_\varepsilon^2 = \text{Var} \left[ y_{it} - \frac{\hat{s}_t \sum_{t'} y_{jt'} \hat{s}_{t'}}{\sum_{t'} \hat{s}_{t'}^2} \right].$$

In this short letter we demonstrate how well the estimator of Equation 5 works through simulations. We simulate  $N \times T$  table  $y_{it}$  according to Equation 1. To obtain imprecise measured values  $X_t$ , we add  $\eta_t$  to  $x_t$  where the  $\eta_t$  are generated as independent draws from a mean zero normal distribution of varying width. The results are plotted in Figure 1.

## 2.2 A naive estimator of the noise $\sigma_\eta^2$

In linear models where in addition to Equation 1,  $E[\sum_i \beta_i e_{it}] = 0$  holds for all observations  $t$ , one may estimate the amount of measurement error  $\sigma_\eta^2$  by an alternative method. This relies on the observation that measurement errors in  $X_t$  will show up as a non-zero correlation between the OLS

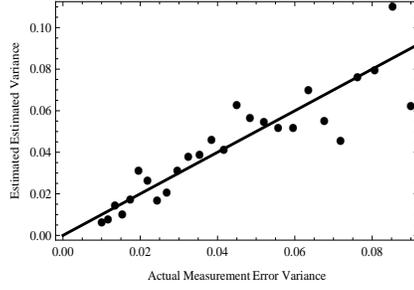


FIGURE 1. Comparison of estimated variance of measurement errors using the estimator of Equation 5 with the actual values. The line  $y=x$  is shown for visual aid to demonstrate how close the estimated values are to the actual values. In this simulation,  $N = 40$ ,  $T = 100$ , and  $\sigma_\varepsilon^2 = 0.25$ . Values of  $x$  are independent draws from normal  $(0,1)$ , while the  $\beta_i$  are independent draws from normal  $(0, 0.5)$ .

estimates  $\hat{\beta}_i^{OLS}$  and the estimated residuals  $\hat{e}_{it} = y_{it} - \left( \frac{\sum_{t'} y_{it'} X_{t'}}{\sum_{t'} X_{t'}^2} \right) X_t$ . For  $\sigma_\eta^2 \ll \sigma_x^2$ , the expectation value of the estimated residuals is:

$$E[\hat{e}_{it}] \approx e_{it} - \hat{\beta}_i^{OLS} \eta_t. \tag{6}$$

The measurement errors can be estimated by regressing the residuals  $\hat{e}_{it}$  against the estimated  $\beta_i$ 's using OLS:

$$\hat{\eta}_t = - \frac{\sum_i \hat{e}_{it} \hat{\beta}_i^{OLS}}{\sum_i (\hat{\beta}_i^{OLS})^2}. \tag{7}$$

The variance of  $\hat{\eta}_t$  is an estimate of  $\sigma_\eta^2$ . For small measurement errors, i.e.  $\sigma_\eta^2 \ll \sigma_\varepsilon^2 \ll \sigma_x^2$ , we now show that this estimate of  $\sigma_\eta^2$  is biased by an amount  $\approx \frac{\sigma_\varepsilon^2}{\sum_i \beta_i^2}$ . To leading order in  $\sigma_\eta^2$ , we may approximate  $\hat{\beta}_i^{OLS} \approx \beta_i + \frac{\sum_t e_{it} x_t}{\sum_t x_t^2} \equiv \beta_i + \vartheta_i$ . Using Equation 8 for  $\hat{\eta}_t$ , we have

$$\hat{\eta}_t = - \frac{\sum_i \hat{e}_{it} \hat{\beta}_i^{OLS}}{\sum_i (\hat{\beta}_i^{OLS})^2} \approx \eta_t - \frac{\vartheta^2}{\beta^2} x_t - \frac{\sum_i e_{it} (\beta_i + \vartheta_i)}{\beta^2 + \vartheta^2}, \tag{8}$$

up to order  $\sigma_\varepsilon^2$ . Noting that the three terms on the right are uncorrelated with each other, the variance of  $\hat{\eta}_t$  is:

$$\text{Var}[\hat{\eta}_t] \approx \sigma_\eta^2 + \text{Var} \left[ \frac{\sum_i e_{it} (\beta_i + \vartheta_i)}{\beta^2 + \vartheta^2} \right]. \tag{9}$$

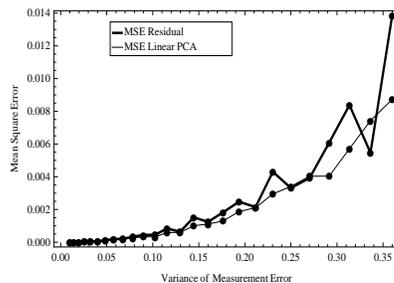


FIGURE 2. Plot of MSE of estimators in Equation 5 and Equation 7 as a function of measurement errors. A single point on the graph is the MSE for 100 repeats.  $N=40$ ,  $T=100$ ;  $x_t$  are i.i.d  $N(0,1)$ ;  $\beta_i$  are i.i.d  $N(0,2)$ , and  $\epsilon_{it}$  are i.i.d  $N(0,0.7)$ . The  $\eta_t$  are i.i.d  $N(0,\sigma_\epsilon)$ .

The expectation value of the above estimator (up to lowest order in  $\sigma_\epsilon^2$ ) is therefore:

$$E[\hat{\sigma}_\eta^2] = \sigma_\eta^2 + \frac{\sigma_\epsilon^2}{\sum_i \beta_i^2}. \quad (10)$$

We include this method of estimating the variance of measurement errors because, where it can be used, it serves as a useful benchmark against which the performance of the previous estimator of Equation 5 can be compared. Indeed, in Figs 2 and 3, we show how the estimators compare against each other as measurement errors and the intrinsic noise increase, respectively. To make sure that we are not trading a smaller bias of the estimator for larger variance, we plot the mean squared errors of the estimators. Note that the naive method of Section 2.2 is inferior to the method which relies on the in-sample correlation of  $X_t$  and  $\hat{s}_t$ ; for low values of measurement noise, it loses out to the linear estimator Equation 5.

### 3 PCA estimates or erroneous measured values?

One common use of linear models of the form Equation 1 is in predicting the returns of risky assets. Model builders estimate their models in one of two ways: either by claiming that  $x_t$  can be measured with a “fair” amount of precision and extracting  $\beta_i$  by least squares, or alternatively, by extracting both  $x_t$  and  $\beta_i$  from the historical time series of returns  $y_{it}$  using PCA. Clearly, it is useful to know which technique to use when  $\sigma_\eta^2 \neq 0$ . A sensible criteria in deciding between techniques is to choose that method which has the highest expected correlation of the input (i.e.  $X_t$  or  $\hat{s}_t$ ) with the true (but unobservable)  $x_t$ . It can be shown that the expected

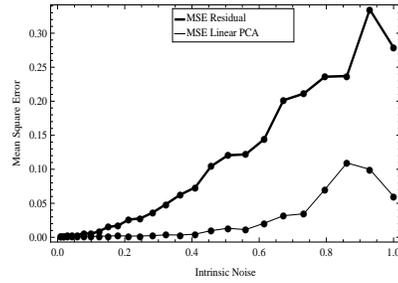


FIGURE 3. Plot of MSE of estimators in Equation 5 and Equation 7 as a function of the strength of shocks. A single point on the graph is the MSE for 100 repeats.  $N=40, T=100; x_t$  are i.i.d  $N(0,1)$ ;  $\sigma_\eta^2=0.16$ .  $\beta_i$  are i.i.d  $N(0, 0.2)$  – a small variance chosen to highlight the bias derived in Equation 10.

correlation of  $\hat{s}_t$  with  $x_t$  (which we denote by  $q_{PCA}$ ) up to first order in noise is

$$E[q_{PCA}] \approx 1 - \frac{\sigma_\varepsilon^2}{2\sigma_x^2 \sum_i \beta_i^2}. \tag{11}$$

The expected correlation of the measured values of the independent variable, i.e.  $X_t$ , with the true  $x_t$  is

$$E[q_X] \approx 1 - \frac{\sigma_\eta^2}{2\sigma_x^2}. \tag{12}$$

Comparing Equations 11 and 12, we find that the measured values  $X_t$  should be discarded in favor of  $\hat{s}_t$  if:

$$\sigma_\eta^2 > \frac{\sigma_\varepsilon^2}{\sum_i \beta_i^2}. \tag{13}$$

Which side of the inequality one is in can be determined by using the estimated version of Equation 13.

**References**

Carroll, R.J., Ruppert, D., Stefanski, L.A. (2006) *Measurement Errors in Nonlinear Models* Chapman & Hall

Jolliffe, I.T. (2002). *Principal Component Analysis* Springer Series in Statistics

Menchero, J., and Mitra, I. (Fall 2008). The Structure of Hybrid Factor Models. *Journal of Investment Management*.

# Estimation of linear errors-in-variables models with error-free covariates: a backfitting approach

Vito M. R. Muggeo<sup>1</sup>

<sup>1</sup> Dip. Scienze Statistiche e Matematiche ‘S. Vianelli’ - Università di Palermo,  
email: vmuggeo@dssm.unipa.it

**Abstract:** We present a backfitting algorithm to estimate linear regression models having both error-prone and error-free covariates as predictors. The algorithm assumes that the variance-ratios are known, and it is particularly efficient when several explanatory variables are included. The resulting estimators are shown to be unbiased and to perform well as compared to method-of-moments estimators which are usually employed when the variance ratio is known.

**Keywords:** measurement error; backfitting; errors-in-variable; variance ratio

## 1 Introduction

In the standard regression framework it is assumed that the measurement error may occur only in the response variable  $Y$ , say. Covariates are assumed fixed or if this is not the case, inference is carried out by conditioning to the observed values: in each case, the explanatory variables are assumed *perfectly* measured. However this assumption sometimes is not met, especially in biology, epidemiology, or psychology, where the explaining variables of main interest,  $X_1, X_2, \dots$ , are actually measured with error: the resulting model is sometimes referred as EIV (‘errors in variables’) model. The effect of ignoring the so-called measurement error (hereafter ME) is well-known and documented in literature; Fuller (1987) and Carroll et al. (2006) are probably the two main references. Although several approaches have been proposed to deal with single EIV models, in practice the actual regression model may include additional truly-measured explanatory variables,  $Z_1, Z_2, \dots$ , whose effects have also to be investigated; relatively few papers are concerned with regression models having multiple error-free and error-prone covariates. In this paper we present a backfitting-type algorithm to estimate multiple linear regression models with true and mis-measured explanatory variables.

## 2 Methods: the model and the estimating algorithm

Let  $(X_i, Y_i)$  the  $n$  observable pairs of random variables; a EIV *functional* model may be written as

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i & \epsilon_i &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ X_i &= x_i + \delta_i & \delta_i &\sim \mathcal{N}(0, \sigma_\delta^2) \end{aligned} \quad (1)$$

where the  $x_i$ s and the  $y_i$ s ( $y_i = \beta_0 + \beta_1 x_i$ ) are *latent*, i.e. unobservable, and  $\epsilon_i$  and  $\delta_i$  are the two independent disturbance terms.

Identification and estimation of such errors-in-variables models is usually performed by setting some constraints on the model parameters, typically known measurement variance, known reliability ratio or known variance ratio (Fuller, 1987). Here we assume that the variance ratio  $\lambda = \sigma_\epsilon^2 / \sigma_\delta^2$  is known, and with a single mis-measured explanatory variable, the objective function is  $S(\beta_0, \beta_1, x_1, \dots, x_n) = \sum_i^n \{(Y_i - \beta_0 - \beta_1 x_i)^2 / \lambda + (X_i - x_i)^2\}$ , which is optimized by

$$\hat{\beta}_1 = \frac{s_Y^2 - \lambda s_X^2 + \{(s_Y^2 - \lambda s_X^2)^2 + 4\lambda s_{XY}^2\}^{1/2}}{2s_{XY}} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad (2)$$

where  $s_X^2$ ,  $s_Y^2$  and  $s_{XY}$  are the usual sample moments. To account for both error-prone and error free covariates, we present an estimating algorithm based on the idea of backfitting. The backfitting is a very efficient algorithm proposed to estimate regression models with several nonparametric smooth terms (Hastie and Tibishrani, 1990): substantially it consists in estimating the nonparametric curves in turn for each variable using some ‘scatterplot smoother’ having the current residuals as response. The process is iterative since the residuals have to be ‘updated’ at each iteration and for each variable. We apply the same idea to estimate EIV models with additional error-free covariates. The algorithm is presented below for the a generic regression model based on the observable variables  $Y$ ,  $\{X_1, \dots, X_k, \dots\}$ , and  $\{Z_1, \dots, Z_j, \dots\}$ , with estimating model  $Y = \beta_0 + \sum_k \beta_k X_k + \sum_j \gamma_j Z_j + \epsilon$ , where it is assumed  $X_k = x_k + \delta_k$ ,  $\delta_k \sim \mathcal{N}(0, \sigma_{\delta_k}^2)$  and such that  $\lambda_k = \sigma_\epsilon^2 / \sigma_{\delta_k}^2$  for each error-prone covariate  $X_k$ .

1. Fit a naive standard regression model to get starting values of regression coefficients;
2. For each explanatory variable:
  - i. compute the current residuals, given by difference between responses and fitted values accounting for the effects of all but the current covariate;

- ii. compute the relevant regression coefficient  $\hat{\gamma}_j$  (via ordinary least squares) or  $\hat{\beta}_k$  (via formula (2)), for error-free or error-prone covariate, respectively.
3. Repeat up to convergence, and then compute the model intercept as  $\hat{\beta}_0 = \bar{Y} - \sum_k \hat{\beta}_k \bar{X}_k - \sum_j \hat{\gamma}_j \bar{Z}_j$ , where the ‘bar’ stands for mean.

The algorithm estimates the effect of each covariate in turn, using the residuals as pseudo-response; note, however, if the current variable is error-prone we estimate the effect via  $\hat{\beta}\hat{x}_i$  (rather than  $\hat{\beta}X_i$ ) where the  $\hat{x}_i$ s are the estimates of ‘true’ values  $x_i$  (Fuller, 1987).

### 3 Application and Simulations

Tomasello et al. (2007) investigate the effect of human impacts on growth performance of *Posidonia oceanica* using data coming from some coastal zones in Sicily, Italy. The response variable is rhizome elongation, and the explanatory variables are the shoot age (AGE, in years) and quality of meadow (MEA, disturbed/non-disturbed); standard least squares yield  $\hat{\beta}_{\text{MEA}} = -5.742$  with  $\text{SE}^* = 0.578$ , and  $\hat{\beta}_{\text{AGE}} = -0.541$  with  $\text{SE}^* = 0.074$ . If it is assumed that AGE is not perfectly measured, a EIV linear regression may be estimated through the aforementioned algorithm; results confirm findings from the standard least squares analysis, however the estimates from the ME model are larger and, as expected, with larger variance ( $\hat{\beta}_{\text{MEA}} = -10.9$  with  $\text{SE}^* = 1.91$  and  $\hat{\beta}_{\text{AGE}} = -1.508$  with  $\text{SE}^* = 0.317$ ). Note that in both cases the starred standard error  $\text{SE}^*$  is the empirical standard deviation coming from 100 bootstrap replicates. We also study the per-

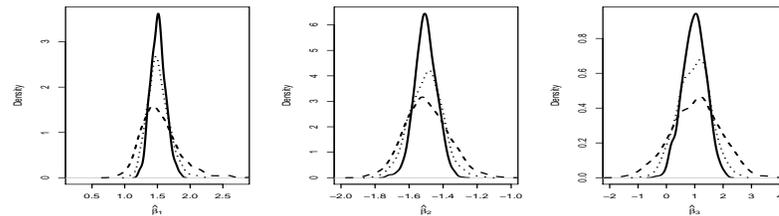


FIGURE 1. Monte Carlo smoothed estimates of sampling distributions of regression parameter estimators in simulation study for three sample sizes:  $n = 50$  (dashed line),  $n = 100$  (dotted line), and  $n = 200$  (continuous line).

formance of the proposed algorithm via a small simulation study: first we assess the sampling distributions of the estimators in a multiple regression model with two error-prone and an additional error-free covariates. Figure 1

TABLE 1. Sampling distribution summary of different estimators (see text).

<i>Sampling distribution</i>	method of moments			backfitting		
	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_0$	$\beta_1$	$\beta_2$
mean	1.982	1.549	-1.510	1.971	1.534	-1.490
sd	0.506	0.285	0.240	0.483	0.258	0.217
mse	0.256	*0.834	*0.576	0.234	*0.678	*0.469

\* values  $\times 10$ 

shows the results for three different sample sizes. Moreover, we consider a regression model with a covariate-by-factor interaction, namely a (continuous) error-prone explanatory variable with separate slopes in two groups whose categorical variable is error-free. Unlike standard (i.e., ordinary least squares) regression framework, modelling interactions within EIV models is not straightforward. The true model is  $Y = 2 + 1.5(x : G_1) - 1.5(x : G_2) + \epsilon$ , with observed  $X = x + \delta$ , where  $G_1$  and  $G_2$  are the two dummies relevant to two groups,  $\epsilon$  and  $\delta$  are independent zero-mean normal variates having  $\sigma_\epsilon = \sigma_\delta = 1.5$ . Table 1 compares our approach with method-of-moments estimators (Fuller, 1987, pag 202-205).

## 4 Conclusions

In this paper we have proposed an efficient backfitting-type algorithm to estimate multiple regression models including both error-prone and error-free covariates: parameter estimation is based on a minimum-distance criterion. Although alternative estimating approaches could be undertaken (e.g., simex), our method appears to be simpler, and moreover it assumes that the variance ratio ( $\sigma_\epsilon^2/\sigma_\delta^2$ ), and not the error variance ( $\sigma_\delta^2$ ), is known. This is the usual and the most widely used approach in EIV modelling of environmental and ecological data (e.g., Chen and Jackson, 2000).

## References

- Carroll, R.J., Ruppert D., and Stefanski L.A., and Crainiceanu C. (2006). *Measurement Error in Nonlinear Models, II ed.* Chapman & Hall.
- Chen, Y., and Jackson, D.A. (2000). An empirical study on estimators for linear regression analyses in fisheries and ecology. *Fisheries Research*, **49**, 193-206.
- Fuller W.A. (1987). *Measurement errors models.* Wiley.
- Tomasello, A., Calvo S., Di Maida, G., et al. (2007). Shoot age as a confounding factor on detecting the effect of human-induced disturbance on *Posidonia oceanica* growth performance. *J Experimental Marine Biology and Ecology*, **343**, 166-175.

# A Generalized Poisson Model for Underreporting

Gerhard Neubauer<sup>1</sup> and Gordana Djuraš<sup>1</sup>

<sup>1</sup> Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria

**Abstract:** Underreporting in register systems can be analyzed using a binomial approach, where both parameters have to be estimated. Parameter estimation fails, whenever the first two sample moments are about equal. Neubauer & Friedl (2006) introduced a regression approach for a binomial overdispersion model. A more general approach is available using the Generalized Poisson distribution, where the binomial, the Poisson and the negative binomial cases are covered. In this paper we propose a regression model that allows to estimate  $\lambda$ , the total number of cases, and  $\pi$ , the reporting probability. Moreover one of the parameters ( $\alpha$ ) carries information on the type of the distribution. It is a real number that is zero in the Poisson case. Thus testing  $\alpha = 0$  can be seen as a test for Poisson over- or underdispersion. The performance of the method is investigated in a simulation study and found to work well, safe for two exceptions. One is the Poisson limit, where  $\lambda \rightarrow \infty$ . The second is the perfect system limit, where  $\pi \rightarrow 1$ . Finally the model is applied to real data from the Austrian crime register.

**Keywords:** Underreporting, Generalized Poisson distribution, testing Poisson under-/overdispersion, regression.

## 1 Introduction

Any register or counting system is prone to errors in recording. The reasons may be quite different in the various fields of application like public health, criminology, animal abundance or production. In public health we have registers for infectious diseases like HIV or chronic diseases like diabetes, and recording failures may occur as result of diagnostic errors or patients avoiding diagnosis. Crimes associated with shame are likely not to be reported to the police, just as theft of low value goods. The same holds for traffic accidents with minor damage. In all cases we are confronted with underreporting and the estimation of the total number of cases is of particular interest. Neubauer & Friedl (2006) addressed this problem by simultaneous estimation of both binomial parameters. They showed that a binomial and a beta-binomial regression model are suited for a wide range of applications. However, both models fail, if the sample variance is considerably larger than the sample mean, i.e a negative binomial type of relation between the first two moments. The negative binomial distribution can be

derived from a conditional binomial distribution, where the parameter  $n$  is random and an appropriate mixing distribution is chosen. Several distributions have the binomial, the Poisson and the negative binomial distribution as special cases. One of them is the Generalized Poisson (GP) distribution (Consul, 1989). We propose a regression approach for the GP distribution that allows to estimate the total number of cases, and moreover allows to identify the type of mean-variance relationship. The performance of the GP model is investigated by a simulation study and in applications to crime data.

## 2 Models for Underreporting

Let  $y_t$  be a sample of counts ( $t = 1, \dots, T$ ), which are the reported cases of some register system. Further let  $\lambda$  denote the total number of cases and  $\pi$  the reporting probability, then  $E(Y_t) = \mu = \lambda\pi$  is the mean model. For the  $Y_t \sim \text{Binomial}(\lambda, \pi)$  we have the binomial model for the estimation of the total number  $\lambda$  with the mean-variance relation  $\text{var}(Y_t) = \mu - \mu^2/\lambda \leq \mu$ . Allowing for larger variability becomes possible by treating parameters as random variables. The counts now have a conditional binomial distribution. For  $Y_t|P \sim \text{Binomial}(\lambda, p)$  and  $P \sim \text{Beta}(\gamma, \delta)$  we obtain the well-known beta-binomial as the marginal distribution of  $Y_t$ , with  $\text{var}(Y_t) = (\mu - \mu^2/\lambda)\phi$ , and  $\phi = (\lambda + \gamma + \delta)/(1 + \gamma + \delta) \geq 1$ .

Assuming  $Y_t|L \sim \text{Binomial}(l, \pi)$  and  $L \sim \text{Poisson}(\lambda)$  we obtain marginally  $Y_t \sim \text{Poisson}(\lambda\pi)$  where the parameters are not identified. Allowing for randomness in  $\lambda$  we state a conditional Poisson model as  $L|K \sim \text{Poisson}(k\lambda)$ . Using  $K \sim \text{Gamma}(\omega, \omega)$  in addition, we obtain a negative binomial marginal distribution for  $Y_t$  with parameters  $\omega$  and  $1 - \pi$ .  $\omega$  is the number of unreported cases and  $\pi$  is the reporting probability. The mean-variance relation is now  $\text{var}(Y_t) = \mu + \mu^2/\omega \geq \mu$ .

The GP distribution is defined by

$$p(Y|\theta, \tau) = \begin{cases} \frac{\theta(\theta+y\tau)^{y-1} \exp[-(\theta+y\tau)]}{y!} & y = 0, 1, 2, \dots \\ 0 & \text{for } y > m, \quad \text{when } \tau < 0 \end{cases} \quad (1)$$

where  $\theta > 0$ ,  $\max(-1, -\theta/m) < \tau \leq 1$  and  $m(\geq 4)$  is the largest positive integer for which  $\theta + m\tau > 0$  when  $\tau < 0$  (Consul, 1989). The parameter  $\tau$  tunes the type of distribution. Obviously for  $\tau = 0$  we get the Poisson distribution. For  $0 < \tau \leq 1$  we have the negative binomial and for  $\tau < 0$  the (positive) binomial distribution. In any case, the first two moments are given as  $E(Y) = \theta(1-\tau)^{-1}$  and  $\text{var}(Y) = \theta(1-\tau)^{-3}$ . This parameterisation does not immediately relate to  $\lambda$  and  $\pi$ . But a simple consideration using  $\lambda\pi = \theta(1-\tau)^{-1}$  and  $\lambda\pi(1-\pi) = \theta(1-\tau)^{-3}$  for the binomial and its analogues for the negative binomial shows, that  $\pi = 1 - (1-\tau)^{2s}$  and  $\lambda = \theta\pi^{-1}(1-\pi)^{-s/2}$ , where  $s = \text{sign}(\tau)$ .

### 3 Regression Models

More flexible models are at hand if  $E(Y_t) = \mu_t$  is allowed and  $\lambda_{t,\beta} = \exp(x'_t\beta)$  is used to make parameters identifiable, in the binomial, the beta-binomial and the negative binomial model. Here  $x_t$  is a  $d$ -vector of known regressors and  $\beta$  is the corresponding vector of unknown parameters. For the GP model we have to regard the requirement  $\mu_t > 0$ . Considering  $\mu = \theta(1-\tau)^{-1}$  and  $\log(\mu) = \log(\theta) + \log[(1-\tau)^{-1}]$  we define  $\log(\theta_{t,\beta}) = \eta_t = x'_t\beta$  and  $\log[(1-\tau)^{-1}] = \alpha$  or equivalently

$$\theta_{t,\beta} = \exp(x'_t\beta), \quad \text{and} \quad \tau_\alpha = 1 - \exp(-\alpha). \quad (2)$$

The likelihood contribution of the  $t$ -th observation is now

$$L(\alpha, \beta | y_t, x_t) = \frac{\theta_{t,\beta}(\theta_{t,\beta} + y_t\tau_\alpha)^{y_t-1} \exp[-(\theta_{t,\beta} + y_t\tau_\alpha)]}{y_t!}, \quad (3)$$

where the restrictions of (1) apply:  $\theta_{t,\beta} > 0$  as required, and also  $-\infty < \tau_\alpha \leq 1$  is fulfilled as implied by  $\max(-1, -\theta/m)$ . The parameter  $\alpha$  is a real number that indicates Poisson overdispersion when it takes values in the interval  $(0, \infty)$ . It indicates Poisson underdispersion for values in the interval  $(-\infty, 0)$ , and not much surprising  $\alpha = 0$  for the Poisson case. Testing  $\alpha = 0$  is therefore a possibility to identify near Poisson data, or in other words to test for Poisson over- or underdispersion.

### 4 Application to simulated data

To investigate the behavior of the GP regression model we performed a simulation study, where two situations were of special interest (i) the Poisson limit ( $\pi \rightarrow 0$ ), and (ii) the perfect reporting system limit ( $\pi \rightarrow 1$ ). Hence we set  $\pi = (0.1, 0.2, 0.5, 0.8, 0.9)$  in the simulations. In all cases we use a model with a trend and a seasonality component, reflecting a typical situation for crime data. For the total number of cases we have  $\lambda_t = \exp(\beta_0 + T_t + S_t)$ , where  $T_t = \beta_1 t + \beta_2 t^2 + \beta_3 \sin(\pi t/\psi)$  is the trend function,  $S_t = \beta_4 \cos(2\pi t/\psi)$  the seasonality,  $\beta = (3, 0.01, -0.00005, 0.2, -0.5)$  and  $\psi = 365.25/7$  tunes the trigonometric functions. Thinking of weekly counts we simulate data for four years ( $T = 209$ ), and for each of the ten settings  $R = 300$  samples are drawn. Figure 1 gives examples of the data, with trend curves for the total number (dashed) and the mean (solid).

Table 1 gives the results for  $\pi$ ,  $\alpha$  and  $\beta_0$ . The point estimates  $\hat{\pi}$  are reasonable in all cases, but the 95% confidence intervals  $CI(\hat{\pi})$  become very large for  $\pi = (0.1, 0.2)$ . For  $\pi = (0.8, 0.9)$  the intervals are narrow, indicating high precision of the estimates.  $T_{\hat{\alpha}}$  is the t-statistic for parameter  $\alpha$ . It gives an impression whether  $\alpha$  differs from zero, i.e. if Poisson over- or underdispersion is present. A value of  $|T_{\hat{\alpha}}| \leq t_{1-c/2, df}$  can be seen as indicator for Poisson distribution. In our case we have  $df = 203$  and thus

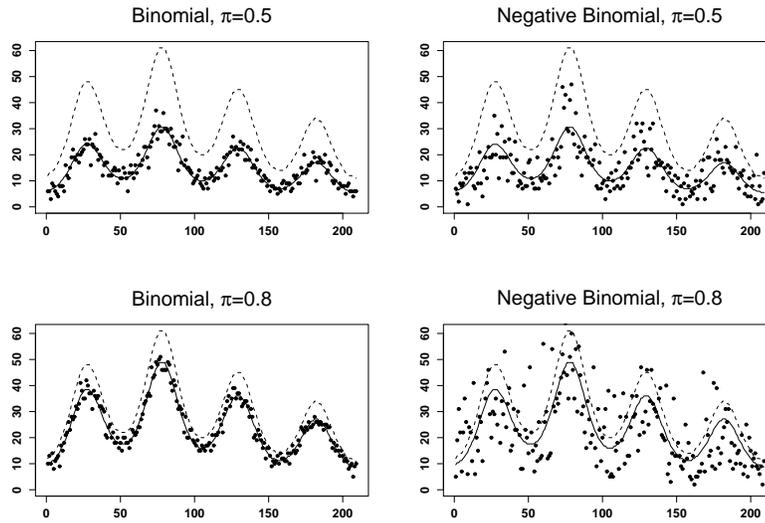


FIGURE 1. Examples of simulated data with true trends for the mean (solid) and the total number of cases (dashed)

for  $c = 0.05$  the critical value is about  $z_{1-c/2} = 1.96$ . Clearly for  $\pi = 0.1$  we have in both cases a Poisson data situation where  $T_{\hat{\alpha}}$  has values of  $-1.283$  and  $0.791$ . The last column of Table 1 gives the mean square error ( $mse$ ) for  $\beta_0$ . For  $\pi \rightarrow 0.5$  we observe the smallest, and for  $\pi = 0.1$  we have the largest values. Also when moving from  $\pi = 0.5$  to  $\pi = 0.9$  the values increase. Thus in either of the limiting situations we have an increase in the  $mse$  when compared to the values for  $\pi = 0.5$ . Considering the values for the bias we find that it largely dominates the values of  $mse$ . The  $mse$  for the slope parameters are not given in detail as they are close to zero in all cases, indicating, that they can be estimated unbiased and with high precision over a wide range of data situations. The overall impression is that the slope parameters are not affected by  $\pi \rightarrow 0$  or  $\pi \rightarrow 1$ .

## 5 Application to Austrian crime data

The real data examples are taken from the Austrian online crime register SIMO. For each of 132 regions in Austria we have weekly counts of different crime categories over the period of 2004-2007. Models similar to the one used in simulations were applied to data of larger regions and crime categories assault, bicycle theft and shop lifting. In most cases we find  $T_{\hat{\alpha}} > 1.96$  indicating a negative binomial distribution. Figure 2 shows the results for

TABLE 1. Results from the simulation

$\pi$	Binomial distribution				Negative Binomial distribution			
	$\hat{\pi}$	$CI(\hat{\pi})$	$T_{\hat{\alpha}}$	$mse_{\hat{\beta}_0}$	$\hat{\pi}$	$CI(\hat{\pi})$	$T_{\hat{\alpha}}$	$mse_{\hat{\beta}_0}$
0.1	0.13	-0.05;0.31	-1.28	5.04	0.10	-0.09;0.28	0.79	5.52
0.2	0.22	0.06;0.38	-2.43	2.19	0.18	0.00;0.35	1.87	2.95
0.5	0.51	0.41;0.61	-7.09	0.12	0.48	0.37;0.60	6.25	1.08
0.8	0.81	0.76;0.85	-15.60	0.36	0.81	0.76;0.86	13.02	1.08
0.9	0.90	0.88;0.93	-21.55	1.16	0.92	0.89;0.95	14.43	1.64

two such cases: shop lifting in an Austrian region and bicycle theft in a city. The estimated reporting probabilities are reasonable for both:  $\hat{\pi} = 0.69$  for shop lifting and  $\hat{\pi} = 0.67$  for bicycle theft.

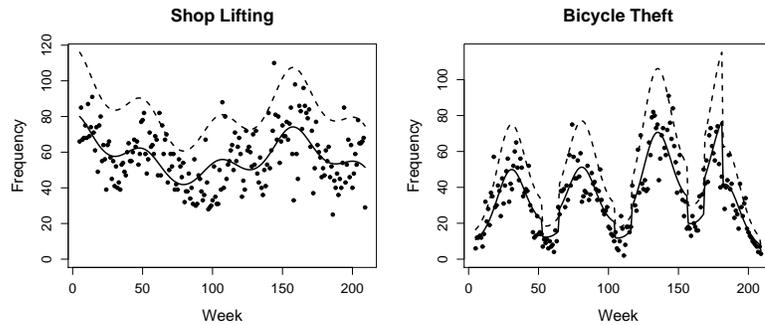


FIGURE 2. Two examples of Austrian crime data with estimated trends for the mean (solid) and the total number of cases (dashed)

## 6 Conclusions

The results from the simulations show that the Generalized Poisson approach works well for simulated data of one of the binomial distributions and also for real data from a crime register. Moreover testing  $\hat{\tau}$  by the usual t-statistic is a good way to identify Poisson under/overdispersion and the Poisson limit, respectively. Near the Poisson limit ( $\pi \rightarrow 0$ ) we observe high variation in the estimated reporting probability and a bias in the intercept estimate. Near the perfect reporting system limit ( $\pi \rightarrow 1$ ) the variability of the estimates approaches zero. This is due to the fact that the binomial models become deterministic for  $\pi = 1$ . Neubauer & Djuraš (2008) suggested a conditional Poisson model where perfect reporting systems are

still stochastic. Experience with this model and comparing it to the Generalized Poisson model are subject of future research.

### References

- Consul, P.C. (1989). *Generalized Poisson Distributions. Properties and Applications*. New York: Marcel Dekker.
- Neubauer, G., and Djuraš, G. (2008). A beta-Poisson approach as a solution for modelling underreporting in register data. Unpublished technical report.
- Neubauer, G., and Friedl, H. (2006). Modelling sample sizes of frequencies. In: Proceedings of the 21st International Workshop on Statistical Modelling, 3-7 July 2006, Galway, Ireland.

# Models for Fluorescence Signals on SNP Arrays

Ralph C.A. Rippe<sup>1</sup>, Paul H.C. Eilers<sup>1,2</sup> and Jacqueline J. Meulman<sup>1</sup>

<sup>1</sup> Data Theory Group, Leiden University, P.O. Box 9555 2300 RB Leiden, The Netherlands ([rrippe@fsw.leidenuniv.nl](mailto:rrippe@fsw.leidenuniv.nl))

<sup>2</sup> Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

**Abstract:** SNP (single nucleotide polymorphism) arrays are used to determine DNA composition, using two-color fluorescence to measure concentrations. In principle the strength of fluorescence should be the same for all SNPs, but strong systematic patterns occur. We develop and apply models for these patterns.

**Keywords:** DNA; fluorescence; polymorphism; genotype.

## 1 Introduction

Single Nucleotide Polymorphisms (SNPs, pronounced as snips) are positions on (human) chromosomes where frequent mutations occur. Several millions of SNP positions are known (Altshuler et al., 2005). A SNP can be in one of two states, called the alleles, which we indicate here by R and G. Chromosomes form pairs, so the possible ordered combinations are RR, RG, GR and GG. However only the unordered pairs RR, RG and GG are observable. These are called genotypes.

Illumina BeadArray (Shen et al., 2005) is a technology for measuring genotypes by two-color fluorescence; we give an overly simplified description. When G (R) is present, green (red) light is observed. So, when the genotype is GG (RR), no red (green) light should be observed, but only green (red) light of double strength. When the genotype is RG, we should observe both red and green light, having equal strengths. The goal of the measurements is to reliably determine genotypes, using the differences in the light intensities. The volume of the data can be enormous: from 1500 to 500 000 SNP measurements on up to many hundreds of samples.

The description given above is too idealistic. In practice the relative strengths of the red and green signals are different and vary systematically between SNPs and biological samples. Variation between samples is unavoidable; it is caused by differences in the quality of the biological material and DNA extraction. Generally a scaling factor per sample is used for correction. Sys-

tematic differences between SNPs are of interest: 1) to improve the measurements themselves by correction, 2) to improve genotype determination, and 3) to look for patterns with biological relevance.

The data have an interesting structure: three matrices, one for red intensity, one for green intensity and one with a discrete code for the genotype (RR, RG or GG). The rows correspond to SNPs and the columns to arrays (the biological samples). We assume here that genotypes have been estimated by the Illumina software that comes with their SNP arrays, or by another specialized algorithm.

We develop an additive model for the logarithm of the light intensity with parameters for SNPs, arrays and genotypes per color (red or green) and apply it to a relatively small scale study (1487 SNPs, 96 arrays). Still we have over 2000 parameters to be estimated. We avoid large systems by using block relaxation which converges in a few iterations (see section 2).

## 2 Models

Let  $X_c$  be the matrix of fluorescence intensities for color  $c$  ( $c = 1, 2$ ). We expect proportionality: if a SNP (array) shows weak signals, this will be the case for all arrays (SNPs), suggesting a multiplicative model. Also, in principle, dependent on the genotypes RR, RG and GG, the intensities of green (red) will be proportional to 0, 1, 2 (2, 1, 0). Due to noise, cross-talk between the colors and a non-zero background, real signals are more complicated.

A multiplicative model would lead to a three-way structure for each color. The dimensions are SNP, sample and genotype. Because at each SNP only one of the three genotypes can occur, 2/3 of the data are structurally missing, complicating the analysis. The cartoon in Figure 1 illustrates the three-way data structure.

We opt for a linear model for the logarithms of the intensities. Let  $Y_c = \log_{10}(X_c)$  contain the intensities for color  $c$  and let the genotypes be coded by the indicator array  $H = [h_{ijk}]$ . The model is:

$$\hat{y}_{ijc} = \mu_c + \alpha_{ic} + \beta_{jc} + \sum_{k=1}^3 \gamma_{ikc} h_{ijk}, \quad (1)$$

where  $\mu$  is the grand mean,  $\alpha_{ic}$  describes the overall effect of SNP  $i$ ,  $\beta_{jc}$  describes the overall level of sample  $j$  and  $\gamma_{ikc}$  is characteristic for genotype  $k$  of SNP  $i$ . To make the model identifiable, the constraints  $\sum_k \gamma_{ikc} \sum_j h_{ijk} = 0$  are enforced.

Both the number of parameters (over  $10^4$ ) and the number of observations (almost  $3 \cdot 10^5$ ) are large, so we use block relaxation to estimate the parameters. Given all  $\alpha_{ic}$  and  $\gamma_{ikc}$ , it is trivial to compute all  $\beta_{jc}$  by weighted averaging. Given all  $\beta_{jc}$ , the same hold for the computation of all  $\gamma_{ikc}$ , if

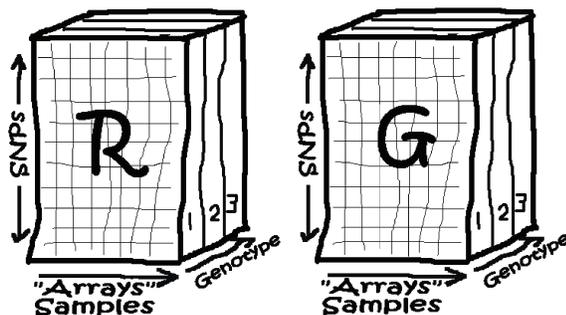


FIGURE 1. Cartoon of the raw data. In each “fiber” of an array only one number is available, because principally there is only one possible genotype at each combination of array and SNP.

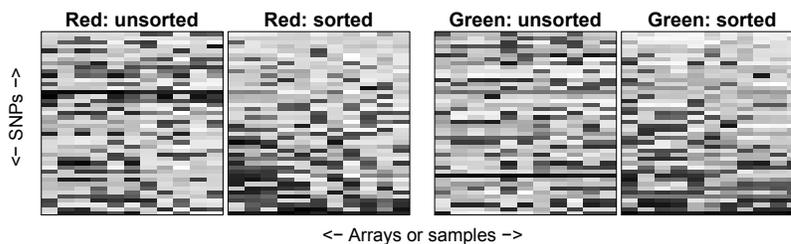


FIGURE 2. Selection of raw data. Left: before sorting. Right, after sorting both matrices by row and column means of red intensities.

the constraints are ignored. It is also trivial to adjust the  $\gamma$ 's and compute all  $\alpha_{ic}$  to conform make the constraints hold. We found that only a handful of iterations were needed to reach convergence.

Figure 2 shows a small selection of the intensities, before and after sorting rows and columns, w.r.t. the average values of the red signal. It appears that after sorting the change in brightness from bottom to top and from left to right is shared by both colors. This suggests to simplify the model by having one vector  $\alpha$  and one vector  $\beta$ . Indeed we found only a small increase of the size of the residuals after this simplification.

### 3 Results

We applied our model on a data set of 1487 SNPs on 96 arrays. The standard deviations of all elements of  $Y$  is approximately 0.6 for both colors.

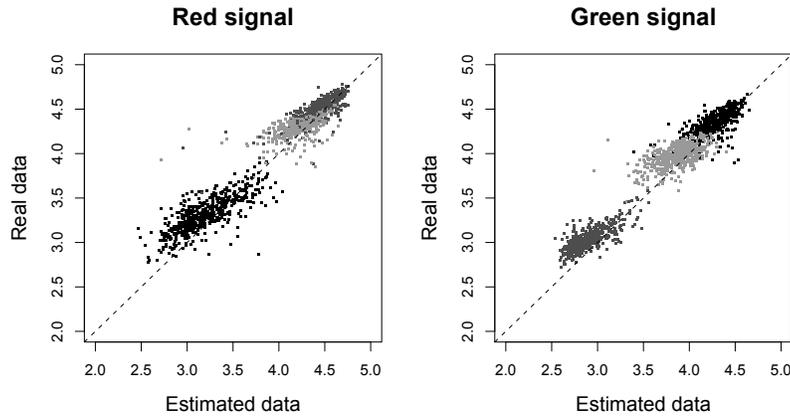


FIGURE 3. Model fit for a typical array.

Those of the residuals are approximately 0.2. The fit of the model is illustrated for a typical array in Figure 3.

According to the idealized description in the introduction the green intensities should have ratios 0, 1, 2 for the genotype RR, RG and GG. For base 10 logarithms this would mean  $-\infty$ , 0 and  $0.3 (= \log_{10} 2)$ . Of course we will never really measure zero, the difference between  $\gamma$  parameters corresponding to RR and RG genotype should be relatively large and 0.3 for the difference between the RG and GG genotypes. This is close to what we actually we find for green, as the left panel of Figure 4. The average difference is 0.34. But for red (see the right panel of the figure) we find the very low value 0.20. Also the patterns in the scatterplots are different. The point clouds for the separate genotypes appear to be tilted. We have no clue yet for an explanation.

## 4 Discussion

We have shown that a linear model for our data makes sense and shows a good fit. The results capture the systematic patterns in variation of signal intensity between SNPs. We also discovered an anomaly in the red intensities where the ratio between RR and RG genotypes is much smaller than the expected value of 2. This ratio also seems to depend on the level of the signal.

One of our other goals is to investigate biological implications. Here we think of relationships with the positions of SNPs on chromosomes and their DNA. We can report no progress on this front yet. We plan to discuss our findings with biologists.

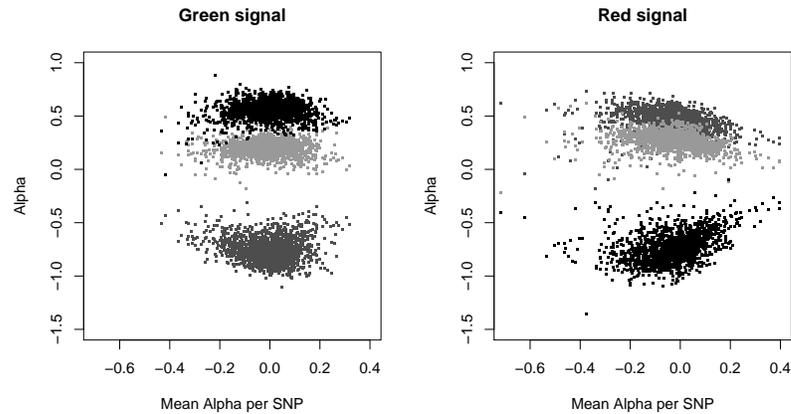


FIGURE 4. Genotype parameters  $\gamma_{ikc}$  plotted against mean SNP effect  $\alpha_{ic}$ , for red and green separately. The genotypes are gray-scale-coded (RR = dark, RG = light, GG = middle).

A third goal is to improve genotyping algorithms. We assumed that error-free genotypes were given, but this is not completely realistic. We estimate genotypes with a specialized algorithm (Eilers et al., 2008) that is illustrated in the left panel of Figure 5. A scatterplot of the logarithms of the ratio of red and green against the log of their average shows three clear clusters representing the three genotypes. A cluster model is fitted, consisting of a mixture of three regression lines, with individual intercept, slope and error variance. From the model follow for each observation (i.e each SNP) the cluster probabilities and hence the genotype (picking the one with the highest probability).

The better the separation between the clusters, the more reliable the genotyping. It is our hope that correction with parameter estimates ( $\alpha$ 's or  $\gamma$ 's) from our model, we can improve the separation. This plan would involve iterations between cluster estimation and model fitting. We have not yet reached that stage, but the right panel of Figure 5 shows correction after fitting the model. It appears that separation between the clusters has improved and that less dots are found in the area between them.

The data set we analyzed here may look large but it is skimpy by modern standard. Nowadays arrays that can analyze over half a million SNPs in one go are being used routinely, increasing the size of the data the size by at least two orders of magnitude. A computation that now take 10 second will take an hour then. Computational efficiency will become an additional goal.

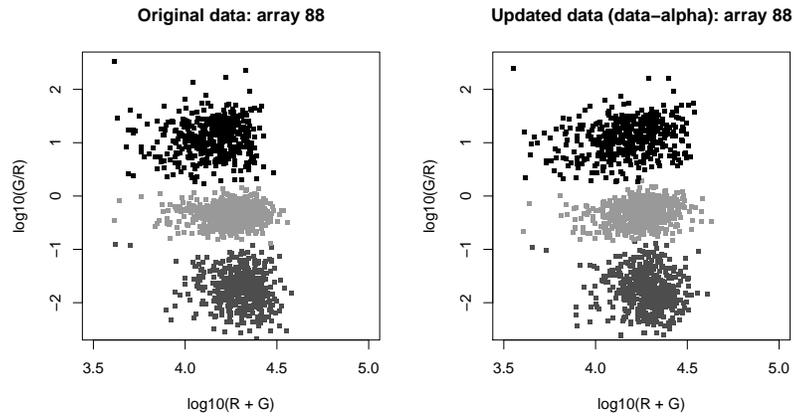


FIGURE 5. Illustration of clustering to estimate genotypes. Left: three clusters formed by original data. Right: after correction with model parameters. See text for explanation.

## References

- Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. et al. (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
- Shen, R., Fan, J. B., Campbell, D., Chang, W., Chen, J., Doucet, D., et al. (2005) High-throughput SNP genotyping on universal bead arrays. *Mutat Res*, **573**, 70–82.
- Eilers, P.H.C. et al. (2008) Reliable Genotyping with Individual SNP arrays. *Manuscript*.

# Quintile stratification on propensity scores and standardization for a T-test

Ismini Savvala<sup>1</sup>, Stef van Buuren

<sup>1</sup> Department of Methodology and Statistics in Behavioral and Social Sciences, Utrecht University, Heidelberglaan 1, Langeveldgebouw, De Uithof, 3584 CS Utrecht, The Netherlands e-mail: I.Savvala@students.uu.nl

**Abstract:** Propensity score is a balancing score and a method used for controlling bias in observational studies. Standardization is an adjustment procedure which estimates what would have been observed given the confounder distributions were the same in the groups being compared. An example estimating the treatment effect by applying propensity quintile stratification and (direct) standardization on the MMCIP data is illustrated.

**Keywords:** propensity score; quintile stratification; standardization; treatment effect.

## 1 Propensity score

Paul R. Rosenbaum and Donald B. Rubin (1983, pp.42) defined that “The propensity score is the ‘coarsest’ function of the covariates that is a balancing score, where a balancing score,  $b(X)$ , is defined as ‘a function of the observed covariates  $X$  such that the conditional distribution of  $X$  given  $b(X)$  is the same for treated ( $Z_i = 1$ ) and control ( $Z_i = 0$ ) units”. The propensity score for subject  $i$  ( $i = 1, 2, \dots, N$ ) with complete data is the conditional probability of being assigned to a particular treatment ( $Z_i = 1$ ) versus a control ( $Z_i = 0$ ) given a vector of observed covariates,  $x_i$ :

$$e(x_i) = \text{pr}(Z_i = 1 | X_i = x_i)$$

The cornerstone result of Rosenbaum and Rubin (1983) is that if the groups of treated-control units have similar propensity score values, then they have the same distribution on the multivariate covariates  $x_i$ , no matter the dimension of  $x_i$ . Consequently, if there is difference between the treatment and control group(s), in a specific value of the  $e(x_i)$ , it is an unbiased estimate of the average treatment effect. This results from the main assumption that the treatment assignment is strongly ignorable given the covariates. It means that the treatment,  $Z$ , and the outcome,  $Y$ , are conditionally independent given the covariates,  $X$  (when  $Y \perp Z | X$ ).

## 2 Quintile Propensity Subclassification

The propensity quintile stratification methodology proposed by Yanovitzky, Zanutto, and Hornik (2005) proposes that the researcher should measure the appropriate confounders. It is essential to determine the initial bias of the confounders. The tests to be performed are two-sample t-test, ANOVA analysis, two-sample test of differences in proportions, logistic regression with outcome the confounder and predictors dummy variables for treatment levels. The measures that will be discussed are the *difference measure* ( $B$ ) and *variance ratio* ( $R$ ):

$$B = \frac{\mu_t - \mu_c}{\sqrt{\frac{\sigma_t^2 + \sigma_c^2}{2}}}, R = \frac{\sigma_t^2}{\sigma_c^2}$$

where  $\mu_i$  and  $\sigma_i^2$  are the mean and the variance in each group. If there is adequate balance in the confounders, propensity score methodology will not provide more than a conventional method of analysis. The estimation of the propensity score, when the treatment exposure is dichotomous, is the logistic regression with treatment as the outcome and all potential covariates as predictors.

Following, the distribution of the estimated propensity score should be checked for adequate overlap by comparing quintiles of estimated propensity scores for both groups. If there are subjects outside the overlapping range, these are discarded. The treatment effect will be estimated in the overlapping range as there are comparable subjects. The remained observations are divided into five equally-sized groups (strata) based on the distribution of the estimated propensity score. The number of strata is defined as five from Cochran's results (1968). He proved that creating five strata at the quintiles of the distribution of the population propensity score removes roughly 90% of the bias in each of the selected covariates.

The researcher should examine whether the confounding distributions are balanced in groups after stratification. Multiple two-way ANOVA with treatment and propensity quintile (categorical variable with 4 degrees of freedom) as factors and each confounder as outcome is performed. In this report, we examine the mean values of difference measure and variance ratio across five strata. If balance is not achieved, then the propensity score should be re-estimated by adding interactions or non-linear functions of imbalanced confounders to the previous propensity model. Thus, the specific steps are repeated, until balance is attained or at least no further improvement can be done.

## 3 Standardization

(Direct) Standardization is a stratified procedure, where the standard population (control group(s)) provides a basis for combining information across

strata, in which confounding distributions of the groups are the same (Statistical Methods, 1980). When the outcome is continuous and the treatment has two levels, the average treatment effect within each propensity quintile is calculated by subtracting the average outcome of treated units from that of control units. An overall estimate is calculated by averaging the differences between the groups across five strata:

$$\hat{\delta} = \sum_{k=1}^5 \frac{n_k}{N} \left( \bar{Y}_{tk} - \bar{Y}_{ck} \right) \quad (1)$$

where  $n_k$  is the number of the overall units in the stratum  $k$ ,  $N$  the total number of units,  $\bar{Y}_{tk}$  and  $\bar{Y}_{ck}$  the average outcome of treated and control units within  $k$  strata. The standard error of the treatment effect is:

$$\hat{s}(\hat{\delta}) = \sqrt{\sum_{k=1}^5 \frac{n_k^2}{N^2} \left( \frac{s_{tk}^2}{n_{tk}} + \frac{s_{ck}^2}{n_{ck}} \right)} \quad (2)$$

#### 4 Application

The MMCIP data is a study that examines the effectiveness of a Multifactor and Multimethod Community Intervention Program to diminish falls among older people by at least 20% in the Netherlands (Wijlhuizen, G.J. et al., 2006). The treatment group was exposed in an intervention program (information and education; training and exercise of older people; volunteers and homecare professionals; environment modifications). The main question is whether the program would decrease fall incidence in and around the house per 1000 persons per year. We chose 21 confounders to participate in the propensity score methodology. In the analysis  $N = 1752$  subjects were included, of which 1122 were from the intervention community.

The most initially imbalanced confounders with their  $B$  and  $R$  values are presented in Table 1. The quintile propensity score methodology was performed. Different models were estimated trying to accomplish better balance in the confounders. The mean values of the two statistics, that were achieved by the best model, show that more balance was attained. In this model, the treatment effect and its standard error were calculated using the formulas (1) and (2).

The average treatment effect  $\hat{\delta}$  was found to be equal to 11.71 and its standard error  $\hat{s}(\hat{\delta})$  equal to 42.64. From these values, we can conclude that the intervention program did not succeed to decrease the falling incidents, as  $CI(\hat{\delta}) = (-71.86, 95.28)$ . The researchers came to the same conclusion, performing an analysis of variance controlling for education, type of the house, dizziness, ability of rising from a chair without help and outdoors physical activity (all  $p > 0.05$ ).

TABLE 1. Difference measure and Variance ratio

	Before stratification		After stratification	
	B	R	$\bar{B}$	$\bar{R}$
	Room	-0.22	1.18	0.001
Income	-0.15	0.96	0.02	1.06
Education	-0.14	1.03	0.01	1.00
Exercise	-0.12	0.96	-0.001	1.18
Chair	0.10	0.78	0.03	0.95
Help out house	0.09	0.87	0.01	0.98
Dress	-0.08	0.73	-0.009	1.02
Family members	-0.07	0.86	0.002	0.99

## 5 Conclusion

The propensity score methodology has limitations. Unmeasured covariates that may bias the treatment effect. The lost information from the discarded treated units results in a lower precision of the estimated treatment effect. There must be adequate amount of imbalance in confounders. In many situations it results to approximately the same conclusions as conventional multivariable methods of analysis.

### References

- Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W., Weisberg, H. I. (1980), Statistical Methods for comparative studies, *John Wiley and Sons*.
- Cochran, W. G. (1968), The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies, *Biometrics*, **24**, 295-313.
- Rosenbaum, P. R., Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55.
- Wijlhuizen, G. J., du Bois, P., van Dommelen, P., Hopman-Rock, M. (2006), Effect evaluation of a multifactor community intervention to reduce falls among older persons. *International Journal of Injury and Safety Promotion*, **00**, 1-9.
- Yanovitzky, I., Zanutto, E. and Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning*, **28**, 209-220.

# A simple method of estimating relative risk from a prevalence study and observations of disease duration

Ramona Scheufele<sup>1</sup> and Ekkehart Dietz<sup>1</sup>

<sup>1</sup> Department of Medical Statistics and Clinical Epidemiology,  
Charite - Universitätsmedizin Berlin, 10098 Berlin, Germany,  
e-mail: ramona.scheufele@charite.de

**Abstract:** Sometimes, prevalence studies have to be used to assess etiologic relationships. To this end, one has to consider estimates of the relative risk from the prevalence study data. Under very restrictive assumptions, an unbiased estimate of the prevalence odds ratio is an unbiased estimate of the incidence density ratio. Otherwise, a prevalence-incidence bias occurs. This paper discusses a computational simple method to adjust for this bias using disease duration data. The method is demonstrated using data from the Saarland cancer registry. Properties of the method are studied by simulation experiments.

**Keywords:** Etiologic study; prevalence-incidence bias; counting process; illness-death model.

## 1 Introduction

An appropriate statistical measure for the exposure effect is the incidence density ratio (IDR), which is estimated from a cohort study, usually. Sometimes, however, it is desirable to get an estimate of relative risk of a disease from cross-sectional studies, because there is not yet a cohort study to the disease and the study factor of interest. Because prevalence studies are less time-consuming and less expensive than cohort studies, they would be clearly preferable, if there was a possibility to obtain unbiased estimates of relative risk from such studies as well. In the literature, you can find several suggestions to use duration of the disease data. Most of them are based on a steady-state condition, which means constant incidence rate, constant recovery/death rate, and constant prevalence over time. In this presentation, a related method is considered. It is a combination of a logistic regression analysis of the prevalence study and an analysis of disease duration data based on an accelerated survival time model.

## 2 The prevalence-incidence bias adjustment

### 2.1 PI-bias adjustment in homogeneous populations

The prevalence-incidence relationship at steady-state is given by

$$ID = PO/D \tag{1}$$

where  $ID$ ,  $PO$ , and  $D$  denote the incidence density, the prevalence odds, and the mean disease duration, respectively. Derivations of the equation (1) can be found in Keiding (1991), e.g..

From this equation, it follows

$$IDR = \frac{ID_e}{ID_{ne}} = \frac{PO_e}{PO_{ne}} * \frac{D_{ne}}{D_e} = POR * \frac{D_{ne}}{D_e}, \tag{2}$$

where  $POR$  denotes the prevalence odds ratio and the indexes  $e$  and  $ne$  indicate, if the respective term refers to the exposed or to the not exposed subpopulation. From (2) it follows, that the estimate of prevalence odds ratio  $POR$  from a cross-sectional study is an estimate of the incidence density ratio  $IDR$ , if the disease duration does not depend on exposure, so that  $D_{ne} = D_e$ . Otherwise, an estimate of  $IDR$  can be obtained from  $POR$  by the correction factor  $\frac{D_{ne}}{D_e}$ . So, if  $D_{ne}$  and  $D_e$  or at least their ratio was known, there would be the potential to estimate  $IDR$  from a cross-sectional study.

Let  $\beta = \log(POR)$  and  $\alpha = \log(\frac{D_{ne}}{D_e})$ . Furthermore, let  $\hat{\beta}$  and  $\hat{\alpha}$  be independent estimates of  $\beta$  and  $\alpha$ , respectively, having standard errors  $s_\beta$  and  $s_\alpha$ , respectively. Then, an estimate of  $\log(IDR)$  can be obtained by

$$\log(\widehat{IDR}) = \hat{\beta} - \hat{\alpha} \tag{3}$$

with standard error

$$s_{\log(IDR)} = \sqrt{s_\beta^2 + s_\alpha^2}$$

In order to get estimates of the correction term  $\alpha$ , observations of the disease duration are required. These could be obtained by a follow up of incident disease cases of both the exposed population and the not exposed population.

### 2.2 PI-bias adjustment in heterogeneous populations

If the population considered is heterogeneous with respect to hazard rate and confounding occurs, the method of PI-bias adjustment described in the previous section fails to produce an unbiased estimates of  $IDR$ . A computational efficient way to obtain confounder adjusted estimates of  $IDR$  is by fitting a certain logistic regression model. Let

$$\text{Log}(PO(x, \vec{Z})) = \beta_0 + \beta_1 x + \vec{B}^t \vec{Z} \tag{4}$$

be the systematic part of a usual logistic model for data of a prevalence study, where  $PO(x, \vec{Z})$ ,  $x$ , and  $\vec{Z}$  denote the conditional prevalence odds, a binary study factor, and a vector of covariables, respectively. Now, this model is slightly modified by taking the logarithm of the conditional expected disease duration  $\log(T_{x,\vec{Z}})$  as offset into the linear predictor:

$$\text{Log}(PO(x, \vec{Z})) = \beta_0^* + \beta_1^* x + \vec{B}^{*t} \vec{Z} + \log(T_{x,\vec{Z}}) \quad (5)$$

The coefficient  $\beta_1^*$  of the study factor  $x$  of this model is just the PI-bias adjusted estimate of  $\log(\text{IDR})$ .

In order to apply the model above one has to have estimates of the expected disease duration for each line-up of the study factor and the covariables. These can be obtained by a survival analysis of a cohort study of incident disease cases, where recovery or death is considered as outcome. At the statistical analysis of a prevalence study with observational data of the disease duration by the method described above, one has to take into account the standard errors of the estimates of the disease duration, obtained by the survival analysis. Therefore, standard software for logistic regression has to be modified to allow for errors in the offset variable, so that valid estimates of the confidence interval become obtainable. To sidestep this problem, an alternative approach is considered now. We will focus on accelerated failure time models. That is, the following model is fitted to the disease duration data of the case cohort study:

$$\log(t) = \alpha_0 + \alpha_1 x + \vec{A}^t \vec{Z} + \sigma e, \quad e \sim \phi(t) \quad (6)$$

where  $\phi(t)$  is a standard distribution. Because some of the components of  $\vec{A}$  can be set to zero, the same vector of covariables  $\vec{Z}$  can be taken for the survival model (6) and the logistic regression model (4) without loss of generality. If model (6) holds, the expected disease duration given the values of the study factor and the covariables is

$$T_{0,\vec{Z}} = e^{\alpha_0 + \vec{A}^t \vec{Z}} * \sigma \int_{-\infty}^{\infty} e^u \phi(u) du$$

and

$$T_{1,\vec{Z}} = e^{\alpha_0 + \alpha_1 x + \vec{A}^t \vec{Z}} * \sigma \int_{-\infty}^{\infty} e^u \phi(u) du,$$

respectively. Thus, the coefficient  $\alpha_1$  of the study factor is just the logarithm of the common PI-bias correction factor:

$$\alpha_1 = \log\left(\frac{T_{1,\vec{Z}}}{T_{0,\vec{Z}}}\right)$$

Let  $\hat{\beta}_1$  and  $\hat{\alpha}_1$  be estimates of the regression coefficients of the study factor  $x$  of the usual logistic regression model for the prevalence and the accelerated

failure time model (6) for the disease duration, respectively. Furthermore, let  $var(\hat{\beta}_1)$  and  $var(\hat{\alpha}_1)$  denote their respective variance estimates. Then, the estimation of the logarithm of the confounder adjusted IDR and the according variance can be displayed as

$$\log(\widehat{IDR}) = \hat{\beta}_1 - \hat{\alpha}_1 \tag{7}$$

and

$$var(\log(\widehat{IDR})) = var(\hat{\beta}_1) + var(\hat{\alpha}_1),$$

respectively. Thus, we have got a generalization of (3).

### 3 An application to registry data

Since the Saarland cancer registry is considered to be the most complete cancer registry in Germany for a long time, it is possible to obtain good annual estimates of both the incidence and the prevalence. Applying the method described above to prevalence allows us to check the validity of the results of the method described. Gender is considered as the study factor and the 5-year-age class was the only covariable available. To estimate the age adjusted relative risk, a poisson regression model was fitted to the age class specific annual incidence using gender and a polynomial of the age-class variable of degree 5 as explanatory variable and the logarithm of the respective age class and gender specific population size as offset. The age adjusted prevalence odds ratio was obtained by fitting a logit model to the age-class and gender specific prevalences at the December 31 in the year considered using the same covariables and using the respective population sizes as Binomial denominator. To compute the PI-correction term, a weibull model was fitted both to the disease duration of incident cases and to the cancer disease duration to date of the prevalent cases of the year considered. Analyses had been done for each of the years from 1981 to 2001 for several cancer sites and for all cancer sites together. The results of the latter analysis is shown in Figure 1. From the results of all analyses, it follows: (1) There is a PI-bias when using cancer prevalence data only. The prevalence odds ratio is larger than the empirical risk ratio for the most cancer sites. This was to be expected, because the mean survival times of cancer cases is mostly larger for females than for males. (2) The PI-adjustment by the method above always goes in the right direction and leads to a better fit of the relative risks, although it is well known, that the steady state assumption is not fulfilled for this data. (3) The PI-adjustment is not perfect and it could not be improved by taking disease duration of the incident cases of the year instead of the disease duration to date of prevalent cases for the estimation of the PI-correction. Actually, in most cases, the deviances of the PI-bias adjusted IDR estimates from the empirical IDR-estimates are smaller if DTD is used.

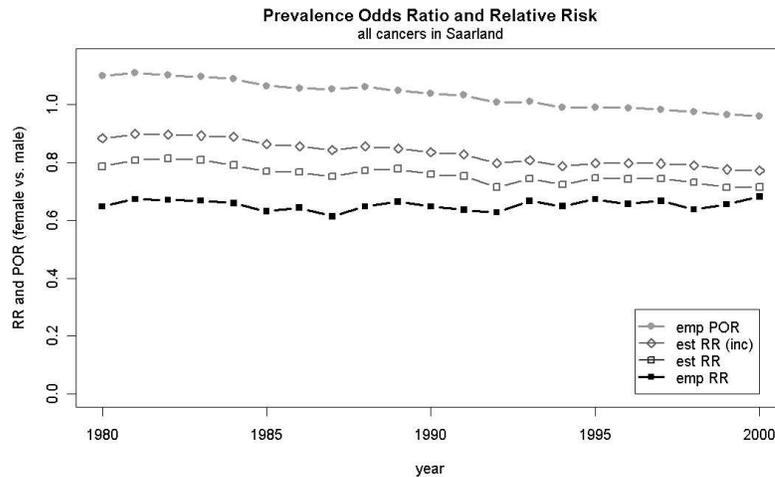


FIGURE 1. Prevalence odds ratio (emp POR), incidence ratio (emp RR), estimation of incidence ratio based on the prevalence odds ratio and observations of the disease duration of the incident cases (est.RR(inc)), and estimation of incidence ratio based on the prevalence odds ratio and observations of the disease duration to date of the prevalent cases (est.RR)

#### 4 Discussion

The remaining deviance of "PI-bias-adjusted estimates" and empirical estimates of cancer incidence ratio can be explained by the fact, that the steady state assumption is not fulfilled. Actually, the PI-bias can be decomposed into one component which is due to unequal distribution of the disease duration and one component which is due to lack of steady state. This became apparent in simulation experiments of some typical situations of instability, which can be characterized by a change of risk and/or a change of mean disease duration in an exposed subpopulation. Figure 2 shows the results of the simulation of situations, where both the risk and the mean disease duration is decreased and where both the risk and the mean disease duration is increased, respectively, at a certain point of time. The results of simulation experiments of instability can be summarized as follows: (1) Also at non-steady-state, both the use of observations of the disease duration of incident cases and the use of observations of DTD of prevalent cases leads to a (incomplete) bias correction, mostly. (2) The bias is smaller, when using observations of DTD, mostly. (3) The bias due to non-steady state may be rather large. Therefore, further research should

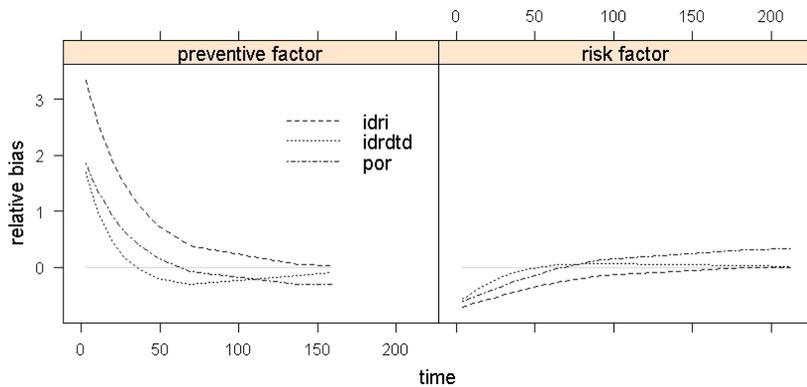


FIGURE 2. Relative bias before and at steady state, if IDR is estimated by the prevalence odds ratio (por), by the prevalence odds ratio and the disease duration of incident cases (idri), and by the prevalence odds ratio and the disease duration to date of prevalent cases (idrtd), respectively.

be done to find a non-steady-state bias correction based on data being available in a prevalence study and in a case-cohort study

**Acknowledgments:** This research is supported by the German Research Foundation and by the Hypatia Program - Promotion of Women in Science at TFH Berlin

**References**

Begg, C.B., and Gray, R.J. (1987). Methodology for case-control studies with prevalent cases. *Biometrika*, **71**, 91-195.

Cox, D.R., and Miller, H.D. (1965). *The Theory of Stochastic Processes*. London: Chapman & Hall.

Freeman, J. and Hutchinson, G.B. (1986). Prevalence, incidence and duration. *American Journal of Epidemiology*, **124**, 134-149.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society A*, **154**, 371-412.

# Optimal expectile smoothing

Sabine K. Schnabel<sup>1</sup> and Paul H.C. Eilers<sup>2</sup>

<sup>1</sup> Max Planck Institute for Demographic Research, Konrad-Zuse-Str.1, 18057 Rostock, Germany; schnabel@demogr.mpg.de (communicating author)

<sup>2</sup> Utrecht University, Faculty of Social and Behavioural Sciences, Postbus 80.140, 3508 TC Utrecht, The Netherlands; p.h.c.eilers@uu.nl

**Abstract:** A combination of asymmetric least squares and P-splines is an attractive alternative to quantile smoothing. We propose two algorithms for optimal smoothing and show results for an experimental data set and compare them to quantile splines.

**Keywords:** Asymmetric least squares, cross-validation, P-splines, quantiles.

## 1 Introduction

When we study a scatterplot of observed (time series) data, we might not only be interested in the trend, but also in the spread around it. Quantile smoothing (QS) (Koenker et al., 1994) is an effective and popular tool for this purpose. QS is based on asymmetrically weighting the sum of absolute values of residuals. We believe we can do better by asymmetrically weighting the sum of squares of residuals, leading to “expectile smoothing”. As in any smoothing problem, it is important to have an automatic method to determine a good value of the smoothing parameter - in our case the weight of the penalty of P-splines. We present two methods: one is based on asymmetric cross-validation, the other adapts Schall’s EM algorithm for estimating variance components. Results for simulated data as well as for two empirical examples are presented and compared to those of the COBS package for QS (He and Ng, 1999).

## 2 Asymmetric smoothing

Asymmetric least squares estimation (ALS) seeks to minimize the following goal function for a range of values  $p$ ,  $0 < p < 1$ :

$$S = \sum_i w_i(p)(y_i - \mu_i(p))^2 \quad (1)$$

with weights 
$$w_i(p) = \begin{cases} p & \text{if } y_i > \mu_i(p) \\ 1 - p & \text{if } y_i \leq \mu_i(p) \end{cases} \quad (2)$$

where  $y_i$  is the response variable and  $\mu_i(p)$  is the estimated value according to a statistical model. The obtained functions are called  $p$ -expectiles as introduced by Newey and Powell (1987). It is extremely easy to fit any ALS model: simply iterate between weighted regression and re-compute the weights. The goal function is convex, so a unique minimum is guaranteed. We combine ALS with P-splines (Eilers and Marx, 1996):  $\mu_i = \sum_j b_{ij} a_j$  where  $B = [b_{ij}]$  is the matrix of B-spline basis functions and  $a$  the coefficient vector. So we are seeking to minimize the penalized ALS function:

$$S^* = (y - Ba)^T W (y - Ba) + \lambda |D_d a|^2 \quad (3)$$

with respect to  $a$ .  $D_d$  is a matrix that forms  $d$ -th order differences of  $a$ . For simplicity we suppress the dependence on  $p$  in the notation. The model parameters are computed iteratively according to

$$\hat{a} = (B^T \tilde{W} B + P)^{-1} B^T \tilde{W} y \quad (4)$$

with current weights in  $\tilde{W} = \text{diag}(\tilde{w})$  and  $P = \lambda D_d^T D_d$ . To optimize  $\lambda$ , we can use leave-one-out cross-validation. The idea is to remove each observation ( $y_i$ ) in turn, predict it ( $\mu_{-i}$ ) from the remaining ones, and measure prediction performance by  $CV = \sum (y_i - \mu_{-i})^2$ . The asymmetric variant introduces the weights according to (2):

$$ACV = \frac{1}{n} \sum_i w_i (y_i - \mu_{-i})^2 = \frac{1}{n} \sum_i w_i (y_i - \mu_i)^2 / (1 - h_{ii})^2, \quad (5)$$

where we have used the fact that  $y_i - \mu_{-i} = (y_i - \mu_i) / (1 - h_{ii})$ , with the hat matrix  $H$  defined as

$$H = W^{1/2} B (B^T W B + P)^{-1} B^T W^{1/2}. \quad (6)$$

We search for the minimum of ACV for a range of values of  $\lambda$  on a grid (linear for  $\log \lambda$ ). This is done for every  $p$  separately. In the ACV we assume that the weight vector is invariant to single missing observations. Simulation studies show that this assumption holds for more than 99 % of all considered cases.

Alternatively, using the formal equivalence between penalized least squares smoothing and mixed models (Pawitan, 2001; Lee et al., 2006), we have that  $\lambda = \sigma^2 / \tau^2$ , with  $\sigma^2$  the variance of the errors and  $\tau^2$  the variance of the contrasts  $D_d a$ . We estimate these variances by

$$\hat{\sigma}^2 = |y - \mu|^2 / (n - ED), \quad \hat{\tau}^2 = |D \hat{a}|^2 / ED \quad (7)$$

where  $n$  is the sample size and  $ED = \text{tr}(H)$  the effective model dimension. We iterate between smoothing asymmetrically with  $\lambda$ , estimating variances (and a new  $\lambda$ ) until convergence. This is a variant of Schall's (1991) method for generalized linear mixed models.

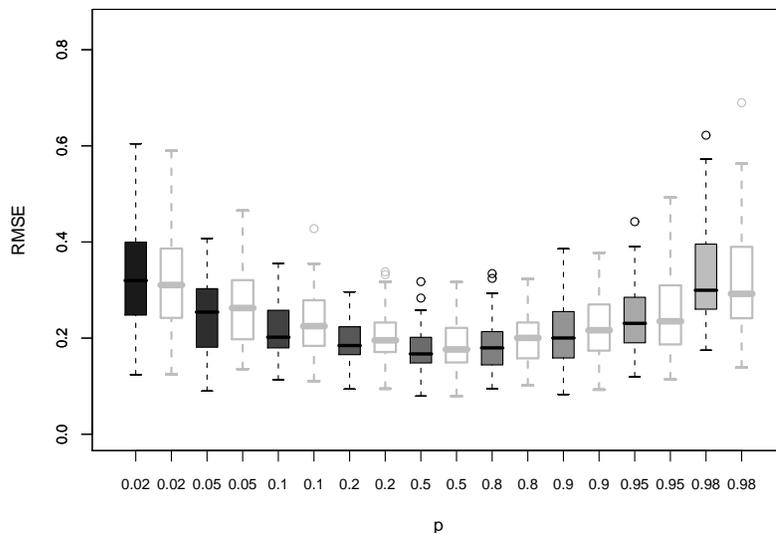


FIGURE 1. Comparison in terms of root mean squared error (RSME) of smoothed curves vs the theoretical curves with selected smoothing parameter according to ACV (blank) and the adapted Schall algorithm (shaded). Simulated data includes a normal error term.

### 3 Simulations and applications

We perform simulations to compare estimated expectile curves with their theoretical values, for different error distributions. The data were of the form  $y = 1.5x^2 + 4 + \cos(3x) + \varepsilon$  for the response  $y$  and  $x$  uniformly distributed on  $[0, 3]$ . With sample size  $n = 200$  we replicated the set-up 50 times. These studies show good results. We also compared the performance of the two smoothing methods in terms of root mean squared error (RSME) as depicted in Figure 1 for data with normal error. Both the amount of smoothing chosen by the asymmetric cross-validation as well as the one iteratively determined by Schall's algorithm lead to a good fit of the data. We use the well-known and widely used light detection and ranging (LIDAR) data set (as described in (Ruppert et al., 2003)) as an example. This textbook data consists of 221 observations of the log ratio of received light from two laser sources vs the distance travelled before the light is reflected back into its source. Figure 3 on page 391 shows the results for

ACV and Schall's method, which are very similar. The expectile curves conform well to our intuitive impression of the data: a smooth downward trend with increasing spread. For comparison we also computed quantile smoothing splines with the package COBS for R, using automatic smoothing. The quantile curves look under-smoothed and, in contrast to the expectile curves, they cross in many places.

Our second example concerns human growth. The Fourth Dutch Growth study collected cross-sectional data on height, weight and head circumference of Dutch children (van Buuren and Frederiks, 2001). We analyze the relation between age and height of about 7000 Dutch boys. The data are available via the first author's web site (van Buuren, 2007).

At low ages the curves are very steep. To follow the local changes there, a relatively small penalty is needed. The opposite is true at high ages. The optimal value of  $\lambda$ , according to ACV or Schall's algorithm, strike a middle ground. It is hard to judge visually whether the steep parts of the expectile curves are true to the data, but in the flat parts we see unrealistic ripples. One possible, but complicated, way out is to combine ALS with ideas for from the literature for non-parametric estimation of curves with variable smoothness. After some experimenting with different transformations, we found an easier solution: the square root of age stretches the scale at low ages and gives pleasing results as is shown in Figure 2.

## 4 Conclusion and outlook

We proposed expectile smoothing with P-splines as a computationally attractive alternative to quantile smoothing and presented two algorithms for optimal smoothing. Good results were obtained in simulation as well as with empirical data as seen from the two examples. In ongoing research we are extending the model to multi-dimensional contexts and mixed models for life expectancy data.

## References

- van Buuren, S. (2007). Worm plot in quantile regression: Code and data [Internet]. Leiden. Available from: <<http://www.stefvanbuuren.nl/wormplot/dutchdata.boys.sdd.txt>>.
- van Buuren, S., and Frederiks, A.M. (2001). Worm plot: A simple diagnostic device for modeling growth reference curves. *Statistics in Medicine*, **20**, 1259-1277.
- Eilers, P.H.C., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Sciences*, **11**, 89-121.
- Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **4**, 673-680.

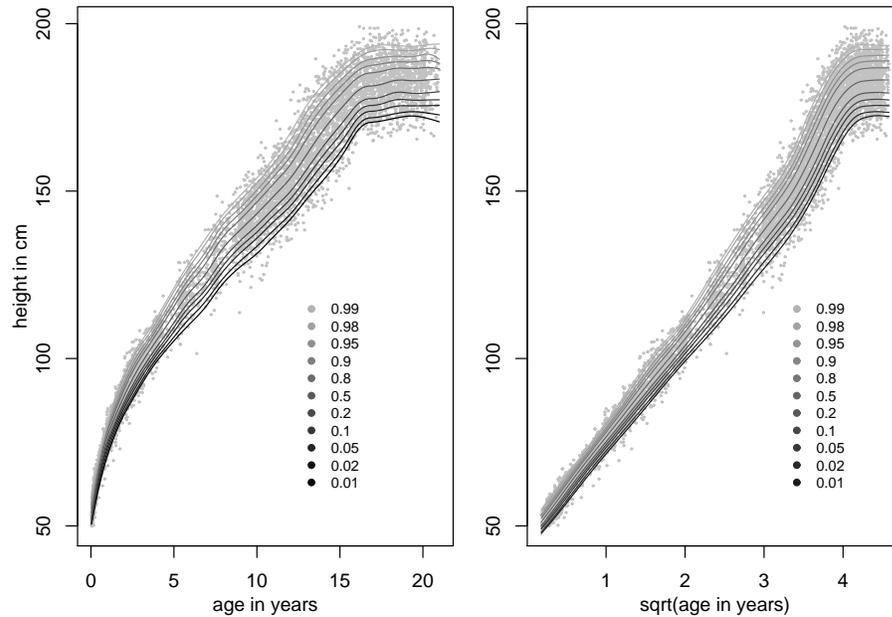


FIGURE 2. Height (in cm) against age (in years) of Dutch boys from the Fourth Dutch Growth Study. Estimated expectiles for  $p=0.01, 0.02, 0.05, 0.10, 0.20, 0.50, 0.80, 0.90, 0.95, 0.98, 0.99$ . Left graph: original data, right graph: transformation of the independent variable.

- He, X., and Ng, P. (1999). COBS: Qualitatively constrained smoothing via linear programming. *Computational Statistics*, **14**, 317-337.
- Lee, Y., Nelder, J.A., and Pawitan, Y. (2006). *Generalized linear models with random effects – unified analysis via H-likelihood*. London: Chapman & Hall.
- Newey, W.K., and Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-847.
- Pawitan, Y. (2001) *In all likelihood*. Oxford: Oxford University Press.
- Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric regression*. New York: Cambridge University Press.
- Schall, R. (1991). Estimation in Generalized Linear Models with random effects. *Biometrika*, **78**, 719-727.

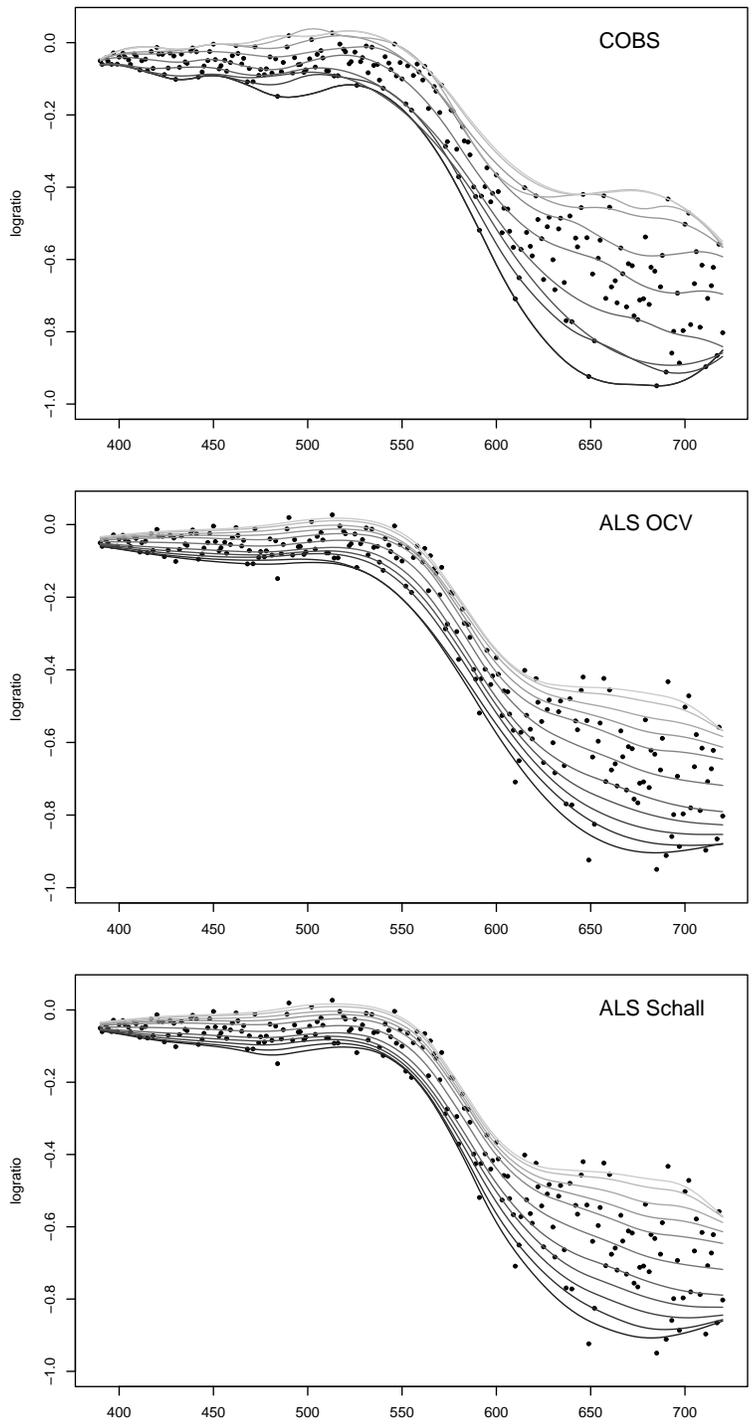


FIGURE 3. Analysis of the LIDAR data set with COBS and ALS,  $p=0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95, 0.98, 0.99$ . Black is indicating the low  $p$ -expectiles, light grey is indicating high  $p$ -expectiles.

# Comparing crossover designs in average bioequivalence studies

Arminda Lucia Siqueira<sup>1</sup> and Daniela Monteiro Braga<sup>1</sup>

<sup>1</sup> Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil. E-mail: arminda@est.ufmg.br

**Abstract:** Bioequivalence studies are necessary for comparing generic drugs to a reference product that has been already placed in the market. In this paper, several crossover designs used in bioequivalence studies are compared in terms of issues related to design and data analysis. Analytical form-based and simulation-based results are presented. The main objective is to assess which design is to be preferred depending on features of the drugs under comparison.

**Keywords:** Bioequivalence study; crossover; high order crossover design.

## 1 Introduction

In order to commercialize a generic drug product, it is necessary to conduct a bioequivalence study. Pharmacokinetic measures must be analyzed, especially the area under the plasma or blood concentration-time curve (AUC) and the maximum concentration (C<sub>max</sub>) of the formulation. Generally, the study is conducted with healthy volunteers following the crossover design. The 2x2 crossover design is the most widely one applied in bioequivalence studies, although higher order crossover designs are sometimes recommended because they allow for estimation of inter- and intra-subject variabilities, as well as subject-by-formulation interaction. Such designs are also used for high variability (intrasubject coefficient of variation larger than 30%), for populational and individual bioequivalence studies and for comparing more than two formulations.

One question of practical importance is how to select a crossover design, taking into consideration the number of recruited volunteers, the number of blood samples taken from each volunteer, the duration of the study, as well as the power of the test used to compare the formulations.

This paper focuses on average bioequivalence studies comparing several crossover designs in order to verify the effect of increasing the number of sequences and/or period, and to evaluate practical issues, such as the effect of variance misspecification. The aim is to determine which design is most appropriate in several situations. We present comparative results obtained from both analytical formulas and a simulation study conducted.

## 2 Methodology

The model commonly used to describe pharmacokinetic data ( $Y$ ), which generally follows the log-normal distribution, obtained through a study with  $L$  formulations conducted under a crossover design with  $J$  periods,  $K$  sequences and  $n_k$  participants in each sequence is given by  $Y_{ijkl} = \mu_l + G_{kl} + S_{ikl} + P_j + F_{(j,k)} + C_{(j-1,k)} + e_{ijkl}$ ,  $i = 1, 2, \dots, n_k$ ;  $j = 1, 2, \dots, J$ ;  $k = 1, 2, \dots, K$ ;  $l = 1, 2, \dots, L$ . In this model, involving fixed and random terms,  $\mu_l$  is the general mean for the  $l$ -th formulation,  $G_{kl}$  is the fixed effect for the  $k$ -th sequence and  $l$ -th formulation ( $G_{kl} = 0$  for the 2x2 crossover design),  $S_{ikl}$  is the random effect related to subject,  $P_j$  and  $F_{(j,k)}$  are the fixed period and formulation effects with the constraint that  $\sum F_{(j,k)} = 0$ ,  $C_{(j-1,k)}$  is the residual fixed effect (*carry-over*) such that  $C_{(0,k)} = 0$  and  $\sum C_{(j-1,k)} = 0$ , and the last term  $e_{ijkl}$  is the intra-subject random error related to the outcome  $Y_{ijkl}$ . The usual assumptions are:  $\{S_{ikl}\} \sim N(0, \sigma_S^2)$ ,  $\{e_{ijkl}\} \sim N(0, \sigma_e^2)$  and  $\{S_{ikl}\}$  and  $\{e_{ijkl}\}$  are mutually independent.

The calculation of the number of volunteers ( $N$ ) and the number of blood samples from each volunteer is important in terms of both cost and ethical point of view. The sample size ( $N$ ) and the power calculations are described in Chen et al. (1997), Chow and Wang (2002), Qu and Zheng (2003) and Siqueira et al. (2005), among others.

Further details for bioequivalence studies can be found for instance in Chow and Liu (2000), Hauschke et al. (2007), Patterson and Jones (2006), and on the Regulatory Agencies web sites.

## 3 Results and discussion

For the comparison of the crossover designs specific to two formulations, including the standard 2x2, and higher-order designs (4x2, 2x4, 2x3 and 4x4), we varied the parameters involved in the calculations, especially the intra-subject coefficient of variation ( $CV$ ) and the expected difference between the means ( $\theta_\gamma$ ). We considered situations with or without carry-over effect. The power and significance level were set at 80% and 5%, respectively.

### 3.1 Formula-based comparison

In general, the smallest total number of volunteers is required by the 2x4, with or without carry-over effect, followed by the 4x4 and 2x3 designs. For fixed power, the design requiring the largest number of volunteers is the 4x2, with or without carry-over effect, followed by the 2x2 (the 4x2 requires twice the number of the 2x2). Compared to the standard 2x2, the difference in terms of number of volunteers is small for  $CV$  less than 0.14, but it increases as the variability and the difference between means ( $\theta_\gamma$ ) increases.

For fixed  $N$ , 4x2 designs (with or without carry-over effect) attain the smallest power, and 4x4 and 2x4 designs (with power curves practically overlaying) have the largest power. The standard 2x2 achieves power larger than that reached in 4x2 and 2x3 designs, but smaller than that found in 2x4 and 4x4 designs. The larger the variability (expressed by  $CV$ ), the larger the difference in powers between the designs. Regarding 4x2 and 2x4 designs for which the carry-over effect affects calculations, the power of the test is always smaller when carry-over effect is taken into consideration. As a result, the number of volunteers needed in a study in which the carry-over effect has been detected is larger.

A decrease in the number of volunteers does not always imply cost reduction. For instance, for a total of 10 blood samples per volunteer in each period, in general the 2x2 crossover design requires the smallest number of measurements. For the selected results presented in Table 1, the numbers of volunteers and blood samples for 4x2 design are twice those of 2x2 design. However, different designs may yield the same number of volunteers and blood samples. In terms of number of blood samples, the designs come in the following increasing order: 2x2, 2x4, 4x4, 2x4 with carry-over effect, 2x3, 4x2, and 4x2 with carry-over effect. In conclusion, 4x2 design requires the largest number of blood samples.

TABLE 1. Number of volunteers and blood samples needed for several crossover designs ( $\theta_\gamma = 0.00$  and power = 80%)

$CV$	Number	Crossover design						
		2x2	4x2	4x2*	2x3	2x4	2x4*	4x4
0.24	Volunteers	22	44	80	16	12	12	12
	Blood samples	440	880	1600	480	480	480	480
0.34	Volunteers	40	80	152	30	20	22	20
	Blood samples	800	1600	3040	900	800	880	800

\* with carry-over effect

### 3.2 Simulation-based comparison

The computational implementation was performed in C language using 10,000 repetitions. The main results are summarized next.

Both 2x4 and 4x4 designs attain the largest percentage of bioequivalence among all cases, followed by 2x3 and 2x2 designs; and 4x2 designs yield the worst results. For high intra-subject variability formulations, 2x4 and 4x4 designs are the most recommended.

In assessing the impact of assuming a wrong  $CV$ , the deficit and excess in number of participants depend on design and actual  $CV$ . The larger the  $CV$  used, the larger the percentage of bioequivalence conclusions. In some situations, one should consider increasing the number of volunteers in order not to compromise the study.

## 4 Conclusions

For choosing the best design in a bioequivalence study, several factors should be considered, such as power, number of volunteers, study cost and duration, and variability. For low CV ( $< 30\%$ ), the 2x2 crossover design, required for most of the regulatory agencies, can be adopted. For high variability products (CV  $> 30\%$ ), 2x4 and 4x4 designs are recommended because they attain the highest power. It is necessary to verify the feasibility of conducting a study with three or four periods. In spite of requiring smaller number of volunteers compared to the standard 2x2, the cost could be higher and dropouts are very likely. It seems that the 2x4 design is more interesting than the 4x4, because two sequences are eliminated and both designs yield very close results. Finally, it is recommended a careful choice of a crossover design for the success of a bioequivalence study.

**Acknowledgments:** This work was sponsored in part by FAPEMIG, a Brazilian Research Agency.

## References

- Chen K-W, Chow S-C and Li G. (1997). A note on sample size determination for bioequivalence studies with higher-order crossover designs. *Journal of Pharmacokinetics and Biopharmaceutical*, **25**, 753-765.
- Chow, S.C., and Liu, J.P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies*. New York: Marcel Dekker.
- Chow S-C, Wang H. (2002). On statistical power for average bioequivalence testing under replicated crossover designs. *Journal of Biopharmaceutical Statistics*, **12**, 265-309.
- Hauschke, D., Steinijans, V., and Pigeot, I. (2007). *Bioequivalence Studies in Drug Development*. Chicester: John Wiley.
- Patterson, S., and Jones, B. (2006). *Bioequivalence and Statistics in Clinical Pharmacology*. London: Chapman & Hall.
- Qu RP, Zheng H. (2003). Sample size calculation for bioequivalence studies with high-order crossover designs. *Controlled Clinical Trials*, **24**, 436-439.
- Siqueira, A. L., Whitehead, A., Todd, S., and Lucini, M. M. (2005). Comparison of sample size formulae for 2 x 2 cross-over designs applied to bioequivalence studies. *Pharmaceutical Statistics*, **4(4)**, 233-243.

# Comparison of Self-Organizing Maps, Mixture, K-means and Hybrid Approaches to Risk Classification of Passive Railway Crossings

Julie Sleep<sup>1</sup> and Irene Hudson<sup>1</sup>

<sup>1</sup> School of Mathematics and Statistics, University of South Australia, Adelaide, Australia (julie.sleep@unisa.edu.au)

**Abstract:** We create factor constructs for the physical characteristics of 864 passive railway crossings in South Australia. Crossings are then classified as dangerous or otherwise by means of self-organizing maps, K-means, mixture models and combinations of the latter. Results are compared using historical accident data. The self-organizing map with mixtures approach is found to be optimal in prediction of dangerous crossings. Results show that there exists two types of dangerous crossings. This has significant implications for prioritisation of crossing upgrades.

**Keywords:** Self-Organizing Maps; Railway Crossings; Mixtures; K-means; Predictive Accuracy.

## 1 Introduction

Railway level crossing accidents are fortunately quite rare, however due to their severe consequences they are an important public safety issue. Railway level crossings fall into two broad classifications - active and passive. Active crossings are those with protections, such as lights or boom gates installed, which inform the road user of the presence of a train. Passive crossings are those that have no such protection systems, only static warning signs. This paper considers only the latter, passive crossings. Previous work by Sleep *et. al.* [in prep.] involved use of expert opinion to construct a Bayesian Belief Network (Pearl, 1988) to model accident risk at crossings with a focus on human factors. The data driven approach presented in this paper focuses on physical characteristics of the crossings, for which, unlike the human side, there are reliable data available.

The aim of this work was to classify the crossings as dangerous or non-dangerous based on new factor constructs of crossings' physical characteristics and then to analyze this classification according to the accident and incident (near miss) history for each crossing. An additional aim of

the work was a comparison of different statistical approaches, in terms of accident prediction.

## 2 Data

The data used in this study were collected by the Department of Transport, Energy and Infrastructure (DTEI) on SA level crossings in 2007; and contained details of 864 passive railway crossings in South Australia. 28 physical characteristics were extracted. Motor vehicle accident and incident counts over a period of ten years (1997-2007) were then matched to these crossings, as a means to assess the validity of our classifications. Approximately 10% of crossings had a history of accidents or incidents.

## 3 Methods

A Factor Analysis (Hair, 1998) of the data set resulted in 10 factors accounting for 63% of the variance in the data. These could all be meaningfully interpreted, as follows: Factor 1 involved variables associated with low-volume crossings and low speed trains; F2 busy crossings; F3 crossing condition; F4 crossing visibility; F5 number of tracks; F6 short stacking and queueing; F7 fatigue and heavy vehicles; F8 train visibility; F9 number of road lanes; F10 road speed and warning sign location.

### 3.1 Self Organising Maps

Self organizing maps (Kohonen, 1997) were run on the crossing data. Each crossing,  $i$ , is represented by a vector of its ‘characteristic’ factor constructs  $x_i = \{c_{i1}, \dots, c_{iN}\}$  where, in our case,  $i = 1, \dots, 864$  and  $N = 10$ . We use the batch learning algorithm from the SOM toolbox (Vesanto et. al., 2000). The self-organizing map is made up of map units, each defined by a map vector  $m_j$ . For several iterations each data vector  $x_i$  is placed on the map according to which map vector  $m_j$  is its best matching unit,  $\min_j \{\|x_i - m_j\|\}$ . Then the map vectors  $m_j$  are recalculated, and so on. Figure 1 gives a SOM with 9 map units for the passive crossing factor scores. The map units show how many crossings had a history of incidents or accidents (shown after the 1 in brackets), and the number which did not (after the 0). The intermediate units give the distance between the map clusters, according to the scale shown on the right of Figure 1.

### 3.2 K-means & SOM/K-means Clustering

When clustering the factors using K-means, 9 clusters were chosen as this was the lowest number for a significant difference in all the factors between

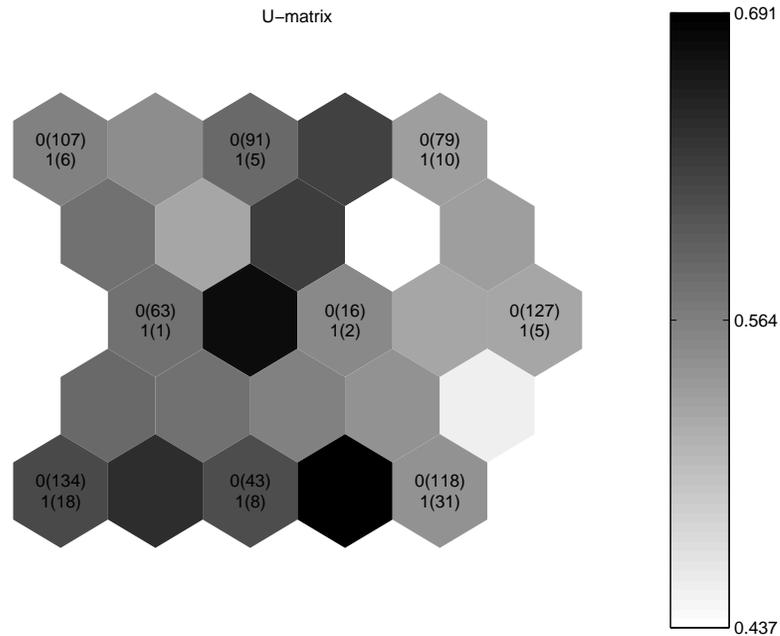


FIGURE 1. 10 Factor SOM.

clusters. This (non-prime) number had the added benefit of being such that SOMs could be run, allowing proper comparison across the different methods. Following Vesanto and Alhoniemi (2000), the SOM units (i.e. the vector representing each cluster of crossings) from Section 3.1 were then clustered using K-means (with 9 clusters) (SOM/K-Means).

### 3.3 Mixture & SOM/Mixture Clustering

The MClust R procedure (Fraley and Raftery, 2006) was also used on the factor scores. The mixture model with the best BIC value gave 7 clusters. However we chose to present the results from 9 mixture clusters, in order to compare all methods equitably. An ellipsoidal distribution with variable volume, shape and orientation was used. We also used MClust on the SOM units (into 9 clusters) (SOM/Mixture).

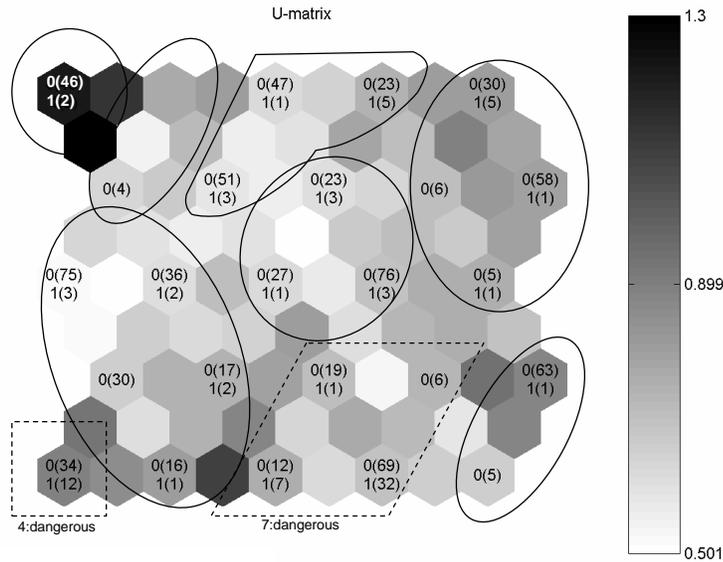


FIGURE 2. 10 Factor SOM using SOM/Mixture clusters. A dashed line indicates a dangerous crossing clusters.

### 4 Discussion, Conclusions and Future Work

The results of the comparison of the 5 methods (SOM, K-Means, Mixture, SOM/K-Means, SOM/Mixture) are given in Table 1. The SOM/Mixture model was chosen to be investigated further as it had the best predictive capability of dangerous crossings; manifested by the highest specificity ( $s$ ), and index of validity ( $I_v$ ). Associated prediction limits were calculated giving a 95% confidence interval for  $s$  of (0.786,0.841). This gives us an specificity for SOM/Mixture significantly better than SOM/K-means (along with all other methods except mixtures).

The counts of crossings falling into the dangerous and regular clusters are given in Table 2, which suggests that those crossings classified as dangerous, without an accident history, are potentially dangerous and at risk of accidents in the future. Those in the regular cluster, which have an accident history, may be those crossings where accidents were not influenced by crossing characteristics, but were due only to driver error.

In the SOM/Mixture model two clusters of crossings are identified as dangerous, 4 and 7. Crossings in cluster 4 scored the highest on Factor 2 (busy), F7 (fatigue and heavy vehicles), F9 (road lanes) and F10 (road speed), but lowest on F4 (queueing). The other dangerous cluster, 7 scored lowest on

TABLE 1. Sensitivity ( $r$ ), specificity ( $s$ ) and index of validity ( $I_v$ ) for the five methods of railway crossing factor clustering.  $c_1$  and  $c_2$  the number of clusters in the 1st and 2nd levels of clustering.

Model	$c_1$	$c_2$	$r$	$s$	$I_v$
SOM	9	-	0.802	0.499	0.529
K-means	9	-	0.616	0.725	0.714
SOM/K-means	25	9	0.616	0.769	0.753
Mixture	9	-	0.698	0.797	0.787
SOM/Mixture	25	9	0.616	0.814	0.794

TABLE 2. Results of Passive Crossing Factor SOM/Mixture Classification.

	Accident History	No History of Accidents
Dangerous	53	145
Regular	33	633

TABLE 3. Factor Means for SOM/Mixture Clusters.

	$\mu_{F1}$	$\mu_{F2}$	$\mu_{F3}$	$\mu_{F4}$	$\mu_{F5}$	$\mu_{F6}$	$\mu_{F7}$	$\mu_{F8}$	$\mu_{F9}$	$\mu_{F10}$
1	0.43	-0.19	-0.41	1.66	0.00	0.04	-0.09	0.42	-0.03	0.18
2	0.54	-0.18	-0.48	0.49	-0.03	0.04	-0.13	0.39	-0.07	0.02
3	0.38	0.14	-0.42	-0.14	0.02	-0.29	0.21	-0.06	0.08	0.37
4	0.53	1.19	-0.35	-0.18	0.14	-0.36	0.40	-0.12	0.42	0.54
5	0.33	-0.16	-0.32	-0.11	0.01	0.39	-0.16	0.32	-0.08	-0.43
6	-0.01	-0.20	-0.26	-0.19	-0.14	-0.05	-0.23	-0.16	-0.08	-0.46
7	-0.83	-0.10	-0.02	-0.17	-0.22	-0.18	0.08	-0.35	-0.03	0.10
8	-0.03	-0.21	0.63	-0.26	0.20	0.24	-0.18	0.10	-0.08	-0.53
9	-0.63	-0.10	0.85	-0.13	-0.20	-0.04	-0.12	-0.40	-0.07	0.18

F1 (low volume and low speed) and F5 (number of tracks) (See Table 3). This may show that there exist two important types of at risk crossings: (1) crossings on multiple lane, high speed roads (especially used by heavy vehicles), and (2) single track high-speed and high volume crossings. This has significant benefit for prioritisation of crossing upgrades. Future work will involve analysis (and comparison) with active crossings.

## References

Fraley, C. and Raftery A.E. (2006) MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report No.

504, Department of Statistics, University of Washington.

- Hair, J.F., Anderson, R.E., Tatham, R.L., and Black, W.C. (1998) *Multivariate Data Analysis* (5th ed.). Upper Saddle River, N.J.: Prentice Hall.
- Kohonen, T. (1997). *Self-Organizing Maps*. Berlin: Springer.
- Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Sleep, J.A., Filar, J.A., and Hudson, I.L. Modelling Driver Behaviour at Passive Railway Level Crossings: A Bayesian Belief Network Approach. [in prep.]
- Vesanto, J. and Alhoniemi, E. (2000) Clustering of the Self-Organizing Map *IEEE Transactions on Neural Networks*, **11**(3), 586-600.
- Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. (2000). SOM Toolbox for Matlab 5. Report A57. Helsinki University of Technology.

# League tables for literacy survey data based on random effect models

Nick Sofroniou<sup>1</sup>, Dominique Hoad<sup>2</sup> and Jochen Einbeck<sup>2</sup>

<sup>1</sup> Educational Research Centre, St Patrick's College, Dublin, Ireland,  
nick\_sofroniou@yahoo.co.uk

<sup>2</sup> Durham University, Department of Mathematical Sciences, Durham, UK

**Abstract:** Data from the International Adult Literacy Survey are used to illustrate how league tables can be obtained from summary data, consisting of percentages and their standard errors, using random effects models estimated by nonparametric maximum likelihood.

**Keywords:** Random effect models; Mixture models, Nonparametric maximum likelihood; Effective sample sizes.

## 1 Introduction

In the years 1994–95 twelve countries participated in the International Adult Literacy Survey (IALS). Literacy is defined as *using printed and written information to function in society, to achieve one's goals and to develop one's knowledge and potential*. The IALS developed a rigorous framework, building on work done in the 1985 Young Adult Literacy Survey (YALS) which consisted of three scales: prose, document and quantitative. It was felt that these three scales were the most significant for measuring literacy and sufficiently practical, with speaking and listening being too costly to measure. We concentrate on the measurement of prose in this paper. The data were reported by giving the percentage of individuals achieving prose level 1, 2, . . . , 5, with level 1 being worst. One way of analyzing these data is to dichotomize the data around the lowest cutpoint (i.e., the threshold between level 1 and level 2) to give percentages of adults in each country who could/could not reach a basic level of literacy. This is of particular interest to educationalists and policy makers concerned with social inclusion and its educational and economic implications. For the prose measure, the data can then be summarized in the form of Table 1.

## 2 Methodology for league table construction

### 2.1 Effective sample sizes

The IALS used complex sample designs that varied with each country and which involved both stratification by factors such as region or school size, and clustering of pupils within schools. This complicates the issue of the

TABLE 1. Percentage of adults not reaching at least Level 2 for 12 countries (Switzerland was split into two parts according to language and is treated as if it were two separate countries. Canada was treated as English-speaking and Belgium (Flanders) as Dutch). SE denotes the standard error of this percentage and  $n$  the sample size.

Country	Male			Female		
	$n$	% Level 1	SE	$n$	% Level 1	SE
Sweden	1289	7.31	0.80	1355	7.18	1.03
Netherlands	1358	10.39	1.08	1479	10.49	0.98
Germany	938	14.31	1.89	1124	13.31	1.85
Australia	3767	18.33	0.85	4437	15.69	0.78
Canada	1979	18.76	2.03	2521	14.44	2.04
New Zealand	1821	19.94	1.28	2402	16.52	1.46
Belgium (Flanders)	1066	15.55	1.69	1180	21.61	2.29
Switzerland (French)	682	17.46	1.88	751	19.44	1.70
Switzerland (German)	659	18.30	1.51	733	20.66	1.66
United Kingdom	1730	21.38	1.26	2081	21.60	1.82
Ireland	1050	24.21	2.91	1319	20.93	1.32
United States	1416	23.00	1.65	1577	18.76	1.45
Poland	1431	43.72	0.91	1569	41.74	1.74

denominator to be used in mixed binomial models as the effective sample sizes will tend to be considerably less than the actual number of students in each country, due to the intra-cluster correlations.

Cochran (1977) states that under simple random sampling the sample proportion  $p = a/n$  is an unbiased estimate of the population proportion  $P = A/N$  and that an unbiased estimate of the variance of  $p$  obtained from the sample is

$$v(p) = s_p^2 = \frac{N - n}{(n - 1)N}pq$$

which simplifies even further assuming large  $N$ , and hence a negligible finite population correction, to

$$v(p) = \frac{pq}{n - 1}.$$

By rearranging these expressions one can obtain the corresponding sample size  $n$  under simple random sampling, e.g.,

$$n = \frac{N(pq + v(p))}{pq + Nv(p)}$$

for the former expression. Thus, it is possible to use the summary information in Table 1, consisting of percentages and their standard errors, to calculate an effective sample size corresponding to the number of independent observations in a theoretical simple random sample. This allows the use of standard mixed binomial modelling software with the effective

sample size as the binomial denominator, reflecting the uncertainty in the percentages of the original table.

## 2.2 Random effect models

Probabilities are commonly either modelled through a binomial logit or a Poisson log model, with the latter one being less adequate in this example as we have relatively large probabilities and small sample sizes involved. The variability of the upper-level units, here countries  $i$ , can be taken into account by adding a random intercept  $z_i$  with *unspecified* distribution  $g(\cdot)$  to the linear predictor, so that the binomial random effect model takes the form

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta' x_{ij} + \gamma' s_i + z_i \quad (1)$$

where  $x_{ij}$  contains the lower-level covariates (here only **gender**),  $s_i$  contains the upper-level covariates (here only the factor for **language** is considered),  $\gamma$  is a non-random parameter and  $\beta$  is a fixed or random parameter. Such variance component models with unspecified random effect distribution can be conveniently fitted using the method of nonparametric maximum likelihood (Aitkin, Hinde & Francis, 2005, p. 440ff), which is implemented in the R package **npmlreg** (Einbeck, Hinde & Darnell, 2007). In short, the density  $g(\cdot)$  is approximated by a discrete distribution with  $K$  mass points, the locations  $z_k$  and masses  $\pi_k$ ,  $k = 1, \dots, K$ , of which are estimated through the EM algorithm. Thereby, the E-step corresponds to updating of the probabilities  $w_{ik} = P(\text{unit } i \text{ comes from mass point } k)$ , and the M-step to a weighted generalized linear model fit with weights  $w_{ik}$ . From the set of weights after the final EM iteration, one computes posterior intercepts  $z_i = \sum_k w_{ik} z_k$  which represent the cluster-level contribution to the response, adjusted by the covariates. As this posterior intercept “sticks” to the cluster for all its lower-level units, it forms a characteristic of the cluster (country). Sofroniou, Einbeck & Hinde (2006) used the posterior intercept for the construction of league tables in the absence of upper-level covariates.

## 3 Results and conclusions for the literacy survey

We considered several additive logistic random effect models of type (1). To keep the model parsimonious (with only 26 observations available), the models considered exclude a **language.gender** interaction term and random coefficients. Fitting gender as a covariate and no language factor requires 5 masspoints for the random intercept distribution and has a disparity of  $-2 \log L = 229.0$  with  $df = 16$ . Table 2 gives posterior probabilities of the membership of each country to a given component. It suggests that there are two main groups of countries, two countries who performed considerably better (Sweden and the Netherlands), and one low scoring outlier

TABLE 2. Posterior probabilities for the IALS data.

	Posterior intercept	Masspoints				
Intercept		-2.602	-2.156	-1.599	-1.379	-0.330
Proportion		0.077	0.093	0.434	0.319	0.077
Sweden	-2.60	1.00				
Netherlands	-2.16		1.00			
Germany	-1.72		0.21	0.79		
Australia	-1.60			1.00		
Canada	-1.59			0.97	0.03	
New Zealand	-1.58			0.92	0.08	
Belgium (Flanders)	-1.58			0.89	0.11	
Switzerland (French)	-1.54			0.72	0.28	
Switzerland (German)	-1.45			0.34	0.66	
Ireland	-1.38				1.00	
United Kingdom	-1.38				1.00	
United States	-1.38			0.01	0.99	
Poland	-0.33					1.00

Posterior probabilities:   $p \geq 0.95$ ,   $0.90 \leq p < 0.95$ ,   $p < 0.90$ .

(Poland). Adding **language** dichotomized into English/non-English speaking required 5-masspoints and reduced the disparity to 223.3 with  $df = 15$ . This was further improved by using all 6 levels of language, with a disparity of 210.3,  $df = 15$ , and 3 masspoints. However, several categories are based on only a single country and so their performance levels become confounded with language spoken. Therefore, we experimented with adding the fitted upper-level contribution to the posterior intercept, yielding a similar league table to the one presented above, but further research on issues such as representing the uncertainty corresponding to each value is required. These last two models provide some evidence in favour of the suggestion that one contribution to the observed differences in performance may be that of the language of testing.

**References**

Aitkin, M., Francis, B. and Hinde, J. (2005). *Statistical Modelling in GLIM 4* (2nd edn.). Oxford, UK.

Cochran, W.G. (1977). *Sampling Techniques* (3rd edn.). Wiley, New York.

Einbeck, J., Hinde, J. and Darnell, R. (2007). A new package for fitting random effect models – The npmlreg package. *R News*, **7**, 26–30.

Sofroniou, N., Einbeck, J. and Hinde, J. (2006). Analyzing Irish Suicide Rates with Mixture Models. In *Proceedings of the 21st International Workshop on Statistical Modelling*, Galway, Ireland, 474–481.

# Power of normality tests under a mixture of gaussian distributions: a simulation study

Júlia Teles<sup>1</sup> and Luzia Gonçalves<sup>2</sup>

<sup>1</sup> CIPER and Departamento de Métodos Matemáticos, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa, Estrada da Costa, 1495-688 Cruz Quebrada-Dafundo, Portugal (email: jteles@fmh.utl.pt)

<sup>2</sup> CEAUL and U. E. I. de Epidemiologia e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Rua da Junqueira 96, 1349-008 Lisboa, Portugal (email: luziaG@ihmt.unl.pt)

**Abstract:** In applied statistics, the normality assumption is an important topic. Although several tests have been developed, the currently available commercial software packages (e.g. SPSS, SAS, STAT) only present two or three tests for normality. In this work, using the R free software, a simulation study was carried out to compare the power of seven goodness-of-fit tests for normality, under a mixture of gaussian distributions.

**Keywords:** Power of test; normality tests; mixture distribution.

## 1 Introduction

Several inferential statistical parametric tests depend on some underlying distributional assumptions, namely the assumption that the sample is derived from a normal distribution. In many practical situations, some variables have a mixture of gaussian distributions, and it is important to assess if normality tests detect this type of departures from normality. Graphical methods provide us some information, but sometimes severe departures from normality can be hidden. This study aims to evaluate the performance of seven tests, namely Shapiro-Wilk (SW), Kolmogorov-Smirnov with Lilliefors correction (KS-L), Cramér-von Mises (CM), Shapiro-Francia (SF), Anderson-Darling (AD), Jarque-Bera (JB) and D'Agostino-Pearson omnibus test (DAP) (*e.g.* Thode, 2002).

D'Agostino et al. (1990) emphasized that despite DAP being a powerful and informative test it is not so frequently used as other normality tests. Bajgier and Aggarwal (1991) conducted a simulation study to compare the performance of several goodness-of-fit tests in detecting balanced mixtures of gaussian distributions. Reviewing several normality tests, Thode (2002) concluded that the SW test and tests which used measures of skewness and kurtosis are, in general, the best methods for assessing normality. Sheskin (2007) mentioned that the JB test, commonly used in econometrics, is more

conservative than the DAP test. Mendes and Pala (2003) studied the type I error and power of SW, KS-L and Kolmogorov-Smirnov tests, for several types of departures from normality, namely for samples generated from Weibull, beta, t, exponential and gamma distributions. These authors concluded that SW was the most powerful test. Öztuna (2006) compared the KS-L, SW, DAP and JB tests in terms of type I error and power, under the same distributions considered by Mendes and Pala (2003). Öztuna pointed out that SW test does not work well when repeated values appear in the data set.

## 2 Materials and methods

In order to compare the power of the tests mentioned before, samples of sizes  $n = 30, 50, 100, 200, 300$  were generated from different mixtures of gaussian distributions. First we considered samples that were taken from a mixture distribution of a standard normal with another normal distribution with the same standard deviation and different means, namely with mean  $\mu = 0.5, 1, 1.5, 2, 2.5, 3, 3.5$ . We also considered samples that were taken from a mixture distribution of a standard normal with another normal distribution with the same mean and different standard deviations, namely with  $\sigma = 1.5, 2, 2.5, 3, 3.5, 4$ . We only considered the case of a balanced mixture of gaussian distributions, i.e., we used a mixing proportion of 0.5. In each case, 10000 replications were run using R software, version 2.6.1 (freely available from <http://cran.r-project.org/>). The performance of all tests were assessed at 5% significance level and the proportion of rejected null hypotheses was used to estimate the power of each test.

## 3 Results and discussion

After 10000 simulation runs, some results are presented in Tables 1 and 2. When samples were drawn from mixtures of  $N(0, 1)$  with  $N(1.5, 1)$  or  $N(2, 1)$  (Table 1) or  $N(0, 1.5)$  (Table 2), all tests reveal low power in rejecting normality hypotheses, even to sample sizes equal to 300.

DAP test is, in general, the most powerful in detecting non-normality for samples that were generated from a mixture distribution of a standard normal with another normal distribution with the same standard deviation and different means. In this case, the results yielded by AD and CM tests were not as good as those attained by DAP test, and SW test presented the lowest performance of these four tests. KS-L, SF and JB tests definitely have the worst performances in rejecting normality hypothesis in the case of mixtures of normal distributions with the same shape but a different location. We could see that, in particular, for the mixture of  $N(0,1)$  with  $N(3,1)$  and  $n = 100$ , the power of KS-L, SF and JB tests are 0.61, 0.57 and 0.11, respectively.



TABLE 2. Power of normality tests under a mixture of gaussian distributions with the same mean and different standard deviations.

$n$	SW	KS-L	CM	SF	AD	JB	DAP
Mixture of $N(0, 1)$ with $N(0, 1.5)$							
30	0.0820	0.0643	0.0712	0.1039	0.0751	0.0738	0.1076
50	0.0953	0.0711	0.0762	0.1193	0.0804	0.1006	0.1161
100	0.1182	0.0753	0.0871	0.1522	0.0938	0.1500	0.1443
200	0.1667	0.0930	0.1186	0.2205	0.1288	0.2309	0.1939
300	0.2232	0.1110	0.1529	0.2761	0.1742	0.2908	0.2470
Mixture of $N(0, 1)$ with $N(0, 2)$							
30	0.1578	0.1120	0.1450	0.2088	0.1517	0.1559	0.1996
50	0.2145	0.1350	0.1828	0.2783	0.2029	0.2315	0.2404
100	0.3468	0.2191	0.3046	0.4269	0.3339	0.3904	0.3518
200	0.5905	0.3903	0.5335	0.6618	0.5780	0.6272	0.5516
300	0.7694	0.5374	0.7143	0.8164	0.7665	0.7901	0.7146
Mixture of $N(0, 1)$ with $N(0, 2.5)$							
30	0.2449	0.1875	0.2469	0.3137	0.2580	0.2210	0.2696
50	0.3690	0.2774	0.3729	0.4560	0.3945	0.3621	0.3633
100	0.6260	0.4843	0.6380	0.6962	0.6644	0.5955	0.5333
200	0.9070	0.8032	0.9157	0.9299	0.9306	0.8715	0.8078
300	0.9825	0.9386	0.9851	0.9879	0.9887	0.9684	0.9396
Mixture of $N(0, 1)$ with $N(0, 3)$							
30	0.3426	0.2955	0.3845	0.4274	0.3907	0.2793	0.3355
50	0.5209	0.4489	0.5863	0.6088	0.5969	0.4509	0.4459
100	0.8244	0.7585	0.8741	0.8687	0.8813	0.7351	0.6604
200	0.9908	0.9749	0.9949	0.9935	0.9957	0.9633	0.9263
300	0.9997	0.9980	0.9998	0.9997	0.9998	0.9955	0.9874

## References

- Bajgier, S.M. and Aggarwal, L.K. (1991). Powers of goodness-of-fit tests in detecting balanced mixed normal distributions. *Educational and Psychological Measurement*, **51**, 253-269.
- D'Agostino, R.B., Belanger, A. and D'Agostino Jr., R.B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, **44**, 316-321.
- Mendes, M., and Pala, A. (2003). Type I error rate and power of three normality tests. *Pakistan Journal of Information and Technology*, **2**, 135-139.
- Öztuna, D. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences*, **36**, 171-176.
- Sheskin, D.J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures (Fourth Edition)*. Boca Raton: Chapman & Hall/CRC.
- Thode Jr., H.C. (2002). *Testing for Normality*. New York: Marcel Dekker.

# Testing Manifest Monotonicity and Weak Item Independence for the Constant Latent Odds-Ratios Model

Jesper Tijmstra<sup>1</sup>, Dave J. Hessen<sup>1</sup>, Peter G.M. van der Heijden<sup>1</sup>

<sup>1</sup> Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University. P.O. Box 80140, 3508 TC Utrecht, The Netherlands

**Abstract:** The Constant Latent Odds Ratios (CLORs) model is a newly developed generalisation of the Rasch model with powerful measurement properties. Two observable consequences follow from the assumptions of the CLORs model: manifest monotonicity of the total score and weak item independence. The property of manifest monotonicity is common to many models within Item Response Theory (IRT), and by testing for this property, a general statement can be made about the applicability of IRT models to the data. By subsequently testing for the property of weak item independence, which only holds for a small selection of IRT models, it becomes possible to see whether the application of a CLORs model instead of e.g. a Birnbaum model would be appropriate. To this purpose, tests for manifest monotonicity and weak item independence have been developed and have been evaluated through simulation study.

**Keywords:** Constant Latent Odds-Ratios model; Rasch model; manifest monotonicity of the total score; weak item independence; Kendall's W.

## 1 Introduction

Within Item Response Theory (IRT), the Rasch model (1960) has received a lot of attention, because of its simplicity and useful measurement properties. Unlike other IRT models, the Rasch model has the property of specific objectivity, meaning that subjects can be ordered independent of items, and that items can be ordered independent of persons. Furthermore, the property of sufficiency of the total score holds within the Rasch model. That is, the total score of the test contains all the available information regarding the latent trait.

Hessen (2004; 2005) has developed the Constant Latent Odds-Ratios (CLORs) model, a promising generalization of the Rasch model for tests consisting of dichotomous items. This CLORs model is less restrictive than the Rasch model, but maintains the useful measurement properties of the latter model. Like the Rasch model, the CLORs model assumes that local independence holds and that therefore scores on items are independent when

the value on the latent trait is taken into account. Unlike the Rasch model, the CLORs model does not require the item response functions (IRFs) to have a lower asymptote of 0 and an upper asymptote of 1, allowing for the possibility of guessing or slipping on an item. Instead, the CLORs model makes the more general assumption of Constant Latent Odds-Ratios of the item passing odds. This means that whenever the odds of passing one item are compared with that of passing another item, the same ratio will be obtained everywhere on the latent trait.

In this context, the need arises to develop tests to assess the appropriateness of applying a CLORs model to a set of data. The present paper focusses on tests for two observable consequence of the CLORs model: manifest monotonicity of the total score and weak item independence. Since these properties follow directly from the assumptions of the CLORs model (Hessen, 2005), they can be used as checks to see whether a CLORs model can reasonably be applied. That is, if a CLORs model holds, then both manifest monotonicity of the total score and weak item independence should be present in the data.

## 2 Manifest monotonicity of the total score

From the assumption of local independence and the assumption of constant latent odds-ratios, the property of manifest monotonicity of the total score follows (Hessen, 2005). If manifest monotonicity of the total score ( $T = \sum_{i=1}^k X_i$ ) holds for a test with  $k$  items, then for each item  $i$

$$P(X_i = 1|T = 1) \leq \dots \leq P(X_i = 1|T = k - 1).$$

(Note that by definition,  $P(X = 1|T = 0) = 0$  and  $P(X = 1|T = k) = 1$ .) Thus, under any CLORs model, the probability of answering an item correctly should not decrease as the level of the total score increases, and hence the proportion of people that answer an item correctly is expected to increase (or stay the same) as the total score increases. Since deviations from manifest monotonicity may occur due to sampling error, the need arises to investigate whether these violations are significant, and two tests will be proposed to this purpose.

### 2.1 Testing manifest monotonicity of the total score using logistic regression

Since manifest monotonicity of the total score implies that the probability of success on an item is nondecreasing over  $T$ , a logistic regression model can be specified for each individual item, where the probability of success on an item  $i$  is predicted by the total score:

$$P(X_i = 1|T = t) = \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} \quad \text{for } t = 1, \dots, k - 1.$$

As long as  $\beta_1$  has a nonnegative value, the probability of success on the item increases over  $t$ . Thus, if this model fits the data, manifest monotonicity of the total score holds for the sample.

It should be noted that this model is restrictive in the sense that manifest monotonicity may still hold if the log odds do not increase by the same amount ( $\beta_1$ ) for every increase in the total score. However, the Type I error rate does not deviate much from the level of significance  $\alpha$ , as long as the sample size remains below  $n = 1000$  (results not included). Still, care should be taken when performing multiple tests, because deviations from  $\alpha$  cumulate and inflate the Type I error rate. Thus, it is inadvisable to use this model to evaluate manifest monotonicity for an entire test. Instead, another overall test of manifest monotonicity is proposed below.

## 2.2 Testing manifest monotonicity of the total score using Kendall's W

Let  $n_t$  denote the number of subjects with a total score  $t$ , and  $s_{t,i}$  the number of subjects with that total score that answer item  $i$  correctly, then

$$\frac{\sum_{i=1}^k s_{t,i}}{n_t} = \sum_{i=1}^k \alpha_{t,i} = t.$$

Therefore, the average proportion of successes ( $E(\alpha_{t,i}) = \frac{t}{k}$ ) increases as the total score increases. From this it follows that if for every item,  $\alpha_{t,i}$  is ordered, and this ordering is the same for *all* items, then this ordering must be the same as the ordering of  $t$ . Therefore, if there is perfect correspondence between the items in their orderings of  $\alpha_{t,i}$  over  $t$ , manifest monotonicity of the total score is known to hold. This may be illustrated by considering Figure 1, where all five items display the same ordering of  $\alpha_{t,i}$ , and manifest monotonicity holds. Likewise, deviations from perfect correspondence indicate a violation of manifest monotonicity, since in that case, at least one item displays an ordering of  $\alpha_{t,i}$  that does not correspond to the ordering of  $t$ .

To obtain a measure of this correspondence, Kendall's W (or Kendall's coefficient of concordance) can be used (Kendall & Babington Smith, 1939). While this W is usually employed to provide a measure of inter-rater reliability, it can just as well be used to evaluate the ordering of proportions between different items. Since Kendall's W can take on values between 0 (no correspondence) and 1 (perfect correspondence), observing a value of 1 implies that manifest monotonicity holds for that set of data. If the observed value of W is significantly lower than 1, then there is deviation from perfect correspondence, and manifest monotonicity has to be rejected.

Because no exact test was available to find out whether a value of W is significantly lower than 1, a simulation study was performed to determine the critical values of W under different conditions (results not included). By

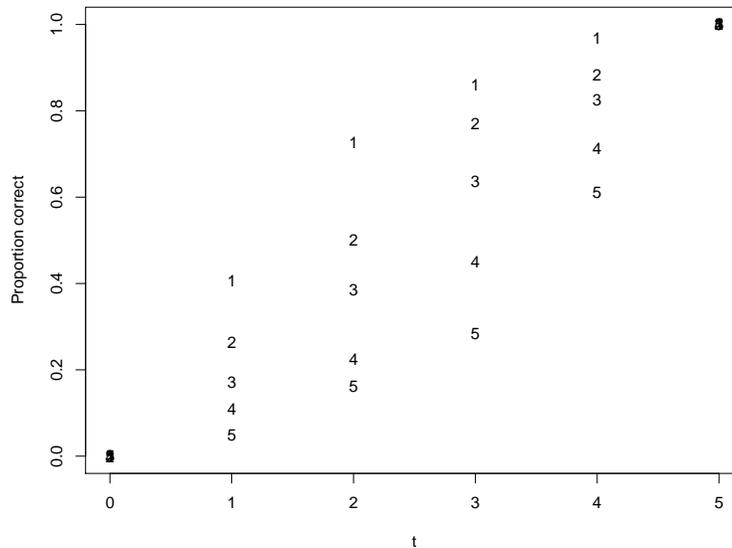


FIGURE 1. The proportion of correct answers on five fictional items, at different levels of the total score. Both manifest monotonicity and weak item independence hold.

repeatedly simulating data under several CLORs models (where manifest monotonicity was known to hold in advance), different distributions of  $W$  were obtained. By taking the fifth percentile of each of these distributions, corresponding critical values of  $W$  were obtained that can be used to test for manifest monotonicity when similar situations apply. By varying the sample size ( $n = 100, 200, 500, 1000$ ), test size ( $k = 5, 10, 15, 20$ ), and the range of the item difficulties, different critical values of  $W$  were obtained. Overall, the test appeared to be quite powerful, rejecting manifest monotonicity even when the observed  $W$  takes on a value of .90, provided that  $n \geq 500$ . Thus, with this test, relative small violations of manifest monotonicity can still be detected.

### 3 Weak item independence

Like manifest monotonicity, weak item independence is an observable consequence of the CLORs model. The property of weak item independence

states that the ordering of the items based on their probabilities should be the same at every level of the observed score. Thus, when item 1 through  $k$  are placed in descending order based on their overall probability of success,

$$P(X_1 = 1|T = t) \geq \dots \geq P(X_k = 1|T = t), \quad \text{for all } t.$$

From this it follows that weak item independence can be tested by investigating whether the ordering of  $\alpha_{t,i}$  over  $i$  is the same at every level of the total score. Significant deviations from perfect correspondence indicate a violation of weak item independence, and Kendall's  $W$  can again be used to obtain a measure of the amount of correspondence, this time between the different levels of the total score.

To obtain the critical values of  $W$  that correspond to different situations, data was generated under several CLORs models, similar to before (results not included). As could be expected, increasing the sample size resulted in an increase of the critical value of  $W$ . On the other hand, increasing the number of items led to a decrease in the critical value, since there were more violations due to chance, caused by the increased number of total scores that were compared.

Since items that are similar in difficulty are relatively hard to order, the situation where the item difficulties were not fixed (an 'optimal' situation in testing for weak item independence), but instead were drawn from a normal distribution, was investigated as well. When the item difficulties were drawn instead of fixed, there were more items with similar difficulties, and the critical values were much lower. Likewise, an increase in the range of the item difficulties resulted in an increase of the critical value, since item were less similar in difficulty. Overall, the critical values of  $W$  were lower than those observed for the manifest monotonicity test, indicating that under CLORs models, violations of weak item independence due to chance may be more common than violations of manifest monotonicity.

To investigate the power of the proposed test for weak item independence in distinguishing the CLORs model from other models, data was simulated under Birnbaum models that were identical to the CLORs models that were used to obtain the critical values of  $W$ , except for the item slope parameters. By varying the latter, the IRFs of several items crossed each other, resulting in violations of weak item independence. By comparing the value of  $W$  observed under these Birnbaum models with the critical value obtained earlier, a judgment could be made whether weak item independence needed to be rejected, and the proportion of rejections could be calculated for these different circumstances (results not included).

The power of the test increased rapidly when the sample size increased, especially when the item difficulties were fixed. When these were fixed, a power of .90 or higher was usually observed when  $n = 1000$ , indicating that in these situations the test is quite powerful. When  $n = 500$ , the power was still lacking somewhat, with values between .40 to .83. When there was no fixed distance between the item difficulties, the power was worse, ranging

from .27 to .63 when  $n = 1000$ . One main finding therefore consisted in the diminishing effect that having items with similar difficulties has on the power with which one can test for weak item independence.

#### 4 Conclusion and prospects

In this paper, tests for manifest monotonicity of the total score and weak item independence have been proposed. Since both properties are observable consequences of the CLORs model, a rejection of either of these properties indicates that the application of a CLORs model to the data would be inappropriate. Hence, the proposed tests can function as a check to see whether the application of a CLORs model could be reasonable, or whether other IRT models (if any) should be applied. By using Kendall's  $W$ , an overall measure of manifest monotonicity of the total score can be obtained. Should this result in a rejection of manifest monotonicity for the test as a whole, then the logistic regression test can be employed to find out which items are the cause of this violation.

When Kendall's  $W$  is used to test for weak item independence, the power of this test is strongly influenced by the item difficulties. In situations where the item difficulties are similar, perhaps in the context of admission testing or examinations, the test may be less useful. However, when the item difficulties cover a broad range, the test quite successfully detects violations of weak item independence.

Although the tests have been evaluated through simulation study, they still need to be tested in practice. To this purpose, data from the field of Pedagogy will be used in the near future. It should also be noted that although critical values of  $W$  were obtained for a broad range of conditions, the most appropriate value can always be obtained through bootstrapping. Whether this results in large differences, compared to using one of the pre-generated critical values of  $W$ , will be investigated as well.

#### References

- Hessen, D.J. (2004). A New Class of Parametric IRT Models for Dichotomous Item Scores. *Journal of Applied Measurement*, **5**, 385-397.
- Hessen, D.J. (2005). Constant Latent Odds-Ratios Models and the Mantel-Haenszel Null Hypothesis. *Psychometrika*, **70**, 497-516.
- Kendall, M.G. (1939). The Problem of  $m$  Rankings. *The Annals of Mathematical Statistics*, **10**, 275-287.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen and Lydiche.

# Haplotype Frequency Estimation with the Penalized Composite Link Model

Hae-Won Uh<sup>1</sup> and Paul H.C. Eilers<sup>2</sup>

<sup>1</sup> Leiden University Medical Centre, Department of Medical Statistics and Bioinformatics, P.O.Box 9600, 2300 RC Leiden, The Netherlands; h.uh@lumc.nl

<sup>2</sup> Utrecht University, Faculty of Social and Behavioural Sciences, & Data Theory Group, Leiden University, The Netherlands; p.h.c.eilers@uu.nl

**Abstract:** The composite link model is a natural choice for haplotype frequency estimation from observed genotype counts. To stabilize the ill-posed problem we add a ridge penalty. We apply the model to a small data set from the literature and compare results with the Bayesian program PHASE.

**Keywords:** haplotype, genotype, ill-posed, penalized likelihood, AIC.

## 1 Introduction

Chromosomes are paired. In many places single nucleotide polymorphisms (SNPs) occur, which have two possible states. This can be coded as a 0/1 vector for each chromosome, which are called the haplotypes. We cannot observe the individual haplotypes, only the sum of the two vectors, the genotype. So for each SNP we observe either 0, 1 or 2. We want to estimate probabilities of haplotypes, for the sampled population as well as for individual with a given genotype. Several algorithms are in use, some of them based on the EM algorithm, others based on Bayesian model (Excoffier and Slatkin (1995), Stephens *et al.* (2001)). We present a novel approach, based on the composite link model (CLM) of Thompson and Baker (1981).

We extend the CLM with a penalty on the parameters, with two goals in mind: 1) to stabilize and speed up the estimation; 2) to get positive probabilities. EM will always give zero probability to unobserved (compatible) haplotypes. This may be reasonable in very large samples, but otherwise it is not correct, because it equates unobserved to impossible. The weight of the penalty is optimized by searching for a minimum of AIC.

We illustrate our model on a small data set from the literature and compare the estimated haplotype probabilities with the results of the program PHASE (Stephens *et al.* (2001)), which represents the current state of the art in haplotype estimation.

## 2 Theory

A haplotype is a 0/1 vector of length  $L$ , representing the two alleles of SNPs;  $K = 2^L$  haplotypes are possible. They combine into  $J = K^2$  ordered pairs, called diplotypes. A genotype is a ternary (0, 1, 2) vector of length  $L$ , the sum of the vectors in the diplotype;  $I = 3^L$  genotypes are possible. Compatibility between genotypes and diplotypes is coded by the  $I$  by  $J$  composition matrix  $C = [c_{ij}]$ . When genotype  $i$  is compatible with diplotype  $j$ ,  $c_{ij} = 1$ ; otherwise  $c_{ij} = 0$ .

Let  $q_k = \exp(\beta_k)$  be the probability of haplotype  $k$ . Under random mating, the probability of diplotype  $(k, k')$  is  $q_k q_{k'}$ . Introducing the  $J$  by  $K$  matrix  $X = [x_{jk}]$ , allows us to write  $\gamma = \exp(X\beta)$ , with  $\gamma_j$  the probability of diplotype  $j$ . If diplotype  $j$  corresponds to haplotype pair  $(k, k')$ , with  $k \neq k'$ , then columns  $k$  and  $k'$  of  $X$  contain a one in row  $j$ . If  $k = k'$ ,  $x_{jk} = 2$ . All other elements in that row are zero.

Combining mating and composition gives  $p = C \exp(X\beta)$  for the genotypes probabilities. Let  $n$  genotypes be observed, with counts in a vector  $y$ . The expected values are  $E(y) = \mu = np = nC \exp(X\beta)$ . Note that the composition matrix  $C$  has a row for each possible genotype, whether it was observed or not. Also some elements of the count vector  $y$  may be zero, reflecting unobserved genotypes.

We extend the CLM with a ridge-type penalty, by forming the penalized Poisson log-likelihood

$$l^* = \sum_i (y_i \log \mu_i - \mu_i) - \kappa \sum_k (\beta_k - \alpha_k)^2 / 2. \quad (1)$$

The penalty pushes the solution, more or less gently, depending on the value of  $\kappa$ , towards a pre-specified distribution  $\exp(\alpha)$ . The choice of  $\alpha$  will be discussed later. The scoring algorithm gives

$$(U' \tilde{W} U + \kappa I) \hat{\beta} = U'(y - \tilde{\mu} + \tilde{W} U \tilde{\beta}) + \kappa \alpha. \quad (2)$$

where a tilde, as in  $\tilde{\mu}$  indicates an approximation to the solution,  $U = M^{-1} C \Gamma X$ , with  $M = \text{diag}(\mu)$ ,  $\Gamma = \text{diag}(\gamma)$  and  $W = \text{diag}(\mu)$ . At convergence, standard errors can be obtained for  $\hat{\beta}$  by computing  $\text{cov}(\hat{\beta}) = (U' W U + \kappa I)^{-1}$ .

To simplify the presentation, we ignored one important practical detail. It is desirable and reasonable to have  $\sum \mu_i = \sum y_i$ . We found that this condition did hold for very high and very low values of  $\kappa$  (for our choice of  $\alpha$ , based on independence), but not for values in between. Our solution is to add an offset  $\delta$ , so that the haplotype probabilities are  $q = \exp(\beta + \delta)$ . There is no penalty on  $\delta$ .

In our experience the scoring algorithm is not always stable. Therefore we check whether the proposed update for  $\beta$  indeed lowers the penalized likelihood. If it does not, we halve the step in the direction  $\hat{\beta} - \tilde{\beta}$ . This correction is repeated if needed.

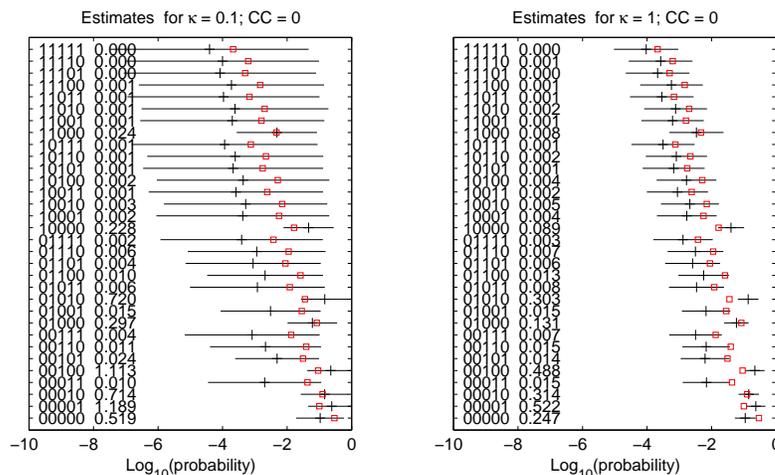


FIGURE 1. Estimated probabilities and error bars for 32 haplotypes of 5 SNPs. The small squares show the prior probabilities  $\alpha$ . Haplotypes and numerical values of probabilities are shown at the left.

The matrices  $C$  and  $X$  are extremely sparse and in any problem of realistic size they quickly would become too large to fit in computer memory. In our Matlab implementation we use sparse matrix facilities. In other languages more work might be needed. One possible approach is to store lists of the indices of the non-zero elements and compute indexed sums to get at  $X\beta$ ,  $C\gamma$  and  $U'WU$  (generally  $U$  is non-sparse). The system of scoring equations contains  $2^L$  equations, with  $L$  the number of SNPs. A practical limit lies at 10 to 12 SNPs, if these equations are formed explicitly.

Natural starting values for  $\beta$  can be based on the assumption that the SNPs are independent. It is also a natural choice for the vector  $\alpha$  in the penalty. We select an “optimal” value for  $\kappa$  using AIC. It is defined as  $AIC = -2l + 2ED$  where  $ED = \text{trace}[(U'WU + \kappa I)^{-1}U'WU]$ , is the effective model dimension. On a linear grid for  $\log \kappa$  we search for the minimum of AIC.

Unfortunately, the penalty does not remove problems with multiple local maxima of the (penalized) likelihood. We could construct artificial data sets for which different starting values of  $\beta$  gave different optima. We also experienced that the algorithm of Thompson and Baker does not guarantee convergence; a line search may be necessary.

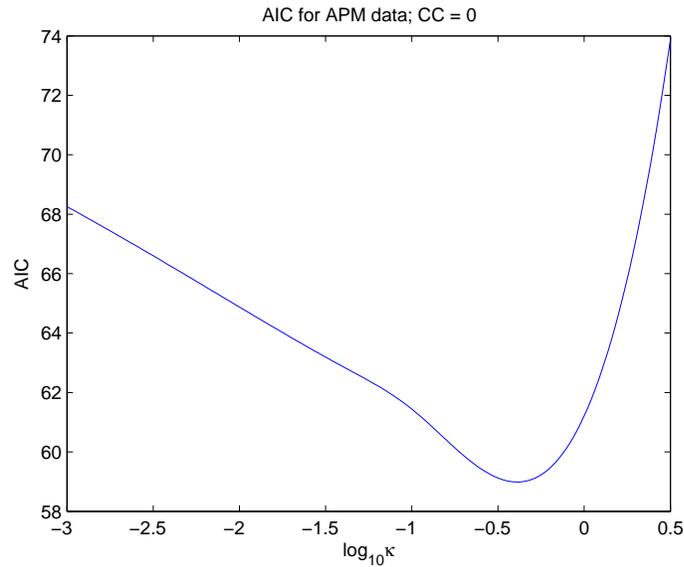


FIGURE 2. Graph of AIC as a function of  $\kappa$ .

### 3 An application

We illustrate our model with data from a case-control study on cervical carcinoma (Mehta et al., 2007). We select 5 SNPs on chromosome 5 and use the control group (122 persons). Figure 1 shows estimated haplotype probabilities, standard errors and the prior estimates  $\alpha$ . As  $\kappa$  increases,  $\hat{\beta}$  gets nearer to  $\alpha$ , and standard errors decrease. In fact, standard errors go to zero for very large  $\kappa$ . Then we have eliminated all uncertainty at the cost of a possibly large bias (forgetting the uncertainty in  $\alpha$  for the moment). This is where AIC comes in: it is an estimate of predictive performance. As Figure 2 shows a clear minimum is indicated near  $\log_{10} \kappa = -0.39$ , between the two values used for  $\kappa$  in Figure 1.

We analyzed the same data with the PHASE program and found very good correspondence between the two sets of results. This is illustrated in Figure 3. Because PHASE uses Monte Carlo-based computations, the logarithms of smaller probabilities will vary appreciably with the length of Markov Chain and the random starting seed.

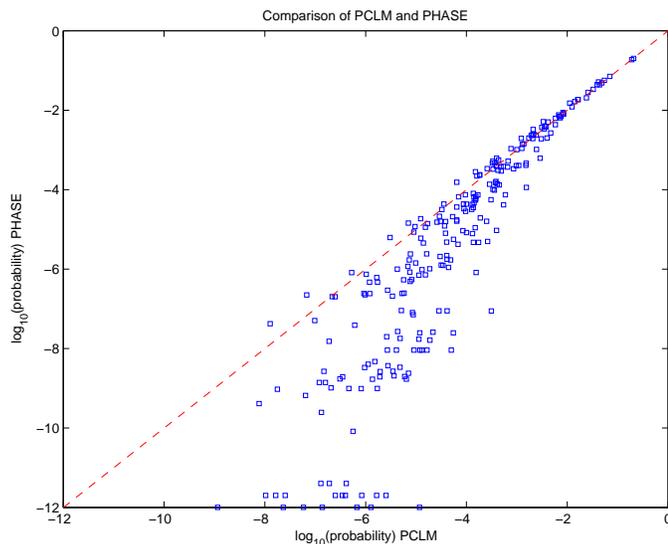


FIGURE 3. Comparison between results of the present model and the PHASE program. The broken line represents equality.

#### 4 Discussion

We introduced a new approach to haplotype estimation, based on penalized composite link model (CLM). The structure of the CLM matches wonderfully with the nature of the genetic problem. A penalty improves the numerical condition of the model and allows prior information to be brought in. The model works well on an illustrative data set (and on many others we investigated). The results are essentially identical to those of PHASE program, the current state of the art.

The matrix  $M^{-1}CT$  has a useful interpretation. Row  $i$  gives the probabilities of the diplotypes that are compatible with genotype  $i$ . These can be used as weights in regression models for case-control studies (French et al., 2006).

Missing data can be handled elegantly by combining rows of the composition matrix. Genotypes are incomplete, their frequencies are written in extra elements of the vector  $y$  and a new composition matrix  $C^*$  is computed with extra rows that are sums of the rows of  $C$  that are compatible with the uncertain genotypes.

The CLM considerably simplifies notation when discussing haplotype-based genetic models. The ubiquitous sums of probabilities over compatible sets, that are characteristic for most of the literature in this field, are replaced by concise expressions with matrices and vectors.

**References**

- Eilers, P.H.C (2007). Ill-posed Problems with Counts, the Composite Link Model, and Penalized Likelihood. *Statistical Modelling* **7**, 239–254.
- Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Molecular Biology and Evolution* **12**, 921–927.
- French, B. *et al.* 2006 Simple Estimates of Haplotype Risks in Case-Control Data. *Genetic Epidemiology* **30**, 485–494.
- Mehta, A.M. *et al.* (2007.) Genetic Variation of Antigen Processing Machinery Components and Association with Cervical Carcinoma. *Genes, Chromosomes & Cancer* **46**, 577–586.
- Stephens, M., Smith, N.J. and Donnelly, P. (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *American Journal of Human Genetics* **68**, 978–989.
- Thompson, R. and Baker, R.J. (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics* **30**, 125–131.

# Analysis of labor spells in Social Security contributors

O. Valero<sup>1</sup>, A. Espinal<sup>1</sup>, P. Puig<sup>2</sup>

<sup>1</sup> Servei d'Estadística, Universitat Autònoma de Barcelona

<sup>2</sup> Dpt. de Matemàtiques, Universitat Autònoma de Barcelona

**Abstract:** In order to analyze the number of labor spells in Social Security contributors a Poisson regression model was established. Data has been taken from the continuous sample of labor histories, a representative sample of all people paying contributions in the Spanish Social Security.

**Keywords:** Poisson regression; offset; overdispersion; NB1.

## 1 Introduction

The continuous sample of working histories<sup>1</sup> (Durán, 2007) is a collection of anonymous microdata extracted from administrative files of the Spanish Social Security. It is a representative sample of all the people having a relationship with the Social Security during a certain year, and it makes reference to the population who pay contributions or are pensioners in the year referred to. It has been done continuously since year 2004 with annual actualization, and for every person there is information of the complete labor history since his first working day. Data previous to year 1980 has a lot of deficiencies and has not been considered. The study has been carried out with contributors only, pensioners have not been considered.

Different studies might be done using this sample. The main objective of the paper is to analyze the number of labor spells in the province of Barcelona.

---

<sup>1</sup>Data comes from the Muestra Continua de Vidas Laborales (a sample of working histories) that has been provided to the Labour Market Service of the Economic Development Area of Diputació de Barcelona (the Administration's Barcelona Province) coming from the Administration's Spanish Social Security, tied to the Ministry of Labour and Immigration. Data have been used as part of a concerted agreement between the Labour Market Service of the Diputació de Barcelona and Servei d'Estadística (Statistical Consulting) of the Universitat Autònoma de Barcelona (UAB).

## 2 Data

The population is composed of every person paying contributions to the Spanish Social Security during the year 2006. The sample is 4% of this population, giving 111,816 people for the province of Barcelona.

The measured variables used as covariates were: gender (male, female), age (<25 years, 25-39 years, 40-54 years, 55-64 years, >64 years), origin (Catalan, Spanish, other) and number of labor spells since 1980. Previous information has not been considered due to the bad quality of the data, but years since 1980 (or since the first labor spell for people who started to work more recently) and a dummy variable indicating spells previous to this year (i80) have been taken into account.

## 3 Poisson regression

A Poisson regression model (McCullagh and Nelder, 1989) was established to analyze the number of labor spells. The set of covariates used were the socio-demographic variables gender, age, nationality, and also the variable i80.

Due to the fact that labor spells happen in a fixed period of time we have established a model for the proportion of events (Agresti, 2002). The variable years since 1980 was used as an offset (McCullagh and Nelder, 1989). The offset variable is used to normalize the number of events in the studied period of time.

A log linear relationship between the mean and the factors was specified by the log link function. The model is:

$$\log(\mu_i) = \log(n_i) + X_i' \beta$$

where  $\mu_i$  is the mean of number of labor spells,  $\log(n_i)$  is the the logarithm of the offset variable,  $X_i'$  is the matrix of known covariates (gender, age, nationality and the indicator of having previous spells), and  $\beta$  is a vector containing the fixed effects to be estimated.

The response variable was overdispersed with respect to the Poisson distribution (deviance/df = 7.987). To model this situation we allowed the variance function to have a multiplicative overdispersion factor  $\phi$ :  $Var(y) = \phi\mu$ . This is the NB1 model described by Cameron and Trivedi (1998). The standard errors of the parameters fitted by the Poisson regression model were corrected by using the estimated value of  $\phi$ .

The NB1 assumption has been checked graphically and also by using one of the overdispersion tests of Dean (1992).

## 4 Results

All the covariates included in the model had a p-value  $<0.001$ , this high signification is mainly due to the large sample size. To analyze the effect of each covariate the parameters have been estimated by maximum likelihood using proc genmod of SAS. The exponential of the estimated parameter can be interpreted as the ratio of the expected number of labor spells in one year.

Table 1 shows the expected mean number of labor spells per year and the 95% confidence limit interval when comparing gender, age, origin and indicator of previous spells' categories.

TABLE 1. Expected mean ratio.

	Categories		$\exp(\beta)$	LCL	UCL
Gender	Female	Male	1.13	1.12	1.14
Age	<25 years	25-39 years	1.90	1.86	1.94
	<25 years	40-54 years	3.40	3.32	3.48
	<25 years	55-64 years	4.56	4.42	4.71
	<25 years	>64 years	6.88	6.30	7.52
	25-39 years	40-54 years	1.79	1.76	1.82
	25-39 years	55-64 years	2.40	2.34	2.46
	25-39 years	>64 years	3.62	3.32	3.95
	40-54 years	55-64 years	1.34	1.31	1.37
Origin	40-54 years	>64 years	2.02	1.86	2.21
	55-64 years	>64 years	1.51	1.38	1.65
	Spain	Catalonia	1.11	1.09	1.12
i80	Other	Catalonia	1.86	1.82	1.90
	Other	Spain	1.68	1.64	1.72
	Before 1980	After 1980	1.24	1.21	1.26

The expected number of spells is 1.13 when comparing females with males, which indicates a number a little bit higher in females.

Comparing age categories major differences are found. In one year people under 25 have 1.9 more labor spells in average than people between 25 and 39 years. The ratio is 3.4 compared with the category of 40-54 years, 4.56 for 55-64 years and 6.88 for >64 years. Other comparisons show the same effect, younger age categories have a major proportion of spells.

Considering the origin we found a proportion of 1.86 spells higher comparing people from other countries with people from Catalonia. The ratio is 1.68 when compared with people from Spain and a very little difference was found comparing Catalonia and Spain, 1.11.

People who don't have relationships before year 1980 have 1.24 more labor spells in one year than people who started to work before that year.

## 5 Conclusions and future work

Using a Poisson regression, the number of labor spells has been analyzed considering gender, age, origin and an indicator of previous labor spells as covariates. The interest of this variable is that it can be considered as an indicator of labor instability.

We found that the age was the most relevant covariate: people under 25 have more labor spells than people within other categories. These relationships are known to be shorter and to have worse labor conditions.

Further analysis might be done including covariates related with the present occupation as contribution group (engineers, administrative managers, administrative personnel, manual workers), type of contract (indefinite, temporary, others), and economical activity sector (industry and energy, construction, commerce, hotel and catering trade, communication and transports, other market services, public administration, education and sanitary activities, other non-market activities).

## References

- Agresti, A. (2002). *Categorical Data Analysis, Second Edition*. Wiley.
- Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.
- Dean, C.B. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models, *Journal of the American Statistical Association*, Vol. 87, 418, 451-457.
- Durán, A. (2007). La Muestra Continua de Vidas Laborales de la Seguridad Social, *Revista del Ministerio de Trabajo y Asuntos Sociales*, ISSN 1137-5868, 1, 231-240.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*. London: Chapman and Hall.

# A Bayesian three-state model to estimate life expectancies

Ardo van den Hout<sup>1</sup> and Fiona E. Matthews<sup>1</sup>

<sup>1</sup> MRC Biostatistics Unit, Institute of Public Health, Cambridge University, U.K.

**Abstract:** Longitudinal data for individuals with Parkinson's disease are investigated with respect to the onset of dementia. Of interest is the subdivision of total life expectancy into life expectancy with and without dementia. A three-state illness-death Markov model is presented in a Bayesian framework. In state one individuals with Parkinson's disease are dementia free, whereas state two indicates the presence of dementia. State three is the death state. Intensities of moving between the states are allowed to change in a piecewise-constant fashion by linking them to age as a time-dependent covariate. Relaxing the assumption of constant intensities makes the model applicable in many situations where individuals are followed over a long time period. In describing how a disease develops over time, the model can help to predict future need for health care. The Bayesian framework is well suited to allow for heterogeneity via random effects, and to investigate additional computation using model parameters. The latter is of importance in the estimation of life expectancies.

**Keywords:** Bayesian Markov model; Dementia; Healthy life expectancy.

## 1 Introduction

Individuals with Parkinson's disease are more likely to develop dementia than individuals without the disease. In studies of individuals with Parkinson's disease the onset of dementia is an important predictor of health and need for care. This paper presents a Bayesian illness-death Markov model with three states where the third state is the death state. The model is applied to a Norwegian longitudinal study where individuals with Parkinson's disease are followed up and the presence of dementia is monitored. Of interest are the risk factors of dementia and the subdivision of total life expectancy into life expectancy with and without dementia.

The Markov model in this paper is a continuous-time model where transition intensities depend on covariates and random effects. To allow intensities to change over time, a piecewise-constant approach is applied to take age into account as a time-dependent covariate. Relaxing the assumption of constant intensities makes the model applicable in many situations where individuals are followed over a long time - especially in the case of the older population, where, for example, mortality increases with time.

The Bayesian framework is well suited to allow for unobserved heterogeneity via random effects, and to investigate derived variables from model parameters estimates. The latter is of importance since life expectancies can be computed using the parameters of the Markov model. Our models are an extension of the models in Sharples (1993) and Pan et al. (2007), with the ability to take an external time-dependent covariate such as age into account. To our knowledge, the estimation of life expectancies using Bayesian multi-state models has not been investigated before. An additional advantage of the Bayesian framework is the straightforward propagation of the variance of model parameters into the variance of life expectancies.

## 2 Bayesian Markov model

The Markov assumption implies that the probability of moving to another state only depends on the current state. Our model assumes that known transition times are only available for transitions into the death state and that there is no recovery from state two to state 1.

Let  $t$  denote time since start of the study. The state of an individual is  $X_t \in \{1, 2, 3\}$ . A transition at  $t$  from state  $r$  to state  $s$ ,  $r \neq s$ , occurs with intensity  $q_{rs}(t)$ , where  $q_{12}(t), q_{13}(t), q_{23}(t) \geq 0$  and  $q_{21}(t), q_{31}(t), q_{32}(t) = 0$ . Intensities are regressed on covariates and a shared random effect by  $q_{rs}(t) = \exp\{\mathbf{b}_{rs}^T \mathbf{z}(t) + \tau\}$ , where  $\tau$  is normally distributed with mean zero and variance  $\sigma$ . The model is flexible to other random-effect structures. The change of intensities over time is approximated by using age as a time-dependent covariate: intensities are constant within individually observed time intervals but may vary between intervals due to changing age. The constant intensities for an observed interval  $(t, u]$  are defined using the values midway, i.e.,  $\mathbf{z}((t + u)/2)$ .

Transition probabilities of an observed time interval  $(t, u]$  are given by  $\mathbf{P}(t, u) = \exp\{(u - t)\mathbf{Q}((t + u)/2)\}$  where for  $t^* \geq 0$  we define

$$\mathbf{Q}(t^*) = \begin{pmatrix} -\{q_{12}(t^*) + q_{13}(t^*)\} & q_{12}(t^*) & q_{13}(t^*) \\ 0 & -q_{23}(t^*) & q_{23}(t^*) \\ 0 & 0 & 0 \end{pmatrix},$$

and entry  $\mathbf{P}(t, u)[r, s]$  is  $\mathbb{P}(X_u = s | X_t = r)$ . For the three-state illness-death model without recovery matrix  $\mathbf{P}(t, u)$  is available in closed form.

Assume that an individual has observations at times  $t_1 = 0, t_2, \dots, t_M$ . Using the Markov assumption, the probability of the observed trajectory  $x_{t_1}, x_{t_2}, \dots, x_{t_M}$  conditional on  $x_{t_1}$  is given by

$$\mathbb{P}(X_{t_2} = x_{t_2} | X_{t_1} = x_{t_1}) \times \dots \times \mathbb{P}(X_{t_M} = x_{t_M} | X_{t_{M-1}} = x_{t_{M-1}}),$$

where the covariates and random effect are ignored in the notation. Next, intervals  $(t_1, t_2], \dots, (t_{M-1}, t_M]$  are modelled independently by using the

multinomial distribution for the transitions to state  $X_{j+1}$  given state  $X_j$ ,  $j = 1, \dots, M - 1$ . Possible right-censored states (denoted by  $c$ ) are taken into account.

Transitions to the states 1, 2, 3, and  $c$  are coded  $(1,0,0,0)$ ,  $(0,1,0,0)$ ,  $(0,0,1,0)$ , and  $(0,0,0,1)$  respectively. At time  $t_{j+1}$ , let  $\mathbf{Y}_{j+1}$  be the variable with the four indicator vectors as sample space. For example, the trajectory  $x_{t_1}, x_{t_2}, x_{t_3} = 1, 1, 3$ , corresponds with  $\mathbf{y}_2, \mathbf{y}_3 = (1, 0, 0, 0), (0, 0, 1, 0)$ . Define  $\mathbf{Y}_{j+1} \sim \text{Multinomial}(\mathbf{p}_{j+1}, 1)$ . If the state observed at  $t_{j+1}$  is 1 or 2, it follows that

$$\mathbf{p}_{j+1} = \left( \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 1], \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 2], \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 3], 0 \right).$$

If the state observed at  $t_{j+1}$  is 3, define

$$\begin{aligned} p_d &= \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 1] \times \mathbf{P}(t_{j+1}, t_{j+1} + \epsilon)[1, 3] \\ &\quad + \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 2] \times \mathbf{P}(t_{j+1}, t_{j+1} + \epsilon)[2, 3], \end{aligned}$$

so that  $\mathbf{p}_{j+1} = (0, 0, p_d, 1 - p_d)$ . Thus we approximate known death times by assuming an unknown state just before death and then death within a small time interval  $\epsilon$ . If the state observed at  $t_{j+1}$  is right censored, it follows that

$$\mathbf{p}_{j+1} = \left( 0, 0, \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 3], 1 - \mathbf{P}(t_j, t_{j+1})[x_{t_j}, 3] \right).$$

We use the logistic regression model for the baseline state. Let  $\theta = \mathbb{P}(X_{t_1} = 2 | \mathbf{z}(t_1))$ . The model is  $\text{logit}(\theta) = \mathbf{a}^T \mathbf{z}(t_1)$  and  $X_{t_1} \sim \text{Bernoulli}(\theta)$ .

Regarding the priors, for regression coefficients  $\mathbf{b}$  in the Markov model and for regression coefficients  $\mathbf{a}$  in the model for the baseline state we use vague univariate normal distributions with mean zero and large variance, and for  $\sigma$  we use a truncated normal distribution. Markov Monte Carlo Chain (MCMC) methods are applied to estimate the parameters. The encompassing model was programmed in the BUGS language.

Expected life expectancy (LE) in state  $s$  given initial state  $r$  is

$$e_{rs}(\mathbf{z}_0) = \int_0^\infty \mathbb{P}(X_t = s | X_0 = r, \mathbf{z}_0) dt, \quad r, s \in \{1, 2\} \quad (1)$$

where  $\mathbf{z}_0 = \mathbf{z}(t_1)$  and where we assume  $\mathbf{z}(t)$  to be deterministic. Expected LE in state  $s$  irrespective of the initial state (marginal LE) is given by  $e_s(\mathbf{z}_0) = (1 - \theta)e_{1s}(\mathbf{z}_0) + \theta e_{2s}(\mathbf{z}_0)$ . Expected total LE is given by  $e_{\text{tot}}(\mathbf{z}_0) = e_1(\mathbf{z}_0) + e_2(\mathbf{z}_0)$ .

### 3 Application

We analysed data taken from a Norwegian study of individuals with Parkinson's disease where 233 individuals have had up to seven interviews in the

TABLE 1. Posterior means of life expectancies (LEs) and 95%-credible intervals for men with specified age at baseline and an eight-year Parkinson’s disease duration in 1993.

Age	LE without dementia	LE with dementia	Total LE
60	8.31 (6.72; 9.90)	3.55 (2.65; 4.54)	11.86 (10.15; 13.58)
70	4.35 (3.56; 5.29)	3.26 (2.61; 4.00)	7.61 ( 6.64; 8.63)

period 1993 to 2002. Data are kindly provided by the Norwegian Centre for Movement Disorders. State one denotes dementia free, state two denotes demented, and state 3 is the death state. In total there were 897 observations (total number of interviews, censored states, and observed deaths). During study time, 187 individuals died. Three observations are right censored.

The loglinear model for the intensities and the logistic regression model for the baseline state use age at time of interview, sex, and Parkinson’s disease duration in years before 1993 as covariates. A shared random effect can be interpreted as general susceptibility to ill-health.

Given a model, MCMC consisted of two chains with different starting values. For both chains, the first 50000 iterations were ignored (burn-in) and the additional 50000 iterations were used to compute the posterior means, the 95% credible intervals, and the Deviance Information Criterion (DIC, Spiegelhalter *et al.* 2002). The BUGS code was run via the R interface `BRugs`. DIC is used to compare models. Models with smaller DIC are better supported by the data. We fitted four models. The model with a shared random effect and with no restriction on the regression parameters for the intensities has DIC = 2428. This model is slightly better than the model where the random effect is not included: DIC = 2429. However, assessing the 95% credible intervals and restricting parameters to zero when their credible intervals include zero yields smaller DICs: for the model with the random effect DIC = 2424 and for the model without the random effect DIC = 2422. The latter model is the final model with constraints  $b_{13}^{age} = b_{13}^{sex} = b_{13}^{hist} = b_{23}^{hist} = 0$  and without a shared random effect.

In the absence of a shared random effect, the integral (1) for the LEs can be estimated numerically using the simulated parameters in the MCMC runs. We used the trapezoidal rule to do this. In the estimation of the integrand we used a piecewise-constant approach. Say we want to compute the LEs of an individual 70 years old in state 1 at baseline. We create a trajectory for this person with times  $u_1 = 0, u_2, \dots, u_M$ , where for each time interval  $(u_j, u_{j+1}]$  we specify  $\mathbf{z}((u_j + u_{j+1})/2)$ . The time between two time points in this trajectory is fixed to be one year, so  $70 + u_M$  is the assumed maximum age. Time-dependent age is included, so  $\mathbf{z}(u_j) \neq \mathbf{z}(u_k)$  for  $j \neq k$ . To compute

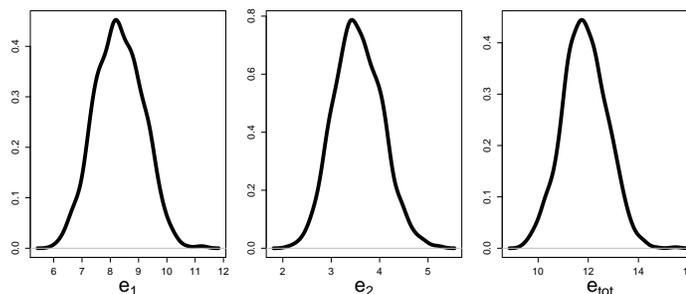


FIGURE 1. Posterior densities of marginal and total life expectancies for men aged 60 at baseline with with an eight-year Parkinson's disease duration in 1993.

the integrand, we use the multiplication  $\mathbf{P}(0, u_2)\mathbf{P}(u_2, u_3) \times \dots \times \mathbf{P}(u_{j-1}, u_j)$ , to approximate  $\mathbf{P}(0, u_j)$ ,  $j = 2, \dots, M$ . Computationally, it is efficient to use one grid both for the piecewise-constant approximation and the trapezoidal rule.

Table 1 presents LEs for men aged 60 and aged 70 at baseline with median duration (eight years). These LEs are for Norwegians with clinically diagnosed Parkinson's disease by 1993. The figures show that although total LEs for the difference ages differ about four years, the LEs spent in the dementia state differ by less than half a year. The effects for duration of Parkinson's disease prior to 1993 (not shown) indicate that duration accelerates the onset of dementia.

Because of the use of MCMC, we readily obtain not only the LEs but also the estimated distribution of the LEs. Figure 1 shows the posterior densities of  $e_1$ ,  $e_2$  and  $e_{\text{tot}}$  for men aged 60 at baseline and with an eight-year duration of Parkinson's disease in 1993.

Using the trapezoidal rule is only possible if the covariate vector  $\mathbf{z}(t)$  in (1) is deterministic in the sense that conditional on knowing  $\mathbf{z}(0)$ , we know  $\mathbf{z}(t)$  for  $t > 0$ . An alternative is micro-simulation where individual trajectories are simulated and corresponding individual survival is used to estimate LEs. Micro-simulation is more flexible in that it can also deal with random effects and with time-dependent covariates such as information about visited states. Micro-simulation has been used in, for example, cost-effectiveness modelling (Spiegelhalter and Best, 2003).

## References

- Pan, S.L., Wu, H.M., Yen, A.M.F., and Chen, T.H.H. (2008). A Markov regression random-effects model for remission of functional disability in

patients following a first stroke: A Bayesian approach. *Statistics in Medicine*, **26**, 5335-5353.

Sharples, L.D. (1993). Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation. *Statistics in medicine*, **12**, 115-1169.

Spiegelhalter, D.J., and Best, N.G. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine*, **22**, 3687-3709.

Spiegelhalter, D., Best, N., Carlin, B., and Van der Linde A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583-640.

# Modeling Spatial Effects in Field Trials

Fred A. van Eeuwijk<sup>1</sup>, Marcos Malosetti<sup>1</sup>, Martin P. Boer<sup>1</sup> and Paul H.C. Eilers<sup>1,2</sup>

<sup>1</sup> Biometris, Wageningen, The Netherlands

<sup>2</sup> Department of Methodology and Statistics, Utrecht University

<sup>3</sup> Data Theory Group, Leiden University (email: p.h.c.eilers@uu.nl)

**Abstract:** We model two-dimensional spatial effects in a large plant breeding field trial. In addition to parameters for rows, columns and genotypes, we have a parameter per plot, constrained by spatial penalties. A trial on barley is used for illustration. Computational strategies for large data sets are proposed.

**Keywords:** Mixed model, penalty, cross-validation.

## 1 Introduction

Spatial variation in plot errors is a common phenomenon in field trials. A popular mixed model methodology to deal with spatial variation was introduced by Gilmour et al. (1997). In this approach the plot error variation is composed of global and local variation. The global component includes design features like, for example, blocks, incomplete blocks, and, if necessary, features related to rows and columns. The local component involves spatial models, where a typical choice is a separable row and column autoregressive process of order 1 (AR1). The inclusion of a spatial model for the error component imposes a structured correlation matrix on the plots. This matrix has a relatively simple structure that can be summarized by a few parameters. For the separable AR1xAR1 process we would require just three parameters; two autocorrelation parameters along the row and column direction, and a scaling parameter.

An alternative approach to model the spatial component of plot error variation is to add parameters, one for each plot, constrained by a penalty that enforces spatial smoothness. This model was pioneered by Green et al. (1985), who reported good results in a one-dimensional setting. They also sketched how to extend their model to two dimensions, but they did not actually implement it. Here we report our experiences in two dimensions and we discuss the advantages offered by recent developments (P-splines and array algorithms) for large-scale spatial smoothing.

We illustrate our ideas on a wheat experiment laid out on a 15 rows by 10 columns grid of plots. The experiment consisted of six full replicates,

where each replicate contained 25 varieties (genotypes), whose performance needed to be evaluated.

For the analysis, we define a model with the following components:

- 25 random effects for the varieties;
- 15 random row effects;
- 10 random column effects;
- 150 random spatial effects.

The total number of parameters is much larger than the number of observations, but penalties on the random effects (corresponding to components of variance in mixed model jargon) reduce the effective model dimension enormously. To the 150 spatial parameters difference penalties are applied (details below), while the other random effects are implemented as ridge penalties on parameters.

## 2 Theory

In one dimension the model can be summarized as

$$y = X\alpha + v + e, \quad (1)$$

where  $X = [X_{ij}]$  is the design matrix assigning the effects of global trends (i.e. blocks) and genotypes to observations;  $\alpha = [\alpha_j]$  contains the corresponding model parameters,  $v = [v_i]$  represents the spatial trend and  $e$  additive noise. To estimate the parameters we minimize

$$S = \sum_i (y_i - \sum_j x_{ij}\alpha_j - v_i)^2 + \lambda \sum_i (\Delta^d v_i)^2 \quad (2)$$

$$= \|y - X\alpha - v\|^2 + \lambda \|D\alpha\|^2, \quad (3)$$

where  $\Delta^d v_i$  denotes the  $i$ th component of the vectors of  $d$ th order differences of  $v$ . To simplify the discussion we initially, and quite unrealistically, assume only fixed effects for  $\alpha$ . In practice all or most factors will be treated as random.

An attractive property of this model is that the strength of the spatial correlation is modeled by one parameter,  $\lambda$ . By writing (3) as

$$S^* = \sum_i (y_i - \sum_j x_{ij}\alpha_j - v_i)^2 / \sigma^2 + \sum_i (\Delta^d v_i)^2 / \tau^2 \quad (4)$$

$$= \|y - X\alpha - v\|^2 / \sigma^2 + \|Dv\|^2 / \tau^2, \quad (5)$$

we recognize a mixed model structure, where  $\sigma^2$  is the variance of the additive noise and  $\tau^2$  the variance of  $d$ th order difference contrasts.

In a two-dimensional setting with  $m$  rows and  $n$  columns the data and the trend are contained in  $m$  by  $n$  matrices  $Y$  and  $V$ . The spatial penalty is

$$\text{Pen}_s = \lambda \|D_m V\|_F + \lambda \|V D'_n\|_F. \quad (6)$$

Here  $D_k$  is a  $k-d$  by  $k$  matrix that forms  $d$ th order differences of a vector of length  $k$ , and  $\|\cdot\|_F$  denotes the Frobenius norm (the sum of the squares of all elements of a matrix).

Unfortunately there is no elegant notation to describe regression models for two-dimensional data, so we have to vectorize  $Y$  and  $V$  (column-wise) to the vectors  $y$  and  $v$  of length  $N = mn$ . The model now becomes

$$y = Z_g b_g + Z_c b_c + Z_r b_r + v + e, \quad (7)$$

where  $Z_g$  is the design matrix for the genotypes and  $Z_c$  ( $Z_r$ ) the design matrix for the column (row) effects.

In addition we have to introduce Kronecker products to describe the spatial penalty. Let

$$P_s = \lambda I_n \otimes D'_m D_m + \lambda D'_n D_n \otimes I_m, \quad (8)$$

with  $I_k$  the  $k$  by  $k$  identity matrix. To model the other random effects, ridge penalties are used:

$$\text{Pen} = \kappa_d \|b_d\|^2 + \kappa_c \|b_c\|^2 + \kappa_r \|b_r\|, \quad (9)$$

where  $\kappa_k = \sigma^2/\tau_k^2$ , where  $\tau_k^2$  is the (unknown) variance of the distribution of  $b_k$ .

Let  $P$  be the matrix that, given the variances, contains all the penalty matrices for the spatial and random components and let  $C = [Z_g : Z_c : Z_r : I_N]$  and  $a' = [\alpha' : b'_g : b'_c : b'_r]$ . Then we solve

$$(C'C + P)[\hat{a}' : \hat{v}'] = C'y \quad (10)$$

to find the coefficients and spatial component. Here  $P$  is a block-diagonal matrix representing the penalties. The components of the diagonal are:  $\kappa_g I_g$ ,  $\kappa_c I_c$  and  $\kappa_r I_r$  for the penalties on genotype, column and row effects in (9) and finally  $P_s$  as defined in (8).

With  $\hat{y} = Z\hat{a} + \hat{v}$  we update the variances by

$$\tilde{\sigma}^2 = \|y - \hat{y}\|^2 / (N - \text{ED}_0), \quad \tilde{\tau}_k^2 = \|\hat{b}_k\|^2 / \text{ED}_k \quad (11)$$

for the rows, columns and varieties. The variance of the spatial contrasts is estimated as

$$\tilde{\tau}_s^2 = \hat{v}' P_s \hat{v} / \text{ED}_s. \quad (12)$$

Here  $\text{ED}_k$  stands for the effective dimension of model component  $k$  (with subscript  $s$  for the spatial component). It is computed as the sum of that part of the trace of  $G = (C'C + P)^{-1} C'C$  that corresponds to the columns

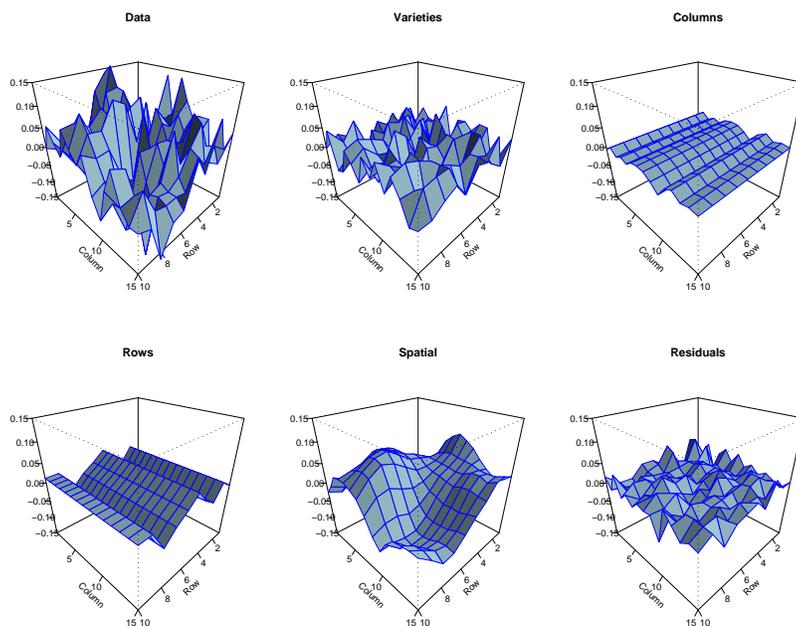


FIGURE 1. Slate Hall wheat trial. Data, model components and residuals, as indicated by the captions.

in  $C$  that represent a component of the model. We repeatedly solve the penalized regression equations and re-estimate the variances, until convergence.

Starting values for the variances of the parameters for varieties row and column effects generally are not hard to set, because they can be based on field experience. For the variance of the spatial differences a small starting value, say  $10^{-6}$ , is advisable.

### 3 An application

Figure 1 shows data from the “Slate Hall” wheat trial (Kempton et al., 1994) on a 15 by 10 grid of plots and the fit of the model. Figure 2 shows the estimated model components in more detail. Second order differences were used in the spatial penalty.

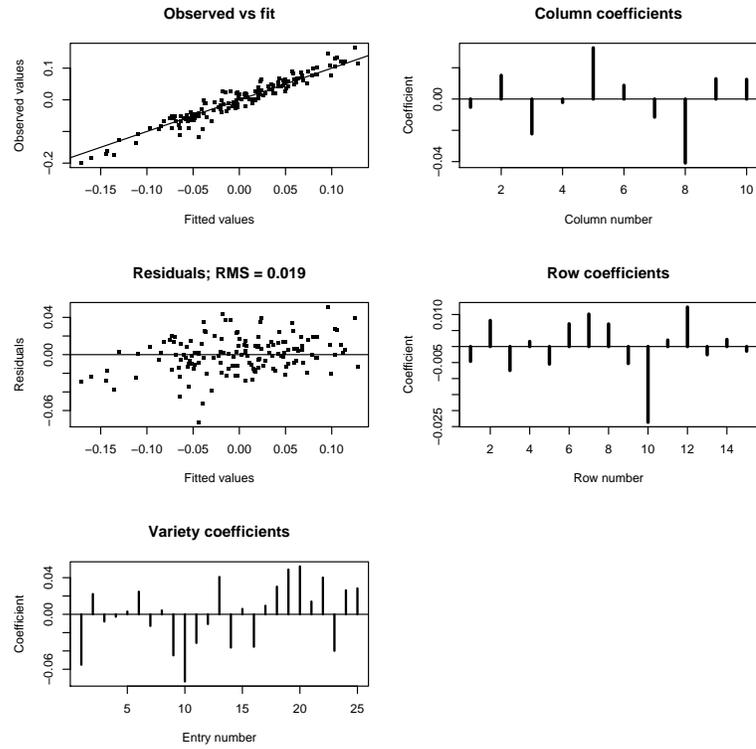


FIGURE 2. Slate Hall wheat trial. Model results, as indicated by the captions.

## 4 Large-scale applications

The number of parameters is  $200 = 25$  (genotypes) +  $15$  (rows) +  $10$  (columns) +  $150$  (spatial). For modern computers this is not a large model, but the size can grow rapidly with larger trials. The spatial parameters form the majority. They can be reduced in number by the use of tensor products of B-splines. With one B-spline knot at every three or four plots, the number of spatial parameters is reduced 10 times.

Array algorithms for tensor-product models avoid the inefficient procedure of first forming the Kronecker products and then computing their inner products (Eilers et al., 2006; Currie et al., 2007). On large data sets this can speed up computations by another order of magnitude.

Efficient use of memory and CPU time also becomes an issue when analyzing multiple trials, sharing crop varieties, but each having their own spatial effects.

### References

- Currie, I.D., Durban, M. and Eilers, P.H.C. (2006) Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B* **68**, 259–280.
- Eilers, P.H.C., Currie, I.D. and Durban, M. (2006) Fast and Compact Smoothing on Multi-dimensional Grids. *Computational Statistics and Data Analysis* **50**, 61–76.
- 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**, 269–293.
- Green, P., Jennison, C. and Seheult, A. (1985) Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society, Series B*, **47**, 299–315.
- Kempton, R.A., Seraphin, J.C. and Sword, A.M (1994) Statistical Analysis of Two-Dimensional Variation in Variety Yield Trials. *Journal of Agricultural Science, Cambridge* **122**, 335–342.

# Bayesian estimation of random effects models for multivariate responses of mixed data

Helga Wagner<sup>1</sup> and Regina Tüchler<sup>2</sup>

<sup>1</sup> Johannes-Kepler-University Linz

<sup>2</sup> University of Economics and Business Administration, Augasse 2-6,  
1090 Vienna, Austria,  
e-mail: [regina.tuechler@wu-wien.ac.at](mailto:regina.tuechler@wu-wien.ac.at) (corresponding author)

**Abstract:** In longitudinal studies often different response variables are measured repeatedly for one subject. In this paper we present a random effects model to estimate multivariate data that consist of normal and count responses. The random effects are included to account for within subject dependencies. Auxiliary mixture sampling leads to a Gibbs sampling type scheme which is easy to implement. The method is illustrated by transaction data of a consumer cohort acquired by an apparel retailer.

**Keywords:** Auxiliary mixture sampling; Generalized linear models; Mixed data types; MCMC; Random effects model

## 1 Random Effects Model for Mixed Data

Let  $\mathbf{Y} = (Y^n, Y^c)'$  denote a bivariate response variable, where the first component is normal and the second is Poisson count. We observe a normal component  $y_{it}^n$  and a count component  $y_{it}^c$  for  $i = 1, \dots, N$  subjects and  $t = 1 \dots T_i$  time points per subject  $i$ .

The sequence of observations of the two components for subject  $i$  are denoted by  $\mathbf{y}_i^n$  and  $\mathbf{y}_i^c$ , respectively. The vector of all observations of subject  $i$ ,  $\mathbf{y}_i$ , is obtained by stacking the two sequences,  $\mathbf{y}_i = ((\mathbf{y}_i^n)', (\mathbf{y}_i^c)')'$ . Note that  $\mathbf{y}_i$  is a vector of  $T_i \cdot 2$  values.

To relate the mean  $\mu_{it}^c = E(y_{it}^c)$  to the linear predictor  $\eta_{it}^c$  we introduce a log-link-function

$$\mu_{it}^c = \exp(\eta_{it}^c),$$

for the Poisson component. We use the identical link

$$\mu_{it}^n = \eta_{it}^n$$

for the normal component  $y_{it}^n$  and assume a constant variance  $y_{it}^n \sim \mathcal{N}(\mu_{it}^n, \sigma^2)$ . We include the following random effects specification for the linear predictors

$$\boldsymbol{\eta}_i^n = \mathbf{X}_i \boldsymbol{\beta}_i^n \quad \text{and} \quad \boldsymbol{\eta}_i^c = \mathbf{X}_i \boldsymbol{\beta}_i^c,$$

where  $\mathbf{X}_i$  is a design matrix of dimension  $T_i$  times  $d$  and  $\beta_i^n$  and  $\beta_i^c$  are normally distributed random effects.

We assume that the same covariates are used for both response components, whereas the random effects are allowed to differ between the components. Dependency between repeated measurements is described by the random effects  $\beta_i^n$  and  $\beta_i^c$  shared for all measurements of one response component. To take into account dependency between the components we assume that for each subject the random effects follow a multivariate normal distribution

$$\beta_i = ((\beta_i^n)', (\beta_i^c)')' \sim \mathcal{N}_{2d}(\beta, \mathbf{Q}),$$

where  $\mathbf{Q}$  is an unknown variance-covariance matrix. Note that assuming pairwise independence between the random effects

$$\text{Cov}(\beta_i^n, \beta_i^c) = \mathbf{0} \quad i = 1, \dots, N$$

would correspond to separate modeling of the two components using a linear random effects model for the normal and generalized random effects model for the count component.

## 2 Data Augmentation

To obtain a Gibbs sampling type scheme for the model two steps of data augmentation are needed for the count response part. We use the ideas of Frühwirth-Schnatter and Wagner (2006) and Frühwirth-Schnatter et al. (2007) who specified an auxiliary mixture sampler for models with count data. This method was also presented at the 22nd IWSM (Wagner et al. 2007). Mixed discrete and continuous data are estimated within a ML-framework for binary data by Gueorguieva and Agresti (2001) and for Poisson data by Yang et al. (2007).

We regard the count responses  $y_{it}^c$  as the number of jumps of an unobserved Poisson process with intensity  $\mu_{it}^c$  in the time interval  $[0,1]$ . In a *first data augmentation step* the inter-arrival time between the last jump before and the first jump after  $t = 1$ , denoted by  $\tau_{it,1}$  and for  $y_{it}^c > 0$  the *arrival* time of the last jump before 1, denoted by  $\tau_{it,2}$  are introduced as missing data. As  $\tau_{it,1}$  is distributed as  $Exp(\mu_{it}^c)$  and  $\tau_{it,2}$  follows the  $\Gamma(y_{it}^c, \mu_{it}^c)$  distribution, the original Poisson regression model can be transformed into the linear model

$$\begin{aligned} -\log \tau_{it,1} &= \mathbf{x}_{it} \beta_i^c + \varepsilon_{it,1}, \\ -\log \tau_{it,2} &= \mathbf{x}_{it} \beta_i^c + \varepsilon_{it,2}, \end{aligned}$$

where the distribution of  $\varepsilon_{it,1}$  is a type I extreme value distribution and  $\varepsilon_{it,2}$  is distributed as the negative logarithm of a Gamma random variable with integer shape parameter  $\nu = y_{it}^c$ . For  $y_{it}^c = 0$  we are dealing only with  $\tau_{it,1}$ .

To obtain a model that is conditionally Gaussian, these non-normal densities can be approximated by a mixture of normal components

$$p_\varepsilon(\varepsilon; \nu) = \frac{\exp(-\nu\varepsilon - e^{-\varepsilon})}{\Gamma(\nu)} \approx \sum_{r=1}^{R(\nu)} w_r(\nu) f_N(\varepsilon; m_r(\nu), s_r^2(\nu)),$$

see Frühwirth-Schnatter et al. (2007) for details. The number of components  $R(\nu)$  as well as the weights  $w_r(\nu)$ , means  $m_r(\nu)$  and variances  $s_r^2(\nu)$  depend on  $\nu$ , but are fixed for a given  $\nu$ . Therefore we only have to introduce the component indicators  $r_i$  for each auxiliary observation in the *second data augmentation step*. The non-normal, non-linear Poisson regression model reduces to the linear Gaussian model

$$-\log \tau_{it,j} = \mathbf{x}_{it} \boldsymbol{\beta}_i^c + m_{r_{it,j}} + \varepsilon_{r_{it,j}}, \quad \varepsilon_{r_{it,j}} \sim N(0, s_{r_{it,j}}^2).$$

We stack the elements  $-\log \tau_{it,j}$  and  $m_{r_{it,j}}$  for each subject to obtain vectors  $\tilde{\mathbf{y}}_i^c$  and  $\mathbf{m}_i$ , respectively.

After data augmentation we combine the normal and the count part of the model and obtain a normal linear model:

$$\tilde{\mathbf{y}}_i = \begin{pmatrix} \mathbf{y}_i^n \\ \tilde{\mathbf{y}}_i^c - \mathbf{m}_i \end{pmatrix} = \begin{pmatrix} \mathbf{X}_i & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{X}}_i \end{pmatrix} \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,$$

where  $\tilde{\mathbf{X}}_i$  is chosen to match  $\tilde{\mathbf{y}}_i^c$  and the errors are normally distributed  $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$  with diagonal matrix  $\boldsymbol{\Sigma}_i$ . The first  $T_i$  entries of  $\boldsymbol{\Sigma}_i$  correspond to the normal response  $\mathbf{y}_i^n$  and are equal to  $\sigma^2$ , whereas the remaining entries are equal to the variances  $s_{r_{it,j}}^2$ .

### 3 Prior

We assume conditionally conjugate priors for the model parameters:  $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0)$ ,  $\sigma^2 \sim \text{InvGamma}(c_0, C_0)$  and  $\mathbf{Q} \sim \text{InvWishart}(s_0, \mathbf{S}_0)$ .

### 4 MCMC

The MCMC scheme to sample the model parameters  $\boldsymbol{\beta}$ ,  $\mathbf{Q}$ ,  $\sigma^2$ , the random effects  $\boldsymbol{\beta}_i$  the indicators  $\mathbf{R} = \{r_{it,j}\}$ , and the inter-arrival times  $\boldsymbol{\tau} = \{\tau_{it,j}\}$  consists solely of standard densities.  $\mathbf{y}^n, \mathbf{y}^c, \tilde{\mathbf{y}}^c, \tilde{\mathbf{y}}$  summarize the observations and the auxiliary data for all subjects  $i$ .

1. Sample  $\boldsymbol{\beta}$  from the multivariate normal distribution  $p(\boldsymbol{\beta} | \tilde{\mathbf{y}}, \mathbf{R}, \mathbf{Q}, \sigma^2)$ .
2. Sample  $\boldsymbol{\beta}_i$  for each subject  $i$  from the multivariate normal distribution  $p(\boldsymbol{\beta}_i | \tilde{\mathbf{y}}, \mathbf{R}, \boldsymbol{\beta}, \mathbf{Q}, \sigma^2)$ .
3. Sample  $\mathbf{Q}$  from the inverted Wishart distribution  $p(\mathbf{Q} | \boldsymbol{\beta}_i, \boldsymbol{\beta})$ .

4. Sample  $\sigma^2$  from the inverted Gamma distribution  $p(\sigma^2|\mathbf{y}^n, \boldsymbol{\beta}_i)$ .
5. Sample the augmented data  $\tilde{\mathbf{y}}^c$  conditional on  $\mathbf{y}^c, \boldsymbol{\beta}_i^c$ .
6. Sample  $\mathbf{R}$  from the discrete density  $p(\mathbf{R}|\tilde{\mathbf{y}}^c, \boldsymbol{\beta}_i)$ .

## 5 Application

Our data come from an apparel retailer who collected information about the buying behaviour of 2157 costumers in his Austrian stores. In our study we investigated the profit values and the number of different items purchased. These two response variables are related to 14 covariates. These covariates include for example the number of shopping trips in the time period, the spendings on promotion, the spendings on weekend, and different product categories purchased, like e.g. accessories, trousers, suits,... . The goal from the marketing point of view is to learn about the importance of marketing mix variables, like promotions or mailings. Since typical data consist of mixed response variables our new method is especially valuable for these kind of studies.

In our analysis we we obtained a substantial random effects correlation between the two response variables for all our covariates. Linking the two response variables leads to an improved understanding of the buying behaviour.

## 6 Conclusions

In this paper we model bivariate responses of a normal and a count outcome. The two outcomes are linked by a random effects specification. As auxiliary mixture sampling is easily implemented for categorical data (Frühwirth-Schnatter and Frühwirth 2007) extended sampling schemes for mixed data with normal, count and categorical data are readily available. Since the number of covariates is typically large variable selection is of special interest for applications in marketing. New variable selection methods are easily available via auxiliary mixture sampling and may be included in our method (Frühwirth-Schnatter and Tüchler 2008, Tüchler 2008).

## References

- Frühwirth-Schnatter, S. and Frühwirth, R. (2007). Hierarchical generalized linear models. *Computational Statistics and Data Analysis*, **51**, 3509–3528.
- Frühwirth-Schnatter S., Frühwirth R., Held L. and Rue H. (2007). Improved Auxiliary Mixture Sampling for Hierarchical Models of Non-Gaussian Data. *IFAS Research Report*, <http://www.ifas.jku.at>.

- Frühwirth-Schnatter, S. and Tüchler, R. (2008). Bayesian Parsimonious Covariance Estimation for Hierarchical Linear Mixed Models. *Statistics and Computing*, **18**, 1–13.
- 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.
- Gueorguieva, R. and Agresti, A. (2001). A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses. *Journal of the American Statistical Association*, **96**, 1102–1112.
- Tüchler, R. (2008). Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling. *Journal of Computational and Graphical Statistics*, **17**, 76–94.
- Wagner, H., Tüchler, R. and Frühwirth-Schnatter, S. (2007). Auxiliary Mixture Sampling for Non-normal data. In *Proceedings of the 22nd IWSM*, Barcelona, July 2–6, 591–596.
- Yang, Y., Kang, J. and Zhang, J. (2007). Regression Models for Mixed Poisson and Continuous Longitudinal Data. *Statistics in Medicine*, **26**, 3782–3800.

# The Dragnet Test: A New Approach to Choosing Between Models

Paul Wilson<sup>1</sup>

<sup>1</sup> Department of Mathematics, National University of Ireland, Galway  
Paul.Wilson@nuigalway.ie

**Abstract:** Traditional log-likelihood based methods for choosing between models, be they nested or non-nested, all concentrate on log-likelihoods evaluated at the maximum likelihood estimates of the model parameters. The true model parameters may in fact differ considerably from their maximum likelihood estimates. We propose a method that examines the relative fits of the models over a cross-section of likely parameter values; this method is based upon testing simple hypotheses, and hence avoids pitfalls associated with compound null hypotheses such as biased estimation of  $p$ -values.

**Keywords:** Model Discrimination, Hypothesis Testing, Cox’s test, nested models, non-nested models, hybrid test.

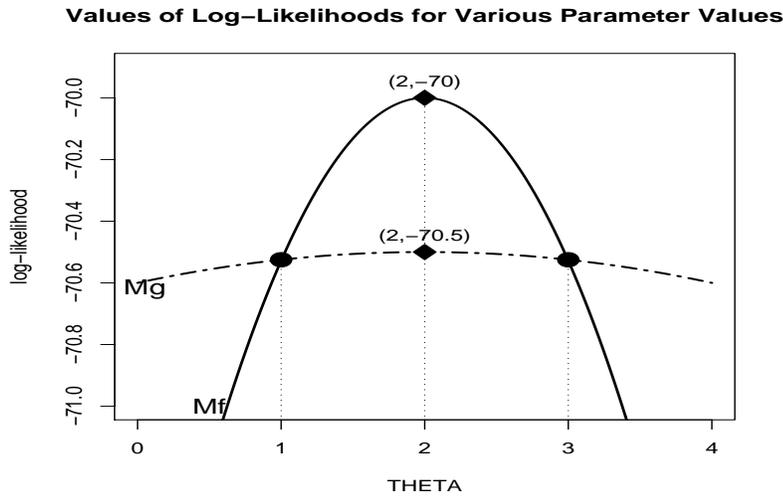
## 1 Introduction

In statistical analysis the substitution of a maximum likelihood estimator,  $\hat{\tau}$ , for the true, but unknown, parameter  $\tau_0$  of a model for given data, is so commonplace that possible consequences of the fact that the values of  $\tau_0$  and  $\hat{\tau}$  may differ considerably are often overlooked.

Consider the situation illustrated in Figure 1, which plots, for some (fictional) data, the log likelihoods of two models  $M_f$  and  $M_g$ , where both models are functions of the same single-valued parameter,  $\theta$ . We see that, for both models, the maximum-likelihood estimate of  $\theta$  occurs at  $\hat{\theta} = 2$ , thus as  $\ell_{M_f}(2) > \ell_{M_g}(2)$   $M_f$  is to be preferred over  $M_g$  *assuming that  $\hat{\theta}$  is the “true” value of  $\theta$* . Say the “true” value of  $\theta$  is 0.9, we see from Figure 1 that  $\ell_{M_g}(0.9) > \ell_{M_f}(0.9)$ , and hence, for  $\theta = 0.9$ ,  $M_g$  is to be preferred to  $M_f$ . Thus, whilst  $M_f$  is clearly the “better” model if it is highly likely that the true parameter value lies between 1 and 3, it is far from clear which model is “better” if there is a reasonable chance that the true parameter value lies outside of this interval.

Here, when we say that we “prefer”, say,  $M_f$  to  $M_g$  at  $\theta = \theta_*$ , we simply mean that the log likelihood of the former, evaluated at  $\theta_*$  is greater than that of the latter, (also evaluated at  $\theta_*$ ). For any test to be of practical use we need to determine criteria that determine whether we may reject:

FIGURE 1. Possible log likelihood Values



$$H_0 : M_g(\theta_*) \text{ is a suitable model for the data} \quad (1)$$

against:

$$H_1 : M_f(\theta_*) \text{ is a suitable model for the data} \quad (2)$$

The criteria we adopt for the dragnet test are basically those of the standard Cox test for non-nested models, (Cox (1962)): we reject the null hypothesis if the observed log likelihood ratio is inconsistent with what would be expected if the null hypothesis were true, rejection being possible both towards and away from the alternative hypothesis. We then reverse the hypotheses, and repeat the procedure. This results in two  $p$ -values, one for each null hypothesis, from which we may classify the models as illustrated in Table 1. Thus, unlike conventional log-likelihood based methods, or score tests, which merely determine if one model is significantly better than another, not whether it is suitable, the Cox test determines whether one or other, either, or both of the two models under consideration are appropriate. This desirable property is also incorporated into the dragnet test.

The test proposed in Cox (1962) was analytic. Following Williams (1970) and Hinde (1992), who proposed simulation based analogues of Cox's test, the various  $p$ -values of the dragnet test are estimated by bootstrap methods. Whereas the Cox test only evaluates "inconsistency" at the maximum likelihood estimate of the parameters of the models concerned, the dragnet test

TABLE 1. Possible outcomes of Cox's test

		$H_0 : M_f$ is the true model		
$p$ -value		<i>small</i>	<i>medium</i>	<i>large</i>
$H_0 : M_f$	<i>small</i>	Neither	$M_f$	Neither
$H_0 : M_g$	<i>medium</i>	$M_g$	Both	$M_g$
	<i>large</i>	Neither	$M_f$	–

evaluates it at  $S$  possible *fixed* parameter values determined by sampling from the parameter spaces of both models, thus obtaining a weighted cross section of possible parameter values. With regard to the testing of models at fixed parameters, the dragnet test may be viewed as an extension of the hybrid test proposed in Wilson (2007), which could be regarded as a dragnet test where the dragnet consists solely of the maximum likelihood estimate of the model parameters. Hence, as the hypotheses of the dragnet test are simple, i.e. they specify the parameters of  $M_f$  and  $M_g$ , problems with bias estimation of  $p$ -values are avoided. (See Wilson (2008)). This fixing of parameters also enables Cox's method to be extended to nested or overlapping models.

## 2 Example: Zero-Inflated Poisson versus Poisson Models

The dragnet test, with and  $S = 1,000$  was used to analyse the random sample of data summarised in Table 2. This sample was drawn from  $ZIP(0.1, 2)$  data.

Value	0	1	2	3	4	5	6	$\geq 7$	Total
Count	9	8	13	9	9	0	2	0	50

When a zero-inflated Poisson model is fitted to these data, parameter estimates  $\hat{\gamma} = 0.100$  and  $\hat{\lambda} = 2.423$  are obtained. A score test returns a  $p$ -value of 0.078, not enabling the rejection of  $H_0 : Poisson$  at  $\alpha = 0.05$ . Table 3 describes the overall classification, at  $\alpha = 0.05$ , by a  $ZIP$  dragnet test (i.e. where the cross-section of parameter values used to determine the dragnet assumes a  $ZIP$  distribution), and a Poisson dragnet test.

We see that the  $ZIP$  dragnet favours the  $ZIP$  model to the Poisson at 0.780 of likely parameter values, and indicates that if a zero modifiedinflated

TABLE 3. Classification of the Table 2 data,  $\alpha = 0.05$ 

$S = 1,000$	ZIP	Poisson	Both	Neither
ZIP dragnet	0.780	0.018	0.110	0.092
Poisson Dragnet	0.188	0.070	0.025	0.217

Poisson distribution is not suitable, then probably neither is a Poisson distribution. If we examine the classification of the Poisson dragnet not only is there is little support for the Poisson model, but there is no particular support for any classification. Overall, the evidence appears to support the ZIP model, but is not conclusive.

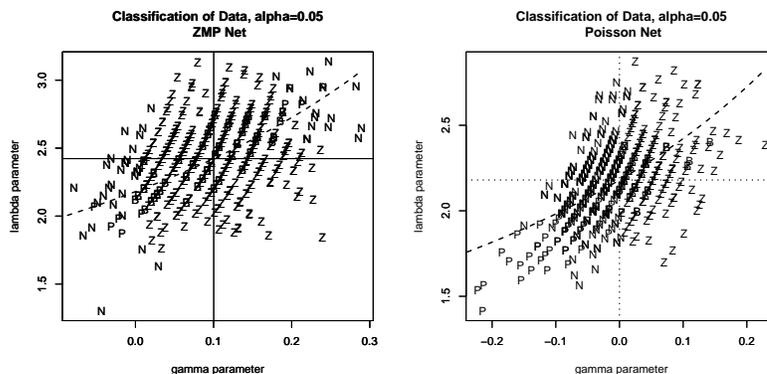
### 2.1 Dragnet Classification Diagrams

The *dragnet classification diagrams* of Figure 2 illustrates the classification, at  $\alpha = 0.05$ , of the data at the various parameter values. Letter Z's indicate a "ZIP" classification, P's a "Poisson" classification, B's a classification of "both", N's a "neither" classification. The "intermingling" of "ZIP" and "both" classifications in some regions, and of "Poisson" and "neither" classifications in others indicates that these classifications are borderline in these regions. The solid black lines correspond to the maximum likelihood estimators for the ZIP model, and the dotted black lines to those of the Poisson. We see that in the vicinity of the ZIP maximum likelihood estimator for the ZIP model the data is classified as ZIP/both, (indicating that  $H_0 : ZIP$  is not rejected in this vicinity, but that for  $H_0 : Poisson$ ,  $p \approx 0.05$  and hence  $H_0 : Poisson$  is borderline accepted or rejected) whereas in the vicinity of the maximum likelihood estimators of the Poisson model all four possible classifications occur, indicating that *both*  $H_0 : Poisson$  and  $H_0 : ZIP$  are borderline accepted/rejected in the immediate vicinity of the Poisson maximum likelihood estimator. Given that if a model is true one would expect stability in the vicinity of the maximum likelihood estimator for the model, this lends support to the ZIP model. Also, along the dashed line representing the locus of the data mean both Poisson and ZIP classification is supported, except at extreme points. As one would expect support for Poisson classification to be strongest along this line, this is further evidence in favour of the ZIP model.

## 3 A Zero-Inflated Negative Binomial versus a Zero-Inflated Generalised Poisson models

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*

FIGURE 2. ZIP versus Poisson Classification of the Table 2 data for both dragnets at  $\alpha = 0.05$



at 4. Ridout et al. fit various standard and zero-inflated Poisson and negative binomial models to the Trajan data, and show that a zero-inflated negative binomial model where both the mean and the zero-inflation parameters are modelled by *period* fits the data well, with a BIC of 1,271.9, compared to a BIC of 1,283.7 for the zero-inflated Poisson model. An alternative, not considered by Ridout et al., is a *zero-inflated generalised Poisson model* based upon the generalised Poisson distribution:

$$f_Y(y; \mu, \phi) = \frac{\mu(\mu + (\phi - 1)y)^{y-1}}{y!} \phi^{-y} \exp\left(-\frac{1}{\phi}(\mu + (\phi - 1)y)\right) \quad (3)$$

Such a model, (fitted using the R package *ZIGP*, Erhardt (2007)), has a BIC of 1270.0, indicating a slightly better fit than the ZINB model. Table 4 presents the results of zero-inflated generalised Poisson and zero-inflated negative binomial dragnet tests.

TABLE 4. ZIGP versus ZINB Classification of Trajan data, ZIGP net,  $\alpha = 0.05$ .

$S = 100$	ZIGP	ZINB	Both	Neither
ZIGP dragnet	0.56	0.01	0.39	0.04
ZINB dragnet	0.19	0	0.81	0

We see that the ZIGP dragnet tends to prefer a “ZIGP” to a “Both” classification, but not overwhelmingly so, whereas the ZINB dragnet strongly favours a “Both” classification, and interesting, favours a “ZIGP” otherwise. This indicates that if a ZINB model is suitable, then so also is a

ZIGP model, but not necessarily vice-versa. We may conclude that, except possibly at some outlying parameters, the ZIGP model is to be preferred.

## 4 Conclusion

The dragnet test is an exciting new approach to choosing between models. Unlike score-tests or standard (log) likelihood based methods it may be applied to nested, non-nested or overlapping models, and it is not dependent upon the maximum likelihood estimates of the model parameters being close approximations to the true parameters. Unlike analytic or simulation-based Cox tests it is free of bias, but it retains the desirable property of being able to accept or reject both models, as opposed to determining the relative merits of one model in relation to the other.

## Acknowledgement

The author wishes to thank Prof. John Hinde for his constructive criticism and feedback concerning the development of the dragnet test.

## References

- Cox DR. (1962). Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society. Series B* **24**, 406–423.
- Erhardt V. (2007), ZIGP: Zero Inflated Generalized Poisson regression models.  
[www.m4.ma.tum.de/Papers/Czado/Czado-Erhardt-Min-Wagner.pdf](http://www.m4.ma.tum.de/Papers/Czado/Czado-Erhardt-Min-Wagner.pdf)
- Hinde JP. (1992). Choosing Between Non-nested Models: a Simulation Approach. In Fahrmeir L et al. eds. *Advances in Glim and Statistical Modelling: Proceedings of the Glim92 Conference and 7th International Workshop on Statistical Modelling*. New York: Springer.
- Williams DA. (1970), Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures, *Biometrics*, **28**, 23–32.
- Wilson P. (2007). A Hybrid Test for Non-Nested Models. In *Proceedings of the 22nd International Workshop on Statistical Modelling*, Barcelona:Universitat Autònoma de Barcelona.
- Wilson P. (2008). Bias estimation of  $p$ -values in analytic and simulated Cox Tests for non-nested models. In *Proceedings of the 23rd International Workshop on Statistical Modelling*, Utrecht: Universiteit Utrecht.

# Bias estimation of $p$ -values in analytic and simulated Cox Tests for non-nested models

Paul Wilson

<sup>1</sup> Department of Mathematics, National University of Ireland, Galway  
Paul.Wilson@nuigalway.ie

**Abstract:** In this paper we show that the estimation of  $p$ -values in both Cox's test for non-nested models and its simulation based analogues is biased and that whilst simulation based Cox tests may be extended to nested models the consequent level of bias is so large as to render the test useless, but that this bias may be removed by adapting the null hypothesis to be simple.

**Keywords:** Bias, Cox's Test, Hybrid Test

## 1 Introduction

One of the few tests for comparing non-nested models is the analytic test of Cox (1962), simulation based variants of which have been proposed by Williams (1970) and Hinde (1993). Cox's analytic test determines  $p$ -values for  $H_0 : M_f \text{ is the correct model}$  and  $H_0 : M_g \text{ is the correct model}$  from which we may classify the data. Whilst Hinde focusses on model discrimination, he shows that  $p$ -values may be derived from his test by standard procedure. In developing his test Cox makes various approximations, and notes that these introduce bias, which could be reduced if these approximations were not made. Whilst simulation based versions of Cox's test avoid his calculations and thus his approximations, problems arise with the estimation of  $p$ -values as the underlying test statistic,  $T_f$ , the difference between the observed log-likelihood ratio and its expected value, depends upon the parameters of the data in question. Thus the null hypotheses are composite, and biased estimation of  $p$ -values may occur.

## 2 Bias and the Distribution of $p$ values

A  $p$ -value may be defined by (see Davison and Hinkley (1999)):

$$p_{\text{obs}} = Pr_0 (T \geq t_{\text{obs}}) \quad (1)$$

where  $T = t(Y)$  is some function of data,  $Y$ , and  $Pr_0$  indicates probability under the null hypothesis.

Assuming that  $F_0$ , the null distribution function of  $T$  is known and continuous, we may regard  $p_{obs}$  as an instance of a random variable  $\mathcal{P} = 1 - F_0(T)$ , thus, for  $0 \leq \xi \leq 1$ :

$$\Pr_0[1 - F_0(T) \leq \xi] = \Pr_0[T \geq F_0^{-1}(1 - \xi)] = 1 - F_0[F_0^{-1}(1 - \xi)] = \xi \quad (2)$$

i.e. the null distribution of  $\mathcal{P}$  is the uniform distribution. Whilst the above only exactly holds if  $F_0$  is continuous, it still holds to a great extent for discrete  $F_0$ . If  $F_0$  is not known, as frequently occurs when the underlying null hypothesis is composite, then the distribution of the estimated  $p$  values will fail to be uniform, and the interpretation (1) will fail to hold, i.e.  $E(\hat{p}) \neq p'$ , where  $p'$  is the true value of  $p$ , i.e. the estimation process is biased. Reversing the argument of (2) we see that the converse also holds, i.e. if the distribution of the  $p$ -values is uniform, then their estimation is unbiased.

### 3 Bias in Cox's Analytic Test

Cox shows that  $L_f(\alpha) - E\{L_f(\hat{\alpha})\}$  is normally distributed with mean approximately zero and a stated variance term, but proceeds to acknowledge that a correction for bias could be obtained by taking the mean to be

$$-\frac{E_\alpha\{f(f_\alpha^2 + f_{\alpha\alpha})\}}{2E_\alpha(f_\alpha^2)} - \frac{1}{2}, \quad (3)$$

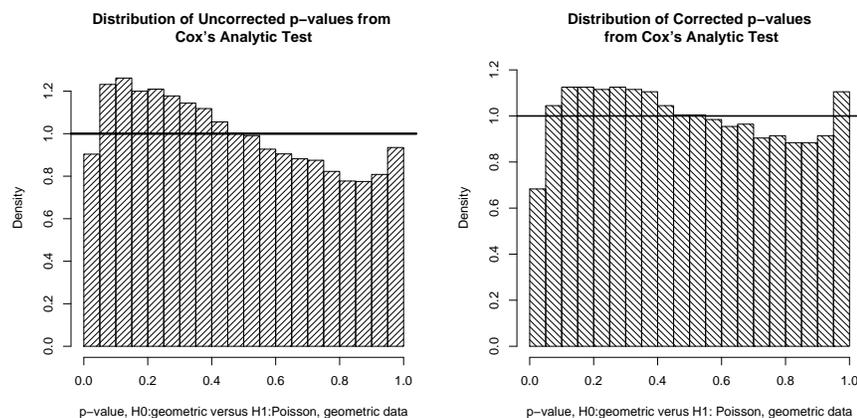
Cox (1962) makes no attempt to derive an analogue of (3) for  $T_f$ . Such an analogue requires the evaluation of complicated integrals which due to the incorporation of terms involving the expectation over  $M_f$  of functions of the "competing model"  $g(y, \beta_\alpha)$ , does not "reduce" to a term of the relative simplicity of (3). We may however determine the departure from zero of the mean of  $T_f$  by simulation. The second column of Table 1 summarises  $E_\alpha(T_f)$  under  $H_0$ :Poisson versus  $H_1$ : geometric, where the data is in fact Poisson; in the third column the data are geometric and the hypotheses are reversed. These expected values are each based upon the results of 100,000 simulations. Clearly the approximation to zero is non-negligible.

Figure 1 illustrates the distributions of unadjusted  $p$ -values and  $p$ -values that have been adjusted by taking the mean to be as in Table 1. These mean values were obtained from repeated application of Cox's test to data drawn from *geometric(1)* data, with the hypotheses  $H_0$ :*geometric* versus  $H_1$ :*Poisson*. We see that the correction reduces, but does not eliminate, bias. Indeed the bias actually increases at the lower tail.

TABLE 1. Expected values of  $T_f$  and  $T_g$ ,  $n = 100$ , based upon 100,000 simulations.

Mean	$H_0$ :Poisson (Poisson Data)	$H_0$ :geometric (geometric data)
0.6	0.15	-0.37
1.0	0.25	-0.63
1.4	0.29	-0.82
1.8	0.32	-1.14
2.2	0.36	-1.24

FIGURE 1. Distribution of uncorrected and corrected  $p$ -values geometric(1) data,  $H_0$ :geometric,  $H_1$ :Poisson,  $n = 100$



### Bias in Simulation-Based Cox Tests

The left hand diagram of Figure 2 illustrates that bias occurs when a simulation based Cox test is used to determine  $p$ -values. To emphasise that this bias is due to the composite nature of the null hypothesis, the right hand diagram shows that if the parameters of the null hypothesis are fixed, the resultant  $p$ -values are uniformly distributed. Whilst the level of bias may be reduced by multiple bootstrapping, this is usually not practical.

It is possible to extend simulation based Cox tests to nested models, however bias in the estimation of  $p$ -values is enormous, as illustrated by Figure 3. The left-most diagram shows the distribution of the unadjusted  $p$ -values, the center diagram that of  $p$ -values adjusted by a double bootstrap, and the right-hand diagram the distribution of the  $p$ -values under the *simple* hypothesis where the model parameters are fixed.

FIGURE 2. Distribution of  $p$ -values Poisson(0.8) data,  $H_0$ :Poisson,  $H_1$ :geometric,  $n = 50$ . Parameters varying and fixed

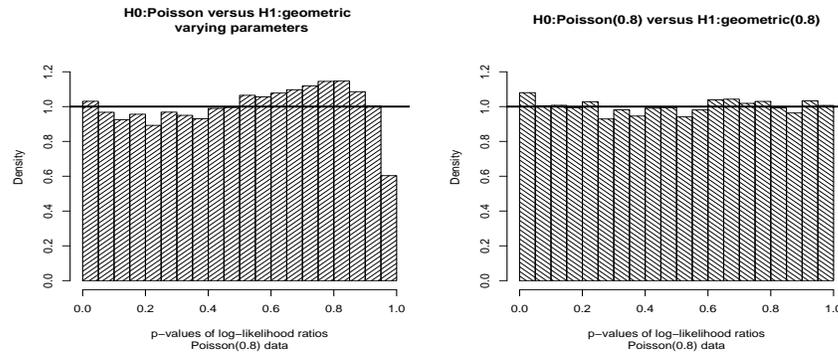
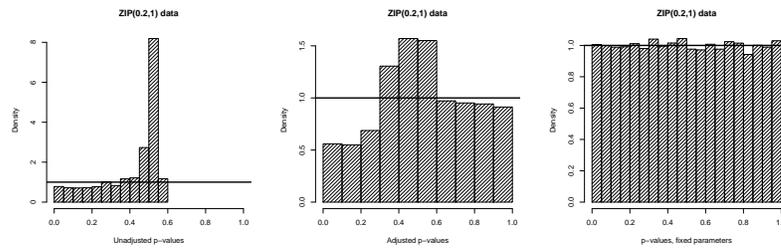


FIGURE 3. Distribution of  $p$ -values ZIP(0.2,1) versus Poisson data,  $H_0$ :Poisson,  $H_1$ :geometric,  $n = 50$ . Parameters varying and fixed



## 4 Conclusion

Bias is evident in all forms of Cox's test. Bias is absent in tests based upon simple null hypotheses incorporating fixed model parameters such as the Hybrid test of Wilson (2007) and the proposed "Dragnet Test", discussed elsewhere at this conference.

**References**

- Cox DR. (1962). Further Results on Tests of Separate Families of Hypotheses. *Journal of the Royal Statistical Society. Series B* **24**, 406–423.
- Davison AC and Hinkley DV (1999), *Bootstrap Methods and their Applications*, 3rd ed. *Cambridge University Press*.
- Hinde JP. (1992). Choosing Between Non-nested Models: a Simulation Approach. In Fahrmeir L et al. eds. *Proceedings of the Glim92 Conference and 7th International Workshop on Statistical Modelling*. New York: Springer.
- Wilson P. (2007). A Hybrid Test for Non-Nested Models. In del Castillo J, Espinal A, Puig P eds. *Proceedings of the 22nd International Workshop on Statistical Modelling*, Barcelona:Universitat Autònoma de Barcelona.

# A physiological application of Bayesian linear regression with a change-point

Jason Wyse<sup>1</sup>

<sup>1</sup> Room 533, School of Mathematical Sciences, Library Building, University College Dublin, Belfield, Dublin 4, Ireland, e-mail:jason.wyse@ucdconnect.ie

**Abstract:** We consider a simple linear regression model with one unknown change-point. Data to measure the gas exchange threshold (GET) in subjects undertaking an incremental exercise test to exhaustion is analyzed. We consider normal models that may be appropriate for this data, including AR(1) errors, and variance change at the change point. A Markov Chain Monte Carlo (MCMC) technique- the Metropolis-Hastings algorithm is used for model estimation.

**Keywords:** Change-point; Metropolis-Hastings algorithm; gas exchange data.

## 1 Introduction

The change-point problem receives considerable attention in the statistics and econometrics literatures. Bayesian MCMC techniques are the basis for some modern approaches to the problem. Some key references for the particular problem we consider are Carlin et al. (1992) and Ferreira (1975). We apply some common Bayesian techniques to analyse data sets collected to measure the GET for three subjects undertaking an incremental exercise test to exhaustion. This data is known to form a two segment linear regression, with a change-point at the GET. The GET may be used as a measure of fitness, and hence as an alternative to more invasive fitness testing procedures, such as measurement of the lactate threshold using a blood sample. A credible interval (CI) for the time at which the GET occurs is of particular use in order to monitor when the real change has occurred.

## 2 Gas exchange data

The onset of metabolic acidosis during exercise can be detected by an increase in the rate of carbon dioxide output ( $\dot{V}CO_2$ ) relative to the rate of oxygen uptake ( $\dot{V}O_2$ ). Breath-by-breath values of  $\dot{V}CO_2$  are plotted against the corresponding breath's value of  $\dot{V}O_2$ . The series of points are related by a simple linear regression with two-regimes. The transition between the two regimes, the GET, is the point of change (on  $\dot{V}O_2$ ) in the rate of carbon dioxide output relative to the rate of oxygen uptake. It makes more sense

physiologically to find the GET on time, rather than trying to find the change point on  $\dot{V}O_2$  directly, as  $\dot{V}O_2$  is not monotone over time. The data analysed here is given a likelihood treatment in Kelly et al. (2004).

### 3 Models and estimation

The time at which each breath was taken was recorded and the breaths were then numbered 1 to  $n$ . The notation is as follows:  $\tau$  will refer to change point in breath number (index) and  $T$  will refer to the time at which the subject took that breath (the time gap between breaths varies). Let the  $\dot{V}O_2$  values be  $\mathbf{x} = (x_1, \dots, x_n)$  and the  $\dot{V}CO_2$  values be  $\mathbf{y} = (y_1, \dots, y_n)$ . Then

$$y_i = \begin{cases} \alpha_1 + \beta_1(x_i - \bar{x}_1) + \varepsilon_i & i = 1, \dots, \tau \\ \alpha_2 + \beta_2(x_i - \bar{x}_2) + \varepsilon_i & i = \tau + 1, \dots, n \end{cases}$$

where  $\bar{x}_1 = \sum_{i=1}^{\tau} x_i / \tau$ ,  $\bar{x}_2 = \sum_{i=\tau+1}^n x_i / (n - \tau)$  and  $(\varepsilon_1, \dots, \varepsilon_n) \sim N_n(\mathbf{0}, \mathbf{\Sigma})$ , an  $n$ -variate normal distribution with mean  $\mathbf{0}$  and variance-covariance matrix  $\mathbf{\Sigma}$ . Let  $\boldsymbol{\theta}$  contain the unknown parameters, e.g.  $\tau, \alpha_1, \beta_1, \dots$ .

The likelihood of  $\mathbf{y}$ ,  $p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ , depends on  $\mathbf{x}$  and  $\boldsymbol{\theta}$ . The posterior density of  $\boldsymbol{\theta}$  is given by the relation

$$\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

where  $\pi(\boldsymbol{\theta})$  denotes the assumed prior density of  $\boldsymbol{\theta}$ . Various structures of  $\mathbf{\Sigma}$  are considered below.

#### 3.1 Various error structures

- $\mathcal{M}_0$ :  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, \dots, n$  with  $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$  for all  $i \neq j$ .
- $\mathcal{M}_1$ :  $\varepsilon_i \sim N(0, \sigma_1^2)$ ,  $i = 1, \dots, \tau$  and  $\varepsilon_i \sim N(0, \sigma_2^2)$ ,  $i = \tau + 1, \dots, n$  with  $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$  for all  $i \neq j$ .
- $\mathcal{M}_2$ :  $\varepsilon_i = \phi\varepsilon_{i-1} + z_i$  and we define  $\varepsilon_1 = z_1$ , where  $z_i \sim N(0, \sigma^2)$ ,  $i = 1, \dots, n$  and  $\text{cov}(z_i, z_j) = 0$  for all  $i \neq j$ .
- $\mathcal{M}_3$ :  $\varepsilon_i = \phi\varepsilon_{i-1} + z_i$  and we define  $\varepsilon_1 = z_1$ , where  $z_i \sim N(0, \sigma_1^2)$ ,  $i = 1, \dots, \tau$  and  $z_i \sim N(0, \sigma_2^2)$ ,  $i = \tau + 1, \dots, n$  where  $\text{cov}(z_i, z_j) = 0$  for all  $i \neq j$ .

Note that  $\mathcal{M}_2$  and  $\mathcal{M}_3$  have AR(1) errors which models the repeated measures nature of the data.  $\mathcal{M}_1$  and  $\mathcal{M}_3$  have a change in variance which models the subjects' breathing becoming less stable and more variable as they approach exhaustion.

We assume independent, ignorant, flat priors for the regression parameters. We also assume  $\phi$  is uniform on  $(-1, 1)$  *a priori* and  $\pi(\tau)$ , the prior for the change-point, is discrete uniform on  $2, \dots, n - 2$ . For  $\mathcal{M}_0$ ,

$$\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) \propto (\sigma^2)^{-\frac{n+2}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\} \pi(\tau)$$

where  $\boldsymbol{\Sigma} = \text{diag}(\sigma^2, \sigma^2, \dots, \sigma^2)$  and  $\mu_i = \begin{cases} \alpha_1 + \beta_1(x_i - \bar{x}_1) & i \leq \tau \\ \alpha_2 + \beta_2(x_i - \bar{x}_2) & i > \tau \end{cases}$ .  
Posterior for other models are derived similarly.

### 3.2 Estimation of the models using the Metropolis-Hastings algorithm

We direct the reader who is unfamiliar with MCMC techniques, and the Metropolis-Hastings algorithm (M-HA) in particular, to Gilks et al. (1996). We used the M-HA to generate samples from  $\pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y})$  in the statistical package R. Post burn-in, we have samples  $\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(N)}$  of the change-point. The approximate posterior of  $\tau$  is

$$\tilde{\pi}(\tau = k|\mathbf{x}, \mathbf{y}) = \sum_{i=1}^N I(\tau^{(i)} = k)/N$$

where  $I(\cdot)$  is the indicator function. Then  $\tilde{\pi}(\tau|\mathbf{x}, \mathbf{y})$  gives the CIs in Table 1 below. CIs are chosen to give good coverage probability and short interval length. Parameter estimates are taken as the mean of the posterior samples of each parameter. In the case of the change-point, the estimate is the closest integer to the mean.

## 4 Results

Estimates and CIs for the change-point are given in Table 1, where  $\Delta$  gives the number of breaths in the interval. Incorporation of AR(1) errors and variance change affects estimates for Subjects 1 and 2, as is seen by comparison of the models. This is not the case for Subject 3; however,  $\mathcal{M}_1$  gave  $\sigma_1^2 = 0.0021$  and  $\sigma_2^2 = 0.0023$  suggesting that  $\sigma_1^2 \approx \sigma_2^2$ . This is most likely due to the particular Subjects' physiological attributes. Overall, interval lengths are short and gave good coverage.

## 5 Final remarks

We have presented a physiological application of Bayesian linear regression with a change-point. It is hoped to extend this analysis to further models in the future.

Model	$T$	CI $T$	$\Delta$	Coverage
Subject 1 $n = 150$				
$\mathcal{M}_0$	240	[232, 246]	6	94.99%
$\mathcal{M}_1$	240	[232, 246]	6	96.50%
$\mathcal{M}_2$	237	[232, 249]	7	91.29%
$\mathcal{M}_3$	195	[186, 200]	7	94.54%
Subject 2 $n = 194$				
$\mathcal{M}_0$	314	[303, 319]	8	97.49%
$\mathcal{M}_1$	316	[305, 321]	8	96.81%
$\mathcal{M}_2$	305	[296, 316]	10	98.92%
$\mathcal{M}_3$	300	[294, 316]	11	93.52%
Subject 3 $n = 206$				
$\mathcal{M}_0$	361	[355, 363]	4	99.09%
$\mathcal{M}_1$	361	[355, 363]	4	99.24%
$\mathcal{M}_2$	361	[352, 361]	4	99.11%
$\mathcal{M}_3$	361	[352, 361]	4	98.48%

TABLE 1. Results of analysis on the three Subjects.  $T$  = time at which breath was taken,  $\Delta$  = number of breaths in the interval

**Acknowledgments:** Special Thanks to SFI for funding this project, Dr. Alasdair Thin for use of his data and my supervisor Dr. Gabrielle Kelly.

## References

- Carlin, B. P., Gelfand, A. E., Smith, A. F. M. (1992). Hierarchical Bayesian Analysis of Change-point Problems. *Applied Statistics*, **41**, 389-405.
- Ferreira, P. E. (1975). A Bayesian Analysis of a Switching Regression Model: Known Number of Regimes. *Journal of the American Statistical Association*, **70**, 370-4.
- Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In: Gilks, W. R., Richardson, S., Spiegelhalter, D. J., eds. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall, 1-19
- Kelly, G. E., Lindsey, J. K., Thin, A. G. (2004). Models for estimating the change-point in gas exchange data. *Physiological Measurement*, **25**, 1425-36.
- R Development Core Team (2007). 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>.

# Reducing Error by Increasing Focus: Multivariate Monitoring of Biosurveillance Data

Inbal Yahav<sup>1</sup> and Galit Shmueli<sup>1</sup>

<sup>1</sup> Department of Decision, Operations and Information Technologies , Smith School of Business, University of Maryland, College Park, MD 20742, USA.  
E-mail: {iyahav,gshmueli}@rhsmith.umd.edu.

*The work was partially supported by NIH grant RFA-PH-05-126.*

**Abstract:** Pandemic outbreaks have threatened worldwide population throughout history, and more so today with increased mobility. Several historic disease outbreaks resulted in extremely high death tolls, yet in some cases special early public health measures reduced disease spread. Early detection of disease outbreaks, therefore, plays a major goal in preventing disease transmission and reducing the size of the affected population. In modern biosurveillance a variety of diagnostic and pre-diagnostic data are monitored. Current detection algorithms suffer from several limitations, causing high false alert rates. One of the main limitations is that they monitor each of the many temporal data streams univariately.

In this work we deal with monitoring multivariate time series of pre-diagnostic daily counts, for the purpose of detecting an *increase* in one or more of these series. We consider two approaches for computing directionally-sensitive Hotelling  $T^2$  charts, and generalize the approaches to obtain directionally-sensitive Multivariate EWMA charts. We illustrate their performance using both large scale simulated data and authentic biosurveillance data. We then show that multivariate monitoring is Pareto efficient compared to univariate.

**Keywords:** Multivariate Control Charts; Sensitivity Analysis; Biosurveillance

## 1 Introduction

In 1918, one of the deadliest Influenza pandemics in history erupted, called *The Spanish Flu*. Approximately 20 to 40 percent of the worldwide population fell ill and over 50 million people died. Outbreaks followed shipping routes from North America through Europe, Asia, Africa, Brazil and the South Pacific. The pandemic reached its peak after 5-6 months (see Figure 1). Nearly 40 years later, in February 1957, the *Asian Influenza* pandemic erupted in the Far East. Unlike the *Spanish Flu*, the *Asian Influenza* pandemic virus was quickly identified and vaccines were available 6 months later. Approximately 2 million people died in this outbreak (compared to

the 50 million in the *Spanish Flu*). Other known outbreaks in history, such as the *Hong Kong Flu* (1968-69), the *Avian Flu* (1997) and *SARS* (2003) also resulted in high death tolls over the years. Unfortunately the threat of new pandemic outbreaks is still looming.

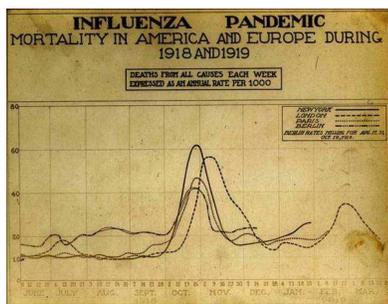


FIGURE 1. Spanish Flu 1918-1920

A major goal of public health is to figure out whether and how transmission of diseases can be diminished. Researchers at the *Center for Humanitarian Logistics* at Georgia Tech ([www.tli.gatech.edu/research/humanitarian/projects.php](http://www.tli.gatech.edu/research/humanitarian/projects.php)) have shown that pandemic outbreak effect can be greatly reduced if quarantine is imposed at the early stages of the disease. The US *Centers for Disease Control & Prevention* (CDC) layout guidelines and strategies for reducing disease transmission, including use of personal protective equipment (e.g., masks and gloves), hand hygiene, and safe work practices. The CDC recommendations also outline actions that might be taken during the earliest stage of a pandemic, when the first potential cases or disease clusters are detected. These include individual-level containment measures such as patient isolation and identification, monitoring, and quarantine of contacts ([www.hhs.gov/pandemicflu/plan/appendixf.html](http://www.hhs.gov/pandemicflu/plan/appendixf.html)).

The early detection of disease outbreaks therefore plays a major role in preventing disease transmission and reducing the size of the affected population. In modern biosurveillance a wide range of pre-diagnostic and diagnostic data are monitored for the purpose of alerting public health officials when there is evidence of a disease outbreak. Unfortunately, current detection methods are still far from achieving their intended purpose. One main reason is that the algorithms used rely heavily on simplifying assumptions about the monitored data that are often not met in practice. These include assumptions of *iid* normal data, the availability of sufficient training data, and knowledge about the outbreak signature in such data.

Another important limitation of current detection algorithms is that they

monitor each data stream univariately, when in practice the number of data streams is usually large. The multiple univariate testing results in an extremely high level of false alerts, and also ignores sources of multivariate information. A major feature of biosurveillance data is multiplicity in several dimensions. A multiplicity of data sources (e.g., hospitals, pharmacies, nurse hotlines, and outpatient clinics); multiple locations (e.g., multiple hospitals in a certain region), a variety of diseases of interest, and multiple series from a single source (e.g., sales of remedies for different symptoms). In contrast to current univariate temporal monitoring, we illustrate the power of monitoring multiple series that arrive from a single data source, or from multiple data sources in a multivariate fashion.

One of the central tools in classic disease surveillance is the statistical control chart. Statistical control charts have been a central tool in classic disease surveillance and have also migrated into modern biosurveillance. An alternative to using multiple univariate control charts is to use multivariate control charts, which have traditionally been used in industry for monitoring multiple series simultaneously. This alternative helps avoid the multiple testing phenomenon (It is known that the combination of  $p$  independent tests, each with significance level  $\alpha$ , has a total significance level equal to  $1 - (1 - \alpha)^p$ ). Furthermore, multivariate control charts take advantage of the correlation structure between individual series, thereby having a higher potential of detecting small signals that are dispersed across series. The challenge that arises, however, is that in the biosurveillance context we are only interested in detecting an *increase* in one or more of the daily counts that indicate disease-related behavior. In other words, we need directionally-sensitive multivariate control charts. Several such charts have been derived, and among them only a few are computationally feasible.

In this paper, we tackle the challenge of directional-sensitivity in multivariate control charts, and compare and evaluate the practical usefulness of several multivariate monitoring methods for detecting outbreak signatures in multivariate biosurveillance-type data. We focus on two approaches that are useful for practical implementation: Follmann (1996) provides a correction for the ordinary Hotelling chart (Hotelling, 1947) and Testik and Runger (2006) present a quadratic-programming approach to estimate the in-control mean vector. These two approaches yield directionally-sensitive Hotelling charts. We describe each of these in detail and then generalize them to obtain directionally-sensitive Multivariate Exponentially Weighted Moving Average (MEWMA) charts. Using a large array of simulated data, we compare the performance of the directionally-sensitive Hotelling and MEWMA charts (and a variation that includes a restart) as a function of the number of monitored series, the cross-correlation structure, and the amount of training data required for estimating the covariance matrix. We then evaluate the robustness of the charts to underlying assumptions of normality and independence. Finally, we apply the methods to a set of authentic biosurveillance data.



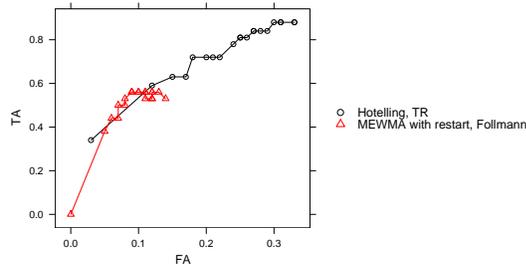


FIGURE 3. True vs. false alert rates for TR's Hotelling chart vs. Follmann's MEWMA chart with restarts

suggested by Burkom, Murphy, and Shmueli (2007).

We compute the false alert (FA) rate by applying each control chart to the preprocessed data. To evaluate the true alert (TA) rate and time to detection, we inject 32 spikes of magnitude  $o$ , where  $o \sim u[1, 4] \times \bar{\sigma}$ , into a random subset of the 13 series. The covariance structure is estimated from the first year of data. Applying each chart to the data with injected signatures we compute its FA and TA rates. We find that: (1) Follmann's MEWMA chart with restart alerts the least, whether or not there are outbreak signatures, and (2) Testik & Runger's (TR) Hotelling chart is most sensitive: it has the highest TA rate, but also the second highest FA rate. The highest FA rate is obtained with TR's MEWMA, but this is likely due to the lack of restart after an alert. Note also that the FA rate computed before and after the signature injections are similar.

To further explore these results and the relationship between TR's Hotelling and Follmann's MEWMA (with restart), we examine their performance across a range of FA rates ( $[0, 0.2]$ ). For a higher sensitivity comparison, we examine only outbreaks of smaller magnitude ( $o \sim u[0.5, 2.5] \times \bar{\sigma}$ ). Results are shown in Figure 3. We see that the low FA rate is controlled by Follmann's MEWMA chart and the high TA rate is controlled by TR's Hotelling chart. These results are inline with those obtained from the simulated data. The conclusion is therefore that the choice of chart should be driven by the tradeoff between *actual* true and false alerts required by the user. We also observe that the FA rate set by the theoretical threshold and the actual FA rate can be quite different. This phenomenon is illustrated clearly in Figure 3, where both methods were set to have the same FA rate but they end up with different actual FA rates. This result is not surprising, given that the theoretical threshold determination is based on normality and independence assumptions. Both assumptions tend to be violated in authentic biosurveillance data. In the simulated study we show that the

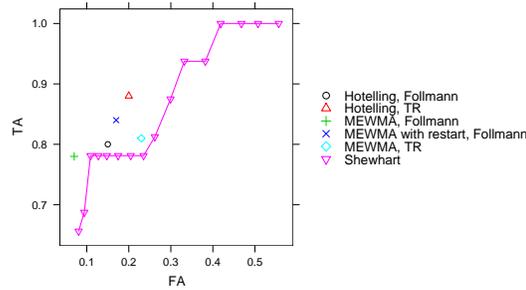


FIGURE 4. True vs. false alert rates; comparing multivariate control charts with multiple-univariate Shewhart charts

actual FA rate increases in the presence of true alert. This result is more pronounced in the MEWMA charts.

Finally, to evaluate the advantages of each of the four multivariate control charts, we compare them to combined univariate monitoring where univariate Shewhart charts are applied simultaneously to each of the 13 series. The rule for alerting is when at least one of the charts alerts (this is the rule currently used in biosurveillance systems). We vary the *actual* FA rate of the multiple-univariate charts between  $[0, 0.6]$  and observe the TA rate. Results are shown in Figure 4. We can see that all of the multivariate charts are Pareto efficient compared to the multiple-univariate Shewhart.

## References

- Burkom, H., Murphy, S., and Shmueli, G (2007). Automated Time Series Forecasting for Biosurveillance. In: *Statistics in Medicine*, **26(22)**, 4202-4218.
- Follmann, D. (1996) A Simple Multivariate Test for One-Sided Alternatives. In: *Journal of the American Statistical Association*, **91**, 434
- Hotelling, H. (1947) Multivariate quality control-illustrated by the air testing of sample bombsights. In: *Techniques of Statistical Analysis*, 111-184.
- Testik, M. C. and G. C. Runger (2006) Multivariate one-sided control charts. In: *IIE Technometrics*, **38**, 635-645.
- Yahav, I. and G. Shmueli (2008) Evaluating Directionally-Sensitive Multivariate Control Charts with an Application to Biosurveillance. Available online: *SSRN*, *abstract id=1119279*

# Index

- Aerts M., 89, 125  
Agrawal M., 287  
Anzai T., 93  
Armero C., 99  
Artacho A., 99
- Böckenholt U., 188  
Bailey T.C., 216  
Baio G., 104  
Bartolucci F., 109  
Beekman M., 265  
Belitz C., 115  
Bhulai S., 281  
Black K., 252  
Blangardio M., 104  
Boer J.M., 349  
Boetzer M., 349  
Boland F., 121  
Bollaerts K., 89, 125  
Boomsma D.I., 281  
Boone I., 89  
Bowman A., 226  
Braumann C.A., 232  
Brentnall A.R., 131  
Brinkhuis M.J.S., 137  
Brion V., 143  
Burnett R., 143
- Cadarso-Suárez C., 317  
Cakmak S., 143  
Callegaro A., 265  
Camarda C.G., 149  
Canas Rodrigues P., 155  
Canto e Castro L., 250  
Carita A.I., 159  
Cecere S., 163
- Chua S.J., 173  
Coffey N., 178  
Conde S., 184  
Connolly J., 252  
Crowder M.J., 131  
Cruyff M.J.L.F., 188, 255, 271  
Currie I., 194
- Dalrymple M., 277  
de Menezes R.X., 349  
de Rooi J.J., 204  
Dejardin D., 200  
Dewulf J., 125  
Dietz E., 380  
Djuraš G., 364  
Do Ha I., 327  
Donoghue O., 178  
Draghicescu D., 210  
Durbán M., 311
- Economou T., 216  
Eilers P.H.C., 149, 204, 238, 335, 370,  
386, 416, 432  
Einbeck J., 402  
Endo T., 93  
Espinal A., 221, 422
- Faes C., 125  
Filipe P., 232  
Fodde R., 349  
Frank L.E., 204  
Freguson C., 226
- Gampe J., 149, 238, 335  
Gaspar C., 349  
Gilchrist R., 244  
Goldberg M., 143

Gomes D., 250  
Gonçalves L., 406  
Grijspeerdt K., 125  
Gultekin T., 143  
  
Hand D.J., 131  
Harrison A.J., 178  
Hawkins M.J., 252  
Hayes K., 178  
Hendrick T.A.M., 255  
Hessen D. J., 259, 410  
Hoad D., 402  
Houwing-Duistermaat J.J., 265  
Hox J.J., 299  
Hubregtse M., 271  
Hudson I.L., 277, 287, 396  
  
Jonker M.A., 281  
  
Kamara A., 244  
Kapelán Z., 216  
Keatley M.R., 287  
Kim S.W., 287  
Klingenberg B., 293  
Korendijk E.J.H., 299  
Kou C., 305  
  
López-de-Ullibarri I., 317  
López-Quílez A., 99  
Lang S., 115  
Lee D.-J., 311  
Lesaffre E., 163, 200  
Ligthart R.S.L., 281  
Lima A.T., 155  
Lynch J., 321  
  
Maas C.J.M., 299  
MacKenzie G., 184, 321, 327  
Majumdar A., 354  
Mallick R., 143  
Malosetti M., 432  
Maris G., 137  
Marques P.L., 159  
Martín N., 331  
Marx B.D., 335  
Matthews F.E., 426  
  
Mejza I., 345  
Mejza S., 341, 345  
Meulman J.J., 370  
Mexia J., 341  
Mintiens K., 89, 125  
Mitra I., 354  
Moerbeek M., 299  
Monteira Braga D., 392  
Muggeo V.M.R., 360  
  
Neubauer G., 364  
Nunes C., 250  
  
Pan J., 305  
Pereira D., 341  
Posthuma D., 281  
Puig P., 221, 422  
  
Rau R., 335  
Rea A., 277  
Ribbens S., 89  
Rippe R.C.A., 370  
Rocha Chellini P., 169  
Rodríguez-Álvarez M. X., 317  
Rudge J., 244  
  
Savvala I., 376  
Scheufele R., 380  
Schnabel S.K., 386  
Shimadzu H., 93  
Shmueli G., 458  
Sieswerda M., 349  
Siqueria A.L., 169, 392  
Slagboom E., 265  
Sleep J., 396  
Sofroniou N., 402  
Solis-Rapala I.L., 109  
  
Tüchler R., 438  
Teles J., 406  
Tijmstra J., 410  
  
Uh H.-W., 416  
  
Valero O., 422  
van Buuren S., 376

van den Hout A., 426  
van der Heijden P.G.M., 188, 255,  
410  
Van der Stede Y., 89  
van der Vaart A.W., 281  
van Eeuwijk F., 432  
van Houwelingen H.C., 265  
van Ommen G.-J., 349  
Verbeke G., 200  
Vieira F., 159  
  
Wagner H., 438  
Watkins A.J., 173  
Wilson P., 443, 448  
Wyse J., 454  
  
Yahav I., 458