LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Andreas Groll & Trevor Hastie & Gerhard Tutz

Regularization in Cox Frailty Models

Technical Report Number 191, 2016
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Regularization in Cox Frailty Models

Andreas Groll[a,*] & Trevor Hastie[b] & Gerhard Tutz[c]

[a] Ludwig-Maximilians-University Munich, Theresienstraße 39, 80333 Munich,

[b] University of Stanford, Department of Statistics, 390 Serra Mall, Sequoia Hall, California 94305,

[c] Ludwig-Maximilians-University Munich, Akademiestraße 1, 80799 Munich

February 25, 2016

Abstract

In all sorts of regression problems it has become more and more important to deal with high dimensional data with lots of potentially influential covariates. A possible solution is to apply estimation methods that aim at the detection of the relevant effect structure by using penalization methods. In this work, the effect structure in the Cox frailty model, which is the most widely used model that accounts for heterogeneity in survival data, is investigated. Since in survival models one has to account for possible variation of the effect strength over time the selection of the relevant features has to distinguish between several cases, covariates can have time-varying effects, can have time-constant effects or be irrelevant. A penalization approach is proposed that is able to distinguish between these types of effects to obtain a sparse representation that includes the relevant effects in a proper form. It is shown in simulations that the method works well. The method is applied to model the time until pregnancy, illustrating that the complexity of the influence structure can be strongly reduced by using the proposed penalty approach.

Keywords: Variable selection; LASSO; Cox frailty model; Time-varying coefficients; Penalization.

1 Introduction

Proportional hazards (PH) models, and in particular the semi-parametric Cox model (Cox, 1972) play a major role in the modeling of continuous event times.

[*] Corresponding author. Tel.: +49 89 2180 4697; fax:+49 89 2180 4452.
E-mail addresses: groll@math.lmu.de, hastie@stanford.edu,gerhard.tutz@stat.uni-muenchen.de

The Cox model assumes the semi-parametric hazard

$$\lambda(t|\boldsymbol{x_i}) = \lambda_0(t)\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}),$$

where $\lambda(t|\boldsymbol{x_i})$ is the hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x_i}^T = (x_{i1},\dots,x_{ip})$. $\lambda_0(t)$ is the shared baseline hazard, and $\boldsymbol{\beta}$ the fixed effects vector. Note that for continuous time the hazard rate $\lambda(t|\boldsymbol{x}_i)$ is defined as

$$\lambda(t|\boldsymbol{x}_i) = \lim_{\Delta t \to 0} P(t \leq T < t + \Delta t | T \geq t, \boldsymbol{x}_i)/\Delta t,$$

representing the instantaneous risk of a transition at time t. Inference is usually based on maximization of the corresponding partial likelihood. This approach allows estimation of $\boldsymbol{\beta}$ while ignoring $\lambda_0(t)$ and performs well in classical problems with more observations than predictors. As a solution to the $p > n$ problem, Tibshirani (1997) proposed the use of the so-called least absolute shrinkage and selection operator (LASSO) penalty in the Cox model. Since then, several extensions have been proposed. In order to fit the penalized model, Gui and Li (2005) provided an algorithm using Newton-Raphson approximations and the adjusted LARS solution. Park and Hastie (2007) applied the elastic net penalty to the Cox model and proposed an efficient solution algorithm, which exploits the near piecewise linearity of the paths of coefficients to approximate the solution with different constraints. They numerically maximize the likelihood for each constraint via a Newton iteration. Also Goeman (2010) addressed this problem and developed an alternative algorithm based on a combination of gradient ascent optimization with the Newton-Raphson algorithm. Another fast algorithm to fit the Cox model with elastic net penalties was presented by Simon et al. (2011), employing cyclical coordinate descent.

Frailty models aim at modeling the heterogeneity in the population. They can be used to account for the influence of covariates that have not been observed. They are especially useful if observations come in clusters, for example, if one models survival of family members or has repeated events for the same individual as in unemployment studies. The extreme case occurs if each individual forms its own cluster. For a careful investigation of identifiability issues see Van den Berg (2001). Parameter estimation in frailty models is more challenging than in the Cox model since the corresponding profile likelihood does not have a closed form solution. In the Cox PH frailty model also known as mixed PH model the hazard rate of the j-th subject belonging to cluster i, conditionally on the covariates $\boldsymbol{x}_{ij}$ and the shared frailty b_i, is given by

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}, b_i) = b_i\lambda_0(t)\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}), \quad i = 1,\dots,n, j = 1,\dots,N_i$$

where the frailties b_i are frequently assumed to follow a gamma distribution because of its mathematical convenience. The R package `frailtypack` (Rondeau et al., 2012) allows to fit such a Cox frailty model, covering four different types of frailty models (shared, nested, joint and additive frailties). In the R package `survival` (Therneau, 2013) a simple random effects term can be specified, following a gamma, Gaussian or t-distribution. A different fitting approach based on hierarchical likelihoods allowing for log-normal and gamma frailty distributions is implemented in the R package `frailtyHL`, see Do Ha et al. (2012). A first approach to variable selection for gamma frailty models was proposed by Fan and Li (2002). They used an iterative, Newton-Raphson based procedure to find the penalized maximum likelihood estimator and considered three types of penalties, namely the LASSO, the hard thresholding and the smoothly clipped absolute deviation (SCAD) penalty. However, no software implementation is available yet. The penalized gamma frailty model methodology of Fan and Li (2002) was extended to other frailty distributions, in particular to inverse Gaussian distributed frailty by Androulakis et al. (2012). They imposed the penalty term on a generalized form of the full likelihood function designed for clustered data, which allows the direct use of different distributions for the frailty term and which includes the Cox model and the gamma frailty model as special cases. For the gamma frailty case they modified the likelihood presented by Fan and Li (2002). However, again, no corresponding software package is available yet.

While some multiplicative frailty distributions, such as, for example, the gamma and the inverse Gaussian, have already been extensively studied (compare Androulakis et al., 2012) and closed form representations of the log-likelihoods are available, in some situations the log-normal distribution is more intuitive and allows for more flexible and complex predictor structures though the corresponding model is computationally more demanding. The conditional hazard function of cluster i and observation j with multiplicative frailties following a multivariate log-normal distribution has the general form

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t)\exp(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{u}_{ij}^T\boldsymbol{b}_i),$$

where $\boldsymbol{u_{ij}}^T = (u_{ij1}, \ldots, u_{ijq})$ is the covariate vector associated with random effects and the random effects $\boldsymbol{b}_i$ follow a multivariate Gaussian distribution, i.e. $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta}))$, with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{Q}(\boldsymbol{\theta})$, which is depending on a vector of unknown parameters $\boldsymbol{\theta}$. Ripatti and Palmgren (2000) show how a penalized quasi-likelihood (PQL) approach based on the Laplace approximation can be used for estimation. The method follows the fitting approach proposed by Breslow and Clayton (1993) for the generalized linear mixed model (GLMM). If, additionally, penalization techniques are incorporated into the procedure, it becomes especially important to provide effective estimation algorithms, as standard procedures for the choice of tuning parameters such as cross validation are usually very time-consuming.

2 Cox Frailty Model with Time-Varying Coefficients

While for Cox frailty models with the simple predictor structure $\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{u}_{ij}^T\boldsymbol{b}_i$ in the hazard function some solutions have already been given (Fan and Li, 2002, and Androulakis et al., 2012), often more complex structures of the linear predictor need to be taken into account. In particular, the effects of certain covariates may vary over time yielding time-varying effects $\gamma_k(t)$. A standard way to estimate the time-varying effects $\gamma_k(t)$ is to expand them in equally spaced B-splines yielding $\gamma_k(t) = \sum_{m=1}^{M} \alpha_{k,m}B_m(t;d)$, where $\alpha_{k,m}, m = 1,\ldots,M$, denote unknown spline coefficients that need to be estimated, and $B_m(t;d)$ denotes the m-th B-spline basis function of degree d. For a detailed description of B-splines, see for example Wood (2006) and Ruppert et al. (2003).

At this point, we address the specification of the baseline hazard $\lambda_0(t)$. In general, for the cumulative baseline hazard $\Lambda_0(\cdot)$ often the “least informative” nonparametric modeling is considered. More precisely, with $t_1^0 < \ldots. < t_N^0$ denoting the observed event times, the least informative nonparametric cumulative baseline hazard $\Lambda_0(t)$ has a possible jump h_j at every observed event time t_j^0, i.e. $\Lambda_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. The estimation procedure may be stabilized, if, similar to the time-varying effects, a semi-parametric baseline hazard is considered, which can be flexibly estimated within the B-spline framework. Hence, in the following we use the transformation $\gamma_0(t) := \log(\lambda_0(t))$ and expand $\lambda_0(t)$ in B-splines.

Let now $\boldsymbol{z_{ij}}^T = (1, z_{ij1},\ldots,z_{ijr})$ denote the covariate vector associated with both baseline hazard and time-varying effects and let $\boldsymbol{\alpha}_k^T = (\alpha_{k,1},\ldots,\alpha_{k,M}), k = 0,\ldots,r$, collect the spline coefficients corresponding to the baseline hazard or the k-th time-varying effect $\gamma_k(t)$, respectively. Further, let $\boldsymbol{B}^T(t) := (B_1(t;d),\ldots,B_M(t;d))$ represent the vector-valued evaluations of the M basis functions in time t. Then, with $\boldsymbol{v}_{ijk} := z_{ijk}\cdot\boldsymbol{B}(t)$, one can specify the hazard rate as

$$\lambda(t|\boldsymbol{x}_{ij},\boldsymbol{z}_{ij},\boldsymbol{u}_{ij},\boldsymbol{b}_i) = \exp\left(\eta_{ij}(t)\right),$$

with

$$\eta_{ij}(t) := \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sum_{k=0}^{r}\boldsymbol{v}_{ijk}^T\boldsymbol{\alpha}_k + \boldsymbol{u}_{ij}^T\boldsymbol{b}_i. \tag{1}$$

In general, the estimation of parameters in the predictor (1) can be based on Cox's well-known full log-likelihood, which is given by

$$l(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{b}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i} d_{ij}\eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds, \tag{2}$$

where n denotes the number of clusters, N_i the cluster sizes and the survival times t_{ij} are complete if $d_{ij} = 1$ and right censored if $d_{ij} = 0$.

As mentioned in the introduction, a possible strategy to maximize the full log-likelihood (2) is based on the PQL approach, which was originally suggested for GLMMs by Breslow and Clayton (1993). Typically, the covariance matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ of the random effects $\boldsymbol{b}_i$ depends on an unknown parameter vector $\boldsymbol{\theta}$. Hence, the joint likelihood-function can be specified by the parameter vector of the covariance structure $\boldsymbol{\theta}$ and parameter vector $\boldsymbol{\delta}^T := (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{b}^T)$. The corresponding *marginal* log-likelihood has the form

$$l^{mar}(\boldsymbol{\delta},\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\int L_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{b}_i)p(\boldsymbol{b}_i|\boldsymbol{\theta})d\boldsymbol{b}_i\right),$$

where $p(\boldsymbol{b}_i|\boldsymbol{\theta})$ denotes the density function of the random effects and the quantities $L_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{b}_i) := \prod_{j=1}^{N_i} \exp(\eta_{ij}(t_{ij}))^{d_{ij}} \exp\left(-\int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right)$ represent the likelihood contributions of the single clusters $i, i = 1, \ldots, n$. Approximation along the lines of Breslow and Clayton (1993) yields

$$\begin{aligned} l^{app}(\boldsymbol{\delta},\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log L_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{b}_i) - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{b} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left(d_{ij}\eta_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right) - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{b}, \qquad (3) \end{aligned}$$

with the penalty term $\boldsymbol{b}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{b}$ resulting from the approximation based on the Laplace method. The PQL approach usually works within the profile likelihood concept. It is distinguished between estimation of $\boldsymbol{\delta}$, given the plug-in estimate $\hat{\boldsymbol{\theta}}$ and resulting in profile likelihood $l^{app}(\boldsymbol{\delta},\hat{\boldsymbol{\theta}})$, and estimation of $\boldsymbol{\theta}$.

3 Penalization

In general, the roughness or "wiggliness" of the estimated smooth functions can be controlled by applying a difference penalty directly on the spline coefficients, see, for example, Eilers (1995) and Eilers and Marx (1996). However, with potentially varying coefficients in the predictor, model selection becomes more difficult. In particular, one has to determine which covariates should be included in the model, and, which of the covariates included have a constant or time-varying effect. So far, in the context of varying coefficient models in the literature only parts of these issues have been addressed. For example, Wang et al. (2008) and Wang and Xia (2009) used procedures that simultaneously select significant variables with (time-)varying effects and produce smooth estimates for the nonzero coefficient functions, while Meier et al. (2009) proposed a sparsity-smoothness penalization for high-dimensional generalized additive models. Also for functional regression models several approaches to variable selection have been proposed, see, for example, Matsui and Konishi (2011), Matsui (2014) and Gertheiss et al. (2013). On the other hand, for example, Leng (2009) presented a penalty approach that automatically distinguishes between varying and constant coefficients.

The objective here is to develop a penalization approach to obtain variable selection in Cox frailty models with time-varying coefficients such that single varying effects are either included, are included in the form of a constant effect or are totally excluded. The choice between this hierarchy of effect types can be achieved by using a specifically tailored penalty. We propose to use

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi \left(\zeta \sum_{k=1}^{r} \psi\, w_{\Delta,k} ||\boldsymbol{\Delta}_M \boldsymbol{\alpha}_k||_2 + (1-\zeta) \sum_{k=1}^{r} \phi\, w_k ||\boldsymbol{\alpha}_k||_2 \right), \qquad (4)$$

where $||\cdot||_2$ denotes the L_2-norm, $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\boldsymbol{\Delta}_M$ denotes the $((M-1)\times M)$-dimensional difference operator matrix of degree one, defined as

$$\boldsymbol{\Delta}_M = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{pmatrix}. \qquad (5)$$

The first term of the penalty controls the smoothness of the time-varying covariate effects, whereby for values of ξ and ζ large enough, all differences $\alpha_{k,l} - \alpha_{k,l-1}, l = 2, ..., M$, are removed from the model, resulting in constant covariate effects. As the B-splines of each variable with varying coefficients sum up to one, a constant effect is obtained if all spline coefficients are set equal. Hence, the first penalty term does not affect the spline's global level. The second term penalizes all spline coefficients belonging to a single time-varying effect in the way of a group LASSO and, hence, controls the selection of covariates. Both tuning parameters ξ and ζ should be chosen by an appropriate technique, such as, for example, K-fold cross validation (CV). The terms $\psi := \sqrt{M-1}$ and $\phi := \sqrt{M}$ represent weights that assign different amounts of penalization to different parameter groups, relative to the respective group size. In addition, we use the adaptive weights $w_{\Delta,k} := 1/||\boldsymbol{\Delta}_M \hat{\boldsymbol{\alpha}}_k^{(ML)}||_2$ and $w_k := 1/||\hat{\boldsymbol{\alpha}}_k^{(ML)}||_2$, where $\hat{\boldsymbol{\alpha}}^{(ML)}$ denotes the corresponding (slightly ridge-penalized) maximum likelihood estimator. Within the estimation procedure, i.e. the corresponding Newton-Raphson algorithm, local quadratic approximations of the penalty terms are used following Oelker and Tutz (2013). Note that the penalty from above may be easily extended by including a conventional LASSO penalty for time-constant fixed effects $\beta_k, k = 1, ..., p$.

Since the baseline hazard in the predictor (1) is assumed to be semi-parametric, another penalty term that controls the roughness of the baseline should be included. If the smooth log-baseline hazard $\gamma_0(t) = \log(\lambda_0(t))$ is twice differentiable, one can, for example, penalize its second order derivatives, similar to Yuan and Lin (2006). Alternatively, if $\gamma_0(t)$ is expanded in B-spline basis functions, i.e. $\gamma_0(t) = \sum_{m=1}^{M} \alpha_{0,m} B_m(t; d)$, one can simply penalize the second order differences of adjacent spline weights $\alpha_{0,l}, l = 1, \ldots, M$. Hence, in addition to $\xi \cdot J_\zeta(\boldsymbol{\alpha})$, the penalty term

$$\xi_0 \cdot J_0(\boldsymbol{\alpha_0}) = \xi_0 ||\boldsymbol{\Delta}_M^2 \boldsymbol{\alpha}_0||_2^2 \tag{6}$$

has to be included. Although this adds another tuning parameter ξ_0, it turns out that in general it is not worthwhile to also select ξ_0 on a grid of possible values. Similar findings with regard to penalization of the baseline hazard have been obtained for discrete frailty survival models, see Tutz and Groll (2014). While some care should be taken to select ξ and ζ, which determine the performance of the selection procedure, the estimation procedure is already stabilized in comparison to the usage of the "least informative" nonparametric cumulative baseline hazard $\Lambda_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$ for a moderate choice of ξ_0.

4 Estimation

Estimation is based on maximization of the penalized log-likelihood, which is obtained by expanding the approximate log-likelihood $l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta})$ from (3) to include the penalty terms $\xi_0 \cdot J_0(\boldsymbol{\alpha_0})$ and $\xi \cdot J_\zeta(\boldsymbol{\alpha})$, i.e.

$$l^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta}) = l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) - \xi_0 \cdot J_0(\boldsymbol{\alpha_0}) - \xi \cdot J_\zeta(\boldsymbol{\alpha}). \tag{7}$$

The estimation procedure is based on a conventional Newton-Raphson algorithm, while local quadratic approximations of the penalty terms are used, following Fan and Li (2001).

4.1 Fitting Algorithm

In the following, an algorithm is presented for the maximization of the penalized log-likelihood $l^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta})$ from equation (7). For notational convenience we omit the argument $\boldsymbol{\theta}$ in the following description of the algorithm and write $l^{pen}(\boldsymbol{\delta})$ instead of $l^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta})$. For fixed penalty parameters ξ_0, ξ and ζ, the following algorithm can be used to fit the model:

Algorithm `PenCoxFrail`

---

1. *Initialization*
   Choose starting values $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$ (see Section 4.2.3).

2. *Iteration*
   For $l = 1, 2, \ldots$ until convergence:

   (a) *Computation of parameters for given* $\hat{\boldsymbol{\theta}}^{(l-1)}$
   Based on the penalized score function $\boldsymbol{s}^{pen}(\boldsymbol{\delta}) = \partial l^{pen}/\partial\boldsymbol{\delta}$ and the penalized information matrix $\boldsymbol{F}^{pen}(\boldsymbol{\delta})$ (see Section 4.2.1) the general form of a single Newton-Raphson step is given by

$$\hat{\boldsymbol{\delta}}^{(l)} = \hat{\boldsymbol{\delta}}^{(l-1)} + (\boldsymbol{F}^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\boldsymbol{s}^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}).$$

   Because the fit is within an iterative procedure it is sufficient to use just one single step.

   (b) *Computation of variance-covariance components*
   Estimates $\hat{\boldsymbol{Q}}^{(l)}$ are obtained as approximate EM-type estimates (see Section 4.2.2), yielding the update $\hat{\boldsymbol{\theta}}^{(l)}$.

4.2 Computational Details of `PenCoxFrail`

In the following we give a more detailed description of the single steps of the `PenCoxFrail` algorithm. First, we describe the derivation of the score function and the information matrix. Then, an estimation technique for the variance-covariance components is given. Finally, we give details for the computation of starting values and the determination of optimal tuning parameters.

4.2.1 Score Function and Information Matrix

In this section we specify more precisely the single components which are derived in Step 2 (a) of the `PenCoxFrail` algorithm. Based on the B-spline design vector $\boldsymbol{B}(t)$, we define $\boldsymbol{\Phi}^T(t) := (z_{ij0} \cdot \boldsymbol{B}^T(t), z_{ij1} \cdot \boldsymbol{B}^T(t), \ldots, z_{ijr} \cdot \boldsymbol{B}^T(t))$. Then, the penalized score function $\boldsymbol{s}^{pen}(\boldsymbol{\delta}) = \partial l^{pen}(\boldsymbol{\delta})/\partial\boldsymbol{\delta}$, obtained by differentiating the log-likelihood from equation (7), has vector components

$$\begin{aligned}
\boldsymbol{s}^{pen}_{\boldsymbol{\beta}}(\boldsymbol{\delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{N_i}\boldsymbol{x}_{ij}\left(d_{ij} - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right),\\
\boldsymbol{s}^{pen}_{\boldsymbol{\alpha}}(\boldsymbol{\delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left(d_{ij}\boldsymbol{\Phi}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))\boldsymbol{\Phi}(s)ds\right) - \boldsymbol{A}_{\xi_0,\xi,\zeta}\,\boldsymbol{\alpha},\\
\boldsymbol{s}^{pen}_{i}(\boldsymbol{\delta}) &= \sum_{j=1}^{N_i}\boldsymbol{u}_{ij}\left(d_{ij} - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right) - \boldsymbol{Q}^{-1}(\boldsymbol{\theta})\boldsymbol{b}_i, \quad i = 1, \ldots, n.
\end{aligned}$$

Note here that the linear predictors $\eta_{ij}(t)$ depend on the parameter vector $\boldsymbol{\delta}$, compare equation (1). This is suppressed here for notational convenience. The vectors $\boldsymbol{s}_{\boldsymbol{\beta}}^{pen}$ and $\boldsymbol{s}_{\boldsymbol{\alpha}}^{pen}$ have dimension p and $(r+1)M$, respectively, while the vectors $\boldsymbol{s}_{i}^{pen}$ are of dimension q.

The penalty matrix $\boldsymbol{A}_{\xi_0,\xi,\zeta}$ is a block-diagonal matrix of the form $\boldsymbol{A}_{\xi_0,\xi,\zeta} = diag(\boldsymbol{A}_{\xi_0}, \boldsymbol{A}_{\xi,\zeta})$. The first matrix $\boldsymbol{A}_{\xi_0} = \xi_0 \boldsymbol{\Delta}_M^T \boldsymbol{\Delta}_M$ corresponds to the penalization of the squared differences between adjacent spline coefficients $\boldsymbol{\alpha}_0$ of the baseline hazard from equation (6), with $\boldsymbol{\Delta}_M$ denoting the $((M-1)\times M)$-dimensional difference operator matrix of degree one from equation (5). The second matrix $\boldsymbol{A}_{\xi,\zeta}$ results from a local quadratical approximation of the penalty in equation (4), following Oelker and Tutz (2013). It is a block-diagonal penalty matrix $\boldsymbol{A}_{\xi,\zeta} = diag(\boldsymbol{A}_{1,\xi,\zeta}, \ldots, \boldsymbol{A}_{r,\xi,\zeta})$, where for $k = 1, \ldots, r$ the single blocks have the form

$$\boldsymbol{A}_{k,\xi,\zeta} = \xi \left(\zeta \psi_k (\boldsymbol{\alpha}_k^T \tilde{\boldsymbol{\Delta}}_M^T \tilde{\boldsymbol{\Delta}}_M \boldsymbol{\alpha}_k + c)^{-1/2} \tilde{\boldsymbol{\Delta}}_M^T \tilde{\boldsymbol{\Delta}}_M + (1-\zeta) \phi_k (\boldsymbol{\alpha}_k^T \boldsymbol{\alpha}_k + c)^{-1/2} \right),$$

where c is a small positive number (in our experience $c \approx 10^{-5}$ works well) and the matrix $\tilde{\boldsymbol{\Delta}}_M$ is equal to $\boldsymbol{\Delta}_M$, except that its first row consist of zeros only.

The penalized information matrix $\boldsymbol{F}^{pen}(\boldsymbol{\delta})$, which is partitioned into

$$\boldsymbol{F}^{pen}(\boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{F}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{F}_{\boldsymbol{\beta}\boldsymbol{\alpha}} & \boldsymbol{F}_{\boldsymbol{\beta}1} & \boldsymbol{F}_{\boldsymbol{\beta}2} & \ldots & \boldsymbol{F}_{\boldsymbol{\beta}n} \\ \boldsymbol{F}_{\boldsymbol{\alpha}\boldsymbol{\beta}} & \boldsymbol{F}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \boldsymbol{F}_{\boldsymbol{\alpha}1} & \boldsymbol{F}_{\boldsymbol{\alpha}2} & \ldots & \boldsymbol{F}_{\boldsymbol{\alpha}n} \\ \boldsymbol{F}_{1\boldsymbol{\beta}} & \boldsymbol{F}_{1\boldsymbol{\alpha}} & \boldsymbol{F}_{11} & 0 & \ldots & 0 \\ \boldsymbol{F}_{2\boldsymbol{\beta}} & \boldsymbol{F}_{2\boldsymbol{\alpha}} & 0 & \boldsymbol{F}_{22} & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \\ \boldsymbol{F}_{n\boldsymbol{\beta}} & \boldsymbol{F}_{n\boldsymbol{\alpha}} & 0 & 0 & & \boldsymbol{F}_{nn} \end{bmatrix}, \tag{8}$$

has single components

$$\begin{aligned}
\boldsymbol{F}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^T \int_0^{t_{ij}} \exp(\eta_{ij}(s)) ds, \\
\boldsymbol{F}_{\boldsymbol{\beta}\boldsymbol{\alpha}} &= \boldsymbol{F}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^T = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^T} = -\sum_{i=1}^{n} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij} \int_0^{t_{ij}} \exp(\eta_{ij}(s)) \boldsymbol{\Phi}^T(s) ds, \\
\boldsymbol{F}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} &= -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} = -\sum_{i=1}^{n} \sum_{j=1}^{N_i} \int_0^{t_{ij}} \exp(\eta_{ij}(s)) \boldsymbol{\Phi}(s) \boldsymbol{\Phi}^T(s) ds + \boldsymbol{A}_{\xi_0,\xi,\zeta}, \\
\boldsymbol{F}_{\boldsymbol{\beta}i} &= \boldsymbol{F}_{i\boldsymbol{\beta}}^T = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{b}_i^T} = -\sum_{j=1}^{N_i} \boldsymbol{x}_{ij} \boldsymbol{u}_{ij}^T \int_0^{t_{ij}} \exp(\eta_{ij}(s)) ds,
\end{aligned}$$

$$
\begin{aligned}
\boldsymbol{F}_{\boldsymbol{\alpha} i} &= \boldsymbol{F}_{i\boldsymbol{\alpha}}^T = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{b}_i^T} = -\sum_{j=1}^{N_i} \boldsymbol{u}_{ij}^T \int_0^{t_{ij}} \exp(\eta_{ij}(s)) \boldsymbol{\Phi}(s) ds, \\
\boldsymbol{F}_{ii} &= -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial \boldsymbol{b}_i \partial \boldsymbol{b}_i^T} = -\sum_{j=1}^{N_i} \boldsymbol{u}_{ij} \boldsymbol{u}_{ij}^T \int_0^{t_{ii}} \exp(\eta_{ii}(s)) ds + \boldsymbol{Q}^{-1}.
\end{aligned}
$$

4.2.2 Variance-Covariance Components

Variance estimates for the random effects can be derived as an approximate EM algorithm, using the posterior mode estimates and posterior curvatures. If we define $\tilde{\boldsymbol{\beta}}^T := (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)$, we get the following simpler block structure for the information matrix from equation (8):

$$
\boldsymbol{F}^{pen}(\boldsymbol{\delta}) = \begin{bmatrix} \boldsymbol{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} & \boldsymbol{F}_{\tilde{\boldsymbol{\beta}}1} & \dots & \boldsymbol{F}_{\tilde{\boldsymbol{\beta}}n} \\ \boldsymbol{F}_{1\tilde{\boldsymbol{\beta}}} & \boldsymbol{F}_{11} & & 0 \\ \vdots & & \ddots & \\ \boldsymbol{F}_{n\tilde{\boldsymbol{\beta}}} & 0 & & \boldsymbol{F}_{nn} \end{bmatrix}.
$$

If the cluster sizes N_i are large enough, the estimator $\hat{\boldsymbol{\delta}}$ becomes approximately normal,

$$
\hat{\boldsymbol{\delta}} \overset{a}{\sim} N(\boldsymbol{\delta}, \boldsymbol{F}^{pen}(\hat{\boldsymbol{\delta}})^{-1}),
$$

see Fahrmeir and Tutz (2001). Hence, the (expected) curvature of $l^{pen}(\hat{\boldsymbol{\delta}})$ evaluated at the posterior mode, i.e. $\boldsymbol{F}^{pen}(\hat{\boldsymbol{\delta}})^{-1}$, is a good approximation to the covariance matrix. Then, using standard formulas for inverting partitioned matrices (see, for example, Magnus and Neudecker, 1988), the required posterior curvatures $\boldsymbol{V}_{ii}$ can be derived via the formula

$$
\boldsymbol{V}_{ii} = \boldsymbol{F}_{ii}^{-1} + \boldsymbol{F}_{ii}^{-1} \boldsymbol{F}_{i\tilde{\boldsymbol{\beta}}} (\boldsymbol{F}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}} - \sum_{i=1}^{n} \boldsymbol{F}_{\tilde{\boldsymbol{\beta}}i} \boldsymbol{F}_{ii}^{-1} \boldsymbol{F}_{i\tilde{\boldsymbol{\beta}}})^{-1} \boldsymbol{F}_{\tilde{\boldsymbol{\beta}}i} \boldsymbol{F}_{ii}^{-1}.
$$

Now, $\hat{\boldsymbol{Q}}^{(l)}$ can be computed by

$$
\hat{\boldsymbol{Q}}^{(l)} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\boldsymbol{V}}_{ii}^{(l)} + \hat{\boldsymbol{b}}_i^{(l)} (\hat{\boldsymbol{b}}_i^{(l)})^T). \tag{9}
$$

4.2.3 Starting Values

For fixed penalty parameters ξ_0 and ζ, we propose to first fit the model with a moderate choice of the parameters $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{u}}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ (typically $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\alpha}}^{(0)} = \hat{\boldsymbol{u}}^{(0)} = \boldsymbol{0}$; $\hat{\boldsymbol{\theta}}^{(0)}$ such that $\boldsymbol{Q}^{(0)}$ is moderate) and a high value for the penalty parameter ξ, such that all spline coefficients $\hat{\boldsymbol{\alpha}}$ are shrunk down to zero. Next, the penalty parameter ξ is successively decreased and for each new fit of the algorithm the previous parameter estimates serve as suitable starting values.

4.2.4 Determination of Optimal Tuning Parameters

As we have already mentioned in Section 3, the tuning parameter ξ_0, which controls the smoothness of the log-baseline hazard $\gamma_0(t) = \log(\lambda_0(t))$, in general needs not to be selected by a complex procedure but the estimation procedure is already stabilized for a moderate choice of ξ_0. However, some care should be taken to select ζ and especially ξ, which essentially determine the performance of the selection procedure.

A possible strategy is to specify for both ζ and ξ suitable grids of possible values and then use K-fold CV to select optimal values. As the tuning parameter ξ controls the overall amount of penalization, and hence, both smoothness and variable selection, it is of particular importance and we recommend to use a fine grid for this parameter. On the other hand, it turned out that for the second tuning parameter ζ, which controls the apportionment between smoothness and shrinkage, a rougher grid is sufficient. A suitable CV error measure on the test data is given by the model's log-likelihood (2), evaluated on the test data, i.e.

$$cve(\hat{\boldsymbol{\delta}}^{\text{train}}) = \sum_{i=1}^{n^{\text{test}}} \sum_{j=1}^{N_i^{\text{test}}} d_{ij}\hat{\eta}_{ij}(t_{ij}) - \int_0^{t_{ij}} \exp(\hat{\eta}_{ij}(s))ds,$$

where n^{test} denotes the number of clusters in the test data and N_i^{test} the corresponding cluster sizes. The estimator $\hat{\boldsymbol{\delta}}^{\text{train}}$ is obtained by fitting the model to the training data, resulting in the linear predictors $\hat{\eta}_{ij}(t)$. As K-fold CV can generally be time-consuming, it is again advisable to successively decrease the penalty parameter ξ and use the previous parameter estimates as starting values for each new fit of the algorithm while fixing the other penalty parameters ξ_0 and ζ. This strategy can considerably save computational time.

5 Simulation Studies

The underlying models are random intercept models with balanced design

$$\begin{aligned}\lambda_{ij}(t|\boldsymbol{z}_{ij}, u_i) &= \exp\left(\eta_{ij}(t)\right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, N_i \\ \eta_{ij}(t) &= \gamma_0(t) + \sum_{k=1}^{r} z_{ijk}\gamma_k(t) + b_i\end{aligned}$$

with different selections of (partly time-varying) effects out of the set:

$$\begin{array}{ll}\gamma_0(t) = 5 \cdot f_\Gamma(t) + 0.1, & \gamma_1(t) \equiv 1.2, \\ \gamma_2(t) \equiv -1.4, & \gamma_3(t) \equiv -0.8, \\ \gamma_4(t) \equiv 0.7, & \gamma_5(t) \equiv 0.8, \\ \gamma_6(t) \equiv -0.7, & \gamma_7(t) = (t+1)^{1/10} - 2, \\ \gamma_8(t) = 0.3 \cdot \sin(0.25t) + 0.4 + 0.03t, & \gamma_9(t) = -15 \cdot g_\Gamma(t) + 1, \\ \gamma_{10}(t) = \sqrt{t} - 2, & \gamma_{11}(t) = 1/(t+0.5), \\ \gamma_{12}(t) = 1.5 \cdot \sin(0.25t) - 1 + 0.2t, & \gamma_{13}(t) = \gamma_{14}(t) = \gamma_{15}(t) = \gamma_{16}(t) \equiv 0,\end{array}$$

where $\exp(\gamma_0(t))$ reflects the baseline hazard and f_Γ denotes the density of a Gamma distribution $\Gamma(\zeta, \theta)$. Shape and scale parameter were chosen as $\zeta = 4, \theta = 2$. Also g_Γ denotes the density of a Gamma distribution with shape and scale parameter chosen to be 5 and 2, respectively. So $\gamma_1(t)$ to $\gamma_6(t)$ represent time-constant and $\gamma_7(t)$ to $\gamma_{12}(t)$ time-varying effects, while the covariates corresponding to the remaining effects are noise variables. All covariates $z_{ijk}, k = 1, \ldots, 16$ have been drawn independently from a uniform distribution on $[-0.5; 0.5]$. The number of observations is either fixed by $n = 100$ or $n = 500$ clusters, each with $N_i \equiv 5$ or $N_i \equiv 1$ replicates, respectively. The random effects are specified by $b_i \sim N(0, \sigma_b^2)$ with three different scenarios $\sigma_b \in \{0, 0.5, 1\}$. In the following, we consider three different simulation scenarios:

$$\begin{aligned}
\textbf{Scenario A}: \quad \eta_{ij}(t) &= \gamma_0(t) + \sum_{k\in\{1,2,3,4,7,8,13,14,15,16\}} z_{ijk}\gamma_k(t) + b_i\,, \\
\textbf{Scenario B}: \quad \eta_{ij}(t) &= \gamma_0(t) + \sum_{k\in\{5,6,9,10,11,12,13,14\}} z_{ijk}\gamma_k(t) + b_i\,, \\
\textbf{Scenario C}: \quad \eta_{ij}(t) &= \gamma_0(t) + \sum_{k\in\{1,2,3,4,13\}} z_{ijk}\gamma_k(t) + b_i\,.
\end{aligned}$$

For the three scenarios, the performance of estimators is evaluated separately for the structural components and the random effects variance. In order to show that the penalty (4), which combines smoothness of the coefficient effects up to constant effects together with variable selection, indeed improves the fit in comparison to conventional penalization approaches, we compare the results of the `PenCoxFrail` algorithm with the results obtained by three alternative penalization approaches. The first approach, denoted by `Ridge`, is based on a penalty similar to the first term of the penalty (4), but with a ridge-type penalty on the spline coefficients, that is $\xi \cdot \tilde{J}(\boldsymbol{\alpha}) = \xi \left(\sum_{k=1}^{r} ||\boldsymbol{\Delta}_M^2 \boldsymbol{\alpha}_k||_2^2\right)$. Hence, smooth coefficient effects are obtained, though neither constant effect estimates are available nor variable selection is performed. The alternative competing approaches, denoted by `Linear` and `Select`, are obtained as the extreme cases of the `PenCoxFrail` algorithm, by setting the penalty parameter ζ either to 1 or 0, respectively. The former choice yields a penalty that can choose between smooth time-varying and constant effects, while the latter one yields a penalty that simultaneously selects significant variables with time-varying effects and produces smooth estimates for the nonzero coefficient functions.

In addition, we compare the results of the `PenCoxFrail` algorithm with the results obtained by using the R functions `gam` (Wood, 2011) and `coxph` (Therneau, 2013), which are available from the `mgcv` and `survival` library, respectively. However, it should be noted that although both functions can in principle be used to fit Cox frailty models with time-varying effects, the use of these packages is not straightforward.

Even though the gam function has recently been extended to include the Cox PH model, the estimation is based on penalized partial likelihood maximization and, hence, no time-varying effects can be included in the linear predictor. However, Holford (1980) and Laird and Olivier (1981) have shown that the maximum likelihood estimates of a piece-wise PH model and of a suitable Poisson regression model (including an appropriate offset) are equivalent. In the piece-wise PH model time is subdivided into reasonably small intervals and the baseline hazard is assumed to be constant in each interval. Therefore, after construction of an appropriate design matrix by "splitting" the data one can use the gam function to fit a Poisson regression model with time-varying coefficients and obtains estimates of the corresponding piece-wise PH model. In the gam function an extra penalty can be added to each smooth term so that it can be penalized to be zero. This means that the smoothing parameter estimation that is part of the fitting procedure can completely remove terms from the model. Though, in general, the equivalence between the piece-wise PH model and the offset Poisson model is well-known, to the best of our knowledge the concept of combining it with the flexible basis function approach implemented in the gam function, including time-varying effects, has not been exploited before.

In order to fit a time-varying effects model with coxph, we first constructed the corresponding B-spline design matrices. Next, we reparametrized them following Fahrmeir et al. (2004), such that the spline coefficients are decomposed into an unpenalized and a penalized part, and then incorporated the transformed B-spline matrices into the design matrix. Finally, to obtain smooth estimates for the time-varying effects, we put a small ridge penalty on the penalized part of the corresponding coefficients. However, for this fitting approach no additional selection technique for the smooth terms, in our case the time-varying coefficients, is available. The fit can be considerably improved if the data set is again enlarged by using a similar time-splitting procedure as for the gam function.

By averaging across 50 data sets we consider mean the squared errors for the baseline hazard, the smooth coefficient effects and σ_b given by:

$$\text{mse}_0 := \sum_{t=1}^{T} v_t(\gamma_0 - \hat{\gamma}_0)^2, \quad \text{mse}_\gamma := \sum_{k=1}^{r}\sum_{t=1}^{T} v_t(\gamma_k - \hat{\gamma}_k)^2, \quad \text{mse}_{\sigma_b} := (\sigma_b - \hat{\sigma}_b)^2.$$

To evaluate the estimated and true coefficient functions in the relevant part weights v_t are included that are defined by use of the cumulative baseline hazard $\Lambda_0(\cdot)$. They are given by $v_t = (\Lambda_0(T) - \Lambda_0(t))/\Lambda_0(T)$.

Simulation Study I ($n = 100, N_i = 5$)

Table 1-3 show the results of these quantities for the Ridge, the Linear, the Select, the gam, the coxph and the PenCoxFrail method for the three different simulation Scenarios A, B and C. In Figure 1 the performance of the methods is compared to `PenCoxFrail`.

First of all, it is obvious that the `Ridge` and `gam` method are clearly outperformed by the other methods in terms of mse_0 and mse_γ. It turns out that in terms of mse_0 the `Select` and `PenCoxFrail` procedures always perform very well, while the best performer in terms of mse_γ is changing over the scenarios: in Scenario A and C, where mostly time-constant or only slightly time-varying effects are present, the `Linear` procedure performs very well, while in Scenario B, where several strongly time-varying effects are present, the `Select` and `gam` procedures perform best. Altogether, here, the flexibility of the combined penalty (4) becomes obvious: regardless of how the underlying set of effects is composed of, the `PenCoxFrail` procedure is consistently among the best performers and yields estimates that are close to the estimates of the respective "optimal type of penalization". As this optimal type of penalization can change from setting to setting and is usually not known in advance, its automatic selection provides a substantial benefit in the considered class of survival models. With respect to the estimation of the random effects variance σ_b^2 all approaches yield satisfactory results, with slight advantages for the `Select`, `gam` and `PenCoxFrail` methods.

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	0	185 (205)	27 (27)	36 (51)	30 (47)	559 (2710)	912 (1290)
	0.5	494 (1652)	48 (68)	52 (69)	42 (58)	600 (2034)	817 (589)
	1	391 (452)	96 (119)	66 (84)	66 (78)	187 (307)	583 (554)
B	0	50 (78)	58 (75)	17 (37)	30 (73)	91 (114)	850 (763)
	0.5	90 (92)	67 (74)	34 (44)	35 (41)	89 (43)	688 (660)
	1	180 (403)	100 (117)	46 (55)	48 (57)	350 (1725)	510 (480)
C	0	118 (144)	29 (36)	16 (22)	15 (17)	595 (2524)	650 (477)
	0.5	132 (130)	42 (60)	29 (30)	22 (18)	362 (701)	623 (437)
	1	220 (321)	60 (81)	46 (62)	43 (54)	372 (905)	794 (2240)

TABLE 1: *Results for* mse_0 *(standard errors in brackets).*

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	0	3150 (1984)	239 (136)	673 (408)	411 (496)	548 (337)	4977 (2811)
	0.5	17270 (84558)	312 (329)	891 (380)	621 (448)	811 (656)	4709 (2295)
	1	9894 (20740)	361 (327)	973 (341)	683 (422)	666 (365)	3994 (2457)
B	0	1401 (958)	1178 (966)	675 (891)	793 (991)	586 (865)	2572 (2175)
	0.5	6333 (22140)	1862 (931)	787 (339)	839 (433)	486 (316)	2434 (1579)
	1	27218 (153493)	2837 (1423)	1213 (676)	1307 (765)	761 (541)	2523 (1742)
C	0	1474 (1820)	228 (431)	416 (210)	240 (231)	460 (635)	3030 (2477)
	0.5	1558 (1767)	182 (261)	475 (256)	305 (276)	395 (290)	2420 (1980)
	1	5024 (8620)	220 (318)	640 (393)	474 (418)	399 (265)	1401 (1079)

Table 2: *Results for* mse_γ *(standard errors in brackets).*

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	0	.032 (.031)	.019 (.021)	.009 (.016)	.011 (.017)	.006 (.013)	.002 (.003)
	0.5	.010 (.011)	.007 (.013)	.011 (.026)	.010 (.024)	.008 (.012)	.049 (.041)
	1	.012 (.018)	.012 (.019)	.018 (.025)	.017 (.024)	.008 (.009)	.032 (.040)
B	0	.040 (.041)	.045 (.041)	.021 (.030)	.025 (.034)	.019 (.028)	.003 (.006)
	0.5	.008 (.019)	.008 (.018)	.007 (.013)	.007 (.013)	.007 (.012)	.056 (.037)
	1	.011 (.019)	.009 (.017)	.013 (.014)	.013 (.014)	.012 (.016)	.037 (.045)
C	0	.037 (.030)	.029 (.027)	.014 (.020)	.016 (.021)	.010 (.016)	.002 (.003)
	0.5	.007 (.009)	.007 (.009)	.010 (.014)	.009 (.012)	.009 (.014)	.067 (.043)
	1	.012 (.013)	.013 (.016)	.018 (.021)	.018 (.021)	.014 (.018)	.054 (.062)

Table 3: *Results for* mse_{σ_b} *(standard errors in brackets).*

Next, we investigate the performance of the four different procedures focussing on the time-varying coefficient functions. Exemplarily, Figure 2-4 show the estimated effects of the coefficient functions obtained by all six methods in Scenario B with $\sigma_b = 1$. Though generally capturing some features of the coefficient functions, the `Ridge` and `coxph` methods do not yield satisfying results, in particular with respect to coefficient functions that are zero (compare $\widehat{\gamma_{13}}(t)$ and $\widehat{\gamma_{14}}(t)$ in Figure 4). In the chosen scenario, the `Select`, the `gam` and the `PenCoxFrail` procedure do a very good job in shrinking the coefficients of noise variables down to zero, see again $\widehat{\gamma_{13}}(t)$ and $\widehat{\gamma_{14}}(t)$, as well as in capturing the features of the strongly time-varying coefficient functions $\gamma_9(t)$ to $\gamma_{12}(t)$. With respect to these effects, the three procedures clearly outperform the other approaches. For the time-constant coefficient functions $\gamma_5(t)$ and $\gamma_6(t)$ the `Linear` method yields the best estimates. In general, note that though the `Select` and the `PenCoxFrail`

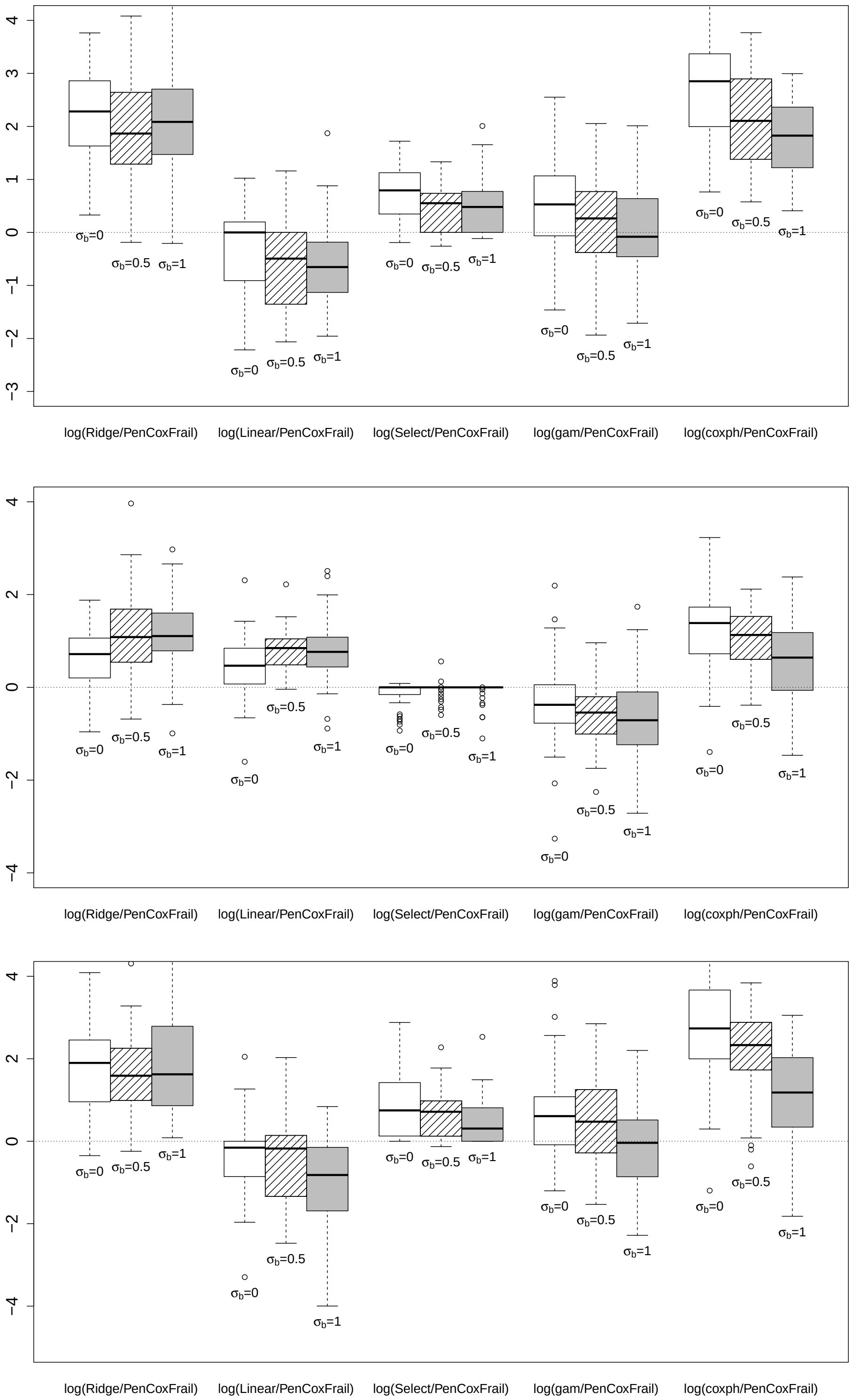

FIGURE 1: *Boxplots of* $\log(mse_\gamma(\cdot)/mse_\gamma(\texttt{PenCoxFrail}))$ *for Scenario A (top), B (middle) and C (bottom)*

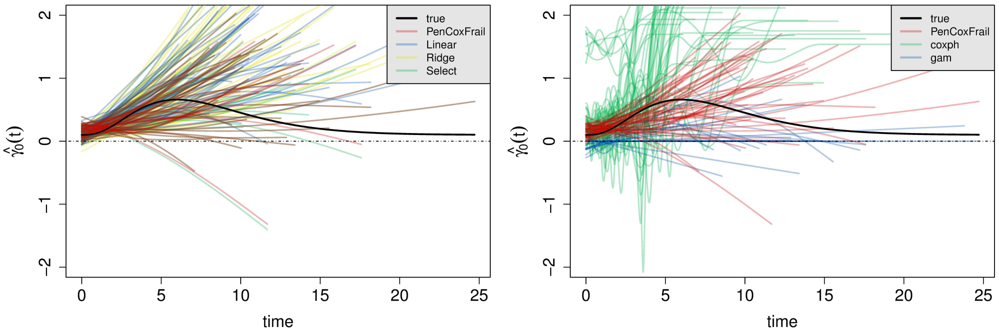

FIGURE 2: *Estimated (log-)baseline hazard $\widehat{\gamma}_0(t)$, exemplarily for Scenario B and $\sigma_b = 1$; left:* `Ridge` *(yellow),* `Linear` *(blue),* `Select` *(green) and* `PenCoxFrail` *(red); right:* `gam` *(blue),* `coxph` *(green) and* `PenCoxFrail` *(red); true effect in black*

method capture the features of the coefficient functions quite well, there is a substantial amount of shrinkage noticeable in the nonzero coefficient estimates, $\widehat{\gamma}_5(t), \widehat{\gamma}_6(t)$ and $\widehat{\gamma}_9(t)$ to $\widehat{\gamma_{12}}(t)$. The resulting bias is a typical feature of LASSO-type estimates and is tolerated in return for the obtained variance reduction.

Simulation Study II ($n = 500, N_i = 1$)

Similar to the simulation study from above, we now investigate classical frailty scenarios, with no cluster structure or repeated measurements, but where each observation obtains its own random intercept for modeling possible unobserved heterogeneity. Hence, the underlying models are basically the same random intercept models from above, but now with the number of observations fixed by $n = 500$ clusters without replicates, i.e. $N_i \equiv 1$. Note that the underlying models fulfill all necessary assumptions from Van den Berg (2001), which guarantee identifiability in the sense that there is a unique choice of the linear predictor and the random effects density that is able to generate these data.

The random effects are again specified by $b_i \sim N(0, \sigma_b^2)$ with three different scenarios $\sigma_b \in \{0, 0.5, 1\}$ and we consider the same three different simulation Scenarios A, B and C from above.

The performance of estimators is again evaluated separately for the structural components and the random effects variance and we again compare the `PenCoxFrail` method with several alternative approaches. In Figure 5, the comparison of the `PenCoxFrail` procedure with the other methods is visualized.

It is obvious that the `Ridge` and `coxph` method are again clearly outperformed

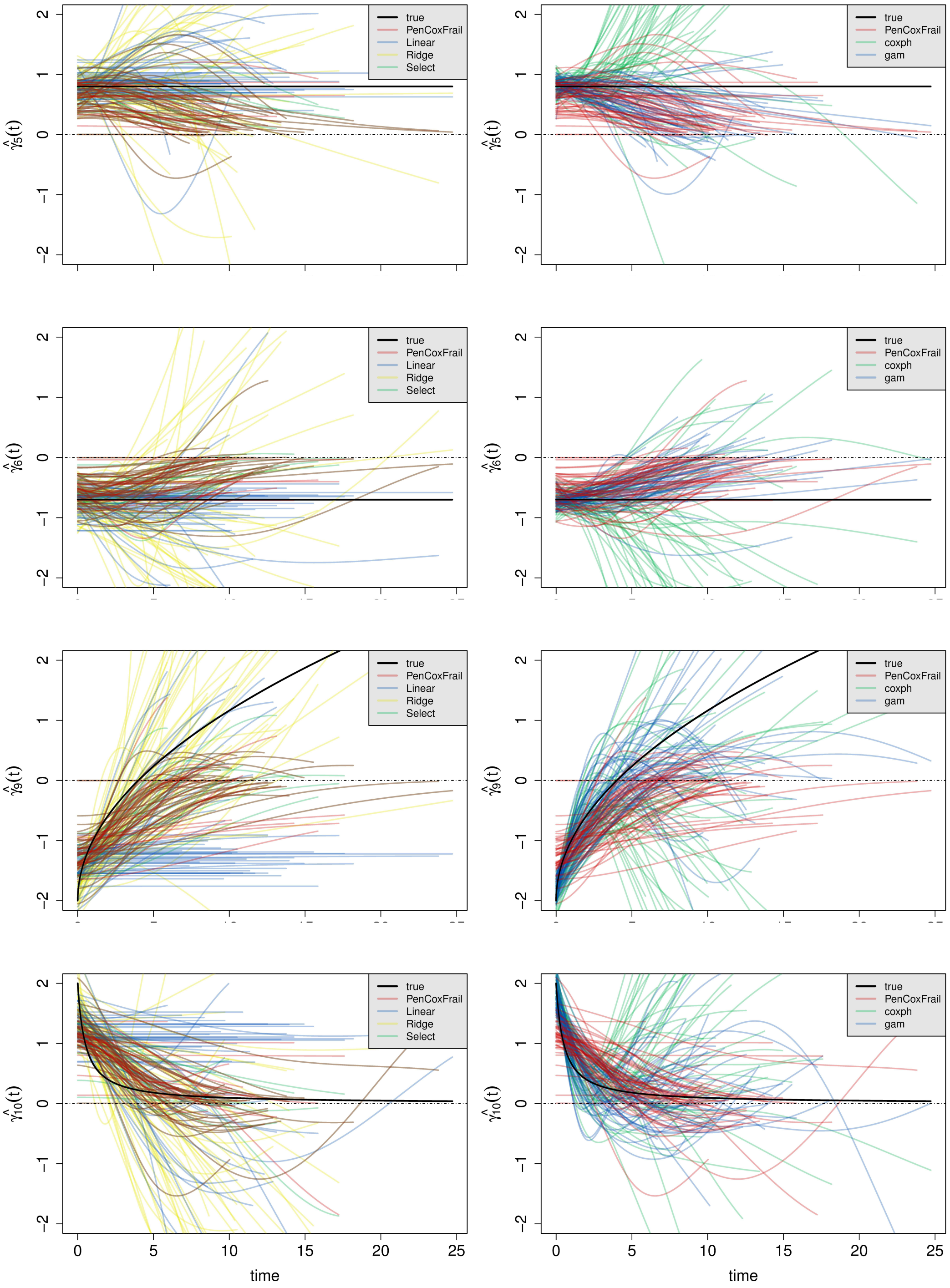

FIGURE 3: *Estimated (partly time-varying) effects $\widehat{\gamma_5}(t), \widehat{\gamma_6}(t), \widehat{\gamma_9}(t), \widehat{\gamma_{10}}(t)$, exemplarily for Scenario B and $\sigma_b = 1$; left:* `Ridge` *(yellow),* `Linear` *(blue),* `Select` *(green) and* `PenCoxFrail` *(red); right:* `gam` *(blue),* `coxph` *(green) and* `PenCoxFrail` *(red); true effect in black*

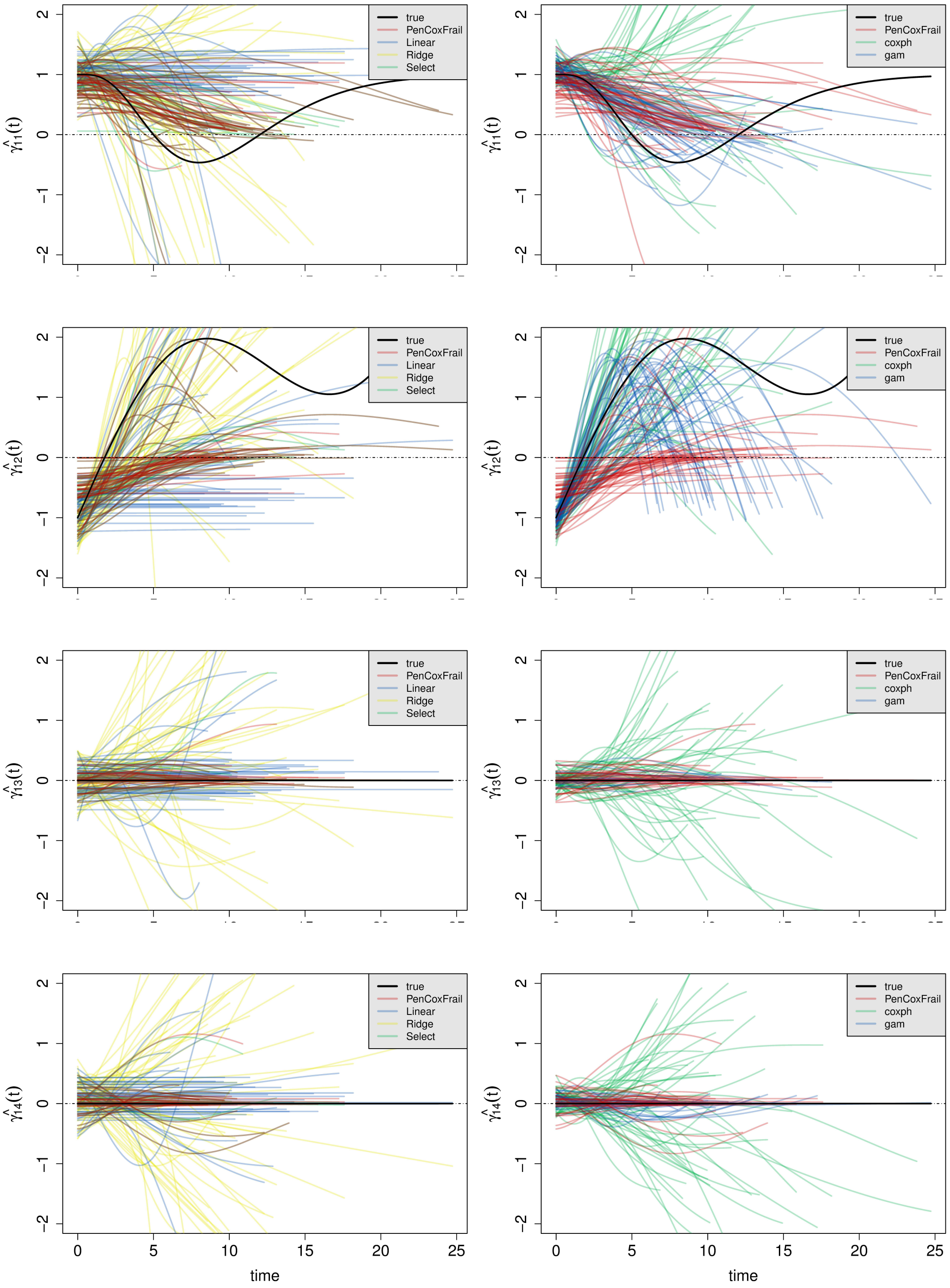

FIGURE 4: *Estimated (partly time-varying) effects $\widehat{\gamma_{11}}(t)$ to $\widehat{\gamma_{14}}(t)$, exemplarily for Scenario B and $\sigma_b = 1$; left:* `Ridge` *(yellow),* `Linear` *(blue),* `Select` *(green) and* `PenCoxFrail` *(red); right:* `gam` *(blue),* `coxph` *(green) and* `PenCoxFrail` *(red); true effect in black*

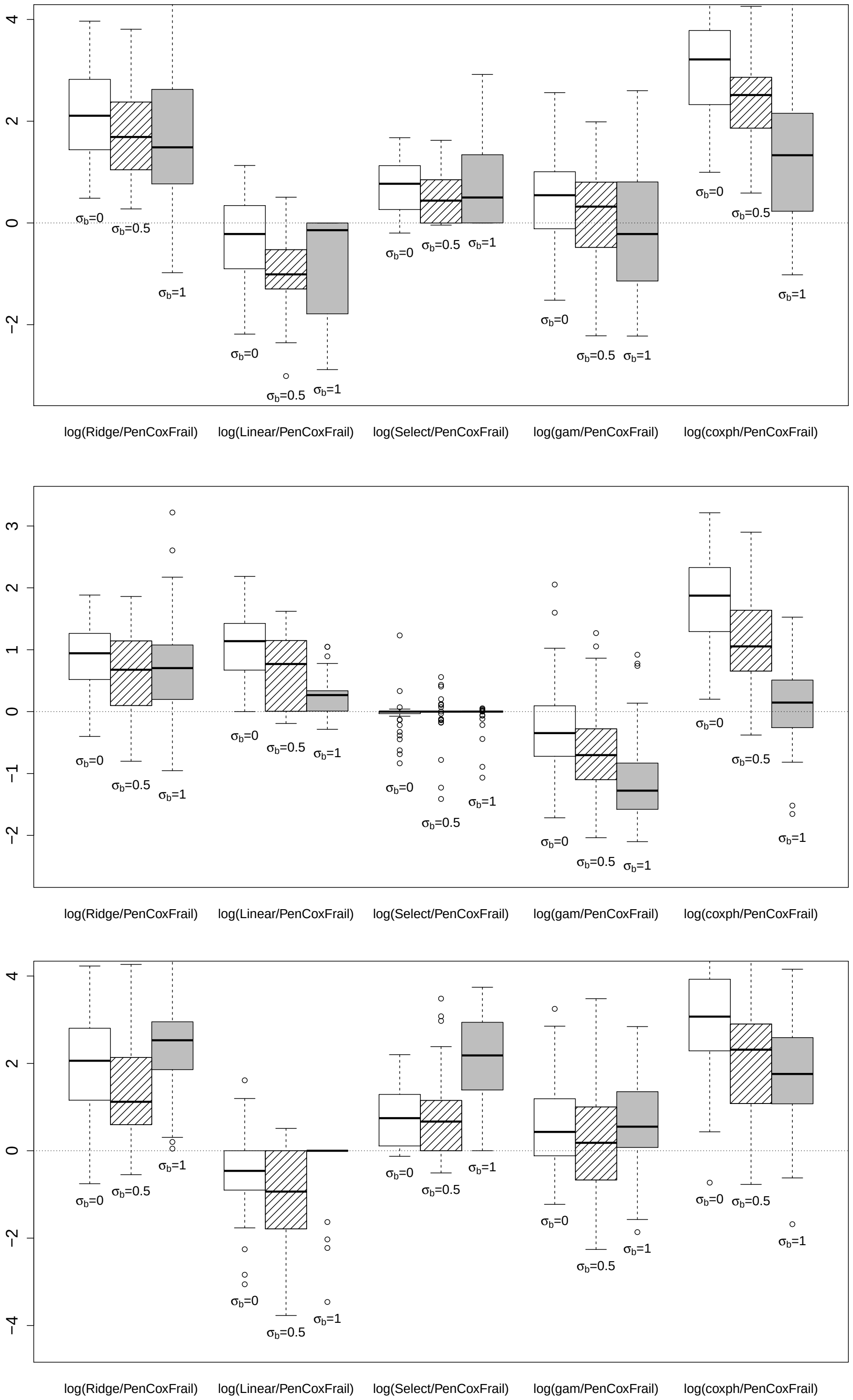

FIGURE 5: *Boxplots of* $\log(mse_\gamma(\cdot)/mse_\gamma(\texttt{PenCoxFrail}))$ *for Scenario A (top), B (middle) and C (bottom)*

by all other methods in terms of mse_0 and mse_γ. In addition, it turns out that in terms of mse_0 all other procedures perform well, but considerably deteriorate for the $\sigma_b = 1$ cases in all scenarios. The best performer in terms of mse_γ is changing over the scenarios, similar to Simulation Study I. Again, the flexibility of the combined penalty (4) becomes obvious: regardless of how the underlying set of effects is composed of, again, the `PenCoxFrail` procedure is consistently among the best performers and yields estimates that are close to the estimates of the respective "optimal type of penalization". With respect to the estimation of the random effects variance σ_b^2 all approaches yield satisfactory results, but have considerably deteriorated in comparison to Simulation Study I as no cluster structure is present, but each observation got its own random intercept.

Altogether, the simulations show that the proposed penalty (4) yields improved estimators in comparison to all conventional penalization approaches, as it can flexibly adopt to the underlying data driving mechanisms.

6 Application

In the following we will illustrate the proposed method on a real data set that is based on Germany's current panel analysis of intimate relationships and family dynamics (pairfam), release 4.0 (Nauck et al., 2013). The panel was started in 2008 and contains about 12.000 randomly chosen respondents from the birth cohorts 1971-73, 1981-83 and 1991-93. Pairfam follows the cohort approach, i.e. the main focus is on an anchor person of a certain birth cohort, who provides detailed information, orientations and attitudes (mainly with regard to their family plans) of both partners in interviews that are conducted yearly. A detailed description of the study can be found in Huinink et al. (2011).

The present data set was constructed similar to Groll and Abedieh (2015) and Schröder and Brüderl (2008). For a subsample of 2,501 women the retention time (in days) until the birth of the first child is considered as the dependent variable, starting at their 14th birthdays. In order to ensure that the independent time-varying covariates are temporally preceding the events, the duration until conception (and not birth) is considered, i.e. the time of event is determined by subtracting 7.5 months from the date of birth, which is when women usually notice pregnancy. For each woman the employment status is given as a time-varying categorical covariate with six categories, compare Table 6. Note that due to gaps in the women's employment histories a category called "no info" is introduced. As in the studies of Schröder and Brüderl (2008) and Groll and Abedieh (2015), for women who belong to this category for longer than 24 months it is set to "unemployed". Besides, several other time-varying and time-constant control variables are included. Table 5 and 6 give an overview of all considered variables together with their proportions in the sample. An extraction of the data set is shown in Table 4.

id	start	stop	child	job	rel.status	religion	siblings	...	federal state
1	0	365	0	school	single	Christian	1	...	Niedersachsen
1	365	730	0	no info	single	Christian	1	...	Niedersachsen
1	730	2499	0	unempl./job-seeking/housewife	single	Christian	1	...	Niedersachsen
1	2499	3261	0	full-time/self-employed	single	Christian	1	...	Niedersachsen
1	3261	3309	1	full-time/self-employed	partner	Christian	1	...	Niedersachsen
2	0	365	0	school	single	none	0	...	Thüringen
2	365	730	0	no info	single	none	0	...	Thüringen
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

TABLE 4: *Structure of the data*

Note that due to the incorporated time-varying covariates, the 2,501 observations have to be split whenever a time-varying covariate changes. This results in a new data set containing 20,550 lines. In order to account for regional fertility differences, we incorporate a random intercept for the German federal state where the women were born. Though this model could generally be fit with the `gam` function by construction of an appropriate design matrix, further splitting the data and then fitting a Poisson regression model with time-varying coefficients, in the present application this strategy would create an extremely large data set, which is not manageable. For this reason, we abstain from using the `gam` function. Moreover, as already pointed out in Section 5, in order to fit a time-varying effects model with `coxph`, again the data would have to be manually enlarged by using a similar time-splitting procedure. So we restrict our analysis to a conventional Cox model with time-constant effects, which we use for comparison with our `PenCoxFrail` approach.

When fitting the data with `PenCoxFrail`, because of the quite large sample size we use an ad-hoc strategy for determining the optimal tuning parameter ξ proposed in Chouldechova and Hastie (2015) and Ravikumar et al. (2007). In addition to considering the original variables in the dataset, we generate 10 noise variables and include them in the analysis. We simply fix the second tuning parameter to $\zeta = 0.5$ and fit `PenCoxFrail` using 5 basis functions for all 16 covariates. Figure 6 shows the regularization plots, which display $||\alpha||_2$ across the sequence of ξ values. It becomes obvious that there are two strong predictors that enter well before the "bulk", namely the "relationship status" (red) and the "education level" (blue).

Figure 7 and 8 show the estimated baseline hazard as well as the time-varying effects of the two strongest predictors, the "relationship status" and the "education level" right before the noise variables enter (i.e. their corresponding spline coefficients excess the threshold $||\boldsymbol{\alpha}||_2 > 0.05$), which corresponds to a tuning parameter choice $\xi_{46} = 7.79$. The baseline hazard exhibits a bell shape, which is in accordance with the typical female fertility curve: it is increasing from early adolescence with a maximum in the early thirties, before it decreases when the female menopause approaches. For the conventional Cox model with simple time-constant effects (red dashed line) the bell shape is more pronounced in

	proportion
Religion	
Christian	0.667
other	0.040
none	0.293
# siblings	
no siblings	0.19
one sibling	0.43
two siblings	0.22
three or more siblings	0.16
Education level of parents	
high	0.271
medium	0.061
low	0.570
no info	0.098
Number of women	2,501
Number of events	1,591

Table 5: *Distribution of the time-constant covariates in the sample*

	# days	proportion
Employment status		
full-time employed/self-employed	3,369,964	0.276
marginal/part-time employed	405,473	0.033
education	187,972	0.015
school	2,832,410	0.232
unempl./job-seeking/housewife	5,023,955	0.412
no info	388,936	0.032
Education level		
high	7,004,695	0.574
medium	4,301,786	0.352
low	837,023	0.069
no info	65,206	0.005
Relationship status		
single	6,463,726	0.529
partner	3,190,299	0.261
cohabitation	1,842,180	0.151
married	712,505	0.058
Number of women	2,501	
Number of events	1,591	
Number of days	12,208,710	

Table 6: *Distribution of the time-varying covariates in the sample*

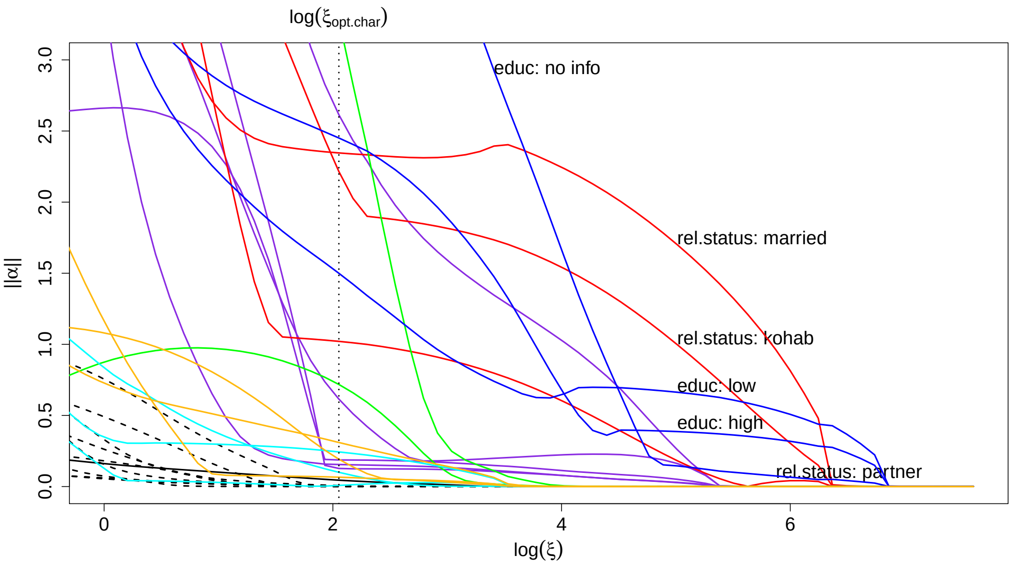

FIGURE 6: *Coefficient built-ups for the pairfam data vs.* $\log(\xi)$*. The colored solid lines correspond to the original 6 variables, black dashed lines to the simulated noise variables; the horizontal dotted line represents the chosen tuning parameter* $\log(\xi_{46}) = \log(7.79)$*.*

comparison to our time-varying effects approach (black solid line), where covariates are allowed to have a more complex effect over time. As to be expected, in comparison to the reference level *single* there is a positive effect on the transition rate into motherhood if the women have a partner in the sense that the closer the relationship the stronger the effect. The strongest positive effect is observed for married women, though this effect clearly declines when women are getting older and approach menopause. Besides, it turns out that a low or a high education level of the women clearly increases or decreases, respectively, the transition rate into motherhood in comparison to the reference level *medium education*. Again, it is remarkable that these effects on the fertility are clearly vanishing when women approach menopause. Furthermore, for young women there is a negative influence on the fertility, if no information regarding their education level is available. However, after a few years, this effect fundamentally changes and becomes positive when women approach menopause. For the remaining covariates we obtained the following results: all levels of the employment status seem to have no effect on the transition into motherhood, with the exception of the category *school* for which the probability of a transition into motherhood is clearly reduced as woman are usually quite young when attending school. Furthermore, we found a clear positive, time-constant effect of women having three or more siblings (reference category: no siblings), a positive effect of the women's parents educational level belonging to the category *no information* (reference

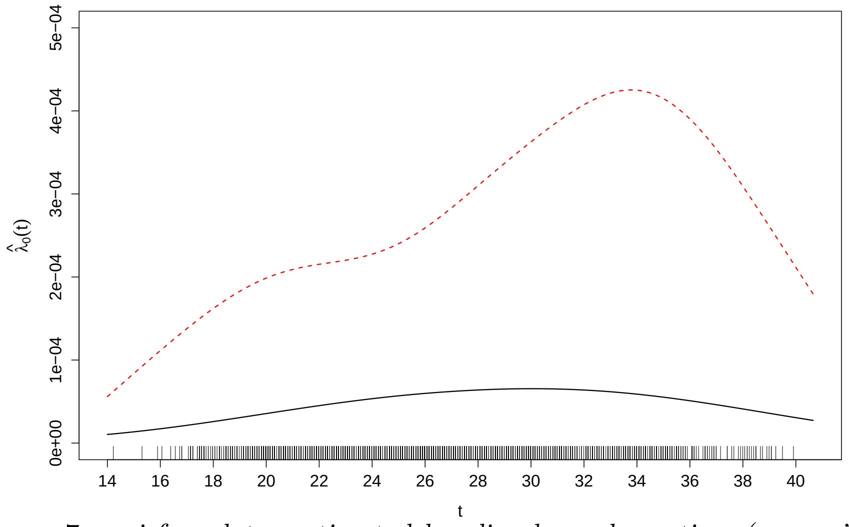

FIGURE 7: *pairfam data: estimated baseline hazard vs. time (women's age in years) at the chosen tuning parameter $\xi_{46} = 7.79$; for comparison, the estimated baseline hazard of a simple Cox model with time-constant effects is shown (red dashed line)*

category: medium education) and a negative effects of the categories *other religion* and *Christian* (reference category: no religion), which both are declining when women approach menopause. Finally, a certain amount of heterogeneity is detected between the German federal states with an estimated random effects variance $\hat{\sigma}_b = 0.179$.

7 Concluding Remarks

It turns out that the combination of the proposed penalization approach for model selection in Cox frailty models with time-varying coefficients with the promising class of multivariate log-normal frailties results in very flexible and sparse hazard rate models for modeling survival data. The conducted simulation study has illustrated the flexibility of the proposed combined penalty: regardless of the underlying set of true effects, the `PenCoxFrail` procedure can automatically adopt to it and yields estimates that are close to the "optimal type of penalization". As this optimal type of penalization can change from setting to setting and is usually not known in advance, its automatic selection provides a substantial benefit in the considered class of survival models.

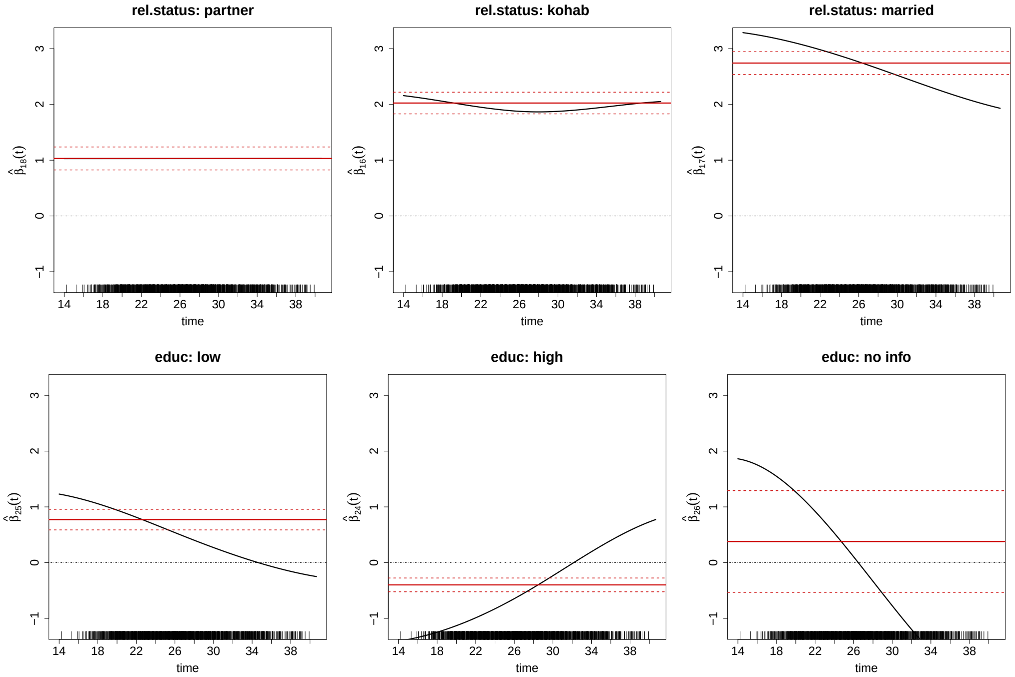

Figure 8: *pairfam data: estimated time-varying effects for the categorical covariates "relation ship status" and "education level" vs. time (women's age in years) at the chosen tuning parameter $\xi_{46} = 7.79$; for comparison, time-constant effects of a conventional Cox model are shown (red solid line) together with 95% confidence interval.*

Acknowledgements

This article uses data from the German family panel pairfam, coordinated by Josef Brüderl, Johannes Huinink, Bernhard Nauck, and Sabine Walper. Pairfam is funded as long-term project by the German Research Foundation (DFG).

References

Androulakis, E., C. Koukouvinos, and F. Vonta (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine 31*(20), 2223–2239.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association 88*, 9–25.

Chouldechova, A. and T. Hastie (2015). Generalized additive model selection. Technical Report, University of Stanford.

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

Do Ha, I., M. Noh, and Y. Lee (2012). frailtyhl: A package for fitting frailty models with h-likelihood. *The R Journal 4*(2), 28–36.

Eilers, P. H. C. (1995). Indirect observations, composite link models and penalized likelihood. In G. U. H. Seeber, B. J. Francis, R. Hatzinger, and G. Steckel-Berger (Eds.), *Proceedings of the 10th International Workshop on Statistical Modelling*. New York: Springer.

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

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective . *Statistica Sinica 14*, 715–745.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*. New York: Springer.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalize likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 74–99.

Gertheiss, J., A. Maity, and A.-M. Staicu (2013). Variable selection in generalized functional linear models. *Stat 2*(1), 86–101.

Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal 52*, 70–84.

Groll, A. and J. Abedieh (2015). Employment and fertility – a comparison of the family survey 2000 and the pairfam panel. In ??? and ??? (Eds.), *???*

Gui, J. and H. Z. Li (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics 2*, 3001–3008.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, 299–305.

Huinink, J., J. Brüderl, B. Nauck, S. Walper, L. Castiglioni, and M. Feldhaus (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Journal of Family Research 23*, 77–101.

Laird, N. and D. Olivier (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association 76*(374), 231–240.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference 139*(7), 2138–2146.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. London: Wiley.

Matsui, H. (2014). Variable and boundary selection for functional data via multiclass logistic regression modeling. *Computational Statistics & Data Analysis 78*, 176–185.

Matsui, H. and S. Konishi (2011). Variable selection for functional regression models via the l1 regularization. *Computational Statistics & Data Analysis 55*(12), 3304–3310.

Meier, L., S. Van de Geer, P. Bühlmann, et al. (2009). High-dimensional additive modeling. *The Annals of Statistics 37*(6B), 3779–3821.

Nauck, B., J. Brüderl, J. Huinink, and S. Walper (2013). The german family panel (pairfam). *GESIS Data Archive, Cologne*. ZA5678 Data file Version 4.0.0.

Oelker, M.-R. and G. Tutz (2013). A general family of penalties for combining different types of penalties in generalized structured models. Technical Report 139, Department of Statistics LMU Munich.

Park, M. Y. and T. Hastie (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B 19*, 659–677.

Ravikumar, P. D., H. Liu, J. D. Lafferty, and L. A. Wasserman (2007). Spam: Sparse additive models. In *NIPS*.

Ripatti, S. and J. Palmgren (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics 56*(4), 1016–1022.

Rondeau, V., Y. Mazroui, and J. R. Gonzalez (2012). frailtypack: an R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *Journal of Statistical Software 47*(4), 1–28.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schröder, J. and J. Brüderl (2008). Der Effekt der Erwerbstätigkeit von Frauen auf die Fertilität: Kausalität oder Selbstselektion? *Zeitschrift für Soziologie 37:2*, 117–136.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software 39(5)*, 1–13.

Therneau, T. M. (2013). *A package for survival analysis in S.* R package version 2.37-4.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine 16*, 385–395.

Van den Berg, G. J. (2001). Duration models: specification, identification and multiple durations. *Handbook of econometrics 5*, 3381–3460.

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*(486).

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* London: Chapman & Hall/CRC.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(1), 3–36.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B 68*, 49–67.