CEM
Center for Empirical Macroeconomics

Working Paper No. 79

Smoothing, Random Effects and Generalized Linear Mixed Models in Survival Analysis

by

Göran Kauermann, Ronghiu Xu and Florin Vaida

University of Bielefeld
Department of Economics
Center for Empirical Macroeconomics
P.O. Box 100 131
33501 Bielefeld, Germany
(http://www.wiwi.uni-bielefeld.de/∼cem)

Smoothing, Random Effects and Generalized Linear Mixed Models in Survival Analysis

Göran Kauermann*, Ronghui Xu† and Florin Vaida‡

10th November 2004

Abstract

In the analysis of censored failure time data, the Cox proportional hazards model assumes that the regression coefficients are invariant in time and across individuals. In this paper we propose an extension of the Cox model for clustered survival data, which allows general, random effects (frailties), and time-varying regression coefficients, which are smooth functions of time. We fit the model using a mixed-model representation of penalized spline smoothing, similar to Cai, Hyndman & Wand (2002). This offers a unified framework for estimation of the baseline hazard, smooth effects and random effects. The estimator is computed using a hybrid Monte Carlo EM algorithm. Variance estimators are calculated in the same hybrid way adapting the classical Louis formula (Louis, 1982) to our model. A marginal Akaike information criterion is developed to assist the selection of an appropriate model. The model is then applied to unemployment data taken from the German Socio-Economic Panel. The duration of unemployment is modelled in a flexible way to depend on a set of covariate and on individual random effects.

*University Bielefeld, Faculty of Economics, Universitätsstraße 25, 33502 Bielefeld

†University of California, San Diego, Departement of Mathematics, 9500 Gilman Drive, San Diego 92093, USA

‡University of California, San Diego, School of Medicine, 9500 Gilman Drive, San Diego 92093, USA

1 Introduction

The Cox (1972) proportional hazards model has been for the last 30 years the most popular regression model for the analysis of censored survival data. In recent years a growing body of work has expanded this model in several ways. One of the most obvious assumptions of the model, as its name suggests, is the proportional hazards (*PH*) assumption. Much work has been done to test or validate the *PH* assumption, see Sasieni (1999) or Therneau & Grambsch (2000) for an overview. In the presence of departure from the *PH* assumption different estimation methods have been proposed which include both smoothed estimates and tree-based parsimonious partition of the time axis; a review of some of the existing literature can be found in Xu & Adak (2002). Another implicit assumption of the partial likelihood inference under the standard Cox model is the independence of observed survival times. This is not fulfilled for clustered data or if there are multiple duration times observed for the individuals. Such dependence can be accommodated by proportional hazards mixed models (*PHMM*) which include random effects, either as frailties which modify the baseline hazard, see e.g. Vaupel, Manton & Stallard (1979), Nielsen, Gill, Andersen & Sørensen (1992), Klein (1992), or as general regression terms, see Ripatti & Palmgren (2000), Vaida & Xu (2000). A general overview is found in Hougaard (2000). For approaches in the econometric area we refer to van den Berg (2001). Frailty models however, still assume proportional hazards among subjects from the same or different clusters, even though Stare & O'Quigley (2004) demonstrate some dualities between frailty models and smooth effects for baseline hazard estimation. In this paper we propose a general model which includes both random effects and smooth time-varying regression effects. The estimation uses the mixed-model representation of penalized splines (P-splines, see Eilers & Marx, 1996, Ruppert, Wand

& Carroll, 2003), thereby combining the estimation of both mixed effects and spline effects in the same framework. Kauermann (2004b) uses P-spline smoothing in a mixed-model representation to estimate time dynamic effects in the fixed-effects only Cox-type regression model, while Cai, Hyndman & Wand (2002) used P-spline estimation of the hazard function with no covariates. The latter idea is extended to proportional hazards models in Cai & Betensky (2003). Recently, Therneau, Grambsch & Pankratz (2003) propose penalized estimation for Frailty models. The combination of frailty and smoothing has been applied by Duchateau & Janssen (2004) using gamma distributed frailties and penalized partial likelihood estimation. We go a similar route but take more advantage of Generalized Linear Mixed Model theory. In our model, the spline effects, treated as random effects, and the random cluster, or frailty effects are two different type of random components, and the marginal likelihood resulting after integrating them out does not have an analytic form. To overcome this problem Laplace approximation seems a natural candidate. However, since Laplace approximation leads to underestimation of the random effects' variance, a more elaborated fitting routine is necessary. We use a Monte Carlo EM algorithm along the lines of Ripatti, Larsen & Palmgren (2002) or Vaida & Xu (2000), see also Booth & Hobert (1999), Dempster, Laird & Rubin (1977) and Klein (1992). Due to the large number of random components, however, sampling of the random effect is not straightforward, and employing the rejection sampler would lead to severe low acceptance rates. Based on asymptotic justifications, we treat the spline and random effects separately, and we apply the numerically intensive sampling step only to the cluster or frailty effects, but estimate spline effects using a Laplace approximation. Such a hybrid Monte Carlo EM algorithm has been suggested and studied by Lai & Shih (2003). The hybrid idea has a number of

advantages. Estimation becomes both numerically more feasible and faster. Moreover, it is possible to extend Louis' (1982) formula to derive confidence bands for the smooth curve.

In multivariable models a particular focus is on model selection. In our context this means, for instance, to determine the set of covariates with time dynamic effects and time constant effects, respectively. This can be carried out in principle by extending the results on testing in Linear Mixed Models provided in Crainiceanu & Ruppert (2004) or Crainiceanu, Ruppert, Claeskens & Wand (2004). However, here we pursue a different route by minimizing an Akaike information criterion based on the marginal likelihood. This likelihood is computed following a similar hybrid algorithm as for estimation. Using this marginal likelihood, an Akaike criterion, called marginal AIC, is easily derived. The marginal likelihood value is also used to control the EM convergence and to observe Monte Carlo variability. In particular, we use the simple idea of increasing the Monte Carlo sample size in the EM steps if the marginal likelihood does not increase and terminate the iteration once the marginal likelihood increments are negligible.

The model is applied to data from the German Socio-Economic Panel. The focus is on the analysis of duration of unemployment for a subsample of individuals reported at least once unemployed between the years 1990 to 2000. Individuals may experience more than one spell of unemployment, which leads to multiple, correlated duration times. Moreover, we consider the possibly time-dynamic effect of covariates gender, nationality and age.

The structure of the paper is as follows: Section 2 presents the model and draws the link to Poisson data and Generalized Linear Mixed Models. In Section 3 we discuss the hybrid version of the Monte Carlo EM including variance calculation. In Section 4 we suggest the use of the Akaike information criterion for model selection. The data example and simulations are included in Section 5.

2 Smoothing mixed-effects survival model

2.1 Model specification

Let t_{ij} be the observed j-th survival time in cluster i. We assume independence between the clusters, but survival times within a cluster are considered as realizations of dependent variables. In our data example, indexes i, j refer to the j-th spell of unemployment of individual i. We denote with δ_{ij} the event indicator taking value 1 if the observed survival time is the true survival time and value 0 if the true survival time exceeds the observed time. The hazard function for each event is modelled as

$$h_{ij}(t) = \exp\{\beta_0(t) + x_{ij}\beta_x(t) + w_{ij}a_i\} \tag{1}$$

where x is a set of covariates and $\beta_0(t)$ and $\beta_x(t)$ are smooth but otherwise unspecified functions, coefficients a_i are cluster-specific random effects with design matrix w_{ij} built from covariates x_{ij} (this is not a technical requirement, but a reasonable assumption in practice). Model (1) extends the classical proportional hazard model in two ways. First, covariate effects $\beta_x(t)$ are allowed to vary with survival time t. Secondly, a random or frailty effect is included to accommodate the dependence of survival times within a cluster. In the paper we will also include semiparametric models where some effects are time dynamic while others are fixed, that is $\beta_x(t) \equiv \beta_x$. To keep the notation simple we will not explicitly emphasize these models in the

estimation step, as they result easily as a special case.

Estimation of the smooth components in (1) is pursued by penalized spline fitting (P-spline). The approach traces back to Eilers & Marx (1996) with a general introduction and recent developments provided in Ruppert, Wand & Carroll (2003). We first demonstrate the recipe of P-spline fitting for the baseline $\beta_0(t)$. We approximate the smooth unknown function $\beta_0(t)$ by some high dimensional basis $B(t)$, say, so that $\beta_0(t) \approx \beta_{00} + B(t)b_0$, where β_{00} gives the intercept listed explicitly here. A convenient choice for $B(t)$ are truncated polynomials, e.g. $B(t) = ((t-t_0)_+, (t-t_1)_+, \ldots, (t-t_p)_+)$ where $(t)_+ = t\, I(t \geq 0)$ with $I(.)$ as the indicator function and $t_0 < t_1 < \ldots < t_p$ are fixed knots covering the range of the observed failure times. One might for instance choose $t_0 = 0$ and t_l as the observed l-th ordered failure time. The dimension p is thereby chosen in a lavish way so that the bias $\beta_0(t) - (\beta_{00} + B(t)b_0)$ has a negligible size compared to the estimation variability. Likewise, we replace $\beta_x(t)$ componentwise by $\beta_l(t) \approx \beta_{0l} + B_l(t)b_l$ for $l = 1, 2, \ldots, q$, where q is the number of covariates.

Denoting with $z_{ij} = (1, x_{ij})$ and $\beta_0 = (\beta_{00}, \beta_{0x})^T$, the log-likelihood conditional on random effects $a = (a_1^T, \ldots, a_n^T)$ takes the form

$$l_c(\beta, b, a) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[z_{ij}\beta_0 + z_{ij}\mathbf{B}(t_{ij})b + w_{ij}\, a_i - \int_0^{t_{ij}} \exp\left\{ z_{ij}\beta_0 + z_{ij}\mathbf{B}(t)b + w_{ij}\, a_i \right\} dt \right], \quad (2)$$

where $\mathbf{B}(t) = I_{q+1} \otimes B(t)$ with $\otimes$ denoting the tensor product and $b = (b_0^T, \ldots, b_q^T)^T$. The second component in (2) is the integrated hazard. In the case of proportional hazards this integrated hazard factorizes into the cummulative baseline hazard times the covariate effects. The resulting cummulative baseline hazard can then be replaced by a step function with jumps at the observed time points, which leads

to the well known partial likelihood function suggested by Cox (1972). We go a similar route here by approximating the hazard in the integral by using Newton's method with knots defined on the observed failure time points, in the following denoted by $0 = \tau_0 < \tau_1 < \ldots < \tau_K$. Then $\int^{t_{ij}} h_{ij}(u)\, du$ is replaced by $\sum_{k:\tau_k \leq t_{ij}} h_{ij}(\tau_k)\, (\tau_k - \tau_{k-1})$. Let now $\mathcal{R}_k$ denote the risk set at time point τ_k, $k = 1, \ldots, K$, that is $\mathcal{R}_k = \{(i,j) : t_{ij} \geq \tau_k\}$. Accordingly, let $\mathcal{F}_k$ denote the failures at time point τ_k, that is $\mathcal{F}_k = \{(i,j) : t_{ij} = \tau_k \text{ and } \delta_{ij} = 1\}$. The likelihood (2) based on the integral approximation is then rewritten as follows. For notational simplicity we define $H_{ijk} = (z_{ij} \otimes B(\tau_k), W_{ij})$, where W_{ij} is the overall design matrix constructed from w_{ij} such that $W_{ij}\, a = w_{ij}\, a_i$. Finally, we set $d = (b^T, a^T)$. This allows to write the approximate likelihood as

$$l_c(\beta, d) = \sum_{k=1}^{K} \sum_{(i,j) \in \mathcal{F}_k} (z_{ij}\, \beta + H_{ijk}\, d) - \sum_{k=1}^{K} \sum_{(i,j) \in \mathcal{F}_k} \exp\{z_{ij}\, \beta + H_{ijk}\, d + o_k\} \qquad (3)$$

with $o_k = \log(\tau_k - \tau_{k-1})$. Note that $l(\beta, d)$ is equivalent to the likelihood of pseudo-Poisson data Y_{ijk} for $(i,j) \in \mathcal{R}_k, k = 1, \ldots, K$ where $Y_{ijk} = 1$ if $(i,j) \in \mathcal{F}_k$ and 0 otherwise. These data follow the model

$$Y_{ijk|d} \sim Poisson\ (\lambda = \exp\{z_{ij}\, \beta + H_{ijk}\, d + o_k\}). \qquad (4)$$

This connection will be exploited subsequently by fitting model (4) to Y_{ijk} (see also Kauermann, 2004b).

2.2 Parameter estimation

The high dimensionality of b forbids simple maximization of the likelihood. Instead, coefficients are penalized in ridge regression style. This can be accommodated by imposing an "a priori" distribution on b:

$$b \sim N\left(0, \text{diag}(\sigma_l\, D^-)\right)$$

where $\sigma_b^2 = (\sigma_0^2, \ldots, \sigma_q^2)$ is the vector of a priori variances and D^- as generalized inverse of some penalty matrix D chosen in accordance to the basis used. For truncated polynomials identity matrices have been proven to work well (see Ruppert, Wand & Carroll, 2003). We also assume that

$$a_i \sim N(0, \Sigma_a),\ i = 1, \ldots, n.$$

The joint normality for $d = (b^T, a^T)$ now provides with (3) a Generalized Linear Mixed Model (GLMM) with the marginal likelihood resulting by integrating out a and b, that is

$$l_m(\beta, \sigma_b^2, \sigma_a^2) = \int\int \exp\{l(\beta, b, a)\}\, \phi(b, \mathrm{diag}(\sigma_l^2\, D^-))\, \phi(a, \mathrm{diag}(\Sigma_a))\, \mathrm{d}a\, \mathrm{d}b \qquad (5)$$

where $\phi(.)$ denotes the normal density. The maximization of (5) can be achieved by a Laplace approximation (see Severini, 2000, Therneau, Grambsch & Pankratz, 2003)

$$\begin{aligned} l_m(\beta, \sigma_b^2, \sigma_a^2) \;\approx\; \exp\Big[& l(\beta, \hat{b}, \hat{a}) - \frac{1}{2}\, \hat{a}^T \Sigma_a^{-1} \hat{a} - \frac{1}{2}\, \hat{b}^T \mathrm{diag}(\sigma_l^{-2} D) \hat{b} \\ & - \frac{1}{2} \log|\Sigma_a| - \frac{1}{2} \log\left|\mathrm{diag}(\sigma_l^2 D^-)\right| \\ & - \frac{1}{2} \log\left\{ \left| \frac{\partial^2\, l(\beta, \hat{b}, \hat{a})}{\partial(a, b)\, \partial(a, b)^T} - \mathrm{diag}\,(\Sigma_a^{-1},\, \sigma_l^{-2}\, D) \right| \right\} \Big] \end{aligned} \qquad (6)$$

where r denotes the dimension of a, and $\hat{a}$ and $\hat{b}$ are the maximizers of

$$l(\beta, b, a) - \frac{1}{2} a^T \Sigma_a^{-1} a - \frac{1}{2} b^T \mathrm{diag}(\sigma_l^{-2} D)\, b. \qquad (7)$$

In particular, (7) plays the role of a penalized likelihood with Σ_a^{-1} and $\sigma_l^{-2}\, D$ as penalty parameters, $l = 0, \ldots, q$. If some of the smooth components are time constant, that is e.g. $\beta_l(t) = \beta_{0l}$, this is easily accommodated by penalizing b_l to zero which leads to reduced likelihood with components related to σ_l^2 excluded.

3 Monte Carlo EM

3.1 Standard Monte Carlo EM

The Laplace approximation can introduce bias (see for instance Breslow & Lin, 1995 or Shun & McCullagh, 1995). A convenient way to circumvent this is to use an EM algorithm, at the price of additional numerical effort. Booth & Hobert (1999) discuss a Monte Carlo version of the EM algorithm which is picked up here (see also Vaida, Meng & Xu, 2004). Let $f(Y|d)$ denote the Poisson density of Y given in (4). With $\phi(d, H)$ we denote the "a priori" distribution of d with mean zero and variance matrix H, where H is block-diagonal with $\text{diag}(\sigma_l^2 D^-)$ and $\text{diag}(\Sigma_a)$ on its diagonal. Fixing β_0, β_x, σ_b^2 and Σ_a at their current estimate, denoted with superscript (s), and considering d unobserved leads to the Expectation step

$$\text{Q}\left(\beta, \sigma_b^2, \Sigma_a | \, \beta^{(s)}, \sigma_b^{2(s)} \Sigma_a^{(s)}\right) = \text{E}\{\log[f(Y|\, d)\; \phi(d, H)] \mid Y, \beta^{(s)}, \sigma_b^{2(s)}, \Sigma_a^{(s)}\}$$

The expectation is now computed by Monte Carlo simulation from the distribution $g\left(d|Y\right) \propto f(Y|d)\; \phi(d, H)$, using rejection sampling. Booth & Hobert (1999) suggest to use as proposal density $h(d) = \phi(d, H)$. However, due to the large dimension of d, in our setting this leads to a very small acceptance rate. A more suitable proposal density $h(d)$, as in Ripatti, Larsen & Palmgren (2002), is the normal density with mean $\hat{d}$ the penalized fit resulting from (6) and variance

$$V_{d|u} = \left\{ \sum_{k=1}^{K} \sum_{(i,j) \in \mathcal{R}_k} H_{ijk}^T \; H_{ijk} \; \exp(\eta_{ijk}) + \text{diag}(\sigma_l^{-2} D, \Sigma_a) \right\}^{-1} \cdot \varrho$$

where η_{ijk} is the linear predictor and $\varrho > 1$ is an inflation factor to be defined below. Let now $\log(c)$ be defined as $\log(c) = max\{\log g(d|Y) - \log h(d)\}$ where the logarithm is used for numerical reasons. Note that the maximum is achieved at $\hat{d}$

and the second order derivative equals

$$\frac{\partial^2 \log g(d|Y)}{\partial d\, \partial d^T} - \frac{\partial^2 \log h(d)}{\partial d\, \partial d^T} = -\left\{ \sum_{k=1}^{K} \sum_{(i,j)\in\mathcal{R}_k} H_{ijk}^T\, H_{ijk}\, \exp(\eta_{ijk}) + \mathrm{diag}(\sigma_l^{-2} D, \Sigma_a) \right\} \left(1 - \frac{1}{\varrho}\right),$$

which is negative definite as long as $\varrho > 1$. A proposed d^* from $h(d^*)$ is now accepted with probability $\pi(d^*) = \exp\{\log g(d^*|Y) - \log(h(d^*)) - \log(c)\}$ where again the logarithm is used for numerical reasons.

3.2 Hybrid Monte Carlo EM

The acceptance probability in the above Monte Carlo EM is low, in particular for a large dimension of the design matrix z_{ij}. A remedy is suggested by the fact the that random coefficients a and b follow two different asymptotic scenarios. For cluster effect a we assume that with growing sample size the number of clusters increases while the number of observations within a cluster is limited. In particular, if the number of replicates within a cluster is small, Laplace approximation of the integral (5) is not advisable (see e.g. Shun & McCullagh, 1995). In contrast, for spline coefficients b we assume that the dimension of the basis is fixed in advance and kept fixed for growing sample size. This in turn implies that information on each coefficient of b is growing with the sample size while the dimension of the integral is kept fixed. In this scenario the Laplace approximation provides satisfactory results and it approximates the integral with order $O(n^{-1})$ (see e.g. Severini, 2000). Along the lines of Lai & Shih (2003) we therefore suggest to use a hybrid routine, with a Laplace approximation for b and Monte Carlo EM for a in (5). Let

$$l_b(\beta, b, \sigma_b^2, \sigma_a^2) = \log \int \exp\{l_c(\beta, b, a)\phi(a, \mathrm{diag}(\Sigma_a))\}\, \mathrm{d}a \tag{8}$$

denote the likelihood after integration with respect to a. The marginal likelihood in

(5) is then approximated by

$$\begin{aligned} l_m(\beta, \sigma_b^2, \sigma_a^2) &\approx l_b(\beta, \hat{b}, \sigma_b^2, \sigma_a^2) + \log \phi(\hat{b}, \mathrm{diag}(\sigma_l^2 D^-)) \\ &\quad -\frac{1}{2} \log \left| -\frac{\partial^2 l_b(\beta, \hat{b}, \sigma_b^2, \sigma_a^2)}{\partial b \, \partial b^T} + \mathrm{diag}(\frac{D}{\sigma_l^2}) \right| \end{aligned} \tag{9}$$

where $\hat{b}$ is the maximizer of the penalized likelihood $l_b(\beta, b, \sigma_b^2, \sigma_a^2) + \log \phi \left(b, \mathrm{diag}(\sigma_l^2 D^-) \right)$. Apparently, $\hat{b}$ is not available analytically since integration over a is involved. However, we can use a Monte Carlo EM, like above, but this time for a only. In principle this means we treat a as random and β and b as parameters where the second is estimated in a penalized manner. The remainder of this section explains this approach in more technical details. Let therefore $b^{(s)}$ be a current maximizer (estimate) of b. Then the hybrid E-step results through the Q Function

$$\begin{aligned} & \mathrm{Q}_b(\beta, b, \sigma_b^2, \Sigma_a^2 |\, \beta^{(s)}, b^{(s)}, \sigma_b^{2^{(s)}}, \Sigma_a^{(s)}) \\ = \; & \mathrm{E}_a \left\{ l_p(\beta, b, \sigma_b^2 | a) + \log \phi(a, \mathrm{diag}(\Sigma_a)) | Y, \beta^{(s)}, b^{(s)}, \sigma_b^{2^{(s)}}, \Sigma_a^{(s)} \right\} \end{aligned} \tag{10}$$

where $l_p(\beta, b, \sigma_b^2 | a)$ is a penalized log likelihood defined through

$$l_c(\beta, b, a) + \log \phi \left(b, \mathrm{diag}(\sigma_l^2 D^-) \right) .$$

The expectation in (10) is carried out with respect to distribution $f(a|b^{(s)}, Y)$ (ignoring the dependence on parameters for notational simplicity). In analogy to above we evaluate it by rejection sampling, using the proposal density

$$a^* \sim h(a|b^{(s)}) = N(\hat{a}, \rho V_{a|b,y}),$$

where $\hat{a}$ is the maximizer of $f(Y|b^{(s)}, a)\phi(a, \mathrm{diag}(\Sigma_a))$ and

$$V_{a|b,y} = \left\{ \sum_{k=1}^{K} \sum_{(i,j) \in \mathcal{R}_k} W_{ij} W_{ij}^T \exp(\eta_{igk}) + \mathrm{diag}(\Sigma_a^{-1}) \right\}^{-1}$$

with W_{ij} as defined in Section 2.1. Defining with $\log c_a = \max\left\{\log g(a|b^{(s)}, Y) - \log h(a|b^{(s)})\right\}$ where $g(a|b^{(s)}, Y) = f(Y|b^{(s)}, a)\phi(b^{(s)}, \text{diag}(\sigma_l^2\, D^-))\ \phi(a, \text{diag}(\Sigma_a))$ provides the acceptance probability

$$\pi_a(a^*) = \exp\left\{\log g(a^*|b^{(s)}, Y) - \log h(a^*|b^{(s)}) - \log c_a\right\}\ .$$

Based on a sample $a^{*^1}, \ldots, a^{*^M}$ the subsequent Maximization Step results by updating estimates $\hat{\beta}^{(s)}$ und $\hat{b}^{(s)}$ simultaneously using a penalized likelihood. Moreover, an estimate for σ_l^2 is available from Laplace approximation (see for instance Kauermann, 2004a)

$$\hat{\sigma}_l^{2^{(s)}} = \frac{\hat{b}^{(s)^T} D\, \hat{b}^{(s)}}{df_l}$$

where df_l is a measure for the degrees of freedom of the fit $\hat{\beta}_l(t) = \hat{\beta}_{0l} + B(t)\hat{b}_l$. This can be calculated, at least approximately, from the smoothing matrix. Let therefore $M_{ijkl} = (z_{ijl}, z_{ijl}B(\tau_k))$, where z_{ijl} denotes the element in z_{ij} corresponding to $\beta_l(t)$. The degree of freedom for the l-th component is then defined through

$$df_l = \text{tr}\ \left[\left\{\sum_{k=1}^{K}\sum_{(i,j)\in\mathcal{R}_k} M_{ijkl}M_{ijkl}^T \exp(\eta_{ijk}) + \sigma_l^{-2}\text{diag}(0, D)\right\}^{-1}\left\{\sum_{k=1}^{K}\sum_{(i,j)\in\mathcal{R}_k} M_{ijkl}M_{ijkl}^T \exp(\eta_{ijk})\right\}\right]$$

A stopping criteria for the iterations of the algorithm is derived later based on the marginal likelihood $l_m(.)$.

3.3 Variance Calculation

Variance calculation in combination with the EM algorithm is primarily focused on variance derivation for the fixed parameters (see Louis, 1982). Here, however, we are interested in the variance of the smooth fit $\hat{\beta}_l(t)$, $l = 0, \ldots, q$. We can interpret the estimate $\hat{\beta}_l(t) = \hat{\beta}_{0l} + B(t)\hat{b}_l$ as predictor, based on the predicted values for b_l. This suggests to focus on $E(\{\hat{\beta}_l(t) - \beta_l(t)\}^2)$ as prediction error, where $\beta_l(t) = \beta_{0l} + B(t)b_l$ with b_l as true but unknown random effect. The expectation in this case is carried

out with respect to both, Y and b. To proceed, we set $\theta = (\beta^T, b^T)$ and treat for simplicity variance components σ_b^2 and Σ_a as known. We show in the Appendix that

$$\mathrm{E}\left\{(\hat{\theta} - \theta)^2\right\} \approx \mathrm{I}_{bp}(\theta|\hat{b})^{-1} \tag{11}$$

with $\mathrm{I}_{bp}(\theta)$ as penalized Fisher matrix defined through

$$\mathrm{I}_{bp}(\theta) = -\mathrm{E}_{Y|b}\left\{\frac{\partial^2 l_b(\theta)}{\partial\theta\partial\theta^T}\right\} + \mathrm{diag}(0, \sigma_l^{-2}D).$$

See also Ruppert, Wand & Carroll (2003) for a similar derivation in the case of normal response.

The next step is to calculate the penalized Fisher matrix based on the Monte Carlo EM for a. Using the same arguments as in Louis (1982) and additionally taking the penalization into account we get the observed Fisher matrix

$$-\mathrm{I}_b(\theta) = \mathrm{E}\left\{\frac{\partial^2 l_b(\theta, a)}{\partial\theta\partial\theta^T}\middle| Y, b\right\} - \mathrm{E}\left\{s_b(\theta)s_b(\theta)^T|Y, b\right\} \tag{12}$$

where $s_b(\theta) = \partial l_c(\theta, a)/\partial\theta$. The components in (12) can now be estimated from the final Monte Carlo simulations of the EM algorithm. In particular, for estimation of the second part of (12) an unpenalized (and hence asymptotically unbiased) estimate for θ is required. This is readily available by fitting a Poisson model with design matrix corresponding to coefficients β and b and then taking empirical expectation with respect to simulated values a^*.

4 Monte Carlo Error and Model Selection

4.1 Computing the marginal likelihood

A simple procedure to compute the marginal likelihood is to take advantage of the Monte Carlo simulations in each EM step. To do so we make use of the reciprocal

importance sampling estimator as introduced by Gelfand & Day (1994). We generalize their idea here by extending it to the hybrid estimation structure from above. With $a^{*^1}, \ldots, a^{*^M}$ as Monte Carlo sample we calculate

$$\hat{A}(b) = \frac{1}{M} \sum_{m=1}^{M} \exp\left\{V(a^{*^m}|b)\right\} \tag{13}$$

with $V(a^*|b) = \log\{h(a^*|b)\} - l_c(\beta, b, a^*) - \log\{\phi(a^*, \mathrm{diag}(\Sigma_a))\}$. In the line of Gelfand & Day (1994) and reflecting that a^{*^m} are drawn from $f(a|b, Y)$ we find $\hat{A}(b)$ as estimate for

$$A(b) = \int \frac{h(a|b)}{f(y|a, b)\,\phi(a)}\, f(a|b, Y)\,\mathrm{d}a = \frac{1}{f(y|b)} = \exp\{-l_b(\beta, b)\}\ .$$

(For numerical reasons it is advisable to add a constant to $V(a^*|b)$ which is then, after taking logarithm, subtracted from $\log(A)$. For simplicity of notation we omit this numerical trick here.) The marginal likelihood is now equal to

$$l_m(\beta) = \log \int \exp\{-\log(A(b))\}\ \phi(b, \mathrm{diag}(\sigma_l^2 D^-))\mathrm{d}b\,. \tag{14}$$

Note that $\hat{b}$ maximizes the integrand of (14) and Laplace approximation leads to

$$l_m(\beta) \approx -\log\left\{A(\hat{b})\right\} + \log\left\{\phi(\hat{b}, \mathrm{diag}(\sigma_l^2 D^-))\right\} - \frac{1}{2}\log\left|\frac{\partial^2 g(\hat{b})}{\partial b \partial b^T}\right| \tag{15}$$

where $g(b) = \log\{A(b)\} - \log\{\phi(b, \mathrm{diag}(\sigma_l^2 D^-))\}$. Plugging in estimate (13) provides an estimate for $l_m(\beta)$ denoted by $\hat{l}_m(\beta)$. Since $l_b(\beta, b) = -\log\{A(b)\}$ the latter component in (15) mirrors the determinant of the Fisher information which is easily estimated from the Monte Carlo sample. In particular we have

$$\frac{\partial^2 g(b)}{\partial b \partial b^T} = -\frac{\frac{\partial A(b)}{\partial b}\frac{\partial A(b)}{\partial b^T}}{A(b)^2} + \frac{\frac{\partial^2 A(b)}{\partial b \partial b^T}}{A(b)} + \mathrm{diag}(\sigma_l^{-2} D)$$

where the separate components can be estimated by

$$\begin{aligned}
\frac{\partial \hat{A}(b)}{\partial b} &= -\frac{1}{M}\sum_{m=1}^{M} \exp(V(a^{*^m}, b))\frac{\partial l_c(\beta, b, a^{*^m})}{\partial b} \\
\frac{\partial^2 \hat{A}(b)}{\partial b \partial b^T} &= -\frac{1}{M}\sum_{m=1}^{M} \exp(V(a^{*^m}, b))\left[\frac{\partial l_c(\beta, b, a^{*^m})}{\partial b}\,\frac{\partial l_c(\beta, b, a^{*^m})}{\partial b^T} - \frac{\partial^2 l_c(\beta, b, a^{*^m})}{\partial b \partial b^T}\right]\ .
\end{aligned}$$

The estimated marginal likelihood can now be used to supervise the convergence of the EM algorithm and for model selection.

4.2 Supervising the convergence of the EM algorithm

The EM algorithm is known to increase the likelihood in each iteration step (Dempster, Laird & Rubin, 1977) and to converge (Wu, 1983; Vaida, 2005). This property is however not guaranteed for the Monte Carlo version, due to Monte Carlo variability, since the marginal likelihood depends on the random sample $a^{*^1}, \ldots, a^{*^M}$. For supervision of the EM convergence it is necessary to assess the variability resulting from the Monte Carlo sample. As in Booth & Hobert (1999) we suggest to start with a small Monte Carlo sample size in the first steps and to increase it successively. Our proposal is thereby to increase the Monte Carlo sample size M_s in the s-th step by a factor $(1+\alpha)$ with $\alpha > 0$ if the marginal likelihood estimate does not increase. We therefore calculate $\hat{l}_m(\hat{\beta}^{(s)})$ and make use of the following model. Denoting with $l_m(\hat{\beta}^{(s)})$ the marginal likelihood without Monte Carlo error, that is for Monte Carlo size $M_s \to \infty$ we can consider $\hat{l}_m(\hat{\beta}^{(s)})$ as noisy version of $l_m(\hat{\beta}^{(s)})$ with noise variance of order $O(M_s^{-1})$. Assuming that $l_m(\hat{\beta}^{(s)})$ is a smooth function in s we have $\hat{l}_m(\hat{\beta}^{(s)})$ as noisy observations for $l_m(\hat{\beta}^{(s)})$. This suggests to plot $\hat{l}_m(\hat{\beta}^{(s)})$ against iteration step s and to use a simple scatterplot smoother and standard software to get the marginal likelihood $l_m(\hat{\beta}^{(s)})$ from the noisy Monte Carlo estimates $\hat{l}_m(\hat{\beta}^{(s)})$. To do so, a weighted smoothing has to be carried out with weights given by M_s. If function $l_m(\beta^{(s)})$ flattens out, this indicates that the EM algorithm has converged.

4.3 The Marginal Akaike Criterion

For model selection between classes of models (1) we use the Akaike information criterion (AIC) based on the marginal likelihood. The AIC is justified from a model

prediction perspective, it is designed to choose the model with the lowest predictive log-likelihood (Akaike, 1973, Burnham & Anderson, 2002), and is related to cross-validation and Mallows' C_p (Hastie & Tibshirani, 1990, p.160). For complex smoothing models, model selection is still in its infancy (Ruppert, Wand & Carroll, 2003, pp.184, 220). In the context of smoothing, AIC has been used mostly for selecting the smoothing parameter (Hurvitch, Simonoff & Tsai, 1998, Simonoff & Tsai, 1999, Ruppert, Wand & Carroll, 2003), and occasionally for choosing between different models (Hastie & Tibshirani, 1990). For mixed models some new results for the AIC are given in Vaida & Blanchard (2005). An alternative approach to the classical AIC is to compute the AIC from the marginal likelihood (5); we will call this the marginal AIC (mAIC). This approach is consistent with the P-spline idea of using the marginal likelihood for estimation of all parameters, including the smoothing parameters. There is no additional difficulty in using mAIC when random effects a are present. Wager, Vaida and Kauermann (2004, unpublished manuscript) showed that in most situations mAIC performs as well for model selection, or better, than the classical AIC, for continuous response. More specifically, mAIC is given by

$$mAIC = -2l_m(\hat{\beta}, \hat{\sigma}_b^2, \hat{\Sigma}_a^2) + 2r \tag{16}$$

where r is the number of parameters to be estimated in the model. For each smooth component there are two parameters, β_{0l} and the related variance σ_l^2, while if the effect is time constant, that is $\beta_l(t) \equiv \beta_{0l}$, the component has one parameter only.

5 Application

5.1 Simulation

To explore the performance of the estimation procedure we run a small simulation study. We simulate data from 100 clusters. The cluster size takes values 1, 2, 4, 6

and 8 and we simulate 20 independent clusters for each size. Data are generated on a discrete grid with survival times taking possible values $t = 0, 1, \ldots, 60$. For each time interval t to $t + 1$ we simulate censoring with probability 0.97. At time point t we simulate data using the model

$$\mathcal{M}_1 : \; h_{ij}(t) = \exp\{\beta_0(t) + x_{1j}\beta_1(t) + a_i\} \,, \tag{17}$$

where cluster effect a_i is drawn from a normal distribution with standard deviation $\sigma_a = 0.5$, and covariates x_{1j} drawn as binary random variates with $P(x_{1j} = 1) = 0.3$. For the functional forms of the effects we use $\beta_0(t) = -4 + 1.5t/60$ and $\beta_1(t) = \sin(1.5t/60\pi)$. We fit the model using a 15-dimensional basis $B(t)$ built from truncated lines for each of the effects. Based on 100 simulations Figure 1 shows for time point t the median (dashed line), upper and lower 25 % (dotted lines) and 10 % quantiles (solid line), respectively. There is a slight bias in the peak of $\beta_1(t)$, which is however moderate. In Figure 2 we show the fitted random effects for clusters with cluster size 1 and 8, respectively, plotted against their true simulated value. These are obtained from the mean of a^* of the Monte Carlo sample in the last iteration step. There is an effect of shrinkage visible, which is reduced if there are more data available in a cluster. The shrinkage is inevitable and has no dominance on the estimation of σ_a, which is estimated with simulation mean 0.53 and simulation standard deviation 0.05.

The next step is to explore the marginal Akaike information criterion. We therefore fit the proportional hazard model $\mathcal{M}_0 : h_{ij}(t) = \exp\{\beta_0(t)+x_{1j}\beta_1+a_i\}$ as competitor. Figure 3 (left boxplot) shows the corresponding difference $\text{mAIC}(\mathcal{M}_0) - \text{mAIC}(\mathcal{M}_1)$ with $\text{mAIC}(\mathcal{M})$ as defined in (16) for the corresponding model. Clearly, model $\mathcal{M}_1$ is preferred in most simulations. In the next step we swap the role of $\mathcal{M}_0$ and $\mathcal{M}_1$,

that is we simulate data from model $\mathcal{M}_0$ and fit again models $\mathcal{M}_0$ and $\mathcal{M}_1$. For $\beta_1(t)$ we set $\beta_1(t) = 1$. The corresponding simulated values of $\text{mAIC}(\mathcal{M}_0) - \text{mAIC}(\mathcal{M}_1)$ are shown in the right hand side boxplot of Figure 3. The preference for model $\mathcal{M}_0$ is now visible.

5.2 Unemployment data

We apply our model to unemployment data from the German Socio-Economic Panel. We analyzed a subsample of $n = 400$ West German individuals who had been registered unemployed at least for one month during 1990 and 2000. (The data set is available for scientific users from the German Institute for Economic Research, see `www.diw.de`.) About 44% of the individuals experienced more than one spell of unemployment, with a maximum of 12 spells for one individual. Table 1 gives the distribution of the number of spells in our sample. The duration of unemployment is defined as censored observation, with an event only if the individual returns to full-time employment. Any other termination like further professional development, (early) retirement, short term or part time contracts are taken as censored observations.

Figure 4 shows the Kaplan-Meier curves for the effects of covariates nationality (1 if German, 0 otherwise), gender (*sex* = 1 for male, 0 otherwise), and age. The age is categorized with two indicator variables. With *under25* we classify individuals which are aged 25 or younger at the beginning of their unemployment (1 if age $\leq$ 25, 0 otherwise). Unemployed individuals aged 50 or older are indicated with covariate *over50* (1 if age $\geq$ 50). In Table 2 we list the distribution of the covariates, with their corresponding values for the whole panel for comparison.

We fitted the following models to the data:

$$\begin{aligned}
\mathcal{M}_1 &= h(t) = \exp\{\beta_0(t) + nat\beta_n(t) + sex\beta_s(t) + under25\beta_u(t) + over50\beta_o(t) + a_i\} \\
\mathcal{M}_2 &= h(t) = \exp\{\beta_0(t) + nat\beta_n + sex\beta_s + under25\beta_u + over50\beta_o(t) + a_i\} \\
\mathcal{M}_3 &= h(t) = \exp\{\beta_0 + nat\beta_n + sex\beta_s + under25\beta_u + over50\beta_o(t) + a_i\} \\
\mathcal{M}_4 &= h(t) = \exp\{\beta_0(t) + nat\beta_n + sex\beta_s + under25\beta_u + over50\beta_o + a_i\} \\
\mathcal{M}_5 &= h(t) = \exp\{\beta_0(t) + nat\beta_n(t) + sex\beta_s(t) + under25\beta_u(t) + over50\beta_o(t)\}
\end{aligned}$$

In model $\mathcal{M}_1$ all effects vary with time. In models $\mathcal{M}_2$ and $\mathcal{M}_3$ only baseline and the effect for *over50*, respectively, vary with time. Finally we exchange the roles of $\beta_o(t)$ and $\beta_0(t)$ by fitting model $\mathcal{M}_4$. Note that model $\mathcal{M}_4$ is a classical proportional hazard model with constant effects but smooth baseline. In the first three models we leave the individual random effect to incorporate frailty effects and to take the clustering of observations into account. Finally, model $\mathcal{M}_5$ lets all effect to vary with time but sets the random individual effect to zero. That is, spells are considered as independent events. In Figure 6 we show the marginal likelihood $mAIC(\mathcal{M}, t) = -2\hat{l}_m(\hat{\beta}^{(s)}) + 2r_{\mathcal{M}}$ for the different models and steps for the EM algorithm. The Monte Carlo sample size M_s starts at value 40 and is increased by 25 % if the likelihood steps $\hat{l}_m(\hat{\beta}^{(s)})$ decrease. For each model we show on the AIC level the estimated marginal likelihood $\hat{l}_m(\hat{\beta}^{(s)})$ including a smooth scatterplot fit of function $\hat{l}_m(\hat{\beta}^{(s)})$ calculated by a simple spline smoother (with 5 degrees of freedom and observations weighted by M_s). For model $\mathcal{M}_1$ we were faced numerical problems since variance estimates σ_l^2 for all covariates except of *over50* converged to zero. The large $mAIC$ for model $\mathcal{M}_5$ clearly showed it inadequate. Based on Figure 6 we selected model $\mathcal{M}_3$. The corresponding estimates are listed in

Table 3. It appears that the baseline is constant (for individuals of age < 50), so that chances of returning to professional life are constant over time, but depend on an unobserved individual heterogeneity. German individuals and males have increased chances to terminate unemployment. Individuals aged 25 or younger are less likely to return to full employment, even though the effect is found not significant. Finally, unemployed aged 50 and older are less likely to return to professional life, with an effect strengthening over time. Figure 5 (right panel) shows the fitted random effects against the mean duration of unemployment for each individual. As expected, individuals with shorter unemployment duration had estimated random effects a_i generally larger than those for individuals with longer duration time.

6 Conclusions

In this paper we extended the classical proportional hazard model simultaneously in two directions, to allow for non proportional hazards and to include frailty or random effects. Even though these two are conceptually different, we demonstrated that a Generalized Linear Mixed Model provides a unifying framework, and it allows for fitting both smooth and random effects. To make the estimation feasible, we proposed a hybrid EM algorithm by combining Laplace approximation and Monte Carlo sampling. The marginal likelihood was used for supervision of the EM convergence, and also for model selection, via the marginal AIC. Such a model expands the scope of statistical modelling, in response to the increasing availability of larger and more complex studies and datasets.

A Technical Details

Derivation of Prediction Error

First we assume that b is a fixed but unknown component, that is, we condition on b. In this case $\theta = (\beta^T, b^T)$ is treated as parameter and like in standard likelihood theory $E_{Y|b} = \{\partial l_b(\beta, b)/\partial\theta|b\} = 0$ with $l_b(.)$ given in (8). A penalized estimate for b and θ, respectively, is defined through $0 = \partial l_{bp}(\hat{\theta}/\partial\theta)$ with l_{bp} as penalized marginalized likelihood $l_{bp}(\theta) = l_b(\theta) - \log\phi\{b, \text{diag}(\sigma_l^2 D^-)\}$. Apparently, the penalization induces a conventional smoothing bias which results to

$$\text{bias}(\sigma_b^2) := \text{E}_{Y|b}\left\{\frac{\partial l_{bp}(\theta)|b}{\partial\theta}\right\} = \begin{pmatrix} 0 \\ -\text{diag}(\sigma_l^{-2} D)b \end{pmatrix}. \tag{18}$$

Note that expectation is taken over a and Y but we condition on b. Expanding the likelihood about θ now provides with classical likelihood arguments by conditioning on b:

$$\hat{\theta} - \theta = \left\{\text{I}_{bp}(\theta|b)^{-1}\left[\frac{\partial l_b(\beta, b)}{\partial\theta} + \text{bias }(\sigma_b^2)\right]\right\}\left\{1 + O_p(n^{-\frac{1}{2}})\right\}$$

where $\text{I}_{bp}(\theta) = \text{I}_b(\theta) + \text{diag}(0, \sigma_l^{-2} D)$ is the penalized Fisher matrix with $\text{I}_b = -\text{E}_{Y|b}\left\{\partial^2 l_b(\theta)/\partial\theta\partial\theta^T\right\}$. Our objective is to derive the prediction error $E\left\{(\hat{\theta} - \theta)^2\right\}$ by taking expectation over both Y and b. Note first, that taking expectation over b cancels out the bias since $\text{E}(b) = 0$, that is $E_b\left(\text{bias}(\sigma_b^2)\right) = 0$. This allows to write the prediction error as

$$\text{E}\left\{(\hat{\theta} - \theta)\right\}^2 = \text{E}_b\left\{\text{Var}_{Y|b}(\hat{\theta} - \theta)\right\} + \text{Var}_b\left\{[\text{E}_{Y|b}(\hat{\theta} - \theta)]^2\right\} \tag{19}$$

where the subscripts indicate the variable we take expectation about (and random effects a are always integrated out). The conditional terms result by standard likelihood theory, since b is treated as parameter (estimated in a penalized manner).

Hence, the conditional variance results to $\mathrm{Var}_{Y|b}(\hat{\theta}-\theta) = \mathrm{I}_{bp}(\theta|b)^{-1}\mathrm{I}_b(\theta|b)\mathrm{I}_{bp}(\theta|b)^{-1}$. Integrating out b is now done by Laplace approximation so that the first component in (19) results to

$$\mathrm{E}_b\left\{\mathrm{Var}_{Y|b}(\hat{\theta}-\theta)\right\} \approx \mathrm{I}_{bp}^{-1}(\tilde{\theta}|\hat{b})\mathrm{I}_b(\tilde{\theta}|\hat{b})\mathrm{I}_{bp}^{-1}(\tilde{\theta}|\hat{b}) \tag{20}$$

where $\tilde{\theta} = (\beta, \hat{b})$. In the same line we now look at the second component in (19). Taking the variance with respect to b over the conditional bias (18) allows by using Laplace approximation to write

$$\mathrm{Var}_b\left\{\mathrm{E}_{Y|b}(\hat{\theta}-\theta)^2\right\} = \mathrm{I}_{bp}^{-1}(\tilde{\theta}|\hat{b})\mathrm{diag}(0, \sigma_l^{-2}D)\mathrm{I}_{bp}^{-1}(\tilde{\theta}|\hat{b})\ . \tag{21}$$

Adding (20) and (21) and reflecting the definition of $\mathrm{I}_{bp}(.)$ provides (11) as simple formula for the prediction error.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics (1992)*, Volume 1, pp. 610–624. Springer-Verlag.

van den Berg, G. J. (2001). Duration Models: Specification, Identification, and Multiple Durations. In J. J. Heckman & E. Leamer (Eds.), *Handbook of Econometrics*, Volume 1. North-Holland.

Booth, J. and Hobert, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **B 62**, 265–285.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information - Theoretic Approach* (second ed.). Springer.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570–579.

Cai, T., Hyndman, R. J., and Wand, M. P. (2002). Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics 11*(4), 784–798.

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

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **66**, 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2004). Exact likelihood ratio tests for penalized splines. Technical Report, Department of Statistical Science, Cornell University.

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

Duchateau, L. and Janssen, P. (2004). Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows. *Biometrics 60*(3), 600–614.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11*(2), 89–121.

Gelfand, A. and Day, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer.

Hurvitch, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271–293.

Kauermann, G. (2004a). A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planning and Inference (in press)*.

Kauermann, G. (2004b). Penalised spline fitting in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis*, (to appear).

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

Lai, T. L. and Shih, M.-C. (2003). A hybrid estimator in nonlinear and generalised linear mixed effects models. *Biometrika* **90**, 859–879.

Louis, T. (1982). Finding observed information using the em algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 98–130.

Nielsen, G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19**, 25–44.

Ripatti, S., Larsen, K., and Palmgren, J. (2002). Maximum likelihood inference for multivariate frailty models using an automated Monte Carlo EM algorithm. *Lifetime Data Analysis* **8**, 349–360.

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

Ruppert, R., Wand, M., and Carroll, R. (2003). *Semiparametric Modelling*. Cambridge University Press.

Sasieni, P. (1999). Cox regression model. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics*, Volume 1, pp. 1006–1020. New York: Wiley.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B* **57**, 749–760.

Simonoff, J. and Tsai, C. (1999). Semiparametric and additive model selection using an improved akaike criterion. *Journal of Computational and Graphical Statistics* **8**, 22–40.

Stare, J. and O'Quigley, J. (2004). Fit and frailties in proportional hazards regression. *Biometrical Journal* **46**, 157–164.

Therneau, T. and Grambsch, P. (2000). *Modelling Survival Data: Extending the Cox Model*. New York, USA: Springer Verlag.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics* **12**, 156–175.

Vaida, F. (2005). Parameter convergence for the EM and MM algorithms. *Statistica Sinica (to appear)*.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed

effects models. *Biometrika (to appear)*.

Vaida, F., Meng, X.-L., and Xu, R. (2004). Mixed effects models and the EM algorithm. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An essential journey with Donald Rubin's statistical family*. Wiley.

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

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.

Wu, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*(1), 95–103.

Xu, R. and Adak, S. (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics* **58**, 305–315.

Number of spells	1	2	3	4	5	7	12
Proportions	61.25	21.0	7.0	5.5	2.25	0.25	0.25

Table 1: Proportion of number of spells for $n = 400$ individuals.

	nation (Germans)	gender (males)	under 25	over 50
sample	81.75	49.25	23.75	20.25
panel	80.65	51.02	24.25	21.13

Table 2: Proportion of individuals for sample and complete panel.

effect	estimate	std dev	p-value
baseline	- 3.55	0.22	< 0.01
nation	0.84	0.15	< 0.01
gender	0.42	0.19	0.02
under 25	- 0.31	0.17	0.07

Table 3: Parametric estimates for unemployment data.

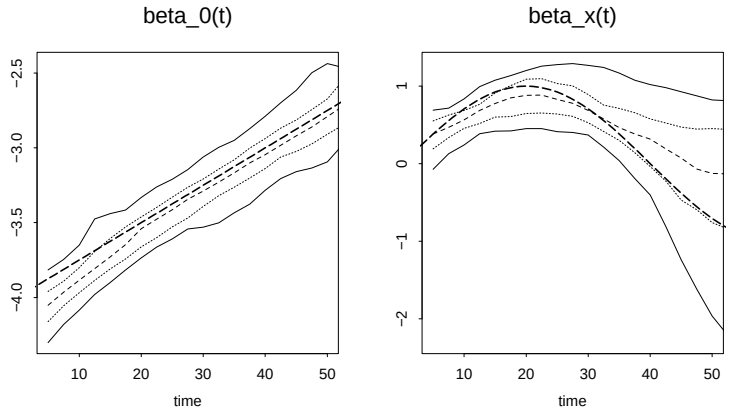

Figure 1: Simulated estimated effects with median (dashed line), lower and upper 25 % (dotted line) and 10 % quantile (solid line). The true curve is shown as bold dashed line.

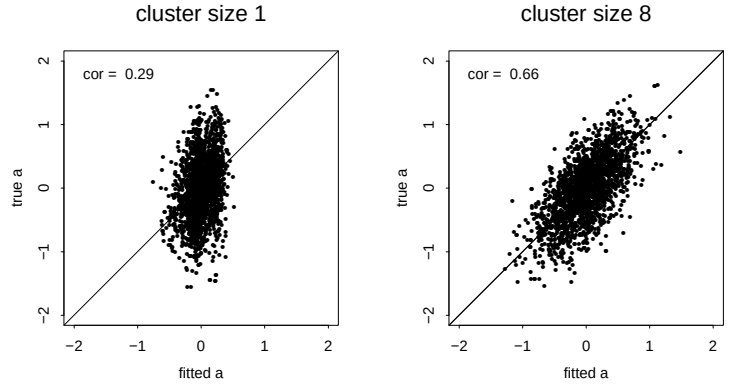

Figure 2: Fitted posterior mean of random effects a_i plotted against their true value for two different cluster sizes. The correlation is given in each plot.

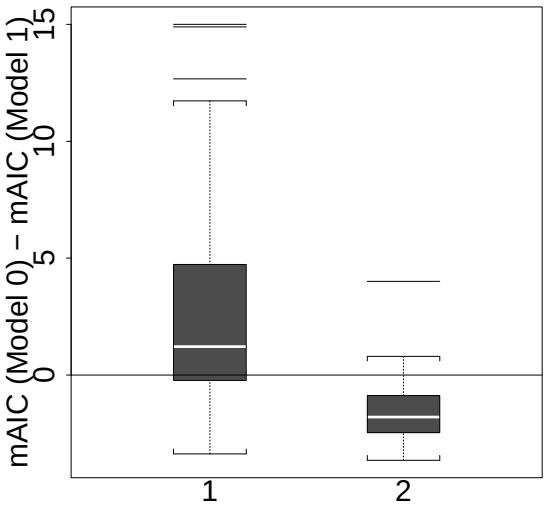

Figure 3: Simulated marginal Akaike criterion if simulations are drawn from model 1 (left boxplot) or model 0 (right boxplot).

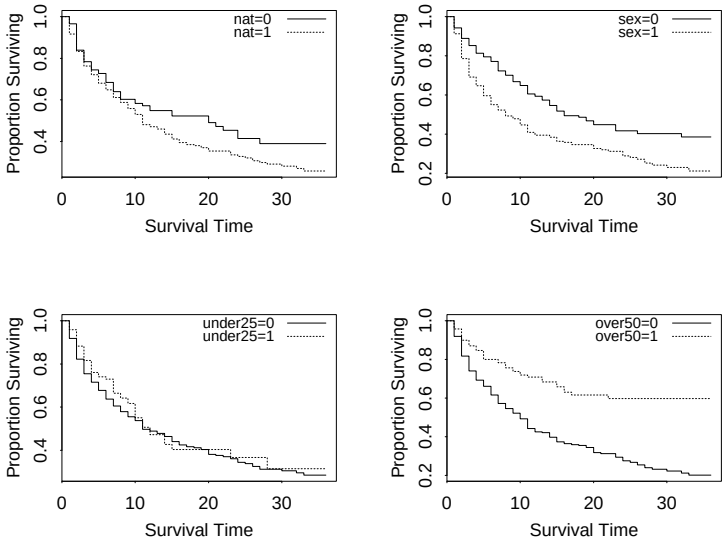

Figure 4: Kaplan-Meier estimates for unemployment data for different covariates.

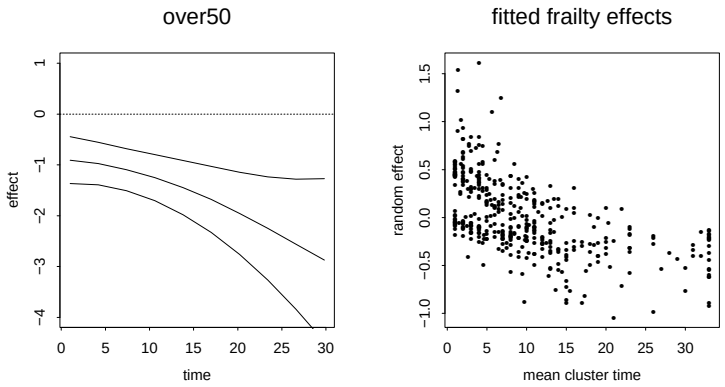

Figure 5: Fitted effect for *over50* (left plot) and fitted random effects (right plot), plotted against the mean survival time for each individual.

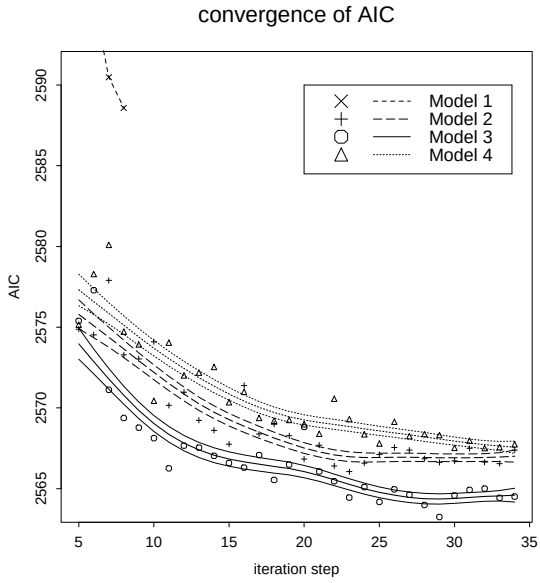

Figure 6: Convergence of EM algorithm shown as AIC value for 4 different models. For model $\mathcal{M}_2$ to $\mathcal{M}_4$ the fitted marginal likelihood values $\hat{l}(\hat{\beta}^{(s)})$ are smoothed as suggested in 4.3 with resulting confidence bands.