Linking Community Air Pollution to Mortality Using Spatial Random Effects Survival Regression Models

Richard Burnett[1,2,7], Renjun Ma[2], Michael Jerrett[3], Mark S. Goldberg[4,5], Sabit Cakmak[1], Arden Pope III[6], Daniel Krewski[2,7]

[1] Healthy Environments and Consumer Safety Branch, Health Canada
[2] Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa
[3] School of Geography and Geology and Institute of Environment and Health, McMaster University
[4] Department of Medicine, McGill University
[5] Joint Departments of Epidemiology, Biostatistics, and Occupational Health, McGill University
[6] Economics Department, Brigham Young University, Provo, Utah
[7] Institute of Population Health, University of Ottawa

Abstract: In 1997, the United States Environmental Protection Agency promulgated new regulations for annual average concentrations of fine particulate matter in ambient air, based, in part, on the somewhat controversial epidemiological evidence that people who lived in areas with elevated particulate levels have elevated mortality rates. This paper addresses one of the most important issues in this controversy, the statistical analyses of the data. We present a new space-time model linking spatial variation in ambient air pollution to mortality. The model incorporates risk factors measured at the individual level, such as smoking, and at the spatial level, such as air pollution. We demonstrate that the spatial autocorrelation in community mortality rates, an indication of not fully characterizing potentially confounding risk factors to the air pollution mortality association, can be accounted for through the inclusion of location in the model assessing the effects of air pollution on mortality. We present a statistical approach that can be implemented using widely available statistical computer software. Our methods are illustrated with an analysis of the American Cancer Society cohort to determine whether all cause mortality is associated with concentrations of sulfate particles.

Keywords: air pollution; cohort; epidemiology; mortality; spatial regression; sulfate particles; survival

1 Introduction

In 1997, the United States Environmental Protection Agency (USEPA) promulgated new regulations for fine particulate matter in ambient air. This decision was based, in part, on the evidence that American citizens had an increased risk of cardiopulmonary mortality if they lived in areas with elevated ambient fine particles as compared to individuals who resided in less polluted areas. Two of the key studies considered by the USEPA in this regard were that of Dockery and colleagues[1] who used data from the Harvard Six-cities study and Pope and colleagues[2] who used data obtained from the American Cancer Society Cancer Prevention II Study (ACS)[3]. A number of criticisms of these two studies[4] have been largely addressed in an extensive reanalysis[5] conducted at the request of the Health Effects Institute, Cambridge, MA.

In both of these cohort studies[1,2], subjects were enrolled from communities with different levels of outdoor air pollution. Subject-specific information on factors such as age, gender, race, health status, tobacco use, alcohol consumption, diet, occupational exposures, education, and residence history were collected by the use of an interview and questionnaire. Subjects were followed over time to assess changes in their health and vital status. Air pollution was measured by fixed-site monitors either prior to enrollment or during follow-up, or both.

The standard Cox proportional hazard model used in these two studies to relate longevity to exposure, assumed that event information (time of death or censoring due to end of study or loss to follow-up) was statistically independent among subjects after controlling for available information on subject-specific mortality risk factors. Such an approach results in at least two, somewhat related concerns. First, health responses can cluster by location[7]. Clustering will cause a positive correlation of the response of subjects in the same location and thus suggests that location is a risk factor or that there are one or more unmeasured or inadequately modeled risk factors specific to the location itself. If this clustering is independent across locations, failure to account for these "random effects" should not result in biased estimates of effect but can lead to an understatement of the uncertainty in these estimates[8,9].

On one hand, clustering may not be entirely independent or random across locations, so that the data are spatially autocorrelated. That is, even after controlling for various subject-specific risk factors, responses of subjects living in communities close together may be more similar than responses of subjects living in cities farther apart. Failure to account for this type of spatial autocorrelation can also lead to misstatement of the uncertainty of the effect estimates[8,9]. Furthermore, if this spatial autocorrelation is due to missing or systematically mis-measured risk factors that are also spatially autocorrelated, then the estimates could be biased. The direction and size of the bias will depend upon the direction and degree of spatial

autocorrelation between the missing risk factors. For example, if there is an important mortality risk factor that is negatively spatially associated with particulate air pollution but missing from the model, then the mortality estimates for particulate air pollution will be biased downward, and the converse is also true. Just as importantly, if the missing risk factor is not spatially associated with particulate air pollution then the estimate will not be biased, nor will this cause spatial autocorrelation in the residuals of the model.

In this paper we present a new statistical approach to deal with these two related methodologic concerns. We present a space-time random effects survival model that links spatial variation in concentrations of ambient air pollution to longevity of cohort subjects, after controlling for temporal effects and individual risk factors for mortality. We will use data from the original ACS study[2] to demonstrate the impact of modeling random location effects and spatial autocorrelation on the estimated air pollution-mortality association and estimates of uncertainty. These results are compared with those obtained using standard methods of survival analysis assuming statistical independence among subjects.

2 The Space-Time Model

The response data, $T^{(l)}$, is the follow-up time defined as the length of time (calendar or age) from the time of enrollment into the study to the time of death or censoring (due to termination of study or loss to follow-up), for a subject in the l^{th} strata. Strata are typically defined by individual characteristics such as gender and age at enrollment. Mortality risk factor information is available at both the individual level, denoted by the vector $\mathbf{x}^{(l)}(t)$ which may vary with time t, and at the spatial level, $\mathbf{z}(s)$, where s denotes an area in space. The purpose of the analysis is to estimate the association between spatial risk factors and longevity, after controlling for relevant individual level risk factors such as smoking and occupation. Spatial risk factors include ambient air pollution, weather, and indicators of the socio-economic status of the community. For the type of epidemiological studies considered here, spatial areas are typically defined in terms of census boundaries, such as metropolitan statistical areas (MSAs).

We propose to analyze these data using a space-time stochastic model which is characterized by the instantaneous probability of death at time t, or hazard function, for a subject residing in area s and a member of stratum l. The hazard for our model is defined by

$$h_0^{(l)}(t) \exp\left\{\mathcal{J}(s) + \boldsymbol{\beta}^\top \mathbf{x}^{(l)}(t) + \boldsymbol{\gamma}^\top \mathbf{z}(s) + \eta(s)\right\}. \tag{1}$$

Here, $h_0^{(l)}(t)$ is the baseline hazard function for the l^{th} strata, $\mathcal{J}(s)$ is the two-dimensional term to account for residual spatial variability, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients linking individual risk factors to the

hazard function, and $\boldsymbol{\gamma}$ is a vector of unknown regression coefficients linking the spatial level risk factors to the hazard function. Covariate information modulates the baseline hazard function with the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ representing the logarithm of the relative risk of death per unit change in the individual and spatial covariates, respectively.
The spatial random effects, $\eta(s)$, or frailties, are shared by all individuals in area s. These random effects reflect the difference between the observed hazard function and the hazard function predicted from a statistical model. We assume that the spatial process $\eta(s)$ has zero expectation, variance $\theta > 0$, and correlation matrix $\boldsymbol{\Omega}(\boldsymbol{\rho})$ with dimension equal to the number of unique observations in space, which is characterized by a vector of unknown correlation parameters $\boldsymbol{\rho}$. The autocorrelation of the random effects between two areas can be modeled by their distance apart, or some other characteristic of their locations. The term "autocorrelation" is used because we are dealing with correlation in the same variable at different distances in space. This process is similar to serial autocorrelation in time series models. Autocorrelation models typically assume that closer locations will have values of the random effects that are more similar than random effect values for locations farther apart. Thus, these models are often characterized by functions that decrease monotonically with distance[10]. Distance alone may not fully describe the correlation structure. Distant communities with similar population sizes, densities, economic activity, and cultural traits may in fact be more alike than more proximal areas. In the absence of prior knowledge about processes that cause spatial autocorrelation, distance-based relationships provide a useful and reasonable metric for operationalizing autocorrelation[11].
Variation at the spatial level (θ) suggests that there is some unexplained (unmeasured or not appropriately modeled) information on mortality at the individual or spatial level. Thus, space (or place location) can be considered a risk factor for survival.
Spatial autocorrelation can be induced in non-infectious health outcomes as a consequence of spatial autocorrelation in mortality risk factors. As a first step, both spatial variation and autocorrelation can be accounted for by individual or spatial risk factors that vary in space. Evidence of spatial autocorrelation in the residuals of the model may indicate the need to account for additional risk factors, which may potentially exert a confounding effect on the air pollution mortality association. An alternate approach to modeling this additional risk factor information, which may be difficult to implement, is to minimize the potential confounding bias arising from spatial contiguous variation by including a term that represents spatial trends $\mathcal{J}(s)$. With large units of analysis such as metropolitan areas, the total impact of these potentially numerous risk factors may vary in a relatively smooth manner over space. Spatial de-trending can remove autocorrelation between geographic areas. In this approach, location and other covariates, such as air pollution, which also vary in space, compete in the regression

model to predict mortality. Thus, the regression coefficients give the effect of these variables adjusted for each other. This approach is analogous to that used in time series studies of mortality and air pollution in which temporal trends in daily mortality rates are jointly modeled with air pollution levels[12].

3 Statistical Estimation and Inference

3.1 The Time-Domain Model

We decompose the estimation procedure into two domains: time and space. In the time domain we consider the hazard model

$$h_0^{(l)}(t) \exp \left\{ \sum_{s=1}^{S-1} \delta(s) I(s) + \boldsymbol{\beta}^\top \mathbf{x}^{(l)}(t) \right\} \tag{2}$$

where $\{I(s), s = 1, \ldots, S-1\}$ are indicator variables taking the value 1 if the subject resides in area s and zero otherwise. One area (S) is (arbitrarily) assigned as a reference. The unknown parameters $\{\delta(s), s = 1, \ldots, S - 1\}$ represent the logarithm of the relative risk of death for those subjects living in area s compared to those subjects in the reference area S, after controlling for the individual risk factors $\mathbf{x}^{(l)}(t)$

Our primary interest focuses on the regression and dispersion parameters, rather than on the shape of the baseline hazard function. In this approach, a procedure has been selected in which the baseline hazard is treated as a nuisance parameter, which need not be parametrically specified or estimated. This approach underlies the familiar class of Cox survival models[6]. We obtain estimates of the area specific parameters, denoted by $\{\widehat{\delta}(s)\}$, and estimates of their statistical uncertainty using the Cox proportional hazards estimation routine available in the statistical computing software package SAS[13].

A limitation of this procedure is that the uncertainty of the estimate of the reference area is not defined. Because these values are based on comparisons with the same reference area, they are correlated, and thus increases the estimated uncertainty in the location-specific log-relative risks $\{\widehat{\delta}(s)\}$. The induced correlation can be removed by methods developed by Easton and colleagues[14]. This procedure eliminates the covariance between the $\{\widehat{\delta}(s)\}$ and defines an associated estimate of uncertainty to the assigned value of zero for $\widehat{\delta}(S)$. If the covariance terms among the $\{\widehat{\delta}(s)\}$ are identical, taking the value c, for example, the adjusted variance is obtained by subtracting c from the unadjusted variance, with the adjusted variance of $\widehat{\delta}(S)$ assigned the value c. The algebra and computer programming effort to implement this adjustment procedure is greatly simplified if the condition of constant covariance of the $\{\widehat{\delta}(s)\}$ holds. A practical consequence of using this procedure is that we are able to use standard statistical computer software for

statistical estimation and inference in the space-domain model. We denote the adjusted statistical estimation uncertainty in the $\{\widehat{\delta}(s)\}$ by $\{\nu(s)\}$.

3.2 The Space-Domain Model

The space-domain model takes the form

$$\widehat{\delta}(s) = \mathcal{J}(s) + \boldsymbol{\gamma}^{\top}\mathbf{z}(s) + \eta(s) + \varepsilon(s) \tag{3}$$

where $\varepsilon(s)$ is a random process with zero expectation, uncorrelated in space, and with variance $\nu(s)$, independent of the spatial random effects process $\eta(s)$. Here, $\widehat{\delta}(s)$ has expectation $\mu(s) \equiv \mathcal{J}(s) + \boldsymbol{\gamma}^{\top}\mathbf{z}(s)$ and variance covariance matrix

$$\boldsymbol{\Sigma} = \theta\boldsymbol{\Omega}(\boldsymbol{\rho}) + \mathbf{V} \tag{4}$$

where $\mathbf{V}$ is a diagonal matrix with entries $\nu(s)$. We have decomposed the variance into a term representing between subject variation within the same area, $\nu(s)$, and variation between areas, θ.

A practical limitation of this error model is that no commercially available software accommodates this stochastic structure (equation 4) when $\boldsymbol{\rho} \neq \mathbf{0}$. We can remove much of this spatial autocorrelation by a judicious choice of the spatial surface $\mathcal{J}(s)$. We consider non-parametric smoothed estimates of $\mathcal{J}$ using the robust locally-weighted regression (LOESS) smoothers[15] within the generalized additive model framework[16]. This method can be implemented in the statistical computing software package S-Plus[17]. The unknown parameter vector $\boldsymbol{\gamma}$ linking the spatial risk factors to the hazard function is also estimated using generalized additive models in S-Plus.

For the case $\boldsymbol{\rho} = \mathbf{0}$, estimation of the space-domain model can proceed by defining a weight function equal to the inverse of the variance of each observation (i.e. $[\theta + \nu(s)]^{-1}$). However, using this approach requires that an estimate, $\widehat{\theta}$, of θ be obtained. Such an estimate is given by the sample variance of the random effects, $S^{-1}\sum_{s=1}^{S}\eta(s)^2$. However, the random effects $\{\eta(s)\}$ are not known and have to be estimated from the data by the iterative procedure[18]

$$\widehat{\eta}(s)^{(\omega+1)} = \frac{\widehat{\theta}^{(\omega)}}{\widehat{\theta}^{(\omega)} + \nu(s)}[\widehat{\delta}(s) - \widehat{\mu}^{(\omega)}(s)], \tag{5}$$

where ω represents the current value of the parameters and $\omega+1$ represents the updated value. Substituting these estimates of the random effects into the sample variance yields a biased estimate of θ (expectation of estimator does not equal true value) because of the statistical uncertainty in the estimated random effects. An unbiased estimator of θ is given instead by the iterative procedure[18]

$$\widehat{\theta}^{(\omega+1)} = \widehat{\theta}^{(\omega)} + S^{-1}\sum_{s=1}^{S}\left\{[\widehat{\eta}^{(\omega)}]^2 - \frac{[\widehat{\theta}^{(\omega)}]^2}{\widehat{\theta}^{(\omega)} + \nu(s)}\right\}, \tag{6}$$

where the last term in the above equation is a bias correction representing the variance of the estimator of the random effects. The estimation procedure is as follows. First, estimate the unknown parameters in the space-domain model (equation 3) using the generalized additive model (GAM) estimation routing in S-Plus with weights specified by $\nu(s)^{-1}$, yielding an initial prediction function $\widehat{\mu}^{(0)}(s)$. Then determine a starting value for $\widehat{\theta}$ by the formula

$$\widehat{\theta}^{(0)} = \frac{\sum_{s=1}^{S} \left\{ [\widehat{\delta}(s) - \widehat{\mu}^{(0)}(s)]^2 \nu(s)^{-2} - \nu(s)^{-1} \right\}}{\sum_{s=1}^{S} \nu(s)^{-2}} \tag{7}$$

which is the penalized least squares estimator of θ using a Fishers scoring algorithm[18] with mean $\widehat{\mu}^{(0)}(s)$ and variance $\nu(s)$. We then obtain updated estimates of the random effects $\eta(s)$ and their variance θ using equations 5 and 6, respectively. Given the current estimate of the random effects variance we obtain updated estimates $\widehat{\mu}^{(\omega+1)}(s)$ using the GAM estimation routine with weights $[\widehat{\theta}^{(\omega)} + \nu(s)]^{-1}$. This procedure is repeated until the relative difference between consecutive estimates of θ is small (in our case $< 10^{-4}$). Estimates of the other parameters will not change if $\widehat{\theta}$ does not change.

The last issue that needs for be addressed is that the variances of $\widehat{\gamma}$ are biased. This is because the GAM estimation routine in S-Plus assumes a variance structure of the form $\sigma^2[\theta + \nu(s)]$, and provides an estimate of σ^2. In contrast, our model assumes a variance of $\theta + \nu(s)$. An unbiased estimate of the standard error of $\widehat{\gamma}$ can be obtained by dividing the standard error provided by the S-Plus routine by the square root of the estimate of σ^2.

The approach described above yields unbiased and fully efficient estimates of the unknown parameters within a generalized estimating equation framework[19] if there is in fact no spatial autocorrelation in the random effects. We have developed a simple method to judiciously select the appropriate span in the LOESS smoother so as to minimize the autocorrelation structure of the random effects. We do this by plotting the correlation of the standardized estimates of the random effects

$$\widehat{\eta}(s) \left(\frac{\widehat{\theta}^2}{\widehat{\theta} + \nu(s)} \right)^{-\frac{1}{2}} \tag{8}$$

versus the distance between areas using the correlogram function in the spatial module of S-Plus[20]. We have standardized the random effects based on their estimation error to meet the assumption of constant variance needed for this procedure. We also determine the spatial autocorrelation of adjacent communities using Moran's I statistic also available in the spatial module of S-Plus. Two areas are considered to be adjacent, or nearest neighbors, if their respective Thiessen polygons share coterminous boundaries. A Thiessen polygon is an area surrounding a location such that all

points within the polygon are closer to the specified location than any other location in the spatial coverage.
We examine the sensitivity of the air pollution association with mortality, the random effects variance, spatial autocorrelation of adjacent communities, and the relation of the spatial autocorrelation with distance between communities to the complexity of the specification of the spatial surface, as measured by the span of the LOESS nonparametric smoother.
Our modeling approach is illustrated with an analysis of the ACS data in the next section.

4 The American Cancer Society Study of Air Pollution and Mortality

Volunteers of the ACS enrolled over 1.2 million people in September of 1982 throughout the United States. Information on history of disease, demographic characteristics, and mortality risk factors was obtained from respondents. Vital status was monitored through the end of 1989.
We obtained information on particulate sulfate levels from the Aerometric Information Retrieval System (AIRS) and the Inhalable Particle Network (IPN) for 1980 and 1981 for 144 Metropolitan Statistical Areas (MSAs) in which ACS subjects were enrolled. Sulfates are secondarily formed particulate aerosols originating from sulfur dioxide emissions and are a major component of fine particulate matter. The sulfate data from AIRS was collected using glass fiber filters, which react in the presence of sulfur dioxide and artifactually inflate the sulfate concentration. The sulfate data obtained from the IPN used teflon filters which are not subject to this artifact problem. Both monitoring networks were operating in 41 MSAs. We calibrated the AIRS sulfate data to the IPN sulfate data using six linear regression models with separate calibrations for three regions of the county and two time periods [April-September and October to March][(5)]. Estimates of exposure were obtained by averaging all available sulfate data from all monitors located in a MSA for the years 1980 and 1981, inclusive. We examined the association between concentrations of sulfate particles and longevity in 144 MSAs for white members of the ACS cohort, totaling 509,292 subjects. The mean age at enrollment was 56.7 years, 5% of subjects were younger than 40 years, 5% were older than 75 years, and 56.3% of subjects were women. During the course of the seven years of follow-up, 39,474 (7.8%) subjects died. The mean concentration of sulfate particles, corrected for the sulfur dioxide artifact, across all 144 cities was 6.4 $\mu g/m^3$, with a minimum value of 1.4 $\mu g/m^3$, an interquartile range of 4.2 $\mu g/m^3$, and a maximum value of 15.6 $\mu g/m^3$.

The first step in the analysis was to use the Cox proportional hazards survival model (equation 2) to identify all relevant individual covariates that were associated with mortality, independent of the city in which subjects

FIGURE 1. Non-parametric smoothed surface of mortality by latitude and longitude, adjusted for individual level covariates in American Cancer Society Study with smoothing parameter of 40 percent (panel a). Non-parametric smoothed surface of particulate sulfate concentrations by latitude and longitude with a smoothing parameter of 40 percent (panel b). Note, z-axis represents residuals from generalized additive model.

lived ($\delta(s) \equiv 0$). As indicated above, this assumes that all observations were statistically independent. The baseline hazard function was stratified by sex and 5-year age groups so that the nuisance baseline hazard functions were estimated separately in each stratum. Twenty risk factors were selected including variables representing tobacco and alcohol consumption, body mass index, education, martial status, passive exposure to tobacco smoke, and exposure to some air toxics[5]. We then added a set of indicator variables, $\{I(s), s = 1, \ldots, S-1\}$, for each MSA with Greenville, South Carolina, assigned the role as the reference area. [Greenville had a sulfate concentration near the median value.] The associated logarithm of the area-specific relative risks $\{\delta(s)\}$ (relative to Greenville) were estimated using the Cox model, adjusted for individual covariates. Then the variances of the $\{\widehat{\delta}(s)\}$ were adjusted by the methods of Easton and colleagues[14]. We

used the simplified version of the method because the covariances of the $\{\widehat{\delta}(s)\}$ were nearly identical.

In the next step, we visualized the spatial association between mortality and sulfate particles using our space-domain model (equation 3). Here, we regressed the area-specific adjusted relative risks $\{\widehat{\delta}(s)\}$ onto the (x, y) coordinates defined by longitude and latitude of the 144 MSAs with a non-parametric smoothed spatial surface $\widehat{\mathcal{J}}$ (Figure 1, panel a), excluding spatial covariate information such as air pollution (i.e. $\mathbf{z}(s) \equiv \mathbf{0}$) using the GAM. We use latitude and longitude for this visualization step since these co-ordinate definitions are more easily interpretable than the Cartesian (x, y) co-ordinate specification. However, we use the Cartesian co-ordinates in all other formal statistical analyses since the examination of spatial autocorrelation usually relies on Euclidian rather than angular distance measures. This procedure produced a three-dimensional surface of $\{\widehat{\delta}(s)\}$ based on our space-domain model, after adjusting for all individual level risk factors. The weighting function $\{\widehat{\theta} + \nu(s)\}^{-1}$ was used in this step so that the estimated spatial surface $\widehat{\mathcal{J}}(s)$ reflected the estimated uncertainty in the data.

We found that adjusted mortality was elevated in the Ohio Valley region south of Lake Erie, diminished in the west and south, and moderately elevated in the mountain states. We also used equation 3 to model concentrations of sulfate particles but with no random effects. The $\{\widehat{\delta}(s)\}$ were replaced by the mean sulfate concentrations for the 144 MSAs, with the weights assigned to unity. The sulfate concentration surface was also modeled by a LOESS smoother using the GAM. Modeled sulfate values centered by their mean concentration are portrayed in panel b of Figure 1. There is a corresponding elevation in concentrations of sulfate particles in the Ohio Valley region, with much lower concentrations in the west. However, sulfate particles were also elevated all along the eastern seaboard, a pattern not found in the analysis of relative mortality risks. This visualization stage suggests, however, that there is a positive association between the two surfaces.

We then fit a space-domain model with no spatial predictors and determined the standardized random effects from this model. The association between the autocorrelation of these standardized estimated random effects (equation 8) and distance is graphically presented in Figure 2 (panel a) using the correlogram function in the Spatial Module of S-Plus[20]. Autocorrelation peaks at a value of 0.40 for communities 100km apart, declines for distances under 1000km, then increases for distances between 1000km and 1200km. No autocorrelation pattern with distance is apparent for communities greater than 1200km apart. This pattern could be due to the two mortality peaks (see Figure 1, panel a). Communities located in regions of elevated (diminished) mortality are 500-1200km away from communities in regions with diminished (elevated) mortality. The inclusion of sulfate

FIGURE 2. Correlation of standardized estimates of random effects by distance between locations for space-domain model with no covariates (panel a), sulfate only (panel b) and sulfate and location with smoothing parameter of 80 percent to 20 percent (panels c-i, respectively). Horizontal line indicates zero values.

particulate matter into the space-domain model dampens the autocorrelations (Figure 2, panel b) but the pattern over distance remains the same compared to the autocorrelation pattern observed using a model with no spatial predictors. Thus sulfate concentrations account for some, but not all, of the spatial autocorrelation. Further inclusion of a non-parametrically estimated surface with LOESS spans of 80, 70, 60, 50, 40, 30 and 20 percent (Figure 2, panels c-i respectively) reduces the autocorrelation as the span of the LOESS smoother decreases. [Estimates of starting values for θ were negative for spans less than 20 percent, indicating the spatial surface was overfitting the data.] However, the pattern with distance is similar for all spans.

The sensitivity of the air pollution association with mortality, γ, the random effects variance, θ, and the spatial autocorrelation of adjacent communities to the LOESS smoothing span are given in Table 1 for the space-time model. The association between sulfates and mortality decreases as the complexity of the surface modeling increases (or decreasing span). The residual variation between mortality rates, θ, in addition to the spatial autocorrelation also decrease with increasing modeling complexity.

5 Discussion and Conclusions

In previous studies using longitudinal cohort designs, statistically significant associations between mortality and combustion-related particulate air

pollution as measured by fine or sulfate particles have been observed[1,2,21]. There are two related concerns about these studies that are directly addressed in this paper. The first concern is that in these studies the data were analyzed using the standard Cox proportional hazard survival model, with the implicit assumption that the observations were statistically independent after controlling for available information on mortality risk factors[6]. If the assumption of statistical independence is not valid, the uncertainty in the estimates of effect may be understated[7,8,9]. The second concern is that missing or systematically mis-measured risk factors that may be correlated with air pollution could confound the pollution-mortality association[4].

With regards to the first concern, our space-time model provides more accurate estimates of the uncertainty of estimates of effect. Based on the analysis of the ACS data, while our model gave similar sulfate-mortality estimates as the standard Cox model, the standard errors of these estimates were somewhat higher than those from the standard Cox model (Table 1).

With regard to the second concern, we have observed a pattern of spatial autocorrelation in mortality that cannot be fully explained by ambient particulate sulfate concentrations, even after controlling for a host of risk factors measured at the individual level. We also found that the association between air pollution and mortality was somewhat sensitive to the specification of the complexity of the spatial surface, with more complex surface specifications resulting in lower estimates of the sulfate effect. These results suggest that there may be some confounding due to missing or systematically mis-measured risk factors that are also spatially correlated with pollution. One approach to deal with this potential confounding problem is to model additional spatially distributed risk factor data[5], but one must be cautious in the selection of these variables, which are often difficult to model and interpret correctly. Furthermore, if the relevant risk factors are not known *a priori*, indiscriminate adding of spatially autocorrelated variables may result in multicolinearity problems and/or serious over-fitting of the models. An alternate approach to minimize the potential confounding bias arising from spatial contiguous variation is to directly model spatial trends, as is done in our space-time model.

While it is difficult to determine with certainty the true association between air pollution and mortality with this type of study design and analysis, our space-time model gives us a realistic way to evaluate how much of the air pollution mortality effects could be explained by missing or systematically miss-modeled risk factors that may be spatially autocorrelated with both mortality and pollution. For example, based on our modeling of the ACS data, the estimated excess mortality risk associated with a change of 4.2 $\mu g/m^3$ in particulate sulfate concentrations (the interquartile range of the data) was 5.5 percent (95 percent confidence interval 3.3-7.7) without modeling of the spatial mortality surface. An excess mortality risk of 3.5 percent (95 percent confidence interval 1.6-5.3) was estimated based on a joint estimate with a spatial surface model using a LOESS span of 20

Model Type	Span (%)	Sulfate Effect (γ) (s.e.)	Relative Risk* (95% C.I.)	Random Effects Variance (θ)	Spatial Auto-correlation+ (p-value)
Cox	NA	0.0118 (0.00177)	1.051 (1.036, 1.066)	NA	NA
Random Effect Cox	NA	0.0125 (0.00252)	1.055 (1.033, 1.077)	0.0027	NA
Space-Time	100	0.0127 (0.00252)	1.055 (1.033, 1.077)	0.0027	0.31 (<0.0001)
Space-Time	90	0.0106 (0.00279)	1.046 (1.022, 1.070)	0.0022	0.20 (<0.0001)
Space-Time	80	0.0106 (0.00277)	1.046 (1.022, 1.070)	0.0021	0.19 (<0.0001)
Space-Time	70	0.0102 (0.00272)	1.044 (1.021, 1.067)	0.0019	0.17 (<0.0001)
Space-Time	60	0.0093 (0.00261)	1.040 (1.018, 1.062)	0.0016	0.15 (0.0026)
Space-Time	50	0.0089 (0.00253)	1.038 (1.017, 1.060)	0.0013	0.13 (0.0089)
Space-Time	40	0.0085 (0.00245)	1.036 (1.016, 1.058)	0.0010	0.10 (0.0334)
Space-Time	30	0.0085 (0.00235)	1.036 (1.017, 1.057)	0.0007	0.07 (0.1338)
Space-Time	20	0.0081 (0.00219)	1.035 (1.016, 1.053)	0.0003	0.04 (0.3670)

TABLE 1. Table 1. Sulfate Effect, random effects variance and spatial autocorrelation by model type and span of LOESS smoother of location surface. NA: not applicable. *: Relative risk evaluated at interquartile range of sulfate concentrations (4.2 μg/m^3). +: Spatial autocorrelation of standardized random effects based on nearest neighbors using Moran's I statistic.

percent.

The above values provide a range in credible estimates obtained from these data and analytical methods. The larger estimate (5.5 percent per 4.2 μg/m^3) should be considered the more accurate one if the broader spatial autocorrelation between mortality and pollution is in fact due to differences in risk posed by different pollution levels across regions. Evidence against this interpretation is found in the presence of spatial autocorrelation in the adjusted community-specific relative mortality rates, even after sulfates are

included in the model, thus suggesting there may be spatially distributed risk factors that have not been fully accounted for, which may confound the observed association between mortality and particulate sulfates. The lower estimate (3.5 percent per 4.2 $\mu g/m^3$), reflects a more micro-scale or within-region association between these variables. This estimate reflects the amount of smoothing used to reduce spatial autocorrelation, both in terms of magnitude and relation to distance. This lower estimate of effect is conservative because any evidence of an association between air pollution and mortality obtained by shared broad-scale spatial patterns has been removed.

The observed association may be attenuated because measures of air pollution are known to miss-represent personal exposure and may not even represent the average of personal exposure for all cohort members within a community. In addition, because location is measured very precisely, further bias could occur because the effect of a variable measured with large error (i.e. air pollution) can be transferred to another variable measured with small error (i.e. location)[22].

We have developed an alternate method for statistical estimation and inference for our space-time random effects model in which we exploited the fact that the partial likelihood function used for parameter estimation in the independent observation Cox Model can be written in terms of a Poisson likelihood. We have shown that our space-time model can also be written as a random effects Poisson likelihood[23]. We then applied the estimation methods of Ma[24] for random effects Poisson models to the suitability transformed space-time model.

We then analyzed the ACS data with this alternative approach without any surface modeling. Here, $\widehat{\gamma} = 0.0125$ (standard error of 0.00252) and $\widehat{\theta} = 0.0027$, values nearly identical to our two-domain estimation procedure. The close correspondence with the two approaches is likely due to the relatively large number of deaths per location (average of 274 deaths per MSA).

We found that the estimates of the association between the individual risk factors and mortality, $\widehat{\boldsymbol{\beta}}$, and their estimates of uncertainty were nearly identical in the Cox survival model and the random effects Cox survival model, thus validating the use of the Cox model to identify the set of individual risk factors for mortality.

There is a substantial computational advantage to decomposing the estimation procedure into time and space domains. However, if there are a few deaths per location, estimates of the location-specific effects from the time-domain model ($\{\widehat{\delta}(s)\}$) are poorly characterized[18]. Areas in which no deaths occurred must be removed from the space-domain portion of the analysis, a limitation not inherent with the Cox random effects modeling approach. A limitation of the latter method is the intensiveness of computer resources. For example, for the ACS study this approach took 37 hours of computing time on a SUN Microsystems ULTRA ENTERPRISE

450 computer. In contrast, the space-time modeling approach took only a few minutes.

Acknowledgments: This research was motivated by a comprehensive reanalysis of the Harvard Six Cities and American Cancer Society (ACS) Studies of particulate air pollution and mortality sponsored by the Health Effects Institute (HEI) in which several of the authors (R. Burnett, R. Ma, M. Jerrett, M. Goldberg and D. Krewski) participated. We are grateful to HEI for their support of the reanalysis, and to the HEI Expert Panel and Review Committee that provided us with many helpful comments during the two year course of the reanalysis. The space-time methods presented in this paper represent extensions of our initial attempts to address spatial patterns in the ACS data in the reanalysis.

References

1. Dockery DW, Pope CA III, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Speizer FE**.** An association between air pollution and mortality in six US cities. *New England J Med* 329:1753-1759 (1993).

2. Pope CA, Thun MJ, Namboodiri, MM, Dockery, DW, Evans, JS, Speizer FE, Heath CW. Particulate air pollution as a predictor of mortality in a prospective study of US adults. *Am J Respir & Crit Care Med* 151:669-674 (1995).

3. Thun MJ, Day-Lally CA, Calle EE, Flanders WD, Health CW. Excess mortality among cigarette smokers: changes in a 20-year interval. *Am J Public Health* 85:1223-1230(1995).

4. Gamble JF. $PM_{2.5}$ and mortality in long-term prospective cohort studies: casuse-effect or statistical associations? *Environ Health Perspect* 106:535-549 (1998).

5. Health Effects Institute. 2000. Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality: A Special Report of the Institute's Particle Epidemiology Reanalysis Project. Health Effects Institute, Cambridge MA.

6. Cox DR. Regression models and life-tables. *J Royal Statist Soc, Series B* 34:187-202 (1972).

7. Ware JH, Stram DO. Statistical issues in epidemiologic studies of the health effects of ambient air pollution. *Can J Statist* 16:5-13 (1988).

8. Miron J. Spatial autocorrelation in regression analysis: a beginner's guide. In: Spatial Statistics and Models. Gaile GL, Willmott CJ eds. D. Reidel Publishing Company, Boston. 1984.

9. Griffin DA, Doyle PG, Wheeler DC, Johnson DL. A tale of two swaths: Urban childhood blood-lead levels across Syracuse, New York. *Ann Assoc Am Geographers* 88:640-645 (1988).

10. Matheron G. Principles of geostatistics. *Eco Geology* 58:1246-1266 (1963).

11. Goodchild MF. *Spatial Autocorrelation*. Norwich: Geo Books. 1986.

12. Cakmak S, Burnett R, Krewski D. Adjusting for temporal variation in the analysis of parallel time series of health and environmental variables. *J Expos Anal Environ Epidemiol* 129-144 (1998).

13. SAS PROC PHREG, SAS/STAT Software: Changes and Enhancements through Release 6.12. SAS Institute Inc., Cary, NC, USA. ISBN 1-55544-873-9. 1997.

14. Easton DF, Peto J, Babiker GAG. Floating absolute risk: an alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Statistics in Medicine* 10:1025-1035 (1991).

15. Cleveland, W.S., and Devlin, S.J. Robust locally-weighted regression and smoothing scatterplots. *J Am Statist Assoc* 74:829-36 (1988).

16. Hastie, T, Tibshirani, R. *Generalized Additive Models*. London: Chapman and Hall, 1990.

17. *S-PLUS 2000 Programmer's Guide*. Data Analysis Products Division, MathSoft, Seattle, WA.

18. Burnett RT, Ross WH, Krewski D. Non-linear mixed regression models. *Environmetrics* 6:85-99 (1995).

19. Zeger, SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimation equation approach. *Biometrics* 73:13-22 (1985).

20. S+SpatialStats: user's manual for Windows and UNIX. Data Analysis Products Division, MathSoft, Seattle, WA. 1997.

21. Abbey DE, Nishino, N, McDonnell, WF, Burchette RJ, Knutsen, SF, Beeson LW, Yang JX. Long-term inhalable particles and other air pollutants related to mortality in nonsmokers. *Am J Respir & Crit Care Med* 159:373-382 (1999).

22. Zidek, JV, Wong, H, Le, ND, Burnett, R. Causality, measurement error and multicollinearity in epidemiology. *Environmetrics* 7:441-451 (1996).

23. Ma R, Krewski D, Burnett, R. Random effects Cox models: a Poisson modelling approach. Technical Report No. 338, Laboratory for Research in Statistics and Probability, Carleton University, 2000.

24. Ma, R. An Orthodox BLUP Approach to Generalized Linear Mixed Models. Ph.D. Thesis. Department of Statistics, The University of British Columbia, 1999.

Realised volatility and estimating stochastic volatility models

Ole E. Barndorff-Nielsen [1], Neil Shephard[2],

[1] The Centre for Mathematical Physics and Stochastics (MaPhySto), University of Aarhus, Ny Munkegade, DK-8000 Aarhus C, Denmark.
[2] Nuffield College, Oxford OX1 1NF, UK

Abstract: The availability of intra-day data on the prices of speculative assets means that we can use quadratic variation like measures of activity in financial markets, called realised volatility, to study the stochastic properties of returns. Here we derive the moments and the asymptotic distribution of the realised volatility error — the difference between realised volatility and the actual volatility. These properties can be used to allow us to estimate the parameters of stochastic volatility models.

Keywords: Kalman filter; Quarticity; Quadratic variation; Realised volatility; Stochastic volatility; Subordination.

1 Introduction

1.1 Stochastic volatility

Here we review some results which we have developed in a recent paper Barndorff-Nielsen and Shephard (2001a) which is available at

`www.nuff.ox.ac.uk/users/shephard/levy.htm`

In the stochastic volatility (SV) model for log-prices a basic Brownian motion is generalised to allow the volatility term to vary over time. Then the log-price $y^*(t)$ follows the solution to the stochastic differential equation (SDE),

$$\mathrm{d}y^*(t) = \left\{\mu + \beta\sigma^2(t)\right\}\mathrm{d}t + \sigma(t)\mathrm{d}w(t), \tag{1}$$

where $\sigma^2(t)$, the *instantaneous* or *spot volatility*, is going to be assumed to (almost surely) have locally square integrable sample paths, while being stationary and stochastically independent of the standard Brownian motion $w(t)$. Over an interval of time of length $\Delta > 0$ returns are defined as

$$y_n = y^*(\Delta n) - y^*((n-1)\Delta), \qquad n = 1, 2, ... \tag{2}$$

which implies that whatever the model for σ^2, it follows that

$$y_n|\sigma_n^2 \sim N(\mu\Delta + \beta\sigma_n^2, \sigma_n^2).$$

where

$$\sigma_n^2 = \sigma^{2*}(n\Delta) - \sigma^{2*}\left\{(n-1)\Delta\right\}, \quad \text{and} \quad \sigma^{2*}(t) = \int_0^t \sigma^2(u)\mathrm{d}u.$$

In econometrics $\sigma^{2*}(t)$ is called *integrated volatility*, while we call σ_n^2 *actual volatility*. Both definitions play a central role in the probabilistic analysis of SV models. Reviews of the literature on this topic are given in Ghysels, Harvey, and Renault (1996). One of the key results in this literature (Barndorff-Nielsen and Shephard (2001b)) is that if we write (when they exist) ξ, ω^2 and r, respectively, as the mean, variance and the autocorrelation function of the process $\sigma^2(t)$ then

$$\mathrm{E}\left(\sigma_n^2\right) = \xi\Delta, \quad \mathrm{Var}\left(\sigma_n^2\right) = 2\omega^2 r^{**}(\Delta) \qquad \text{and} \tag{3}$$

$$\mathrm{Cov}\{\sigma_n^2, \sigma_{n+s}^2\} = \omega^2 \diamondsuit r^{**}(\Delta s), \tag{4}$$

where

$$\diamondsuit r^{**}(s) = r^{**}(s+\Delta) - 2r^{**}(s) + r^{**}(s-\Delta), \tag{5}$$

and

$$r^*(t) = \int_0^t r(u)\mathrm{d}u \quad \text{and} \quad r^{**}(t) = \int_0^t r^*(u)\mathrm{d}u. \tag{6}$$

One of the most important aspects of SV models is that $\sigma^{2*}(t)$ can be exactly recovered using the entire path of $y^*(t)$. In particular, for the above SV model the *quadratic variation* is $\sigma^{2*}(t)$, i.e. we have

$$[y^*](t) = \mathrm{p}\text{-}\lim_{r\to\infty} \sum \{y^*(t_{i+1}^r) - y^*(t_i^r)\}^2 = \sigma^{2*}(t) \tag{7}$$

for any sequence of partitions $t_0^r = 0 < t_1^r < ... < t_{m_r}^r = t$ with $\sup_i\{t_{i+1}^r - t_i^r\} \to 0$ for $r \to \infty$. This is a powerful result for it does not depend upon the model for instantaneous volatility nor the drift terms in the SDE for log-prices given in (1).

In practice, although we often have a continuous record of quotes or transaction prices, at a very fine level the SV model is a poor fit to the data. This is due to market microstructure effects. As a result we should regard the above results as indicating that we can estimate actual volatility, for example over a day, reasonably accurately by sums of squared returns, say, using five, ten or thirty minute periods. Suppose we have M intra-day observations during each day, then the sum of squared intra-day changes over a day is

$$\{y\}_n = \sum_{j=1}^{M} \left\{ y^* \left((n-1)\Delta + \frac{\Delta j}{M} \right) - y^* \left((n-1)\Delta + \frac{\Delta (j-1)}{M} \right) \right\}^2, \tag{8}$$

which is an estimate of σ_n^2. It is a consistent estimate as $M \to \infty$, while it is unbiased when μ and β are zero. In econometrics $\{y\}_n$ has recently been labelled *realised volatility*, and we will follow that convention here. Andersen, Bollerslev, Diebold, and Labys (2001). have empirically studied the properties of $\{y\}_n$ in foreign exchange and equity markets.
In particular the contribution of our paper will be to allow us to:

- understand the exact second order properties of $\{y\}_n$ when $\mu = \beta = 0$.
- use the models for instantaneous volatility to provide *model based* estimates of actual volatility (rather than model free estimates which assume $M \to \infty$) using the series of realised volatility measurements when $\mu = \beta = 0$.
- estimate the parameters of SV models using simple and rather accurate statistical procedures when $\mu = \beta = 0$.
- derive the asymptotic distribution of $\sqrt{M}\left(\{y\}_n - \sigma_n^2\right)$ for large M, showing this does not depend upon μ and β.

2 Relating actual to realised volatility

2.1 Generic results

Actual volatility, σ_n^2, plays a crucial role in SV models. It can be estimated using realised volatility, $\{y\}_n$, given in (8). Here we discuss this in the simplest context where $\mu = \beta = 0$.
In SV models we can always decompose

$$\{y\}_n = \sigma_n^2 + u_n, \qquad \text{where} \qquad u_n = \{y\}_n - \sigma_n^2. \tag{9}$$

Here we call u_n the *realised volatility error*, which has the property that $\mathrm{E}\left(u_n|\sigma_n^2\right) = 0$. We can see that

$$\begin{aligned} \mathrm{E}\left(\{y\}_n\right) &= \Delta\xi, \qquad \mathrm{Var}\left(\{y\}_n\right) = \mathrm{Var}(u_n) + \mathrm{Var}(\sigma_n^2), \\ \mathrm{Cov}(\{y\}_n, \{y\}_{n+s}) &= \mathrm{Cov}(\sigma_n^2, \sigma_{n+s}^2). \end{aligned}$$

Further, writing

$$\sigma_{j,n}^2 = \sigma^{2*}\left((n-1)\Delta + \frac{\Delta j}{M}\right) - \sigma^{2*}\left((n-1)\Delta + \frac{\Delta(j-1)}{M}\right)$$

we have that $u_n \stackrel{\mathcal{L}}{=} \sum_{j=1}^{M} \sigma_{j,n}^2\left(\varepsilon_{j,n}^2 - 1\right)$, where $\varepsilon_{j,n} \overset{i.i.d.}{\sim} N(0,1)$ and independent of $\left\{\sigma_{j,n}^2\right\}$. It is clear that $\{u_n\}$ is a weak white noise sequence which is uncorrelated to the actual volatility series $\left\{\sigma_n^2\right\}$.

Now unconditionally,

$$\begin{aligned} \mathrm{Var}(u_n) &= 2M\mathrm{E}\left\{\left(\sigma_{1,n}^2\right)^2\right\} \\ &= 2M\left\{\mathrm{Var}\left(\sigma_{1,n}^2\right)+\mathrm{E}\left(\sigma_{1,n}^2\right)^2\right\}, \end{aligned} \tag{10}$$

for $\sigma_{1,n}^2$ has the same marginal distribution as each element of $\left\{\sigma_{j,n}^2\right\}$. In general we have, from (3) that

$$\mathrm{E}\left(\sigma_{1,n}^2\right)=\Delta M^{-1}\xi, \quad \mathrm{Var}\left(\sigma_{1,n}^2\right)=2\omega^2 r^{**}\left(\Delta M^{-1}\right). \tag{11}$$

Hence we can compute $\mathrm{Var}(u_n)$ for all SV models when $\mu=\beta=0$. In turn, having established the second order properties of σ_n^2 and u_n, we can immediately use the results in Whittle (1983) to provide best linear prediction and smoothing results for the unobserved actual volatilities σ_n^2 from the time series of realised volatilities $\{y\}_n$.

2.2 Simple example: $r(t)=\exp(-\lambda|t|)$

Suppose the volatility process has the autocorrelation function $r(t)=\exp(-\lambda|t|)$. Here we recall two classes of processes which have this property. The first is the constant elasticity of variance (CEV) process which is the solution to the SDE

$$\mathrm{d}\sigma^2(t)=-\lambda\left\{\sigma^2(t)-\xi\right\}\mathrm{d}t+\omega\sigma(t)^{\eta}\mathrm{d}b(\lambda t), \qquad \eta\in[1,2],$$

where $b(t)$ is standard Brownian motion uncorrelated with $w(t)$. The second process is the non-Gaussian Ornstein-Uhlenbeck, or OU type for short, process which is the solution to the SDE

$$\mathrm{d}\sigma^2(t)=-\lambda\sigma^2(t)\mathrm{d}t+\mathrm{d}z(\lambda t), \tag{12}$$

where $z(t)$ is a Lévy process with non-negative increments. These models have been developed in this context by Barndorff-Nielsen and Shephard (2001b). In Figure 1 we have drawn a curve to represent a simulated sample path of σ_n^2 from an OU process where $\sigma^2(t)$ has a $\Gamma(4,8)$ stationary distribution, $\lambda=-\log(0.99)$ and $\Delta=1$, along with the associated realised volatility (depicted using crosses) computed using a variety of values of M. We see that as M increases the precision of realised volatility increases, while Figure 1.d shows that the variance of the realised volatility error increases with the volatility.
For both CEV and OU models

$$\mathrm{E}\left(\sigma_n^2\right)=\Delta\xi, \qquad \mathrm{Var}\left(\sigma_n^2\right)=\frac{2\omega^2}{\lambda^2}\left(e^{-\lambda\Delta}-1+\lambda\Delta\right)$$

and

$$\mathrm{Cor}\{\sigma_n^2,\sigma_{n+s}^2\}=de^{-\lambda\Delta(s-1)}, \qquad s=1,2,.... \tag{13}$$

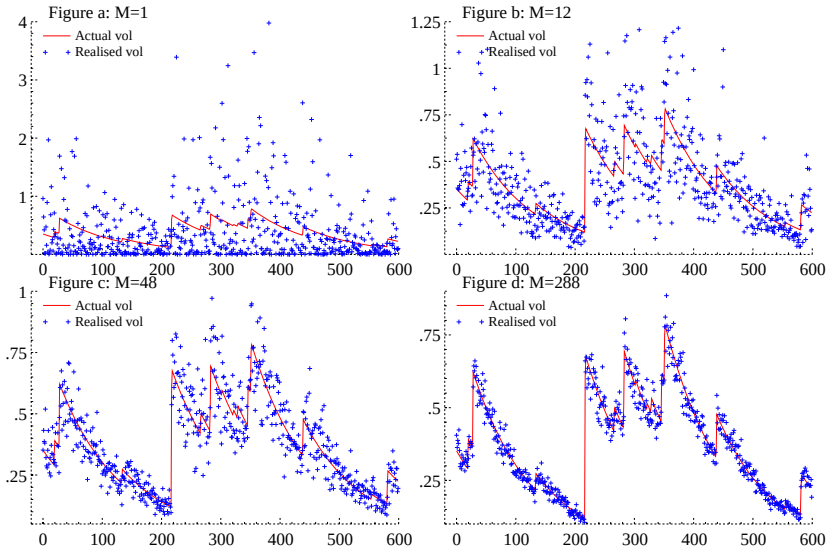


FIGURE 1. *Actual* σ_n^2 *and realised* $\{y\}_n$ *(with* M varying*) volatility based upon a* $\Gamma(4,8)$-OU process with $\lambda = -\log(0.98)$ and $\Delta = 1$. This implies $\xi = 0.5$ and $\xi\omega^{-2} = 8$. *The file containing the code used to carry out these calculations is called* `ssf_mse.ox`.

Finally

$$\begin{aligned}\mathrm{Var}(u_n) &= 2M\left\{\mathrm{Var}\left(\sigma_{1,n}^2\right) + \mathrm{E}\left(\sigma_{1,n}^2\right)^2\right\} \\ &= 2M\left\{2\omega^2\lambda^{-2}\left(e^{-\lambda\Delta/M} - 1 + \lambda\Delta M^{-1}\right) + \left(\Delta M^{-1}\right)^2\xi^2\right\} \quad (14)\end{aligned}$$

2.3 Extension of the example: superpositions

The OU/CEV volatility models are often too simple to accurately fit the types of dependence structures we observe in financial economics. One mathematically tractable way of improving the flexibility of the volatility model is to let the instantaneous volatility be the sum, or superposition, of independent OU or CEV processes. Superpositions of such processes also have potential for modelling long-range dependence and self-similarity in volatility. This is discussed in the OU case in Barndorff-Nielsen and Shephard (2001b) and at more depth by Barndorff-Nielsen (2000) who formalises the use of superpositions as a way of modelling long-range dependence.
Consider volatility based on the sum of J independent OU or CEV pro-

cesses

$$\sigma^2(t) = \sum_{i=1}^{J} \tau^{(i)}(t), \qquad \text{where} \qquad \{w_i \geq 0\} \quad \text{and} \quad \sum_{i=1}^{J} w_i = 1$$

where the $\tau^{(i)}(t)$ process has the memory parameter λ_i. We assume

$$\mathrm{E}(\tau^{(i)}(t)) = w_i \xi \qquad \mathrm{Var}(\tau^{(i)}(t)) = w_i \omega^2,$$

implying $\mathrm{E}(\sigma^2(t)) = \xi$ and $\mathrm{Var}(\sigma^2(t)) = \omega^2$.

3 Efficiency gains: model based estimators of volatility

3.1 State space representation

If $\sigma^2(t)$ is OU or CEV then σ_n^2 has an ARMA$(1,1)$ representation and so it is computationally convenient to place (9) into a linear state space representation (see, for example, Harvey (1989, Ch. 3) and Hamilton (1994, Ch. 13). In particular we write $\alpha_{1n} = \left(\sigma_n^2 - \Delta\xi\right)$ and $u_n = \sigma_u v_{1n}$, then the state space is explicitly

$$\begin{aligned} \{y\}_n &= \Delta\xi + \begin{pmatrix} 1 & 0 \end{pmatrix} \alpha_n + \sigma_u v_{1n}, \\ \alpha_{n+1} &= \begin{pmatrix} \phi & 1 \\ 0 & 0 \end{pmatrix} \alpha_n + \begin{pmatrix} \sigma_\sigma \\ \sigma_\sigma \theta \end{pmatrix} v_{2n}, \end{aligned} \tag{15}$$

where v_n is a zero mean, white noise sequence with an identity covariance matrix. The parameters ϕ, θ and σ_σ represent the autoregressive root, the moving average root and the variance of the innovation to this process, while σ_u^2 is found from (10) and (11). Can use a Kalman filter to unbiasedly and efficiently (in a linear sense) estimate σ_n^2 by prediction (that is the estimate of σ_n^2, using $\{y\}_1, ..., \{y\}_{n-1}$) and smoothing (that is the estimate of σ_n^2, using $\{y\}_1, ..., \{y\}_T$ where T is the sample size). Biproducts of the Kalman filter are the mean square errors of these *model based* (that is they depend upon the assumption that σ_n^2 has an ARMA$(1,1)$ representation) estimators.

4 Empirical illustration

To illustrate some of these results we have fitted a set of superposition based OU/CEV SV models to realised volatility time series constructed from the 5 minute exchange rate return data discussed in the introduction to this paper. Here we use the quasi-likelihood method to estimate the parameters of the model — ξ, ω^2, $\lambda_1, ..., \lambda_J$ and $w_1, ..., w_J$. We do this for a variety of

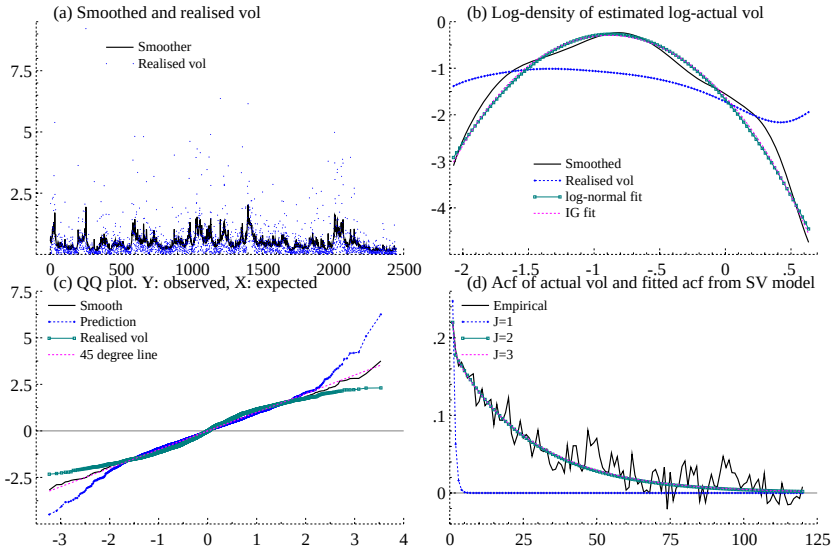


FIGURE 2. *Results from the fit of the SV model using M=6 (4 hour returns) on the 5 minute Olsen data. (a) gives the time series of realised volatilities* $\{y\}_n$ *together with the smoothed estimates of actual volatility. (b) draws the kernel based estimates of the log-density of log-realised volatility and log-smoothed log-actual volatility. Also drawn are the log-normal and inverse Gaussian fits. These lines are so close they appear to be almost on top of one another. (c) QQ plot for returns divided by estimated actual vol, using realised, predicted and smoothed volatility. Perfect fit sits on a 45° line. (d) Acf of realised volatility and the fit of SV model with various superpositions of J processes. Code is available at* `ssf_empirical.ox`.

values of M, starting with $M = 6$, which corresponds to working with 4 hour returns.

To provide a more detailed assessment of the fit of the model we have drawn a series of graphs in Figure 2. Except where explicitly noted we have computed the graphs using the $J = 3$ fitted model, although there would be very little difference if we had taken $J = 2$. Figure 2(a) draws the computed actual volatility $\{y\}_n$, together with the corresponding smoothed estimate of actual volatility using the fitted SV model. We can see that realised volatility is much more jagged than the smoothed quantity. In Figure 2(b) we have drawn a kernel based estimate of the log-density of the log of realised volatility. The bandwidths were taken to be $1.06\widehat{\sigma}T^{-1/5}$, where T is the sample size and $\widehat{\sigma}$ is the empirical standard deviation of the log of realised volatility (this is an optimal choice against a mean square error loss for Gaussian data, e.g. Silverman (1986)) while we have chosen the

range of the display to match the upper and lower 0.05% of the data — so trimming very little of the data. Andersen, Bollerslev, Diebold, and Labys (2001) have suggested that the marginal distribution of realised volatility is closely approximated by a log-normal distribution when M is high, and that this would support a model for actual volatility which is log-normal. However, when we draw on the corresponding fitted (choosing the parameters by using maximum likelihood based upon the estimated smoothed realised volatilities as data) log-normal log-density we can see that the fit is extremely poor. The same holds for the inverse Gaussian log-density which is also drawn in this figure (but is so close to the fit of the log-normal that it is extremely hard to tell the difference between the two curves). Inverse Gaussian models for volatility were suggested by Barndorff-Nielsen and Shephard (2001b). The rejection of the log-normal and inverse Gaussian marginal distributions for realised volatility itself seems conclusive here. However, when we carry out the same action on the smoothed realised volatilities this rejection no longer holds, implying realised volatility error really matters here. The kernel based estimate of the log-density of the log smoothed estimates is very much in line with the log-normal or inverse Gaussian hypothesis.

Figure 2(c) draws a QQ plot of returns y_n divided by a number of estimates of σ_n. The Figure indicates that when we scale returns by realised volatility the returns are highly non-Gaussian, while when we plug in the smoothed estimate then the model seems to fit extremely well. The conclusion does not continue to hold when we use the predictor of actual volatility, rather than the smoothed quantity. This fits as poorly as the plot based on the realised volatility. Overall, the Figure 2(c) again confirms the fit of the model, while suggests when $M = 6$ the difference between realised and smoothed volatility is important.

Figure 2(d) shows the corresponding autocorrelation function for the realised volatility series together with the corresponding empirical correlogram. We see from this figure that when $J = 1$ we are entirely unable to fit the data. A superposition of two processes is much better, picking up the longer-range dependence in the data.

5 Asymptotic distribution of realised volatility error

5.1 The theory

In Section 2 we derived the mean and variance of the realised volatility error for a continuous time SV model when $\mu = \beta = 0$. Although it is possible to derive the corresponding result when $\mu \neq 0$ but $\beta = 0$, adapting to the risk premium case seems difficult. Instead we employ an asymptotic route.

In our paper we obtain a limit theory for

$$\frac{\{y\}_n - \sigma_n^2}{\sqrt{\frac{4}{3}\sum_{j=1}^{M} y_{j,n}^4}} \xrightarrow{\mathcal{L}} N(0,1).$$

which covers the case of a drift and risk premium.

Acknowledgments: This paper gives some of the results reported more fully in "Econometric analysis of realised volatility and its use in estimating stochastic volatility models." Ole E. Barndorff-Nielsen's work is supported by CAF (`www.caf.dk`) and by MaPhySto (`www.maphysto.dk`). Neil Shephard's research is supported by the UK's ESRC through the grant "Econometrics of trade-by-trade price dynamics," which is coded R00023839. We owe a large debt to Michel M. Dacorogna at the Olsen and Associates for allowing us to use Olsen's high frequency exchange rate data in our study.

References

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001). The distribution of exchange rate volatility. *Journal of the American Statistical Association 96*, 42–55.

Barndorff-Nielsen, O. E. (2000). Superposition of Ornstein-Uhlenbeck type processes. *Theory of Probability and Its Applications*. Forthcoming.

Barndorff-Nielsen, O. E. and N. Shephard (2001a). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. Unpublished discussion paper: Nuffield College, Oxford.

Barndorff-Nielsen, O. E. and N. Shephard (2001b). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion). *Journal of the Royal Statistical Society, Series B 63*, 167–241.

Ghysels, E., A. C. Harvey, and E. Renault (1996). Stochastic volatility. In C. R. Rao and G. S. Maddala (Eds.), *Statistical Methods in Finance*, pp. 119–191. Amsterdam: North-Holland.

Hamilton, J. (1994). *Time Series Analysis*. Princeton: Princeton University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Silverman, B. W. (1986). *Density Estimation for Statistical and Data Analysis*. London: Chapman & Hall.

Whittle, P. (1983). *Prediction and Regulation* (2 ed.). Oxford: Blackwell. 1st edition, 1963.

Fitting exponential family mixed models

Juni Palmgren[1,2] and Samuli Ripatti[1,3]

[1] Mathematical Statistics, Stockholm University, S-10691 Stockholm, Sweden. Email: juni@matematik.su.se
[2] Medical Epidemiology, Karolinska Institutet, Sweden
[3] Rolf Nevanlinna Institute, University of Helsinki, Finland

Abstract: The seminal papers by Nelder and Wedderburn (Generalized Linear Models, JRSS A 1972) and Cox (Regression models and life tables, JRSS B 1972) both rely on the assumption that conditionally on covariate information (including time) the observations are independent. The difficulty in identifying and measuring all relevant covariates has pushed for methods that can handle both mean and covariance structures jointly. There has been a parallel development of (i) marginal models and (ii) random effects models as multivariate extensions of the generalized linear model and the multiplicative hazard model, respectively. After a brief review of this development we focus on estimation and computational aspects of fitting random effects models. We discuss the use of penalized likelihood, Monte Carlo EM and MCMC methods using examples involving censored survival time responses and Poisson responses.

Keywords: Frailty; Generalized linear mixed model; Markov chain Monte Carlo; Monte Carlo EM; Penalized likelihood; Random effects.

1 Introduction

Soon after the introduction in 1972 Nelder's and Wedderburn's generalized linear model [23] was recognized as a useful conceptual framework for a wide class of regression models used in biomedical research. Cox's semi-parametric regression model for censored failure time data [9], also published in 1972, has had an equally profound influence on the statistical methodology used in the medical field. Since the 1970's there has been escalating efforts to extend these two families of non-normal non-linear models to allow for between-cluster heterogeneity and within-cluster dependence. Study designs such as group randomization, litter based toxicology studies, longitudinal studies, studies on spatial variation, as well as family studies have pushed for this development.

In Section 2 we account for some of the milestones in the development of multivariate generalized linear models and multivariate hazard regression models. In Section 3 we present the random effects model, and a battery of estimation and inference approaches are outlined in Section 4. Section

5 describes two data analyses examples: a hiearchical survival data problem where the lifetime of roses is assessed, and a Poisson random effects model for spatial smoothing of alcohol related mortality in Finland. We conclude in Section 6 by discussing pros and cons of the different estimation and inference procedures, and we argue that a new layer of unification is emerging for handling the multivariate generalized linear models and multivariate hazard regression models.

2 From univariate to multivariate models

2.1 The generalized linear model

The generalized linear model (GLM) is specified through the probability distribution for the observations, and the link function relating the regression parameters to the means. Conditionally on the means the observations are assumed statistically independent. The standard linear regression model is a GLM with normal distribution and identity link. The log linear model for count data is a GLM with Poisson distribution and logarithmic link. The logistic regression model is a GLM with binomial distribution and logit link. These three special cases are useful standard GLM's with attractive theoretical properties [22].

2.2 Cox regression

Censored failure time data arise in many areas of biomedical research. Early methodology was confined to descriptive life table techniques and to the mathematical formulation of the survival experience over time. Cox's regression model changed the focus to partial likelihood inference for the relative hazard as function of covariate values, while the baseline time dynamics were treated as a secondary feature and modelled non-parametrically. Partial likelihood estimation of relative risk parameters may be viewed as a stratified analysis, in which time is controlled for by matching on the risk set at each time point when a failure occurs. Counting process and martingale theory provide the theoretical basis for the Cox regression model [1]. An important feature of the model, which nicely bridges the gap to the generalized linear model, is that conditional on the past the counting process behaves like a Poisson process, with independent increments and time varying rate function.

2.3 Multivariate responses

Logistic regression, Poisson regression and Cox regression can all be viewed as univariate probability models for a series of binary events [8]. They all share the property that conditional on measured exposures and covariates

the responses are assumed statistically independent across individuals, with a constant event probability. For the Poisson and Cox models this conditional event probability is 'small' and the 'risk sets' are large. However, incomplete covariate information is often a reality, rendering the standard model specification too simplistic.

When no information is available on sources of unobserved heterogeneity, then one single overdispersion parameter may capture the additional component of variation. Compound distributions such as the beta-binomial or gamma-Poisson may be used, or an extra parameter may be multiplied to the Binomial or Poisson variance expressions [31], [5]. When the data involve identified clusters, e.g. repeated measurements on the same individual or clusters of individuals in families, then a structured model can be specified for the between-cluster heterogeneity and the within-cluster dependence.

Multivariate models for the mean and dependence structures for responses measured on a wide variety of scales has been the focus of escalating methodological interest. Two main routes have emerged: (i) the marginal models and (ii) the random effects models. We briefly touch on (i) here and discuss (ii) in detail in Section 3. In 1986 Liang and Zeger proposed a general procedure for multivariate generalized linear models [19]. Their focus was on the estimation of regression parameters that linked covariate effects to population averages. The within-cluster dependence was treated as a nuisance, needing to be accounted for since it affects the power of tests and the precision of regression estimates. Zhao and Prentice [32] extended the Liang and Zeger procedure by setting up two sets of estimating equations jointly, one for the mean parameters and one for the dependence parameters. Wei, Lin and Weissfeld [30] considered semi-parametric regression models in which two or more distinct failure times are recorded on each individual. Each marginal failure time is modelled by a semi-parametric Cox model, and the dependence is accounted for when estimating the parameter uncertainty. The Wei, Lin and Weissfeld model is a multivariate failure time analogue to the Liang and Zeger generalized estimating equation (GEE) approach for the generalized linear model.

3 Random effects models

While the marginal models focus on inference for the fixed regression parameters, the random effects models jointly describe the mean and dependence using fixed and random regression parameters. Stiratelli, Laird and Ware [27] elegantly extend to the multivariate binary setting the Laird and Ware [18] mixed model for normally distributed repeated measures. Breslow and Clayton [6] give a thorough account of random effects generalized linear models, and call them generalized linear mixed models. For right censored failure time data the random effects are referred to as frailties

[29]. There is a rather extensive literature on so called shared frailty models with a simple covariance structure [17], [15], [1]. Below we present the generalized linear mixed model and the frailty model in general terms, and proceed to discuss estimation and inference.

3.1 The model

The generalized linear mixed model: Let Y_i, for $i = 1, \ldots, n$, denote the observation on unit i. Let β denote a p -vector of unknown fixed effect parameters, with an associated known design vector X_i for unit i. Let b_i denote a q -vector of unknown random effect parameters, with associated known design vector Z_i. For given $b = (b_1 \ldots b_n)$, the conditional distribution for Y_i is of exponential family form $p(Y_i \mid \gamma_i) = c_i(y_i) \exp(\gamma_i y_i - a(\gamma_i))$, with γ_i the canonical parameter, $a(.)$ a known monotone differentiable function, $E(Y_i \mid \gamma_i) = \mu_i = a'(\gamma_i)$ the mean parameter and $var(Y_i \mid \gamma_i) = v(\mu_i) = a''(\gamma_i)$ the variance function. Following [6], [22] we write the generalized linear model for unit i in the form

$$\begin{aligned} p(Y_i \mid \mu_i) &= \exp\left[\int_{y_i}^{\mu_i} \frac{y_i - u}{v(u)} du\right] \\ h(\mu_i) &= X_i\beta + Z_i b, \end{aligned} \tag{1}$$

with $h(.)$ a known, monotone, differentiable function linking the regression parameters to the mean. Conditionally on b the observations are assumed independent. At the second stage a distribution is imposed on b, capturing the structure for between cluster heterogeneity and within cluster dependence as defined through the design vectors Z_i. We assume that jointly $b \sim p(b \mid D(\theta))$, with θ a vector of unknown parameters which vary independently of β.

The frailty model: Let T_i, for $i = 1, \ldots, n$, denote the event time, C_i the censoring time, $U_i = \min(T_i, C_i)$ and $\delta_i = I_{\{T_i \leq C_i\}}$. Given the random effects, or frailties $b = (b_1 \ldots b_n)$, the event times are assumed independent and the conditional hazard function $\lambda_i(t)$ for unit i has the form

$$\lambda_i(t) = \lambda_0(t) \exp(X_i\beta + Z_i b), \tag{2}$$

with $\lambda_0(t)$ the baseline hazard and $b \sim p(b \mid D(\theta))$ as before.

For models (1) and (2) the random effects b may be viewed as a set of latent observations, and the model may be characterized as an incomplete data model. Besides making inferences about the regression parameters β and the variance component parameters θ, the purpose of the modelling is often to make predictions for the random effects b.

3.2 The likelihood

Following the missing data terminology, the complete data are (Y, b), but only Y are observed. The observed data likelihood takes the form

$$p(Y \mid \beta, \theta) = \int p(Y, b \mid \beta, \theta) db = \int p(Y \mid \beta, b)\, p(b \mid \theta) db, \tag{3}$$

and we write for model (1)

$$\log p(Y \mid \beta, b) = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{v(\mu_i)} \right], \tag{4}$$

with

$$h(\mu_i) = X_i\beta + Z_i b.$$

For model (2) we denote the data by $Y = (U, \delta)$ and write

$$\log p(Y \mid \lambda_0(t), \beta, b) = \sum_{i=1}^{n} \delta_i [\log \lambda_i(t)] - \exp[\Lambda_i(t)], \tag{5}$$

with

$$\begin{aligned} \lambda_i(t) &= \lambda_0(t) \exp(X_i\beta + Z_i b) \\ \Lambda_i(t) &= \int_0^t \lambda_i(s) ds. \end{aligned}$$

4 Estimation and inference

For given b, the complete data log likelihood in (4) or (5) is easy to maximize, suggesting that the EM-algorithm is a natural choice for computing maximum likelihood estimates based on (3).

4.1 The EM algorithm

The EM algorithm finds the maximum of the observed data likelihood (3) by alternates between finding the expectation of the unobserved part of the data, given the observed data (E-step), and maximizing the complete data likelihood as if the non-observables were observed (M-step) [10]. The random effects b are treated as unobserved data and they are imputed in the E-step. More precisely, for $\psi = (\beta, \theta)$ in model (4) and $\psi = (\lambda_0(t), \beta, \theta)$ in model (5), the E-step in iteration (r) involves the evaluation of

$$\begin{aligned} Q(\psi, \psi^{(r)}) &= E[\log(p(Y, b \mid \psi)) \mid Y, \psi^{(r)}] \\ &= \int \log(p(Y, b \mid \psi)) p(b \mid Y, \psi^{(r)}) db. \end{aligned} \tag{6}$$

In the M-step the Q function is maximized with respect to ψ to obtain $\psi^{(r+1)}$. The M-step equals maximization of the complete data log-likelihood (4) or (5), and standard software for the generalized linear model or the Cox model can be used, treating $Z_i b$ as an offset term. However, the elegance of the simple M-step is shadowed by the fact that the E-step in (6) involves an integral of the same dimension as in the observed data likelihood (3). A computatational problem thus remains, to which several solutions have been suggested, including penalized likelihood methods based on the Laplace approximation to the integral, and simulation based Monte Carlo EM and Markov chain Monte Carlo procedures.

4.2 Penalized likelihood

Breslow and Clayton [6] derive a penalized likelihood solution for the generalized linear mixed model (4) assuming Gaussian random effects. We recapture their argument and present a parallel approximation for the semiparametric frailty model (5) [25]. For Gaussian random effects we have $p(b \mid \theta) \propto |D(\theta)|^{-\frac{1}{2}} \exp[-\frac{1}{2} b' D(\theta)^{-1} b]$, and we write (3) in the form

$$c\,|D|^{-\frac{1}{2}} \int \exp[-\kappa(b)]db.$$

with

$$\kappa(b) = \log p(Y \mid \beta, b) - \frac{1}{2} b' D^{-1} b. \tag{7}$$

Let κ' and κ'' denote the q-vector and the $q \times q$ matrix of first- and second order partial derivatives of κ with respect to b. Ignoring the multiplicative constant c, the Laplace approximation to the marginal log likelihood takes the form

$$l(\beta, \theta) \approx -\frac{1}{2} \log |D(\theta)| - \frac{1}{2} \log \left| \kappa''(\tilde{b}) \right| - \kappa(\tilde{b},) \tag{8}$$

with $\tilde{b} = \tilde{b}(\beta, \theta)$ the solution to $\kappa'(\tilde{b}) = 0$.

For the generalized linear mixed model Breslow and Clayton argue that if the variance function $v(\mu)$ varies slowly (or not at all) as a function of the mean μ, then the first two terms in (8) may be ignored. An approximate solution to the likelihood in (4) is thus obtained by maximizing $\kappa(b)$ in (7), which corresponds to Green's penalized log likelihood [14]. Following the same rationale, Ripatti and Palmgren [25] derive expressions for $\kappa(b)$,$\kappa'(b)$ and $\kappa''(b)$ for the frailty model. They further show that for fixed θ the values $\hat{\beta}(\theta), \hat{b}(\theta)$, which maximize the penalized log likelihood (7) based on $\log p(Y \mid \lambda_0(t), \beta, b)$ in (5) also maximize the penalized partial log likelihood

$$\sum_{i=1}^{n} \delta_i \left((X_i\beta + Z_i b) - \log \sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \right) - \frac{1}{2} b' D(\theta)^{-1} b. \tag{9}$$

For given θ, the estimating equations for $\beta(\theta)$, $b(\theta)$, based on the first partial derivatives of the penalized log likelihood (7) derived from the generalized linear mixed model (4) are of the form

$$\sum_{i=1}^{n} [y_i - \mu_i] X_i = 0 \quad (10)$$

$$\sum_{i=1}^{n} [y_i - \mu_i] Z_i - D^{-1} b = 0, \quad (11)$$

with $h(\mu_i) = X_i\beta + Z_i b$. The corresponding estimating equations for $\beta(\theta)$, $b(\theta)$ for the frailty model derived from (9) are

$$\sum_{i=1}^{n} \delta_i [1 - \nu_i] X_i = 0 \quad (12)$$

$$\sum_{i=1}^{n} \delta_i [1 - \nu_i] Z_i - D^{-1} b = 0, \quad (13)$$

with

$$\nu_i = \frac{\exp(X_i\beta + Z_i b)}{\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b)}.$$

We find $\hat{\beta}(\theta), \hat{b}(\theta)$ by alternating between solving the equations (10) and (11) for the generalized linear mixed model, and between solving (12) and (13) for the frailty model. Note that solving (10) or (12) corresponds to the M-step in the EM-algorithm for β, and can be done with standard software for the generalized linear model or the Cox regression model, using estimated values of the random effects in an offset term. Maximizing the penalized likelihood (7) rather than the marginal likelihood (3) has replaced the awkward integral in the E-step with estimating equations (11) and (13), respectively.

Once $\hat{\beta}(\theta), \hat{b}(\theta)$ are computed, we update θ in $D(\theta)$ by maximizing the approximate profile likelihood derived from (8)

$$l(\hat{\beta}(\theta), \theta) \approx -\frac{1}{2} \log |D(\theta)| - \frac{1}{2} \log \left| K''(\hat{b}) \right| - \frac{1}{2} \hat{b}' D(\theta)^{-1} \hat{b}. \quad (14)$$

For the generalized linear mixed model Breslow and Clayton compute κ'' in (14) from the likelihood (4), both with and without a REML adjustment for the degrees of freedom. For the frailty model Ripatti and Palmgren [25] compute κ'' in (14) from the penalized partial likelihood (9) rather than from the full likelihood (5). The choice is motivated by the former performing better in simulations, and it is obtained as a side product from the previous iteration step.

4.3 Monte Carlo EM

Instead of alternating between (10) – (11) or (12) – (13) and (14) to obtain an approximate solution to (3), samples may be drawn from the predictive distribution $p(b \mid Y, \psi^{(r)})$ in (6), and the sample mean computed instead of the expectation in the E-step of the EM-algorithm. The distribution $p(b \mid Y, \psi^{(r)})$ is not a standard multivariate distribution, but rejection or importance sampling may be used [13], [11]. If enough samples are drawn, then the Monte Carlo EM-iterations converge to the maximum of the marginal likelihood (3). Booth and Hobert [3] and Ripatti, Larsen and Palmgren [24] suggest procedures where the number of samples is automatically increased when approaching the target, thus gaining absolute convergence for the MCEM algorithm.

4.4 Covariances for $\hat{\psi}$

For $\psi = (\beta, \theta)$ in model (4) and $\psi = (\lambda_0(t), \beta, \theta)$ in model (5), we write the Louis' [20] observed information

$$I(\psi) = E\left(-\frac{\partial^2 \log(p(Y, b \mid \psi))}{\partial\psi\partial\psi'} \mid Y, \hat{\psi}\right) - \mathrm{var}\left(\frac{\partial \log(p(Y, b \mid \psi))}{\partial\psi} \mid Y, \hat{\psi}\right), \tag{15}$$

with $\mathrm{cov}(\hat{\psi}) = I^{-1}(\psi)$, evaluated at $\psi = \hat{\psi}$. A discretized baseline hazard with jumps at distinct event times is used for $\lambda_0(t)$. Note that the conveniant procedure of computing the covariance matrix for $\hat{\beta}(\theta)$ from the estimating equations (10) or (12) neglects the additional variation stemming from the uncertainty in the estimated $\hat{\theta}$. This additional variation is captured in the second term in the information matrix (15), and needs to be computed separately when using the penalized likelihood estimating equations. When using Monte Carlo EM both terms in (15) are obtained as a side product from the samples in the last iteration.

4.5 Posterior inference and MCMC

We make a conceptual shift and treat $\psi = (\beta, \theta)$ in model (1) and $\psi = (\lambda_0(t), \beta, \theta)$ in model (2) as random, and the data Y as fixed. We write the observed data posterior

$$p(\psi \mid Y) = \int p(\psi \mid Y, b) p(b \mid Y) db. \tag{16}$$

Using Bayes' theorem the complete data posterior $p(\psi \mid Y, b)p(b \mid Y)$ inside the integral (16) is proportional to the product of a prior distribution $p(\psi)$ and the complete data likelihood $p(Yb \mid \psi)$ in (3). If samples from $p(b \mid Y)$

could be drawn easily, then it would be straight forward to evaluate the observed data posterior (16) as a Monte Carlo mean. We write $p(b \mid Y)$ as

$$p(b \mid Y) = \int p(b \mid Y, \psi) p(\psi \mid Y) d\psi. \tag{17}$$

¿From the symmetry of the expressions in (16) and (17) an iterative two-step algorithm is suggested, involving an imputation step (I-step), with draws $b^{(r)}$ from the conditional predictive distribution $p(b \mid Y, \psi^{(r)})$ in (17), and a posterior step (P-step), with draws $\psi^{(r+1)}$ from the conditional posterior distribution $p(\psi \mid Y, b^{(r)})$ in (16). Under broad regularity conditions the sequence $\{\psi^{(r)}, b^{(r)}, r = 1, 2, \ldots\}$ converges to the joint posterior $p(\psi, b \mid Y)$, and the sequences of the components to their respective marginal posteriors $p(\beta \mid Y)$, $p(b \mid Y)$ and $p(\theta \mid Y)$ [13]. For the hazard model Clayton [7] discusses how to sample from the non-parametric distribution for the conditional baseline hazard $\lambda_0(t)$. Note that sampling in the P-step (I-step) depends on the previous I-step (P-step), but given the previous I-step (P-step) is conditionally independent of the previous P-step (I-step). This motivates the terminology Markov chain Monte Carlo. The I-step and P-step may be seen as stochastic counterparts to the E-step and M-step of the EM-algorithm. For large samples the likelihood will overrule the prior, and the mode and the curvature of the posterior (16) will coincide with the mode and the curvature of the likelihood (3). Note that per definition the credible intervals for $p(\beta \mid Y)$ and $p(b \mid Y)$ include the uncertainty in θ. For specific problems there is an extensive literature on clever choices of conditional distributions that are easy to sample from, and on computational tricks to speed up the sampling process and to ensure that all parts of the parameter space are covered [12].

5 Data analyses

5.1 Lifetime of roses

In the first example, we study data from a greenhouse experiment on vase lifetimes of cv. Frisco rose cuts. This is an incomplete randomized block design with four blocks, three plots in each block and eight plants per plot. From each plant, several rose cuts were picked and put to a vase, and for each cut the lifetime in the vase was recorded. There were total of 716 cuts with 3 censored lifetimes because of bent rose necks. There were four different lighting treatments randomized within blocks, and the primary interest is to study the effects of treatments on the average vase life as well as on the between plant variation. The details of the experiment are reported in [26].

We fit two different models to these data. The first is a shared frailty model

$$\lambda_{ij}(t) = \lambda_0(t) \exp(X_{1ij}\beta_1 + X_{2ij}\beta_2 + b_i), \tag{18}$$

TABLE 1. Parameter estimates and standard errors for two models for the rose survival data based on the MCEM algorithm and penalized partial likelihood (PPL) estimating equations.

	Model 1		Model 2	
Parameter	MCEM	PPL	MCEM	PPL
Treatment A	0.23(0.16)	0.24(0.16)	0.26(0.16)	0.28(0.17)
Treatment B	0.38(0.15)	0.39(0.16)	0.40(0.16)	0.41(0.17)
Treatment D	0.25(0.16)	0.25(0.16)	0.27(0.16)	0.22(0.17)
Block 1	-0.49(0.17)	-0.48(0.17)	-0.51(0.16)	-0.50(0.18)
Block 2	-0.50(0.15)	-0.51(0.16)	-0.52(0.15)	-0.51(0.16)
Block 3	-0.27(0.15)	-0.28(0.15)	-0.29(0.16)	-0.29(0.15)
$\hat{\theta}$	0.18(0.07)	0.22(0.06)		
$\hat{\theta}_A$			0.12(0.10)	0.30(0.13)
$\hat{\theta}_B$			0.15(0.11)	0.27(0.11)
$\hat{\theta}_C$			0.32(0.13)	0.21(0.12)
$\hat{\theta}_D$			0.11(0.10)	0.12(0.09)

where $i = 1, \ldots, 224$ for plant i and $j = 1, \ldots, n_i; 1 \leq n_i \leq 10$ for cut j within plant i, X_{1ij} is a vector indicating which of the four blocks the plant belongs to and X_{2ij} which of the four treatments is allocated to cut j in plant i. The random effects are assumed to be independent realizations from a normal distribution, i.e. $b_i \sim N(0, \theta)$. The second model allows the frailty variances to differ between the four treatments, i.e. the covariance matrix for $b = (b_1, \ldots, b_{224})$ is diagonal with variances $\theta_A, \theta_B, \theta_C, \theta_D$ depending on the treatment allocation for the respective cut.

Table 1 shows estimates and standard errors for the parameters in the two models based on the MCEM algorithm and on the penalized partial likelihood estimating equations. For both models and estimation methods treatment B gives the shortest lifetime and treatment C the longest. When the model allows for differential variability, then the MCEM fit indicates that the lifetimes of roses treated with C vary the most. The difference between the variance component estimates is not, however, significant, and differential variability does not show in the penalized likelihood fit. The rose data are discussed in more detail in [24].

5.2 Alcohol related mortality in Finland

FIGURE 1. Raw standardized mortality ratios in Finnish municipalities.

In the second example we smooth alcohol related mortality rates in 452 Finnish municipalities, using Bayesian GLMM (for details of the study,

FIGURE 2. Posterior modes of the estimated standardized mortality ratios.

see Mäkelä, Ripatti and Valkonen [21]. The observed number of deaths O_i in municipality i are assumed to follow a Poisson distribution with expectation $\mu_i, i = 1, \ldots, 452$. Each μ_i is assumed to depend log-linearly on the logarithm of the expected number of deaths E_i and a municipality specific random effect b_i

$$\log(\mu_i) = \log(\mathrm{E}_i) + b_i. \tag{19}$$

The expected mortality E_i is computed based on the size and structure of the population in the municipality, and it is treated as fixed. Conditionally on E_i and b_i, the observed counts O_i, for $i = 1, \ldots, 452$, are assumed independent. For the random effects b_i, a Markov random field prior [2] is specified, with mean equal to the average of the effects from municipalities immediately adjacent to municipality i. The variance function for the random effects b_i is set to θ/k_i, where k_i is the number of municipalities adjacent to i, and θ is a random term, with $1/\theta$ following a gamma-distribution $\Gamma(1, 1)$. Gibb's sampling is used to draw from the posterior distribution $p(\theta, b \mid Y)$. Raw and smoothed standardized mortality ratios ($\mathrm{SMR}_i = \mu_i/\mathrm{E}_i$) based on posterior means are plotted in Figures 1 and 2, respectively. The more extreme SMR's in Figure 1 are smoothed in Figure 2, and a clear pattern of high mortality is shown in Northern and Eastern Finland, with lower rates in the West.

6 Discussion

We emphasise the parallel approaches to estimation and inference for the generalized linear mixed model and the frailty model. In our treatment of the frailty model the baseline hazard is profiled out. In penalized likelihood this is done following the profiling argument for the partial likelihood in the Cox model [16]. In the MCEM fit the complete data log likelihood is $\log p(Y \mid \lambda_0(t), \beta, \theta)$ in (5), and the M-step involves the standard partial likelihood procedure together with the Breslow estimator [4] for the cumulative hazard. Sampling in the E-step may be done using an independent sampler [24] or a dependent sampler [28]. For the Bayesian Markov chain Monte Carlo procedure an independent increments gamma-process may be used as the conditional distribution for the baseline hazard [7]. Note that although we derive the penalized likelihood estimating equations assuming Gaussian random effects, other distributions may be used for the MCEM and the MCMC procedures. All estimation and inference approaches are computer intensive. None can be singled out as universally best, but they all give acceptable accuracy over a wide range of conditions. The likelihood methods in sections 4.1 – 4.5 are justified by large sample arguments.

The penalized likelihood estimating equations are computationally simpler than the other methods, and they have been shown to perform well in many situations. A separate routine is, however, needed to give standard errors for the estimates, whereas Monte Carlo sampling in the E-step of the EM-algorithm provides the Louis' observed information matrix as a side product. The posterior procedures are conceptually attractive, void of ad hoc fixes. If there are plentiful of data, then likelihood inference and posterior inference will give similar results. The Bayesian approach to inference is, however, valid also when data are sparse, and the possibility to include informative priors allows external information to be added to the model in a coherent way. The overruling difficulty with the posterior MCMC sampling is to assess convergence. In contrast, the likelihood is monotonically increasing in each iteration, and assessing convergence is a non-issue. This applies to the MCEM procedure provided the Monte Carlo error is small.

By adding layers of random effects to the linear predictor of the generalized linear model or the multiplicative hazard model, a large and flexible class of models is offered for empirical use. Complex hierarchical structures and missing data constitute natural parts of the model specification. Although estimation and inferences are not straight forward, a unified and reasonably well understood framework is emerging. When incorporated into the applied statisticians toolbox this large class of models allows increased freedom and flexibility to tailor the statistical framework to the applied problem at hand. Formal or informal procedures to assess the sensitive of results to the model structure and distributional assumptions should be part of the toolbox.

Acknowledgments: This work was partially funded by research grants from the Yrjö Jahnsson foundation in Finland and from the Natural Science Research Council (NFR) in Sweden.

References

[1] Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes.* Berlin: Springer-Verlag.

[2] Besag, J.E., York, J.C., and Mollié (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

[3] Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B* **61**, 265–285.

[4] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89 – 99.

[5] Breslow, N.E. (1984). Extra-Poisson variation in log linear models. *Applied Statistics* **33**, 38 – 44.

[6] Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear models. *Journal of the American Statististical Association* **88**, 9–25.

[7] Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* **47**, 467–485.

[8] Clayton, D. (1994). Some aproaches to the analysis of recurrent event data. *Statistical Methods in Medical Research* **3**, 244–262.

[9] Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* **34**, 187–220.

[10] Dempster, A. P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

[11] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

[12] Gelman, A and Rubin D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

[13] Gilks, W.R., Richardson, S, and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice.* London: Chapman and Hall.

[14] Green, P. J. (1987). Penalized likelihood for general semi-parametric regression model. *International Statististical Review* **55**, 245–259.

[15] Hougaard, P. (1991). Modelling heterogeneity in survival data. *Journal of Applied Probability* **28**, 695 – 701.

[16] Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review* **51**, 158–262.

[17] Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48**, 795–806.

[18] Laird N.M. and Ware J.H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–74.

[19] Liang, K-Y. and Zeger, SL. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 11–22.

[20] Louis, T.A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society B* **44**, 98 – 130.

[21] Mäkelä, P., Ripatti, S, and Valkonen T. (2001). Alue-erot modesten alokoholikoulleisuudessa. *Suomen Lääkärilehti* in press.

[22] McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

[23] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models *Journal of the Royal Statistical Society A* **135**, 370–384.

[24] Ripatti, S., Larsen K., and Palmgren J. (2001) Maximum likelihood inference for multivariate frailty models using a Monte Carlo EM Algorithm. *submitted*.

[25] Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56** 1016-1022.

[26] Särkkä, L. E., Rita, H.J., and Ripatti, S. (2000). Cut rose flower longevity and its variation between plants of cv. Frisco grown in different lighting periods. *submitted*.

[27] Stiratelli R., Laird N.M., Ware H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–71.

[28] Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine* **19**, 3309–3324.

[29] Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454.

[30] Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *Journal of the American Statistical Association* **84**, 1065–1073.

[31] Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics 1982* , 144–148.

[32] Zhao, L.P. and Prentice, R.L. (1989). Correlated binary regession using a quadratic exponential model. *Biometrika* **77**, 642 –28.

Some trends in computational statistics

Peter Dalgaard[1]

[1] Department of Biostatistics, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen N, Denmark

Abstract: I shall review some recent developments and a number of ideas that are currently being actively pursued in the statistical computing community. The R project has had a major impact, not only in making an S-like environment available under Free Software conditions, but also in solidifying the interfaces and quality control mechanisms for user contributions, and thus providing an infrastructure which enables dissemination of ideas embodied in high quality software. The Omegahat project has not had the same visibility among general statisticians as R has, but it is nevertheless an active forum for experiments with many new ideas, notably component based software, inter-language communications, and web-based programming. Multiprocessing and event-based processing are important topics as are issues relating to large scale and computer-intensive problems.

Keywords: Computing; Software development; The R project; The Omegahat project

1 Introduction

Statistics is about dealing with data, and to a large part the science of statistics should deal with devising methods for the analysis of real-world data. This entails mathematical modeling of real-world events and the analysis of the mathematical properties of the models.
However, statistics cannot be a purely mathematical discipline. For the methods we build to have any impact on the fields for which they are designed, there has to be a layer of implementation, specifying how to apply the methods to actual data. In the not-too-distant past that meant setting up calculation schemes for sums of squares and the production statistical tables. Nowadays, it will generally involve software development, sometimes in the form of direct programming in a systems-level language such as Fortran or C, but more often in the form of ad-hoc programming in a general statistical system.
New statistical methods stand a much better chance of obtaining widespread use if they are made available in the form of software. Statistical software should be viewed as a form of scientific publication which is subjected to the strongest peer review when other statisticians use the software in practice.

The development of statistical software poses special challenges to the statistical scientist. Once one goes beyond ad hoc programming the need for proper structuring arises and one is led to consider the models in larger generality. Methods for model specification becomes an important issue.
It has long been recognized that flexible statistical software is facilitated by a true programming system. More recently, it has been realized that many potentially useful features are already provided by non-statistical software. Hence, a substantial effort currently aims at providing methods to tap into these resources from statistical programs.
In recent years, the supply of data has increased explosively, both due to computerized registration of routine data and because of the advent of new technologies, such as scanners and DNA microarrays. A major challenge to computational statistics is to find efficient methods to deal with huge amounts of data. This and other computer-intensive statistical methods make it desirable to make use of new hardware architectures such as networked clusters of large numbers of inexpensive computers.

2 Free software in statistics

Scientific software is a method of communicating methods and ideas. From that perspective, the traditional method of commercial distribution has shortcomings in that it hinders co-development by the scientific community and in many cases does not even allow the computer code to be reviewed. In the fields of numerical analysis and computer science there has always been a tradition of making algorithms available for scrutiny by peers. The free software movement originates in that academic tradition.
The need for free software can be (and is being) argued on principal grounds, but even from a purely pragmatic point the scientific community has gained much from its existence. It is liberating for a researcher to have available a large number of tools, which are often simply distributed with the operating system or can be downloaded from the net and installed in a matter of minutes. Many of the free software tools designed to replace existing tools have surpassed the originals in terms of both features and quality, and several new developments released under free software licences have been of major importance for the computing world at large.
Free software tends to promote the development of and adherence to standards. This allows interoperability between different programs, and gives the standards themselves a much better chance of being sensible and consistent.
One view of the role of public research is that it is a form of infrastructure, supporting industry as a whole. From that point of view it is difficult to see any problem with free software developed in academic settings being used in the in commercial world. On the other hand, the "social contract" of an open development model is one of mutual benefit. Users chip in with

whatever they can, and when a large number of talented people do that, there will be a large pay-back to every one of them. Since it is useful for development to have a wide user base and an open and friendly discussion of ideas, the community will accept an amount of freeloaders, and even people who require more support than they give back. However, developers are not particularly interested in providing free support for larger corporations, especially if same companies bind their employees with contracts preventing them from contributing modifications back to the community.
One potential downside to free software development lies in its anarchistic nature. Projects have become orphaned when the developers lost interest or found themselves with no time for maintaining the software. In some projects there has been a tendency to downplay quality control in favour of a rapid development cycle. For statistical software it is imperative to avoid releasing flawed software, certainly anything that could lead to incorrect analyses, since the results of a statistical analysis can have so far-reaching effects.

3 S and the R project

The S language devised by John Chambers et al. has played a major role in the development of statistical software. In addition to basic statistical functionality, it provides a full programming language with a consistent syntax allowing seamless extension. This was a significant improvement not only over procedurally oriented systems like BMDP, SAS, and SPSS procedural systems but also over earlier programming oriented systems like GLIM and Genstat. Initially, S was weak on code for standard statistical modeling but it eventually gained the functionality of the other systems. Not only that, it extended the functionality through the notion of *model objects* embracing data and parameter estimates, from which one can extract information, make predictions on new data, etc. The S language formed the cornerstone of the widely used commercial S-PLUS package.
In 1994 Ross Ihaka and Robert Gentleman started work on a small statistical system for teaching purposes, with a side view to trying out some ideas about the internal structuring of programming languages. The system, called R, was based on Scheme, but with a syntax "not unlike" S. After a while, they decided on placing their code under the General Public Licence. It quickly turned out to fill a gap for people who were accustomed to using S-PLUS, but needed to use it under more liberal circumstances than the licence could provide. This included teaching, use on private computers (many of which were running the Linux operating system already at that time), as well as research problems. Sometimes, this was to overcome situations where computer-intensive simulations would run for days, reducing the number of simultaneous users allowed, but there were also people interested in the internal workings of statistical programs and wanting to

try out new ideas. In 1996, the first mailing list for discussing the development of R was formed and in 1997 we had the formation of the Core Team, which currently consists of 15 people, each contributing in their own way to the project. E.g. Luke Tierney provided the new memory management which eliminated the fixed-size workspace in R 1.2, and Paul Murrell has done an enormous amount of work on the graphics code.
The functionality of R has advanced to a point where it matches that of S-PLUS in most respects and the books by Venables and Ripley can be used with R with small modifications.

3.1 The R development process

The R team considers it important that R is of a quality and stability at least as good as its commercial competitors. We seek to achieve this goal by thorough testing, a conservative release policy, and a quick response to bug reports from the community.
In connection with the build process, the person installing R has the opportunity to run "make check" which performs a number of test runs. One very basic item of that test suit simply ensures that all the examples on all help pages will run — a simplistic idea, but combined with a policy to have documentation for every function in R and associated meaningful examples, it has proven to be an effective guard against unexpected effects of changes to the internals. It is not quite sufficient for proper regression testing in which it is explicitly ensured that code that once caused problems will not do so ever again, since such examples are often esoteric and conflicts with the tutorial purpose of help page examples. In recent versions of R we therefore include a collection of true regression tests as a separate item. Martin Maechler has been a major player in the enforcement of these policies and also in building the parts of the test suite that run thorough checks for internal consistency throughout major areas of functionality, e.g. the check that the `as.xxx` functions returns values for which `is.xxx` is `TRUE`. No test suite for a project as complex as R will ever be perfect, and we keep striving for improving the one we have, but it has certainly played an important part in making R as stable and reliable as it is today.
Collaborative development with the geographic dispersion of developers of the R team poses particular challenges. It must be ensured that different developers do not make conflicting changes to the code base. We use CVS (Concurrent Version System) as a mechanism for ensuring that any conflicts are detected and resolved. This is in fact a rare occurrence in practice since developers tend to target different areas. It is more difficult to handle issues related to timing of changes. It must be ensured (as far as humanly possible) that all new developments are completed and have received sufficient testing before being officially released. On the other hand, some changes solve high priority problems and we want to be able to release them rather quickly. The

solution that we have adopted involves a branched development schedule corresponding to the three-part version numbers of R releases (e.g. R-1.2.3). Bug reports from users have contributed greatly to improving the quality of R. In order to keep this process alive, it is important to make sure that reports are dealt with adequately. Since developers will often not have time to deal with problems here and now, there is a very real danger that an issue after some initial discussion is simply forgotten. To make sure that a report leads to a solution eventually, we maintain a bug repository on a machine in Copenhagen, using the Jitterbug system originally developed for the Samba Unix/Windows file-sharing software. Jitterbug does have a number of shortcomings, but provides the basic functionality both via a web interface and via email.

3.2 Contributed packages

A major driving force in the early development of R was the possibility of porting code originally written for S. The rapid availability of functionality like Terry Therneau's survival package (ported by Thomas Lumley) was obviously a boost to the development of R.
It was quickly realized that it would be advantageous to have a centralized repository for contributed code as well as the core R distribution. This led to the formation of CRAN (Comprehensive R Archive Network, modeled on the CTAN and CPAN for TEX and Perl) which is a network of computers that mirror a master site in Vienna. Kurt Hornik and Friederich Leisch from the Technical University of Vienna were instrumental in making this happen.
However, we were not satisfied with just storing contributed code. For contributed code to be of maximum use, it is necessary to make it easy to install, make sure that it works and keeps on working as the development of R itself progresses. Experience with other repositories indicated that this could not be left as the responsibility of the authors, who may often be well excused for not foreseeing the kinds of problems that can arise. The measures that we have taken to achieve our goal are: (a) The formulation of an official API (Application Programming Interface) detailing the entry points inside R that user code can access — with the understanding that the API is not to be changed carelessly, (b) the formulation of a standard package format which — although it may be perceived as somewhat rigid by the package authors — is necessary both for making automated and cross-platform installation tools available, and for making sure that there is an official package maintainer and that the licencing situation is clarified, and (c) provide tools to check the consistency and documentation of each package, similar to the checks we apply to R itself.

3.3 Cross-platform compatibility

Software is most useful when it is available for the systems that people own. It has long been a goal to make R available for the three main system architectures around: Unix/Linux, Windows and Macintosh, but this goal has only recently been achieved.
The various variants of Unix and Linux have generally given us little trouble. Even though there is quite a large number of Unix implementations by different vendors, the problems associated with creating portable software for them are rather few, and their solution is mostly known from other free software projects. Tools like `autoconf` are a great help in sorting out the differences. We do see the occasional problem with vendor-supplied compilers on some fairly rare platforms, though.
The Windows platforms are popular and often perceived as user-friendly, but they are certainly not programmer-friendly when it comes to large systems like R. Robert Gentleman made the Windows versions for a while, using Microsoft compiler tools, but it became desirable to have similar tools to those available on Unix. Guido Masarotto enabled us to use the MinGW tools and thereby have a version that could be integrated with the Unix sources. Brian Ripley also contributed much to this port, in particular in the mechanisms for installation and for building packages.
The Macintosh version was initially built by Ross Ihaka for the computer labs in Auckland but lay unmaintained for quite a while. Recently, the port was picked up by Stefano Iacus who very quickly got it brought up to date with the other versions. Jan de Leeuw has put in some effort to make R work with the new Mac OS X, which is much more similar to Unix than earlier versions of Mac OS. If this new operating system becomes popular, we can expect support to be much easier in the future.
We have achieved full integration of the source trees for all platforms. Although there is still of course platform dependent code, this makes it possible to ensure that all binary versions of a given release is based on the same set of source files. Since almost all the files relating to the R interpreter and the statistical computations are common to all platforms, this keeps platform-specific problems to a minimum.

4 The Omegahat project

The Omegahat project grew out of a need to experiment with new ways of working with statistical software. One of the major ideas was to find a way to utilize developments in computer science at large rather than have the limited group of computer-literate statisticians develop features from the ground up.
Many of the things that we want (or might want) to do in statistics have already been done by others. For instance, excellent database programs exist, in both commercial and noncommercial flavours. For a statistician,

there is very little point in re-solving the problems associated with simultaneous transaction processing and security. What we need is to learn how to interface to the existing technologies and how these interfaces must reflect on specifically statistical methodology. Conversely, there is considerable interest in letting statistical functionality be embedded in other software. There is a range of data processing applications with statistical aspects, some of which — for instance computer vision — are exciting new areas of development, others are more straightforward, but in either case it is important to get a non-statistical audience to use statistical methods and software.

Omegahat is not a competitor to R, or any other statistical environment for that matter. In terms of traditional statistical procedures, there is simply nothing there. It is more useful to think of it as a workbench equipped with a set of tools which can be used to try out new ideas. What is currently present is an interactive Java-like language and a set of interface packages to access various kinds of functionality. This is not restricted to Java, although that is an important aspect of component-based programming. Many of the packages that are made available under the Omegahat heading are directly usable from R and S.

Much of the rest of this paper deals with ideas that originates in Omegahat, although there have also been contributions from other sides.

5 Software components

The notion of constructing software in the form of reusable blocks with well-defined interfaces between them has been around in the computer world for quite a while. The immediate benefits of such a structure are obvious: No need to write your own data entry module if you can just call up a spreadsheet to do the job, for instance. The idea is strongly linked with *object-oriented programming* because the concept of invoking methods contained in objects can be extended to active objects that are not necessarily parts of the same program, they might not even reside on the same machine. Thus all sorts of client/server applications can be implemented from basic building blocks.

However, this will not work without communication protocols so that objects can contact each other in an orderly fashion. Notice that several applications might be running and there might be several instances of programs accessing several instances of active objects. The CORBA (Common Object Request Broker Architecture) specification organizes the communication by using an ORB (you guessed it: Object Request Broker) which is a program which takes care of announcing services to the application and routing requests to the relevant objects, activating/deactivating them as necessary.

Another method for object communication is provided by Microsoft's COM and DCOM interfaces. This is essentially limited to Microsoft's own plat-

forms, but within that world there is a substantial activity in providing complex application using Visual Basic to bring different components together. A statistical application of this sort is described by Ripley and Ripley and a bidirectional interface between R and the Excel spreadsheet has been developed by Neuwirth and Baier.
In the more open world of free software and Unix-like systems, CORBA appears to be a more promising direction. In particular, CORBA is the foundation of the Bonobo architecture for the Gnome desktop environment. Bonobo is scheduled to be used for many Gnome applications, and deployment of this software is about to happen at the time of writing.

6 Inter-language interfaces and embeddings

An idea which is related to component reuse is that it can be useful to access software written in different programming languages. For instance, a large amount of code has been written in the Java programming language, and much of this code could be utilized in statistical applications, especially in the graphics area. Also, some languages are specially designed to support certain tasks easily. For accessing code written in other compiled languages, R and S have long had the technique of dynamic loading of C and Fortran routines, but for interpreted languages things get a bit more complicated because both sides of the interface expect to have control over the run-time environment both in terms of event loop and object management.
As part of the Omegahat project, Duncan Temple Land and John Chambers have provided methods to access Java, Python, and Perl from S and R and conversely embedded R in other applications and languages like Netscape, the Postgres database, and XML. My own modest contribution to this area has been an interface between R and the Tcl/Tk language and toolkit for graphical user interfaces. Some very interesting projects concern the embedding of dynamic graphics functionality like `ggobi` (Duncan Temple Lang and Deborah Swayne) and OpenGL (Duncan Murdoch) in R. At a different level, Doug Bates and Saikat DebRoy have worked on accessing `C++` classes from R with a view to accessing more modern numerical libraries.

7 Database interfaces

Not all large data sets are handled well by statistical programs. Consider for instance the output of a medical scanner providing a number of images for each patient. It would be absurd to store such data in a traditional cases-by-variables data matrix whether one takes patients, images, or pixels as the basic observational unit. For longitudinal data one has a similar problem of representing irregular time series.

In many real-world data collection efforts (with or without statistical aspects) one would use relational databases. Typically, this contains several linked "tables" (database parlance for dataframe) where for instance one table records basic information for each patient and another table records visits and just need to store the patient ID. To extract information from such a database, the de facto standard Structured Query Language (SQL) has emerged. Database software also solves many other problems such as multiuser access and atomicity of transactions (making sure that withdrawing money from one account and depositing it into another either both succeed or both fail) as well as computer efficiency. Thus, it is very attractive for the statistician to let the database software do what it is good at and look for ways to have statistical software access existing databases, preferably in ways that are transparent to the user. Several efforts are underway for R/S, see the overview paper by Hothorn, James, and Ripley. The basic mechanics of pulling data to and from databases using SQL and connection methods are well studied by now, but questions remain e.g. whether one can usefully hide the SQL from the user by the use of proxy dataframes.
Another aspect of the use of databases is how it should reflect on the specification of statistical models. It seems to be necessary to find ways of using at least hierarchical database structures in the specification of mixed-effects models (cf. the book by Pinheiro and Bates) so that data duplication can be minimized, and the in-memory storage of entire data sets can be avoided. Robert Gentleman describes some ideas based on allowing elements of data frames to be complex objects.

8 Markup languages

Anyone who has clicked "view source" in a web browser will have seen the Hypertext Markup Language (HTML) that web documents are written in with its `<i>italics</i>` tagged format. HTML is mainly *visual markup* describing font choices and so forth. At a higher level of abstraction there is *logical markup* describing which parts of a document are (say) sections, chapters, etc. This sort of markup can be described in the Standard Generalized Markup Language (SGML) of which HTML is a specific application. Other applications of SGML can be defined through Document type definitions (DTD) which, briefly put, define what each SGML tag is supposed to do.
The eXtensible Markup Language (XML) is a slightly simplified subset of SGML, which has gained considerable momentum in later years. One feature of particular interest, separating it from the fixed-form HTML is the use of *stylesheets* which can be specified in an eXtensible Stylesheet Language (XSL). The point of the latter is that it provides a transformation engine whereby an XML document can be transformed into virtually anything. Although originally intended for rendering of text documents, XML

is proving to be something of a Swiss Army knife and finding its way into a variety of areas of relevance for statistical computing:

Data specification. XML deals with *content description* and thus it is an obvious idea to use it as a method for describing data sets. Such a description could be much better than the traditional "flat file" by including specifications of the encoding of the data and also *metadata* on who collected the data and when and for what purpose.

The StatDataML project aims at turning the XML approach into a standard for data sets that could replace current proprietary and non human-readable formats. Virtanen describes a practical situation in which XML has been used for transporting data between different platforms

Documentation of programs. A powerful aspect of the XML/XSL mechanism is that it is possible to invoke external software on parts of a document. For documentation of software, this allows you to include the source rather than the output of commands. Anyone who has worked on writing tutorials knows how hard it is to ensure that the output is really what the commands generate, especially when dealing with programs that evolve rapidly. In the context of R, XML may help us to overcome shortcomings of the current help system, where in particular the "Examples" sections contain items that serve a variety of purposes: some are pedagogical, some display fine points, some are really regression tests or test for internal consistency, some are "show-offs", and some generate useful displays and tables. With XML coupled to a more advanced viewing system, we could provide a much more flexible way for the user to select the items that are helpful.

Literate statistical analysis. Donald Knuth invented the term "literate programming" and proved its viability by using it to develop his TEX and METAFONT packages. The basic idea is to have one document, embodying both code and the associated description, which you can "tangle" to extract the program and "weave" to view as a pretty-printed document. Anthony Rossini has experimented with using similar techniques to describe the course of a statistical consulting situation, using XML and XSL as a vehicle.

9 Distributed computing

Typical computer installations these days consist of multiple computers of varying capabilities. Usually they spend the most of their time in an idle state, but nevertheless bottlenecks occur when several people decide to use the same server at the same time. New methodologies constantly push the envelope of what can be done with computers, but even for well-established procedures there is a demand for larger and larger simulation studies. Consequently, there is a constant demand for optimal utilization of resources and investigations into improvements of computer architectures. The massively parallel architectures that were much talked about a decade ago seem to be slow in reaching affordable levels. Machines with a few CPUs

operating in parallel are getting more common, but the most immediate potential seems to lie in clustering of many inexpensive computers to form a network with a high total throughput. The Beowulf type cluster of networked PC hardware running Linux is quite popular in high-performance computing.
Such clusters can be used to run parallelized algorithms although the gain there is often lost to communications overhead. More assured success is obtained by having multiple processes work on problems with coarse-grained parallelism, such as Monte Carlo simulations.
The biostatistics department in Seattle has access to a cluster running Mosix and the computer science department in Wisconsin runs a Condor cluster which the statisticians have access to. R has been tried on both of these installations with good results.

10 Concluding remarks

The longevity of programming languages in statistical computing is striking, individual languages come and go, but the concept itself stays on in spite of all attempts to make computer usage non-verbal. This is natural. Like the language of mathematics the languages of computers deal with formal specifications. Computer languages have the unique feature that they can be operationalized and run as programs, which has the gratifying effect of providing some verification of the specifications. However, making a computer program run is only a small part of the work. Proper design of software requires a substantial theoretical effort to understand the principles of what one is trying to build software for, and the conceptual models involved are intellectual constructions in their own right. This applies to statistical computing whether one talks about numerical issues, the process of bringing general mathematical models on a computable form, or tools for graphics or database access.

Acknowledgments: Special thanks to all participants at the Workshop on Distributed Statistical Computing, Vienna, March 2001, who provided a large portion of the material reviewed in this paper.

References

Becker RA, Chambers JM, Wilks AR (1988) *The New S Language*. London: Chapman & Hall.

Chambers JM, Hastie TJ (1992) *Statistical Models in S*. London: Chapman & Hall.

Chambers JM (1998) *Programming with Data* New York: Springer.

Venables WN, Ripley BD (1999) *Modern Applied Statistics with S-Plus*. 3rd Edition. Springer.

Venables WN, Ripley BD (2000) *S Programming*. Springer.

Pinheiro JC, Bates DM (2000) *Mixed-Effects Models in S and S-Plus*. Springer.

Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Hornik K, Leisch F (ed.) (2001) *Proceedings of the 2nd International Workshop on Distributed Statistical Computing (DSC 2001)*. Technische Universität Wien, Austria: Online document, ISSN 1609-395X, http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings.

Bates DM, DebRoy S (2001) `C++` Classes for R Objects. DSC-2001 (ibid.)

Dalgaard P (2001) The R-Tcl/Tk interface. DSC-2001 (ibid.)

Gentleman R (2001) Modeling with Objects. DSC-2001 (ibid.)

Hothorn T, James DA, Ripley BD (2001) R/S Interfaces to Databases. DSC-2001 (ibid.)

Iacus S (2001) R porting for the Macintosh. DSC-2001 (ibid.)

Murdoch D (2001) RGL: An R Interface to OpenGL. DSC-2001 (ibid.)

Neuwirth E, Baier T (2001) Embedding R in Standard Software, and the other way round. DSC-2001 (ibid.)

Ripley BD, Ripley RM (2001) Applications of R Clients and Servers. DSC-2001 (ibid.)

Rossini A (2001) Literate Statistical Analysis. DSC-2001 (ibid.)

Temple Lang D, Swayne DF (2001) GGobi meets R: an extensible environment for interactive dynamic data visualization. DSC-2001 (ibid.)

Temple Lang D (2001) Embedding S in Other Languages and Environments. DSC-2001 (ibid.)

Virtanen M (2001) Distributing data to different platforms using XML. DSC-2001 (ibid.)

The R project http://www.r-project.org

The Omegahat project http://www.omegahat.org

The Comprehensive R Archive Network http://cran.r-project.org

The MinGW project http://www.mingw.org

Analyzing ion channel time series by hidden Markov models

J. Timmer[1], M. Wagner[2]

[1] Center for Data Analysis and Modelling, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany
[2] the scientific consulting group, Bismarkallee 9, 79098 Freiburg, Germany

Abstract: Ion channels are proteins in the membrane of cells. These proteins can change their configuration between different discrete states. The dynamics is believed to be Markovian. In some of the states (open) they conduct ions, in others (closed) not. In experiments, only the currents are observed and, thus, the observed current follows an aggregated discrete state Markov process. The graph of allowed transitions usually comprises loops. We show that in this case the parameters are not identifiable if dwell times of either closed or open states are equal and demonstrate consequences for the estimation in the case of nearly equal dwell times. If the system is in equilibrium, the dynamics has to obey the law of detailed balance. We investigate a likelihood ratio test for detailed balance and derive a scale to judge whether the violation is biologically relevant. We discuss how likelihood ratio tests can be applied for model selection in the case of nested and of non-nested models. Measured ion channel time series usually contain a large amount of additive observational noise calling for a treatment by hidden Markov models. We introduce a generalization of hidden Markov model that is capable to deal with colored observational noise and report the analysis of measured time series from a Na^+ channel.

Keywords: Hidden Markov models; parameter estimation; identifiability

1 Introduction

Biological aspects

Ion channels are proteins in the membrane of cells. Their main physiological task is to regulate the concentration of ions in the cell and the propagation of electrical signals. The protein forms a pore which can open and close by changing the configuration of the aminoacid-chain. In its open state the channel is not permeable for every type of ion but highly selective. A biological fact which leads to interesting mathematical challenges is that different configurations of the protein can lead to a conducting, respectively non-conducting the ion channel. The recording of the current through single channels became possible by the Nobel price honored patch-clamp technique [16]. The structure of the protein forbids certain transitions between

its different configurations. The corresponding graph is called the *gating scheme*. The inference of the gating scheme from measured time series is a prominent task of the data analysis [2]. The dynamics of the transitions is believed to be Markovian [12]. The rate constants of the transitions depend on e.g. the temperature, concentration of certain ligands in the cell, trans-membrane potential or mechanical stress. Some ion channel produce stationary time series. Especially voltage dependent ion channels which are involved in the propagation of electrical signals are usually closed, called inactivated, until they are depolarized, leading to a transient process of some openings and closings and end up in the inactivated state again [10]. Genetic point mutations leading to the exchange of single aminoacids in the protein prohibit the final inactivation and are basis for different diseases. A detailed discussion of ion channels, the measurement procedure and first steps of the analysis is given in [19].

Mathematical aspects

The mathematical challenges arising in the modelling of ion channel data first stem from the aggregation of different protein configurations in the open, respectively the closed state. Second, observational noise introduced nontrivial issues. The observational noise is on the one hand produced by the channel itself. On the other hand it is generated by the amplifier that is necessary to record the small current in the order of pA. The problem becomes even more involved because of the anti-aliasing filter that has to be applied before sampling the time series.

The Markovian dynamics can be described by time-discrete models invoking transition probabilities:

$$a_{ij} = p(x(t+\Delta t) = j | x(t) = i), \quad i,j = 1,\ldots,s, \quad \text{with } \sum_j a_{ij} = 1, \quad \forall i$$

or by time-continuous models invoking rate constants

$$\dot{P}_j = \sum_i P_i q_{ij}, \quad \text{with } q_{ii} = -\sum_{j \neq i} q_{ij}, \quad \forall i \quad .$$

The transition probability matrix A is related to the generator matrix Q by:

$$A = exp(Q\Delta t)$$

Two reasons advocate the use of the time-continuous concept. First, external influence on the dynamics act on the rate constants. Second, assuming that the underlying dynamics is time-continuous or at least orders of magnitude faster than the sampling, the concept of gating schemes, i.e. forbidden transitions between certain states, for aggregated processes can only be formulated in terms of rate constants: The transition probabilities will always be strictly positive.

The aggregation of states leads to the issue of model equivalence and model, resp. parameter non-identifiability. Fig. 1 shows all possible gating schemes for a channel with two open and one closed state.

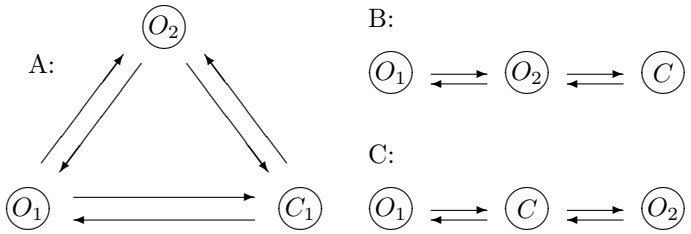


FIGURE 1. Example for model equivalence. All possible Markov models with two open states and one closed states are equivalent.

As proven in [11] these three gating schemes are equivalent with respect to the observed currents, i.e. for each model there are rate constants such that the statistical properties of the other models can be reproduced. Therefore, it can not be discriminated between these models based on measured data (taken at constant conditions. If the conditions, e.g. temperature is varied, the models become identifiable). Furthermore, the loop-model A comprises 6 parameters, while the others comprise only 4. Therefore, two of the parameters of model A can not be identifiable. In models with n_O aggregated open and n_C aggregated closed states the maximum number of identifiable parameters is $2n_O n_C$ [6, 7]. A sufficient condition for non-identifiability if this upper bound is not reached is given in [25].

A first approach to deal with the observational noise was to low-pass filter the measured time series. If the signal-to-noise ratio is low, heavy filtering will also discard short events of openings or closing. For a review of possible treatments of the *missed events* problem, see e.g. [1] and references therein. A more natural approach is to incorporate the observational noise into the model. This leads to hidden Markov models (HMM). In its simplest version, an HMM consists of two parts, a nonobservable Markov chain $\{X(t)\}$ and an observable random variable $\{Y(t)\}$. Given $\{X(t)\}$, $\{Y(t)\}$ is conditionally independent with conditional distribution of $Y(t)$ depending only on $X(t)$. Usually, the conditional distributions of $Y(t)$ given $X(t)$ all belong to a single parametric family, in the case considered here the Gaussian distribution.
Maximum likelihood parameter estimation in HMMs can be performed by the EM-algorithm, which in this frame is called *forward-backward algorithm* [17, 14]. The asymptotic normality of the MLE for HMMs was proven only recently [3]. HMMs were first applied to ion channel data in [4].

2 Parameter identifiability in loop models

All realistic gating schemes comprises at least one loop. A simple, non-trivial example is given in Fig. 2.

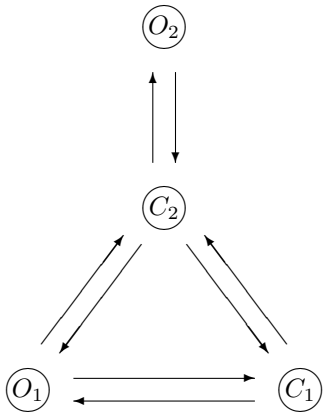


FIGURE 2. Gating scheme with one loop and four states. "O" and "C" denote an open and a closed state, respectively.

If the mean dwell times in the two open, resp. the two closed states are different, the rate constants of this model are identifiable, i.e. the statistical properties of the observed current are different for different parameters or, in other words, the likelihood has a unique maximum. If either the open or the closed dwell times are equal, the parameters are no longer identifiable [23]. The Hessian at the maximum likelihood point becomes singular and it is possible to vary the parameters of the generator matrix Q in a way that the gating scheme remains the loop gating scheme while the likelihood does not chance.
Under realistic conditions the dwell times will not be equal, but might be similar. In this case, the "nearness" to the non-identifiable case effects the finite sample properties of the MLE by enlarging the confidence regions. For realistic amounts of typical data, the confidence regions are effected even for a ratio of the dwell times of 3 [23].

3 Testing for detailed balance

In thermodynamic equilibrium, the dynamics of the gating are subject to the principle of detailed balance. A violation of detailed balance indicates the presence of an external energy source. If detailed balance holds, the clockwise multiplied rate constants in a loop must equal the counterclock

wise multiplied. This reduces the degree of freedoms by one and allows for a likelihood ratio test:

$$\begin{aligned} H_0: &\quad \text{Detailed balance is fulfilled} \\ H_1: &\quad \text{Detailed balance is not fulfilled} \end{aligned}$$

The likelihood ratio :

$$LR(\hat{\gamma}, \hat{\theta}) = 2(L_n(\hat{\gamma}) - L_n^{DB}(\hat{\theta}))$$

should follow

$$LR(\hat{\gamma}, \hat{\theta}) \overset{n\to\infty}{\sim} \chi_1^2$$

under H_0.
In the previous section it was shown that parameters in loop models are not identifiable if open/close dwell times are equal and that confidence regions are large if they are similar. In the present setting non-identifiability in the equal dwell time case means that a model that obeys detailed balance can be transformed in one that not fulfills this condition. Therefore the test has no power. Analogously to the setting of the previous section here there results a loss in power for the nearly equal dwell time case [25].
Due to the dichotomy of Kakutani [21], if a test has any power, asymptotically its power reaches one. Therefore, it is important to judge whether a violation of the null hypothesis of detailed balance is a relevant one in terms of biology. Rate constants are related exponentially to activation energies by Arrhenius' law. Taking the logarithm of the ratio of the rate constants multiplied clock- and counterclockwise transforms the constraint of detailed balance in a relation between activation energies. A natural scale for this energies is given by the work needed to push an elementary charge against the membrane potential of typical 70 mV. Thereby, it is not only possible to test the statistical significance, but also to judge the biological relevance of a violation of the principle of detailed balance [25].

4 Model selection

For the task of model selection, the case of nested and non-nested models have to be distinguished.

4.1 Nested models

The classical result that under the null hypothesis, the twofold log-likelihood ratio is distributed asymptotically as χ^2 with the number of degrees of freedom given by the difference of the numbers of parameters of the model classes, especially depend on the following five assumptions [22]:

1. the model classes are nested,
2. the model classes are not misspecified,
3. the maximum likelihood estimators are asymptotically normally distributed,
4. the true parameters are not part of the boundary of the parameter space,
5. all nuisance parameters are identifiable under the null hypothesis.

Consider the case that one wants to test for an additional open state as depicted in Fig. 3.

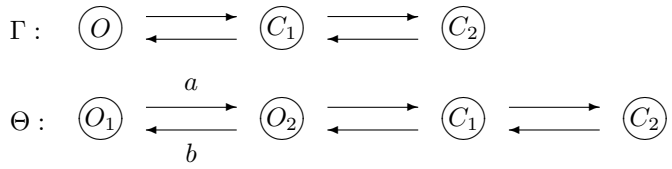


FIGURE 3. Example of two model classes of Markov models. Γ is nested in Θ by requiring $b = 0$. Furthermore, the rate a is only identifiable, if $b > 0$.

For this situation, the hypotheses read:

$$\begin{aligned} H_0 : &\quad \text{true model} \in \Gamma \\ H_1 : &\quad \text{true model} \in \Theta \setminus \Gamma \ . \end{aligned}$$

or, more explicit:

$$\begin{aligned} H_0 : &\quad b = 0 \\ H_1 : &\quad b \neq 0 \end{aligned}$$

Therefore, the last two conditions (and by condition 3 also condition 2) are violated leading to nonstandard LRTs.
As in the example, if exactly one parameter is part of the boundary like a rate constant constraint to be positive, it has been shown [20] for independent random variables, that the likelihood ratio obeys

$$LR_{y_1,\ldots,y_T}(\hat{\theta}) \sim \frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_0^2 \quad \text{for } T \to \infty \ ,$$

where χ_0^2 is the Dirac measure on the point zero. Since not the dependence structure of the Markov process but only the asymptotic normality of the

parameter estimates determine the distribution of the likelihood ratio, this results carries over to (hidden) Markov models.

Since parameter b is constraint to be zero under the null hypothesis, parameter a can not be identified. For the case of non-identifiable parameters under the null hypothesis no analytical results concerning the distribution of the test statistic are known. A procedure to obtain upper bounds for the quantiles of the likelihood ratio in the case of independent random variables was proposed in [9]. Unfortunately, this result is not easily extendable to (hidden) Markov models.
Parametric bootstrap [5] might offer a strategy to treat this problem [24]. Simulation studies indicate that the test statistic, again, follows a mixture of χ^2 distributions and that disregarding the non-identifiability problem leads to conservative tests.

4.2 Non-nested models

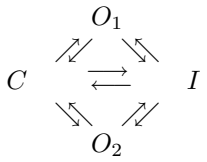
Likelihood ratio testing is not directly applicable for model selection of non-nested gating schemes. In order to use this method from the nested case, the non-nested gating schemes have to be embedded in a general model. Arbitrary complex gating schemes, however, cannot serve for this purpose because firstly, the number of identifiable parameters in aggregated Markov models is limited as mentioned in the Introduction. Secondly, gating schemes are typically embedded in other gating schemes by constraining certain transition rates to zero, so that these transition rates are part of the boundary of the parameter space leading to the challenges discussed in the previous section.
For selecting between different gating schemes it is not necessary that a general model can be interpreted as a gating scheme. It is sufficient that this model provides a parameterization of the likelihood functions of all proposed gating schemes and that these gating schemes are not on the boundary of the parameter space. If the candidate models all comprise the same number of open and of closed states such a parameterization can be obtained [26].
This leads to a two step selection procedure: In the first step, it is tested whether any of the proposed models is consistent with the data. In the second step it is evaluated whether one of the models has to be preferred to the others. The possible outcomes include the case that their might not be enough data to select exactly one model [26]. No analogous procedure is known if the candidate models comprise different numbers of open, resp. closed states.
In the Bayesian framework, another possibility for model selection in (hidden) Markov models is provided by the reversible jump Markov Chain Monte Carlo method [18].

5 Correlated output noise, application to measured time series

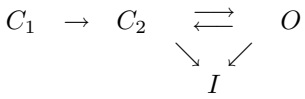
All phenomena and distributional results for aggregated Markov models of the previous section carry over to aggregated HMMs.
In the case of correlated output noise the property conditional independence of the $Y(t)$ given $X(t)$ is lost. Mainly two types of correlation structure has been treated in the literature. The case of autoregressive (AR) processes can be dealt with by state space augmentation [8]. The moving average (MA) case is more involved. An approximative MLE for this and the general ARMA case is given in [15]. Often, because of invertibility of the processes finite order AR processes can well be approximated by finite order MA processes and vice versa. Unfortunately, the usually applied anti-aliasing filters who present a major source of correlation output noise present MA processes that are not invertible. Therefore, the MA process has to be modeled explicitely.
We report an analysis of time series from a voltage dependent Na^+ channel [13]. Fig. 4 shows five of 600 traces of the transient behavior of the channel.

FIGURE 4. Representative raw current traces of single channel recordings. The vertical line marks the beginning of the activating depolarization. Openings (plotted downwards) occur at the beginning of traces 2 and 5.

Starting from the most simple model $C \rightleftarrows O$, we investigated by likelihood ratio tests successively more complex physiological reasonable gating schemes for the unfiltered and the filtered HMM. For the unfiltered HMM the model selection procedure resulted in:



For the filtered HMM the result was:



The result is reasonable: The unfiltered HMM has to propose an additional open state to capture the correlation structure in the time series that is due to the correlated output noise. The forward model selection strategy is not completely satisfactory on theoretical grounds. Therefore, we performed two consistency checks of our result. Fig. 5 compares the theoretical time course of the mean current for the two selected models with the one obtained from the measured time series. The superiority of the filtered HMM is obvious.

FIGURE 5. Theoretical time course of the mean current for the filtered HMM (solid smooth curve) and for the unfiltered HMM (dotted) compared to the mean of the 600 traces. The vertical line marks the beginning of the activating depolarization.

Additionally, we simulated time series from the fitted filtered HMM and repeated the model selection procedure. As result, the same gating schemes as in the original analysis were selected for both types of HMMs, supporting

the conclusion that the unfiltered HMM has to introduce a spurious second open state to deal with the colored output noise.

6 Summary

Interpreting the notion of "non-identifiability" in a wide sense, non-identifiability is omnipresent in aggregated Markov – for the noiseless case – and aggregated hidden Markov models – for the noisy case.
There is non-identifiability of

- different open (closed) states.
- open and closed states, if signal-to-noise ratio is low.
- equivalent models.
- parameters.
- detailed balance models, if dwell times are equal.
- (continuous time) topology if discrete time models are used.

(Hidden) Markov models viewed as dynamical description of ion channel time series offer interesting statistical challenges and allow for applications yielding otherwise unobtainable information about an important physiological system.

Acknowledgments: We would like to thank S. Michalek for fruitful discussion and for providing the analysis of the measured data.

Bibliography

[1] F.G. Ball and S.S. Davies. Statistical inference for a two-state Markov model of a single ion channel, incorporating time interval omission. *J. Roy. Stat. Soc. B*, 57:269–287, 1995.

[2] F.G. Ball and M.S.P. Sansom. Ion-channel ganting mechanisms: model identification and parameter estimation from single channel recordings. *Proc. Roy. Soc. Lond. B*, 236:385–416, 1989.

[3] P.J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals Stat.*, 26:1614–1635, 1998.

[4] S.-H. Chung, J. Moore, L. Xia, L.S. Premkumar, and P.W. Gage. Characterization of single channel currents using digital signal processing techniques based on hidden markov models. *Proc. Roy. Soc. Lond. B.*, 199:231–262, 1990.

[5] A.C. Davison and D.V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, 1997.

[6] D.R. Fredkin, M. Montal, and J.A. Rice. Identification of aggregated Markovian models: apllication to the nicotine acetylcholine receptor. In L.M. Le Cam and R.A. Olshen, editors, *Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer*, volume I, pages 269–289, Belmont, 1985. Wadsworth, Inc.

[7] D.R. Fredkin and J.A. Rice. On aggregated Markov processes. *J. Appl. Prob.*, 23:208–214, 1986.

[8] J.D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.

[9] B. E. Hansen. The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. *J. Appl. Econom.*, 7:61–82, 1992.

[10] B. Hille. *Ionic Channels of Exitable Membranes*. Sinauer, Sunderland, Mass., 1992.

[11] P. Kienker. Equivalence of aggregated markov models of ion-channel gating. *Proc. Roy. Soc. Lond. B*, 236:269–309, 1989.

[12] S.J. Korn and R. Horn. Statistical discrimination of fractal and markov models of single-channel gating. *Biophys. J.*, 54:871–877, 1988.

[13] S. Michalek, H. Lerche, M. Wagner, N. Mitrovic, M. Schiebe, F. Lehmann-Horn, and J. Timmer. On identification of sodium channel gating schemes using moving-average filtered hidden Markov models. *Euro. Biophys. J.*, 28:605–609, 1999.

[14] S. Michalek and J. Timmer. Estimating rate constants in hidden Markov models by the EM algorithm. *IEEE Trans. Signal Proc.*, 47:226–228, 1999.

[15] S. Michalek, M. Wagner, and J. Timmer. A new approximate likelihood estimator for ARMA-filtered hidden Markov models. *IEEE Trans. Signal Proc.*, 48:1537–1547, 2000.

[16] E. Neher and B. Sakmann. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, 260:799–802, 1976.

[17] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257–285, 1989.

[18] C.P. Robert, T. Rydén, and D.M. Titterington. Bayesian inference in hidden Markov models by the reversible jump markov chain monte carlo method. *J. Roy. Stat. Soc. B*, 62:57–75, 2000.

[19] B. Sakmann and E. Neher. *Single-Channel Recording*. Plenum Press, New York, 1995.

[20] S. G. Self and K. Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Ass.*, 82:605–610, 1987.

[21] A. Shiryaev. *Probability*. Springer, Berlin, 1995.

[22] Q. H. Vuong. Likelihood ratio tests for modelselection non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

[23] M. Wagner, S. Michalek, and J. Timmer. Estimating rate constants in hidden Markov models with loops and with nearly equal dwell times. *Proc. Roy. Soc. B*, 266:1919–1926, 1999.

[24] M. Wagner, S. Michalek, and J. Timmer. Testing for the number of states in hidden Markov models with application to ion channel data. In W. Gaul and H. Locarek-Junge, editors, *Studies in Classification, Data Analysis and Knowledge Organization*, pages 260–267, Heidelberg, 1999. Springer.

[25] M. Wagner and J. Timmer. The effects of non-identifiability on testing for detailed balance in aggregated Markov models of ion-channel gating. *Biophys. J.*, 79:2918–2924, 2000.

[26] M. Wagner and J. Timmer. Model selection in non-nested Markov models for ion channel gating. *J. Theo. Biol.*, 208:439–450, 2001.

Event history analysis: overview

Niels Keiding[1]

[1] Department of Biostatistics, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen N, Denmark

Abstract: In event history analysis individuals are assumed to move between states. Simple cases include survival analysis with two states "alive" and "dead"; several types of failure or competing risks with transitions possible from "alive" to "dead of cause i" for $i = 1, \cdots, k$; and illness-death or disability models with three states - transitions are allowed back and forth between "healthy" and "diseased" and from each of these to "dead".
This survey will outline the powerful counting process approach to event history analysis, and it will focus on three features: interaction between life history events; the use of multistate models for prediction, both in this world and in hypothetical worlds where some transition rates are artificially changed (application: bone marrow transplantation); and the role of the sampling design and of unobserved heterogeneity ("frailty") in models for repeated events (application: repeated admissions of psychiatric patients).

Keywords: Survival analysis; Multi-state models; Counting processes; Aalen-Johansen estimator; Markov processes.

1 Introduction

Event history analysis deals with data obtained by observing individuals over time focusing on events occurring for the individuals. Thus, typical outcome data consist on *times of occurrence* of events and on the *types of events* which occurred. Frequently, an event may be considered as a *transition* from one state to another and, therefore, *multi-state models* will often provide a relevant modeling framework for event history data. Multi-state models are discussed from several points of view in the books by Andersen et al., (1993); Blossfeld and Rohwer (1995) and Courgeau and Lelièvre (1992); see Hougaard (1999) and Commenges (1999) for recent surveys.
There are two broad purposes of event history modelling: *analysis* where the interest is in the statistical modelling of each individual transition, including its possible dependence on internal or external covariates, and *synthesis* where the combination of several transitions is described via suitable summary measures.
This presentation summarizes some of my recent work in the area, for further reference see Keiding (1998, 1999), Keiding et al. (2001) and Andersen

and Keiding (2001).

2 Multivariate counting processes and the Aalen-Johansen estimator

In *event history analysis* individuals are assumed to move between states. Simple cases include *survival analysis* with two states "alive" and "dead" and transition only possible from "alive" to "dead"; *several types of failure* or *competing risks* with transitions possible from "alive" to "dead of cause i" for $i = 1, ..., k$; and *illness-death* or *disability* models with three states: transitions are allowed back and forth between "healthy" and "diseased" and from each of these to "dead".
A flexible framework for statistically modelling such problems is given by *multivariate counting processes*, se Andersen et al. (1993) for a comprehensive exposition of the mathematical theory with many worked practical examples. A (univariate) counting process $(N(t), t \in \mathcal{T})$ on an interval $\mathcal{T} = [0, \tau)$ or $[0, \tau], \tau \leq \infty$, is a stochastic process with $N(0) \equiv 0$ and whose sample functions are step functions with steps $+1$. The *multistate models* in event history analysis are then specified by a *multivariate counting process* $(N(t)) = (N_1(t), \ldots, N_k(t), t \in \mathcal{T})$ counting the transitions between each pair of states as just described. Each component of $\mathbf{N}$ is a univariate counting process as defined above, and with probability one, no two components may jump simultaneously.
The "history" $(\mathcal{F}_t, t \in \mathcal{T})$ of the multivariate counting process is mathematically specified as a family of σ-algebras which is increasing: $s < t \Rightarrow \mathcal{F}_s \subset \mathcal{F}_t$ and right continuous: $\mathcal{F}_s = \bigcap_{t \geq s} \mathcal{F}_t$ for all s. The development in time of a multivariate counting process is assumed to be governed by its (random) intensity process $(\lambda(t), t \in \mathcal{T})$ where $\lambda = (\lambda_1, \ldots, \lambda_k)$ and $\lambda_h(t)dt$, heuristically speaking, is the conditional probability of a jump of N_h in $[t, t + dt)$ given the "history" $\mathcal{F}_{t-}$ up to, but not including t.
In this framework techniques based on martingales and stochastic integrals allow for very general modelling and statistical inference, in particular very general censoring patterns are readily handled.
Consider now a nonhomogeneous, time-continuous Markov process $X(t)$ on $\mathcal{T} = [0, \tau]$ or $[0, \tau)$ with finite state space $\{1, 2, ..., k\}$ having transition probabilities $P_{hj}(s, t)$ and transition intensities $\alpha_{hj}(t)$. For n conditionally (given the initial states) independent replications of this process, subject to quite general censoring patterns, the multivariate counting process $\mathbf{N} = (N_{hj}; h \neq j)$, with $N_{hj}(t)$ counting the number of observed direct transitions from h to j in $[0, t]$, has intensity process $\lambda = (\lambda_{hj}; h \neq j)$ of the multiplicative form $\lambda_{hj}(t) = \alpha_{hj}(t)Y_h(t)$. Here $Y_h(t) \leq n$ is the number of sample paths observed to be in state h just before time t.
Nonparametric estimation of the transition intensities turns out to be most conveniently formulated in terms of the integrated transition intensities

$$A_{hj}(t) = \int_0^t \alpha_{hj}(s)ds$$

which are estimated by the *Nelson-Aalen estimator*

$$\hat{A}_{hj}(t) = \int_0^t \frac{1}{Y_h(s)} dN_{hj}(s)$$

$$= \sum_{T_{hjk} \le t} \frac{1}{Y_h(T_{hjk})}$$

where $0 < T_{hj1} < T_{hj2} < \cdots$ are the observed direct transitions $h \to j$.
As we shall see, there will in practice often be a need to combine the thus analysed transition intensities into a *synthesis* describing the *net effect* of the various transitions. The transition probabilities

$$P_{hj}(s,t) = P\{X(t) = j | X(s) = h\}$$

depend on the transition intensities α_{hj} through the Kolmogorov forward differential equations, whose solution may be represented as the *matrix product integral*

$$\mathbf{P}(s,t) = \prod_{(s,t]} (\mathbf{I} + d\mathbf{A}(u))$$

with $\mathbf{I}$ the identity matrix. Aalen and Johansen (1978) used this relation to postulate the estimator

$$\hat{\mathbf{P}}(s,t) = \prod_{(s,t]} (\mathbf{I} + d\hat{\mathbf{A}}(u))$$

which may be given a nonparametric maximum likelihood interpretation. The rather compact notation may not fully reveal that the estimator is really a simple finite product of elementary matrices.
As before, martingales and stochastic integrals are available to derive exact and asymptotic properties and to estimate covariance matrices.

3 Interaction between life history events

A nontrivial example of the application of nonhomogeneous Markov processes is the study of association of occurrence of two life history events A and B. It is a justifiedly strong convention in biostatistics that there is no way to infer on statistical grounds from association to direction (not to speak of causation). However, consider the simple four-state Markov process specified in Figure 1,

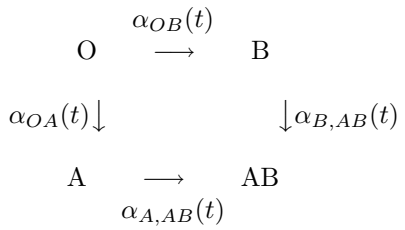


Figure 1. Interaction between life history events

where O means that no event has occurred; A means that the single event A has occurred and similarly for B; AB means that both A and B have occurred. If now $\alpha_{OB}(t) \equiv \alpha_{A,AB}(t)$ (occurrence of B has the same intensity before and after A) but $\alpha_{OA}(t) < \alpha_{B,AB}(t)$ for all t (occurrence of A happens faster after B than before B) then, following Schweder (1970), we say that A is *locally dependent* on B but B is not locally dependent on A and we have an *asymmetrical concept of dependence*. For further general discussion of this idea see Aalen (1987), Courgeau and Lelièvre (1992, Chapter 5), Blossfeld and Rohwer (1995, Section 6.3).
The idea was implemented by Aalen et al. (1980) (cf. Borgan (1980)) to analyse a cross-sectional sample of prevalent cases of women with the chronic skin disease *pustulosis palmo-plantaris*. Using retrospective information on time of (natural or induced) menopause, if yet occurred, and of time of onset of disease, these authors derived a conservative test of the hypothesis of identical incidence of the disease before and after menopause. The test rejected this hypothesis, indicating that menopause may increase the risk of occurrence of this disease. The dual test of comparing occurrence of menopause before and after disease showed (as expected) no indication of a difference.

4 Bone marrow transplantation

Patients receiving bone marrow transplant (BMT) as a treatment for leukaemia frequently develop graft versus host disease (GvHD) wherein the transplanted (grafted) immune cells attack the host tissues. On biological grounds, one expects the development of GvHD to increase the risk of patients dying in remission and to possibly decrease the risk of leukaemic relapse (because the graft's immune cells kill both normal and leukaemic host cells).,
A simple Markov process model is indicated in Figure 2; note that a separate state AC is included in order to allow interaction between acute and chronic GvHD. Keiding et al. (2001) studied two models:

a. Nonparametric Markov model. Here each transition rate is a freely varying nonnegative function of time t since transplantation, leading to the nonhomogeneous Markov process model described above.

b. Cox (semiparametric) Markov model. Because the fit of separate intensities to all eleven permitted transitions require very extensive data, some parsimony is usually required. Following Klein et al. (1993), we assume that the intensities of occurrence of chronic GvHD before and after acute GvHD are proportional, that the four transition intensities into relapse, with obvious notation $\lambda_{0R}(t)$, $\lambda_{AR}(t)$, $\lambda_{CR}(t)$ and $\lambda_{AC,R}(t)$ are proportional, and finally that $\lambda_{0D}(t)$, $\lambda_{AD}(t)$, $\lambda_{CD}(t)$ and $\lambda_{AC,D}(t)$ are proportional.

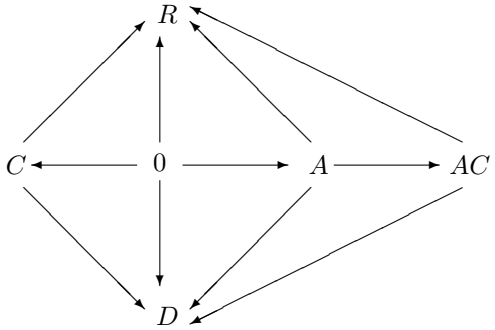


Figure 2. Simple multistate model for events after bone marrow transplantation. All patients start in state 0 at transplantation and will ultimately either relapse (R) or die in remission (D). As intermediate states we include acute (A) or chronic (C) or both acute and chronic (AC) graft-versus-host disease.

Keiding et al. (2001) calculated summary probabilities. A classical summary measure is the probability $P_{0R}(0,t)$ of having relapsed by time t, which may be estimated directly by the relevant Aalen-Johansen estimator. Keiding et al. compared this to the calculated probability of relapse in a hypothetical world where death in remission is no longer possible (all transition rates $\lambda_{0D}(t), \lambda_{AD}(t), \lambda_{CD}(t), \lambda_{AC,D}(t)$ into D are set to zero) but all other transition rates are unchanged. Obviously, this hypothetical relapse probability is somewhat larger than the one estimated when death is possible. This distinction is conceptually identical to the classical choice between (correctly) using the so-called *cumulative incidence* as an estimate of $P_{0R}(0,t)$, as opposed to using 1–the Kaplan-Meier estimate, which will estimate the cumulative relapse probability in a world where nobody dies in remission but all other intensities stay unchanged.
In a similar way Keiding et al. studied what would happen if acute and/or chronic graft-versus-host disease were prevented, that is, setting one or more of $\lambda_{0A}(t), \lambda_{0C}(t)$ and $\lambda_{0AC}(t)$ to 0, with all other transition intensities staying unchanged. In that case the prediction is that the relapse probability $P_{0R}(t)$ would increase towards the upper confidence limit of the estimate from this world, with the effects of removing acute or chronic GvHD being rather similar, and a slightly larger effect of removing both.

On the other hand the death in remission probability would decrease considerably, primarily as a result of the removal of acute GvHD.
Keiding et al. studied several other "hypothetical worlds", using the non-parametric Markov model as well as the Cox semi-parametric Markov model.

5 Repeated events: Example about recurrence of affective disorders

Our final example illustrates models for *repeated events* where individual heterogeneity not directly captured by the Markov models becomes an issue. Kessing et al. (1999) studied 20,350 Danish psychiatric patients who were discharged from their first admission to a psychiatric hospital with a diagnosis of major affective disorder (manic-depression, depression, manic/circular episode). Register follow-up allowed identification of time from each discharge to the next admission (operationally separated by a lag of eight weeks) and the main purpose was to assess whether the admission intensity increased with number of previous admissions.
Similar to the generalization of the Cox regression model to "modulated renewal processes", cf. Cox (1972) and Oakes & Cui (1994), a Cox regression model for the intensity of recurrence for person i at time t after episode k was postulated as

$$\lambda_{ik}(t) = \lambda_0(t)\, \exp(\beta_1 age + \beta_{period} + \beta_k)$$

where *age* is age at first admission and *period* indicates one of five calendar time periods.
Under an assumption of independent censoring, a fit to this model, for bipolar younger men, gave the following estimated relative risks $\exp(\beta_k)$

episode k	1	2	3	4	$\geq$5
exp (β_k)	1	1.18	1.46	1.72	2.35
95% C.I.	-	1.03-1.34	1.27-1.69	1.48-2.02	2.07-2.67

showing a highly significant effect of increasing recurrence intensity with increasing number of previous admissions. For each patient the series of admissions and discharges all had to be contained in the observation window 1971-1993. This of course meant that patients with shorter intervals were more likely to reach a large number of admissions, in other words, the interpretation of the above analysis was critically dependent on the independence between episodes for each patient implied by the above Cox model.
To check this independence a frailty model was specified:

$$\lambda_{ik}(t \mid Z_i) = \lambda(t) Z_i\, \exp(\beta_1 age + \beta_{period} + \beta_k)$$

with Z_i gamma distributed with unit expectation. Note that Z_i depends on patient i but not on episode k. Under this model the relative risks $\exp(\beta_k)$ were estimated as

episode k	1	2	3	4	≥ 5
exp (β_k)	1	0.99	1.10	1.16	1.30
95% C.I.	-	0.84-1.15	0.91-1.33	0.93-1.45	1.04-1.64

These were marginally significantly different (P=0.06) whereas the frailty variance, estimated as 0.45, was highly significantly positive.
The analysis shows that there is a strong intra-person correlation in length of episode between admissions, and that failure to account for this will lead to serious bias because of the sampling frame of the data.
More generally, because many patients have more than one readmission, there is a repetitive structure supporting the Z_i and thereby the separability of frailty distribution and individual hazard rates. In such situations the "individual hazard" $\lambda(t \mid Z_i)$ is rather better identified than for single-spell data.

References

Aalen, O.O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 177-190.

Aalen, O.O., Borgan, Ø., Keiding, N., and Thormann, J. (1980). Interaction between life history events: nonparametric analysis of prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics*, **7**, 161-171.

Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, **5**, 141-150.

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Andersen, P.K. and Keiding, N. (2001). Event history analysis in continuous time. *International Encyclopedia of the Social and Behavioral Sciences* (to appear).

Blossfeld, H. and Rohwer, G. (1995). *Techniques of Event History Modeling*. New Jersey: Lawrence Erlbaum.

Borgan, Ø. (1980). Applications of nonhomogeneous Markov chains to medical studies. In: Victor, N., Lehmacher, W. and van Eimeren, W. (eds.) *Explorative Datenanalyse, Frühjahrstagung München 1980, Proceedings* Medizinische Informatik und Statistik **26**, 102-115. Heidelberg: Springer.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis*, **5**, 315-327.

Courgeau, D. and Lelièvre, E. (1992). *Event History Analysis in Demography.* Oxford, Clarendon.

Cox, D.R. (1972). The statistical analysis of dependencies in point processes. In: Lewis, P.A.W., editor, *Stochastic Point Processes*, 55-66. New York: Wiley.

Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis*, **5**, 239-264.

Keiding, N. (1998). Selection effects and nonproportional hazards in survival models and models for repeated events. *Proc XIXth International Biometric Conference, Cape Town, Invited papers*, 241-250.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine*, **18**, 2353-2363.

Keiding, N., Klein, J.P., and Horowitz, M.M. (2001). Multistate models and outcome prediction in bone marrow transplantation. *Statistics in Medicine* (to appear).

Kessing, L.V., Olsen, E.W., and Andersen, P.K. (1999). Recurrence in affective disorder: Analyses with frailty models. *American Journal of Epidemiology*, **149**, 404-411.

Klein, J.P., Keiding, N., and Copelan, E.A. (1993). Plotting summary predictions in multistate survival models: Probabilities of relapse and death in remission for bone marrow transplantation patients. *Statistics in Medicine*, **12**, 2315-2332.

Oakes, D. and Cui, L. (1994). On semiparametric inference for modulated renewal processes. *Biometrika*, **81**, 83-90.

Schweder, T. (1970). Composable Markov processes. *Journal of Applied Probability*, **7**, 400-410.