



# HHS Public Access

Author manuscript

*Biometrics*. Author manuscript; available in PMC 2018 November 28.

Published in final edited form as:

*Biometrics*. 2017 September ; 73(3): 846–856. doi:10.1111/biom.12637.

## Selection of Effects in Cox Frailty Models by Regularization Methods

Andreas Groll<sup>1,\*</sup>, Trevor Hastie<sup>2,\*\*</sup>, and Gerhard Tutz<sup>1,\*\*\*</sup>

<sup>1</sup>Department of Statistics, Ludwig-Maximilians-University Munich, Akademiestraße 1, 80799 Munich, Germany

<sup>2</sup>Department of Statistics, University of Stanford, 390 Serra Mall, Sequoia Hall, California 94305, U.S.A.

### Summary.

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 article, 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, 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

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

## 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. The Cox model assumes the semi-parametric hazard

---

\* andreas.groll@stat.uni-muenchen.de

\*\* hastie@stanford.edu

\*\*\* gerhard.tutz@stat.uni-muenchen.de

[Correction added on September 11, 2017, after first online publication: page 848, col 2, line 49, m instead of l for the index of alpha]

8. Supplementary Materials

Web Tables and Figures referenced in Sections 5 and 6 are available with this article at the *Biometrics* website on Wiley Online Library. The `PenCoxFrail` algorithm is implemented in the `pencoxfrail` function of the corresponding R-package (Groll, 2016; publicly available via CRAN, see <http://www.r-project.org>).

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

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

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

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 unobserved covariates. 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 has no closed form solution. In the Cox PH frailty model, also known as mixed PH model, the hazard rate of subject  $j$  belonging to cluster  $i$ , conditionally on the covariates  $\mathbf{x}_{ij}$  and the shared frailty  $b_i$ , is given by

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

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 frailties by Androulakis et al. (2012). They imposed the penalty on a generalized form of the full likelihood function designed for clustered data, which allows direct use of different distributions for the frailty term and which includes the Cox and 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, 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

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

where  $\mathbf{u}_{ij}^T = (u_{ij1}, \dots, u_{ijq})$  is the covariate vector associated with random effects  $\mathbf{b}_i$ , which follow a multivariate Gaussian distribution, that is,  $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}(\boldsymbol{\theta}))$ , with mean vector  $\mathbf{0}$  and covariance matrix  $\mathbf{Q}(\boldsymbol{\theta})$ , 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. A preliminary version of this work is available as technical report (Groll et al., 2016).

## 2. Cox Frailty Model with Time-Varying Coefficients

While for Cox frailty models with the simple predictor structure  $\eta_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{u}_{ij}^T \mathbf{b}_i$  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  $\alpha_{k,m}$ ,  $m = 1, \dots, M$ , 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” non-parametric modeling is considered. More precisely, with  $t_1^0 < \dots < t_N^0$  denoting the observed event times, the least informative non-parametric cumulative baseline hazard  $\Lambda_0(t)$  has a possible jump  $h_j$  at every observed event time  $t_j^0$ , that is,  $\Lambda_0(t) = \sum_{j=1}^N h_j I(t_j^0 \leq t)$ . 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  $\gamma_0(t)$  in B-splines.

Let now  $\mathbf{z}_{ij}^T = (1, z_{ij1}, \dots, z_{ijr})$  denote the covariate vector associated with both baseline hazard and time-varying effects and let  $\boldsymbol{\alpha}_k^T = (\alpha_{k,1}, \dots, \alpha_{k,M})$ ,  $k = 0, \dots, r$ , collect the spline coefficients corresponding to baseline hazard or  $k$ -th time-varying effect  $\gamma_k(t)$ , respectively. Further, let  $\mathbf{B}^T(t) := (\mathbf{B}_1(t, d), \dots, \mathbf{B}_M(t, d))$  represent the vector-valued evaluations of the  $M$  basis functions in time  $t$ . Then, with  $\mathbf{v}_{ijk} := z_{ijk} \cdot \mathbf{B}(t)$ , one can specify the hazard rate

$$\lambda(t \mid \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{u}_{ij}, \mathbf{b}_i) = \exp(\eta_{ij}(t)), \text{ with } \eta_{ij}(t) := \mathbf{x}_{ij}^T \boldsymbol{\beta} + \sum_{k=0}^r \mathbf{v}_{ijk}^T \boldsymbol{\alpha}_k + \mathbf{u}_{ij}^T \mathbf{b}_i. \quad (1)$$

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

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

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  $\mathbf{Q}(\boldsymbol{\theta})$  of the random effects  $\mathbf{b}_i$  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  $\boldsymbol{\delta}^T := (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \mathbf{b}^T)$ . The corresponding *marginal* log-likelihood has the form

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

where  $p(\mathbf{b}_i|\boldsymbol{\theta})$  denotes the density function of the random effects and the quantities

$L_i(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{b}_i := \prod_{j=1}^{N_i} \exp(\eta_{ij}(t_{ij}))^{d_{ij}} \exp(-\int_0^{t_{ij}} \exp(\eta_{ij}(s)) ds)$  represent the likelihood contributions

of single clusters  $i, i = 1, \dots, n$ . Approximation along the lines of Breslow and Clayton (1993) yields

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

The penalty term  $\mathbf{b}^T \mathbf{Q}(\boldsymbol{\theta}) \mathbf{b}$  results from Laplace approximation. 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 the profile likelihood  $l^{pp}(\boldsymbol{\delta}, \hat{\boldsymbol{\theta}})$ , and estimation of  $\boldsymbol{\theta}$ .

### 3. Penalization

In general, the roughness or “wiggleness” 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 those 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 non-zero coefficient functions, while Meier et al. (2009) proposed a sparsity-smoothness penalization for high-dimensional generalized additive models. More recently, Xiao et al. (2016) proposed a method that is able to identify the structure of covariate effects in a time-varying coefficient Cox model. They use a penalization approach that combines the ideas of local polynomial smoothing and a group non-negative garrote. It is implemented in the statistical program R and the corresponding code is available on one of the authors’ webpages. 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). Leng (2009), for example, 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

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

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

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

The first term of the penalty controls the smoothness of the time-varying covariate effects, whereby for values of  $\xi$  and  $\zeta$  large enough, all differences  $\boldsymbol{\alpha}_{k,l} - \boldsymbol{\alpha}_{k,l-1}$ ,  $l = 2, \dots, M$ , 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  $\xi$  and  $\zeta$  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, that is, the corresponding Newton-Raphson algorithm, local quadratic approximations of the penalty terms are used following Oelker and Tutz (2015). Note that the penalty from above may be easily extended by including a conventional LASSO penalty for time-constant fixed effects  $\boldsymbol{\beta}_k$ ,  $k = 1, \dots, p$ .

Since the baseline hazard in the predictor (1) is assumed to be semi-parametric, another penalty term should be included controlling its roughness. 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, that is,  $\gamma_0(t) = \sum_{m=1}^M \alpha_{0,m} B_m(t; d)$ , simply second order differences of adjacent spline weights  $\alpha_{0,m}$ ,  $m = 1, \dots, M$ , can be penalized. Hence, in addition to  $\xi \cdot J_{\zeta}(\boldsymbol{\alpha})$  the penalty

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

has to be included. Although this adds another tuning parameter  $\xi_0$ , it turns out that in general it is not worthwhile to also select  $\xi_0$  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 Groll and Tutz (2016). While some care should be taken to select  $\xi$  and  $\zeta$ , which determine the performance of the selection procedure, the estimation procedure is already stabilized in comparison to the usage of the “least informative” non-parametric cumulative baseline hazard  $\Lambda_0(t) = \sum_{j=1}^N h_j J(t_j^0 \leq t)$  for a moderate choice of  $\xi_0$ .

## 4. Estimation

Estimation is based on maximizing the penalized log-likelihood, obtained by expanding the approximate log-likelihood  $l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta})$  from (3) by the penalty terms  $\xi_0 \cdot J_0(\boldsymbol{\alpha}_0)$  and  $\xi \cdot J_\zeta(\boldsymbol{\alpha})$ , that is,

$$l^{pen}(\boldsymbol{\delta}, \boldsymbol{\theta}) = l^{app}(\boldsymbol{\delta}, \boldsymbol{\theta}) - \xi_0 \cdot J_0(\boldsymbol{\alpha}_0) - \xi \cdot J_\zeta(\boldsymbol{\alpha}). \quad (7)$$

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).

The presented algorithm is implemented in the `pencoxfrail` function of the corresponding R-package (Groll, 2016; publicly available via CRAN, see <http://www.r-project.org>).

### 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  $\xi_0$ ,  $\xi$  and  $\zeta$ , 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 (A.3) of supplementary materials).

(2) *Iteration*

For  $l = 1, 2, \dots$  until convergence:

(a) *Computation of parameters for given  $\hat{\boldsymbol{\theta}}^{(l-1)}$*

Based on the penalized score function  $s^{pen}(\boldsymbol{\delta}) = \mathbf{F}^{pen} / \boldsymbol{\delta}$  and the penalized information matrix  $\mathbf{F}^{pen}(\boldsymbol{\delta})$  (see Section (A.1) of supplementary materials) the general form of a single Newton-Raphson step is given by

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

As 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 (A.2) of supplementary materials), yielding the update  $\hat{\boldsymbol{\theta}}^{(l)}$ .

The `PenCoxFrail` algorithm stops if there is no substantial change in the estimated regression parameters collected in the parameter vector  $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{b}^T)$ . More mathematically, for a given small number  $\epsilon$  the algorithm is stopped, if



$$\|\delta^{(l+1)} - \delta^{(l)}\|_2 / \|\delta^{(l)}\|_2 \leq \varepsilon. \quad (8)$$

#### 4.2. Computational Details of PenCoxFrail

In the following, we give a more detailed description of the single steps of the PenCoxFrail algorithm. In Web Appendix A of the supplementary materials, we describe the derivation of score function and information matrix. Then, an estimation technique for the variance-covariance components is given. Finally, we give details for computation of starting values.

As already mentioned in Section 3, the tuning parameter  $\xi_0$  controlling smoothness of the log-baseline hazard  $\gamma_0(t) = \log(\lambda_0(t))$  needs not to be selected carefully. A possible strategy is to specify for both  $\zeta$  and  $\xi$  suitable grids of possible values and then use  $K$ -fold CV to select optimal values. From our experience for the second tuning parameter  $\zeta$ , which controls the apportionment between smoothness and shrinkage, a rough grid is sufficient, whereas a fine grid is needed for  $\xi$ . A CV error measure on the test data is the model's log-likelihood evaluated on the test data, that is,

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

where  $n^{\text{test}}$  denotes the number of clusters in the test data and  $N_i^{\text{test}}$  the corresponding cluster sizes. The estimator  $\hat{\delta}^{\text{train}}$  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  $\xi$  and use the previous parameter estimates as starting values for each new fit of the algorithm while fixing the other penalty parameters  $\xi_0$  and  $\zeta$ . This strategy can considerably save computational time.

### 5. Simulation Studies

The underlying models are random intercept models with balanced design

$$\lambda_{ij}(t | z_{ij}, u_i) = \exp(\eta_{ij}(t)), \quad i = 1, \dots, n, \quad j = 1, \dots, N_i, \quad \eta_{ij}(t) = \gamma_0(t) + \sum_{k=1}^r z_{ijk} \gamma_k(t) + b_i$$

with different selections of (partly time-varying) effects  $\gamma_k(t)$ ,  $k \subset \{1, \dots, r\}$ , and random intercepts  $b_i \sim N(0, \sigma_b^2)$ ,  $\sigma_b \in \{0, 0.5, 1\}$ . We consider three different simulation scenarios, and the performance of estimators is evaluated separately for the structural components and the random effects variance. A detailed description of the simulation study design is found in Web Appendix B. 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,  $\xi \cdot \tilde{J}(\alpha) = \xi(\sum_{k=1}^r \|\Delta_M^2 \alpha_k\|_2^2)$ . 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 extreme cases of the PenCoxFrail algorithm, by choosing a penalty parameter  $\zeta$  equal 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 non-zero 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 the equivalence of piece-wise PH model and the offset Poisson model is generally 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.

Averaging across 50 data sets, we consider mean squared errors for baseline hazard ( $mse_0$ ), smooth coefficient effects ( $mse_\gamma$ ) and  $\sigma_b$  ( $mse_{\sigma_b}$ ), defined in Web Appendix B. In addition, we investigate proportions of correctly identified null, constant and time-varying effects as well as of the correctly identified exact true model structure. Finally, all methods are also compared with regard to computational time and proportions of non-convergent simulation runs.

### Simulation Study I ( $n = 100, N_j = 5$ )

Table 1 and Web Tables S1 and S2 show the three different mean squared errors (B.1) for the Ridge, Linear, Select, gam, 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 coxph method are clearly outperformed by the other methods in terms of  $mse_0$  and  $mse_\gamma$ . 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  $mse_\gamma$  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  $\sigma_b^2$  all approaches yield satisfactory results, with slight advantages for the Select, gam, and PenCoxFrail methods.

Next, we investigate the performance of the four different procedures focussing on the time-varying coefficient functions. Exemplarily, Figure 2 and Web Figures S1 and S2 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 2). In the chosen scenario, the Select, gam, and PenCoxFrail procedures do a very good job in shrinking 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 method capture the features of the coefficient functions quite well, there is a substantial amount of shrinkage noticeable in the non-zero 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.

The `Linear` and `Select` approaches only yield good results either with respect to distinction between constant and time-varying effects or with respect to variable selection, whereas the `PenCoxFrail` method performs well in terms of both criteria. Also `gam` yields good results regarding the distinction between zero and time-varying effects, however, it fails to correctly identify constant effects. While in Scenarios A and B not a single approach is able to correctly identify the exact true model structure, in Scenario C at least the `PenCoxFrail` succeeds in about one fifth of the simulation runs in the three  $\sigma_b$  settings (compare Web Table S5).

In terms of computational time, the `coxph` method is by far the fastest approach, followed by the three methods `ridge`, `linear`, and `select`, which all depend on a single tuning parameter. Together with `gam`, the `PenCoxFrail` approach is the most demanding method because it uses two tuning parameters (compare Web Table S3). In addition, we considered the proportions of non-convergent simulation runs. The methods are considered to have converged, if criterion (8) was reached within 100 iterations. Although the `ridge`, `linear`, `select`, and `PenCoxFrail` methods did not converge in some scenarios, their performance with respect to the three mean squared errors (B.1), including the non-convergent simulation runs, is nevertheless satisfactory, particularly for `PenCoxFrail` (compare Web Table S4). Two additional simulation studies are presented in the supplementary materials, see Web Appendix B.

## 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, that is, 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 (2016). For a subsample of 2501 women retention time (in days) until birth of the first child is considered as dependent variable, starting at their 14th birthdays. To ensure that the (partly time-varying) covariates are temporally preceding the events, duration until conception (and not birth) is considered, that is, the event time 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 covariate with six categories. Note that due to gaps in the women's employment histories, the category "no info" is introduced. As in the preceding studies, for women remaining in this category longer than 24 months it is set to "unemployed." Besides, several other time-varying and time-constant control variables are

included. Web Tables S9 and S10 give an overview of all considered variables together with their sample proportions and Table 2 shows an extraction of the data set.

Note that due to the incorporated time-varying covariates, the 2501 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 cross-validation as the standard procedure for the choice of the tuning parameters turned out to be extremely time-consuming. Hence, we use an alternative ad hoc strategy to determine the optimal tuning parameter  $\xi$  proposed in Liu et al. (2007) and Chouldechova and Hastie (2015). In addition to considering the original variables in the data set, 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 3 shows the regularization plots, which display  $\|\mathbf{a}\|_2$  along the sequence of  $\xi$  values. It turns out that there are two strong predictors that enter well before the “bulk,” namely the “relationship status” (solid) and “education level” (dashed-dotted).

Figure 4 shows the estimated time-varying effects of the two strongest predictors, the “relationship status” and the “education level” right before the noise variables enter (that is, their corresponding spline coefficients exceed the threshold  $\|\mathbf{a}\|_2 > 0.05$ ), which corresponds to a tuning parameter choice  $\xi_{48} = 6.09$ .

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 (compare Web Figure S4). For the conventional Cox model with simple time-constant effects (red dashed line) the bell shape is more pronounced in comparison to our time-varying effects approach (black solid line), where covariates are allowed to have more complex effects over time. As expected, compared to the reference *single* there is a positive effect on the transition rate into motherhood if 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 get older and approach menopause. Besides, it turns out that a low (high) education level of women clearly increases (decreases) the transition rate into motherhood compared to the reference *medium education*. Again, it is remarkable that these effects on the fertility are clearly vanishing when women approach menopause. For the remaining covariates we obtained the following results: no level of the employment status

seems to have an effect on the transition into motherhood, except for the category *school* for which the probability of a transition into motherhood is clearly reduced as women 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 slightly positive effect of the women's parents educational level belonging to the category *no info* (reference: medium education) and a negative effect of the women's parents being highly educated, which is diminishing over time. Moreover, negative effects of the categories *other religion* and *Christian* (reference: no religion) were observed, which both decline when women approach menopause. Finally, a certain amount of heterogeneity is detected between German federal states with an estimated standard deviation  $\hat{\sigma}_b = 0.179$ .

## 7. Concluding Remarks

It turns out that combining 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 simulations illustrate 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 close to the “optimal type of penalization.” As this optimal type of penalization can change from setting to setting and is usually unknown in advance, its automatic selection provides a substantial benefit in the considered class of survival models.

`PenCoxFrail` uses the Laplace approximation, which yields a log-likelihood (3) that already contains a quadratic penalty term. By appending structured penalty terms that enforce the selection of effects one obtains a penalized log-likelihood that contains all the penalties. An alternative strategy would be the indirect maximization of the log-likelihood by considering the random effects as unobservable variables. Then an EM algorithm could be derived, in which the penalization for effect selection is incorporated in the M-step. For the GLMM, the EM strategy has the advantage that one can use maximization tools from the generalized linear models framework, however, this simplification may not work for the Cox model.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## 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). Trevor Hastie was partially supported by grant DMS-1407548 from the National Science Foundation, and grant 5R01 EB 00198821 from the National Institutes of Health.

## References

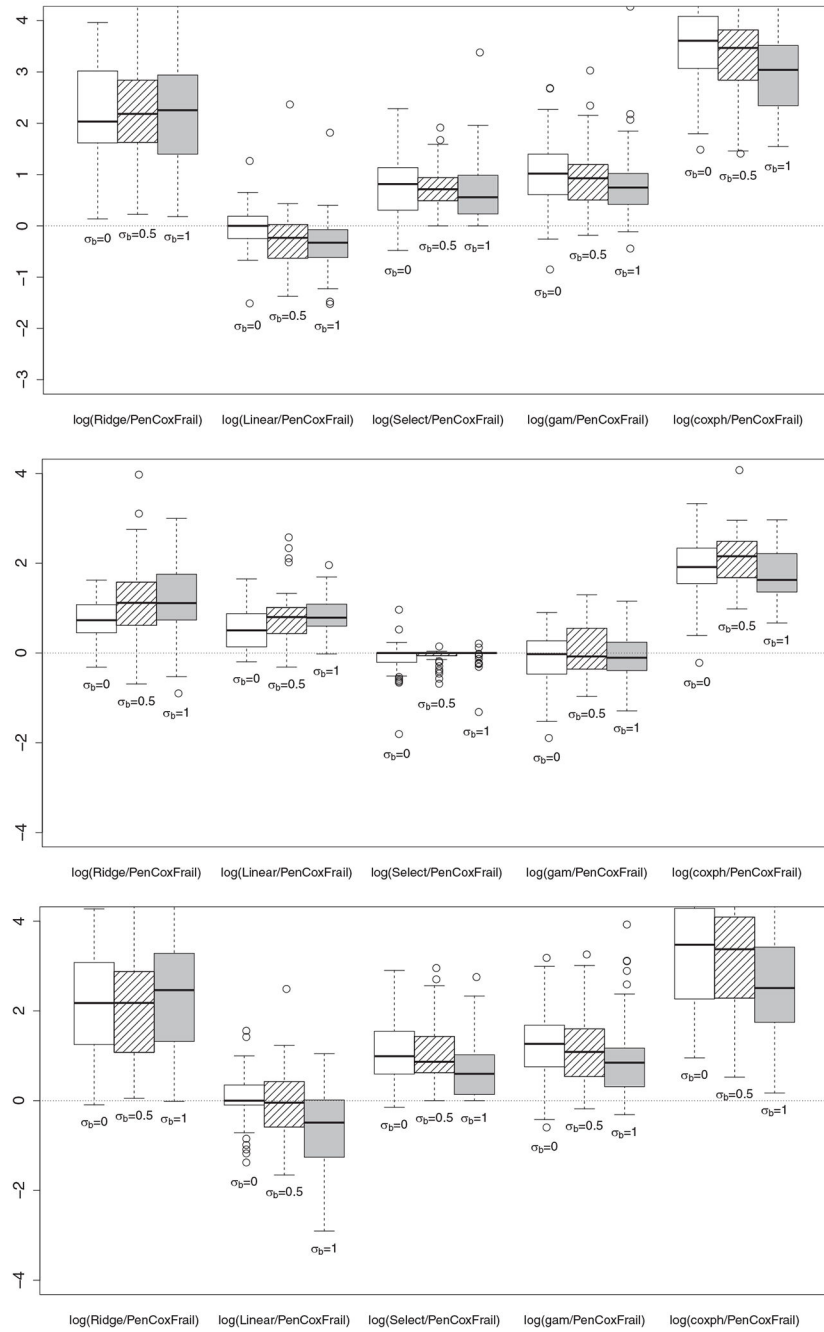
Androulakis E, Koukouvinos C, and Vonta F (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine* 31, 2223–2239. [PubMed: 22419584]

- Breslow NE and Clayton DG (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88, 9–25.
- Chouldechova A and Hastie T (2015). Generalized additive model selection. Technical Report, University of Stanford.
- Cox DR (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 34, 187–220.
- Do Ha I, Noh M, and Lee Y (2012). frailtyhl: A package for fitting frailty models with h-likelihood. *The R Journal* 4, 28–36.
- Eilers PHC (1995). Indirect observations, composite link models and penalized likelihood In Seeber GUH, Francis BJ, Hatzinger R, and Steckel-Berger G (eds), *Proceedings of the 10th International Workshop on Statistical Modelling*. NY: Springer.
- Eilers PHC and Marx BD (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11, 89–121.
- Fahrmeir L, Kneib T, and Lang S (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica* 14, 715–745.
- Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.
- Fan J and Li R (2002). Variable selection for Cox’s proportional hazards model and frailty model. *Annals of Statistics* 74–99.
- Gertheiss J, Maity A, and Staicu A-M (2013). Variable selection in generalized functional linear models. *Stat* 2, 86–101. [PubMed: 25132690]
- Goeman JJ (2010). Li penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 52, 70–84. [PubMed: 19937997]
- Groll A (2016). PenCoxFrail: Regularization in Cox Frailty Models R package version 1.0.1.
- Groll A and Abedieh J (2016). Employment and fertility—A comparison of the family survey 2000 and the pairfam panel In Manca R, McClean S, and Skiadas CH (eds), *New Trends in Stochastic Modeling and Data Analysis*. ISAST to appear.
- Groll A, Hastie T, and Tutz G (2016). Regularization in Cox frailty models. Technical Report 191, Ludwig-Maximilians-University.
- Groll A and Tutz G (2016). Variable selection in discrete survival models including heterogeneity. *Lifetime Data Analysis*, to appear.
- Gui J and Li HZ (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 TR (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* 299–305. [PubMed: 7407317]
- Huinink J, Briiderl J, Nauck B, Walper S, Castiglioni L, and Feldhaus M (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Journal of Family Research* 23, 77–101.
- Laird N and Olivier D (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* 76, 231–240.
- Leng C (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference* 139, 2138–2146.
- Liu H, Wasserman L, Lafferty JD, and Ravikumar PK (2007). SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems* (pp. 1201–1208).
- 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 Konishi S (2011). Variable selection for functional regression models via the l1 regularization. *Computational Statistics & Data Analysis* 55, 3304–3310.
- Meier L, Van de Geer S, Buhlmann P (2009). High-dimensional additive modeling. *The Annals of Statistics* 37, 3779–3821.
- Nauck B, Bruderl J, Huinink J, and Walper S (2013). The german family panel (pairfam). GESIS Data Archive, Cologne ZA5678 Data file Version 4.0.0.

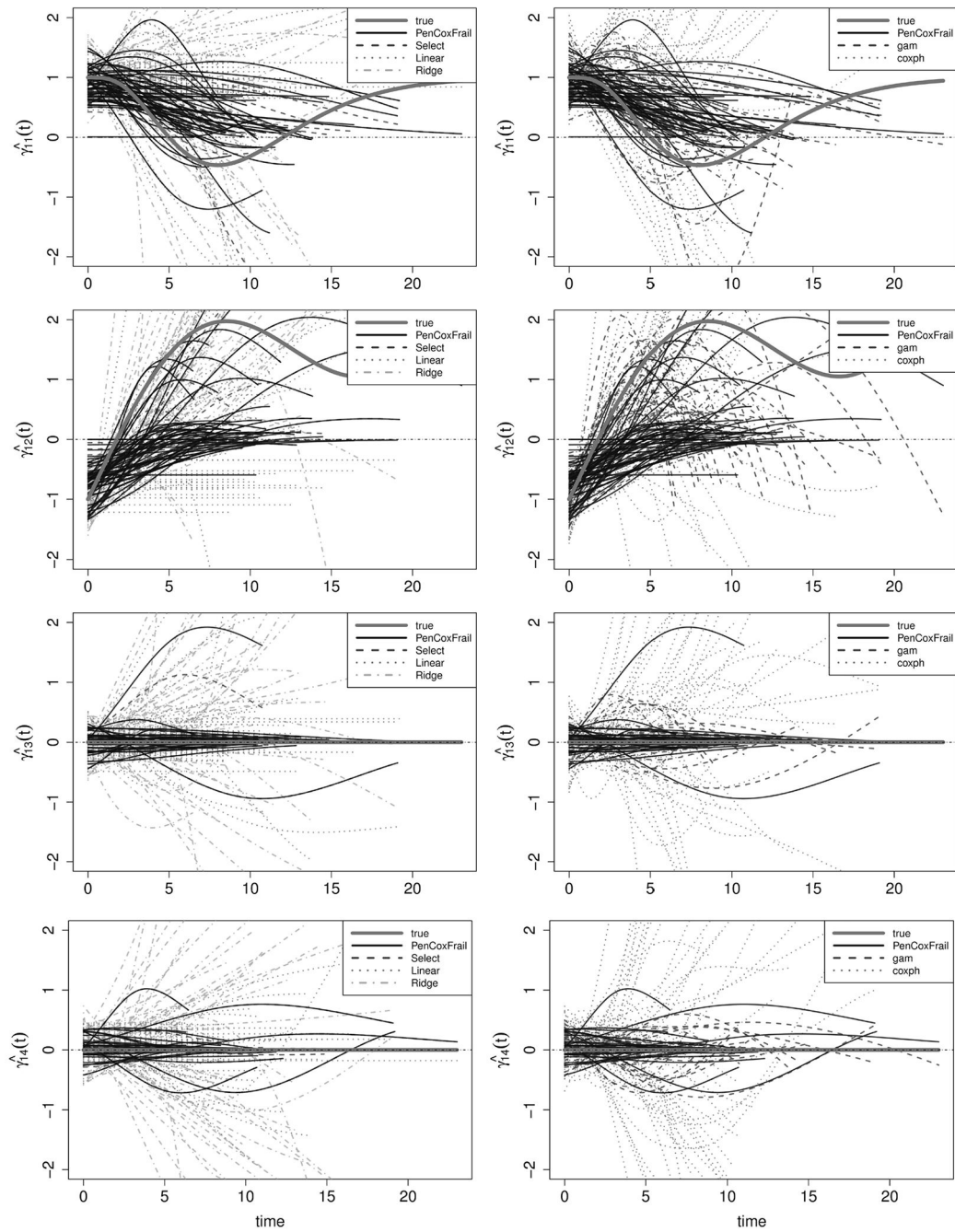


- Oelker M-R and Tutz G (2015). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, 1–24.
- Park MY and Hastie T (2007).  $L_1$ -regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B* 19, 659–677.
- Ripatti S and Palmgren J (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* 56, 1016–1022. [PubMed: 11129456]
- Rondeau V, Mazroui Y, and Gonzalez JR (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, 1–28.
- Ruppert D, Wand MP, and Carroll RJ (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Simon N, Friedman J, Hastie T, and Tibshirani R (2011). Regularization paths for Cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software* 39, 1–13.
- Therneau TM (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. [PubMed: 9044528]
- Van den Berg GJ (2001). Duration models: Specification, identification and multiple durations. *Handbook of Econometrics* 5, 3381–3460.
- Wang H and Xia Y (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* 104, 747–757.
- Wang L, Li H, and Huang JZ (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* 103, 1556–1569. [PubMed: 20054431]
- Wood SN (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall/CRC.
- Wood SN (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, 3–36.
- Xiao W, Lu W, and Zhang HH (2016). Joint structure selection and estimation in the time-varying coefficient Cox model. *Statistica Sinica* 26, 547–567. [PubMed: 27540275]
- Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B* 68, 49–67.

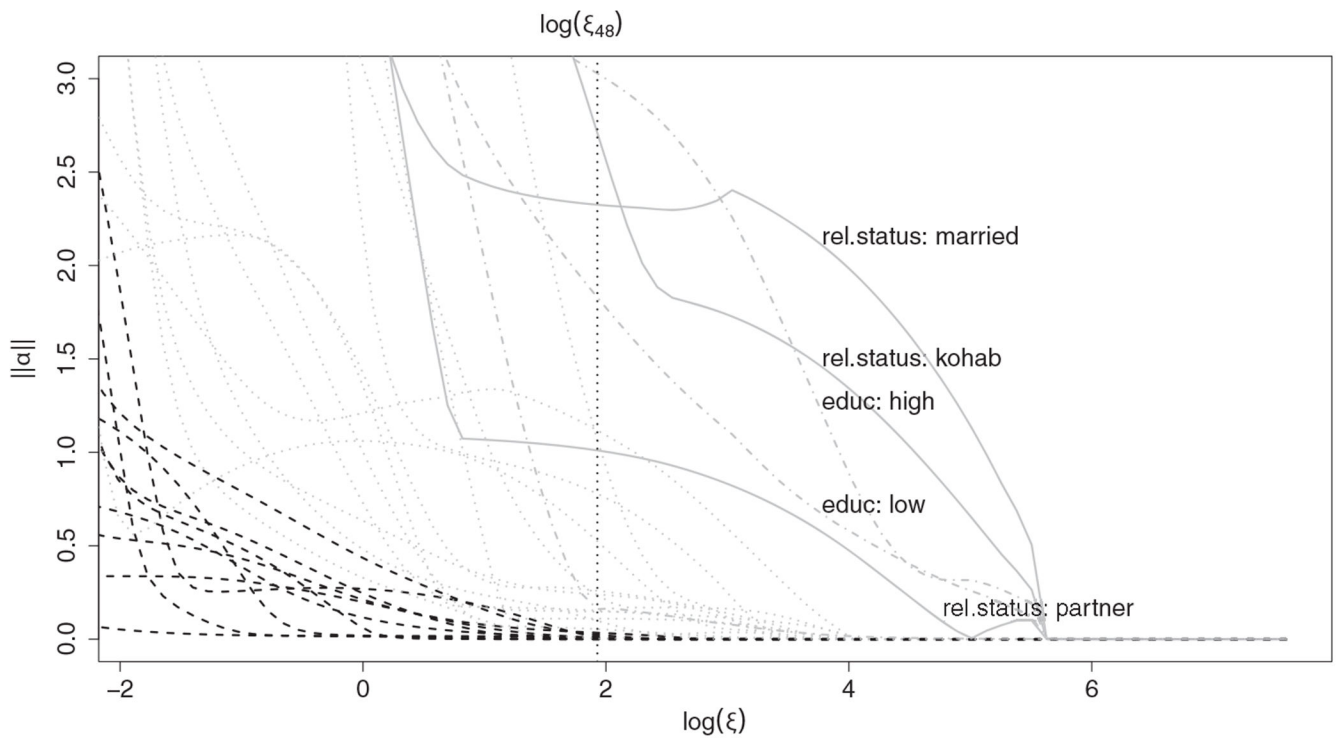




**Figure 1.** Simulation study I: boxplots of  $\log(\text{mse}_{\gamma}(\cdot)/\text{mse}_{\gamma}(\text{PenCoxFrail}))$  for Scenario A (top), B (middle), and C (bottom).

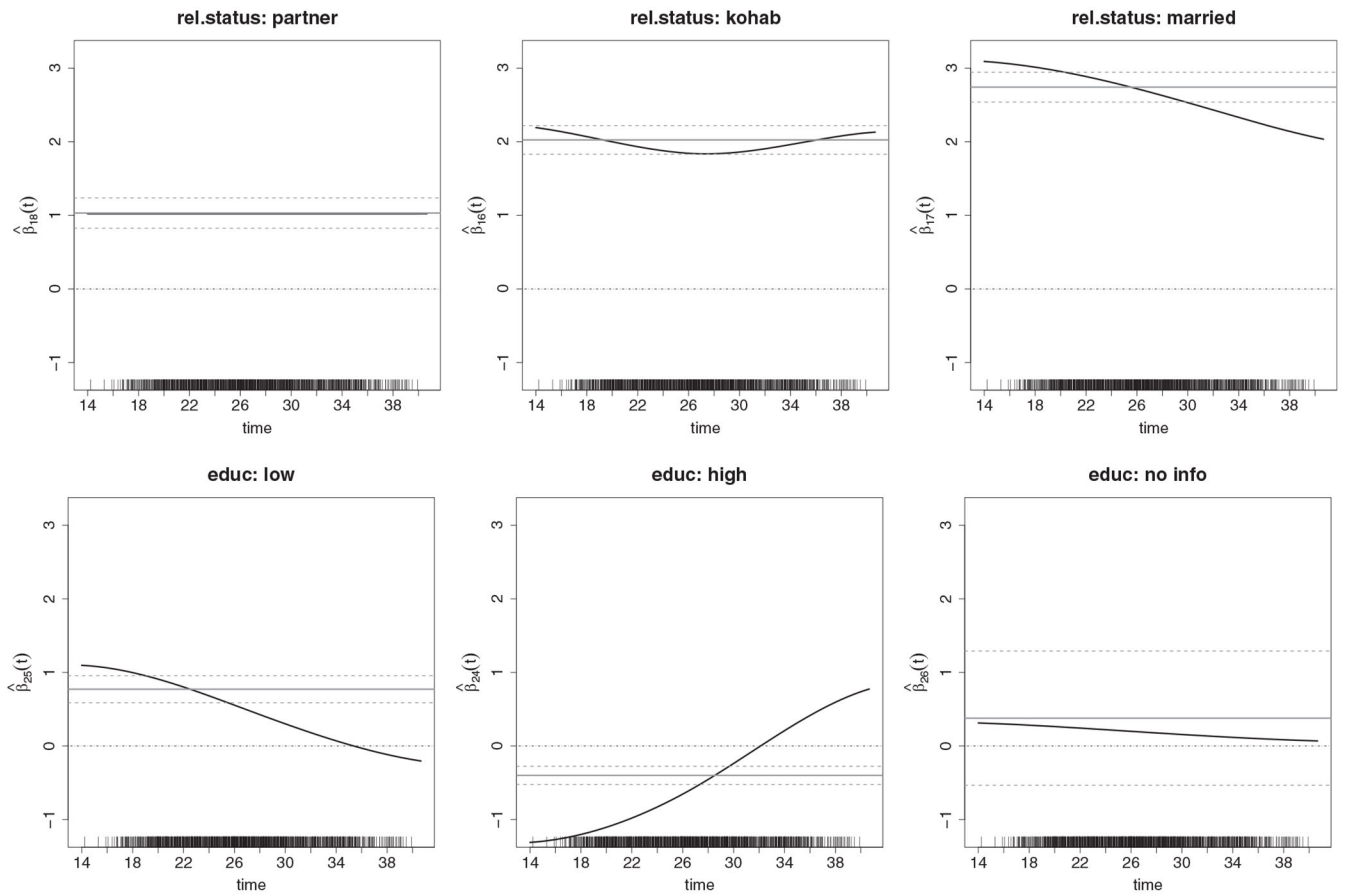


**Figure 2.** Estimated (partly time-varying) effects  $\widehat{\gamma}_{11}(t)$  to  $\widehat{\gamma}_{14}(t)$ , exemplarily for Scenario B and  $\sigma_b = 1$ ; left: Ridge (dashed-dotted light gray), Linear (dotted gray), Select (dashed dark gray) and PenCoxFrail (black); right: gam (dashed dark gray), coxph (dotted gray) and PenCoxFrail (black); true effect in thick gray.



**Figure 3.**

Coefficient built-ups for the pairfam data along a transformation of the tuning parameter, namely  $\log(\xi)$ . The gray solid, dashed-dotted and dotted lines correspond to the original six variables, black dashed lines to the simulated noise variables; the horizontal dotted line represents the chosen tuning parameter  $\log(\xi_{48}) = \log(6.09)$ .



**Figure 4.**

Pairfam data: estimated time-varying effects for the categorical covariates “relationship status” and “education level” versus time (women’s age in years) at the chosen tuning parameter  $\xi_{48} = 6.09$ ; for comparison, time-constant effects of a conventional Cox model are shown (gray solid line) together with 95% confidence intervals.

**Table 1**

Results for  $mse_{\gamma}$  (standard errors in brackets)

Scenario	$\sigma^2$	Ridge	Linear	Select	PenCox	Frail	gam	coxph
A	0	3283 (3772)	337 (427)	562 (210)	305 (276)	760 (524)	10709 (7403)	
	.5	17270 (84558)	307 (247)	769 (265)	406 (221)	992 (489)	11737 (6958)	
	1	9894 (20740)	429 (583)	1044 (1064)	529 (375)	1234 (1212)	10411 (6027)	
B	0	1401 (958)	1196 (918)	775 (1885)	734 (878)	549 (331)	4731 (3396)	
	.5	6333 (22140)	1925 (1369)	744 (346)	835 (483)	944 (750)	7020 (3989)	
	1	27218 (153493)	2798 (1609)	1140 (680)	1205 (648)	1123 (699)	7293 (4382)	
C	0	1474 (1820)	217 (286)	361 (210)	177 (217)	478 (323)	5248 (4937)	
	.5	1558 (1767)	200 (502)	438 (230)	204 (224)	500 (323)	4414 (2897)	
	1	5024 (8620)	179 (212)	516 (236)	316 (256)	773 (735)	4057 (3336)	

**Table 2**

Structure of the data

<b>Id</b>	<b>Start</b>	<b>Stop</b>	<b>Child</b>	<b>Job</b>	<b>Relstatus</b>	<b>Religion</b>	<b>Siblings</b>	<b>...</b>	<b>Federal state</b>
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
:	:	:	:	:	:	:	:	...	: