Web-based Supplementary Materials for Selection of Effects in Cox Frailty Models by Regularization Methods

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

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

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

Web Appendix A: Computational Details of `PenCoxFrail`

A.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 $\boldsymbol{s}^{pen}_{\boldsymbol{\beta}}$ and $\boldsymbol{s}^{pen}_{\boldsymbol{\alpha}}$ have dimension p and $(r+1)M$, respectively, while the vectors $\boldsymbol{s}^{pen}_{i}$ 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

*Corresponding author. Tel.: +49 89 2180 3044; fax:+49 89 2180 5308.
E-mail addresses: andreas.groll@stat.uni-muenchen.de, hastie@stanford.edu, gerhard.tutz@stat.uni-muenchen.de

and Tutz (2016). 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 consists of zeros only.

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

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

has single components

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

Note that the information matrix from equation (A.1) is subject to certain limitations with regard to the number of observations and of time-varying effects that can be considered. Situations of very high dimensions can lead to numerical instability. However, in those scenarios that we have investigated, in particular in the application from Section (6) with more than 20,000 lines and 16 covariates, each endued with a potentially time-varying effect, the method still works well.

A.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 (A.1):

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

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

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

A.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.

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`).

Web Appendix B: Simulation Studies

In the following, we present the details of the simulation study design of the simulation study from Section 5. The underlying models are random intercept models with balanced design

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

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

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

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 a Gamma density 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, which are included into the linear predictors to check the performance with respect to variable selection. All covariates $z_{ijk}, k = 1,\dots,16$, 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.

In order to compare the different approaches' performances, we consider the following mean squared errors for the baseline hazard, the smooth coefficient effects and σ_b, averaging across 50 data sets:

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

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 $v_t = (\Lambda_0(T) - \Lambda_0(t))/\Lambda_0(T)$.

B.1 Additional Results for Simulation Study I

Scenario	σ_b	Ridge	Linear	Select	PenCoxFrail	gam	coxph
	0	163 (182)	33 (40)	26 (47)	23 (39)	72 (48)	1089 (1467)
A	0.5	494 (1652)	51 (91)	47 (56)	35 (38)	108 (109)	614 (390)
	1	391 (452)	98 (141)	74 (93)	69 (93)	172 (378)	693 (676)
	0	50 (78)	59 (59)	14 (18)	20 (24)	43 (21)	758 (1131)
B	0.5	90 (92)	76 (79)	27 (25)	29 (25)	93 (185)	649 (1131)
	1	180 (403)	106 (121)	45 (56)	50 (70)	112 (134)	422 (473)
	0	118 (144)	34 (48)	15 (28)	20 (43)	70 (106)	583 (487)
C	0.5	132 (130)	35 (37)	23 (29)	23 (23)	74 (36)	535 (544)
	1	220 (321)	62 (88)	45 (58)	46 (54)	100 (104)	352 (423)

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

Scenario	σ_b	Ridge	Linear	Select	PenCoxFrail	gam	coxph
	0	.032 (.030)	.021 (.022)	.009 (.017)	.011 (.018)	.006 (.017)	.002 (.003)
A	0.5	.010 (.011)	.007 (.013)	.011 (.026)	.010 (.026)	.008 (.017)	.049 (.042)
	1	.012 (.018)	.012 (.019)	.016 (.024)	.014 (.023)	.012 (.020)	.060 (.085)
	0	.040 (.041)	.046 (.043)	.019 (.028)	.020 (.029)	.015 (.028)	.003 (.008)
B	0.5	.008 (.019)	.008 (.018)	.006 (.010)	.006 (.010)	.007 (.015)	.060 (.038)
	1	.011 (.019)	.009 (.017)	.012 (.014)	.012 (.014)	.012 (.023)	.062 (.220)
	0	.037 (.030)	.029 (.025)	.013 (.021)	.017 (.024)	.009 (.016)	.002 (.003)
C	0.5	.007 (.009)	.007 (.009)	.009 (.013)	.007 (.010)	.009 (.012)	.064 (.042)
	1	.012 (.013)	.013 (.016)	.017 (.021)	.016 (.020)	.016 (.017)	.047 (.043)

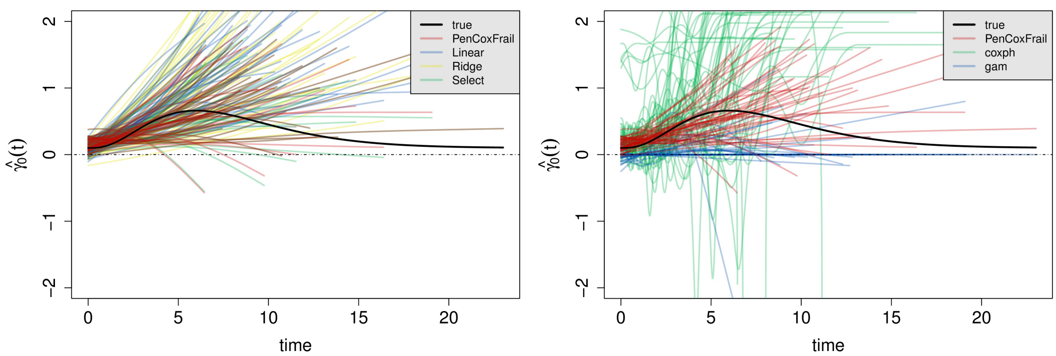
Web Table 2: *Results for* mse_{σ_b} *(standard errors in brackets).*

Scenario	σ_b	Ridge	Linear	Select	PenCoxFrail	gam	coxph
	0	1152 (346)	1770 (542)	3656 (1173)	13746 (4799)	1491 (316)	12 (1)
A	0.5	909 (289)	1438 (675)	3110 (1243)	12469 (9409)	1781 (731)	10 (1)
	1	1072 (409)	1634 (1270)	2869 (1665)	11166 (7530)	2256 (938)	15 (3)
	0	541 (125)	976 (290)	2282 (420)	6736 (2202)	1065 (344)	8 (1)
B	0.5	714 (170)	851 (234)	1746 (377)	5454 (1997)	1271 (534)	7 (1)
	1	767 (234)	1029 (466)	1786 (728)	5638 (2050)	1510 (523)	10 (2)
	0	371 (84)	619 (202)	1372 (331)	4398 (1542)	622 (207)	6 (1)
C	0.5	272 (47)	430 (120)	1009 (394)	3502 (1462)	680 (298)	5 (1)
	1	304 (93)	430 (120)	818 (342)	1929 (2468)	1044 (524)	4 (1)

Web Table 3: *Results for average computational times (standard errors in brackets).*

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
	0.0	0.14	0.24	0.00	0.14	0.00	0.00
A	0.5	0.00	0.00	0.00	0.02	0.00	0.00
	1.0	0.00	0.00	0.00	0.00	0.00	0.00
	0.0	0.10	0.12	0.06	0.08	0.00	0.00
B	0.5	0.00	0.00	0.00	0.00	0.00	0.00
	1.0	0.00	0.00	0.00	0.00	0.00	0.00
	0.0	0.34	0.28	0.06	0.10	0.00	0.00
C	0.5	0.00	0.00	0.00	0.02	0.00	0.00
	1.0	0.00	0.00	0.00	0.00	0.00	0.00

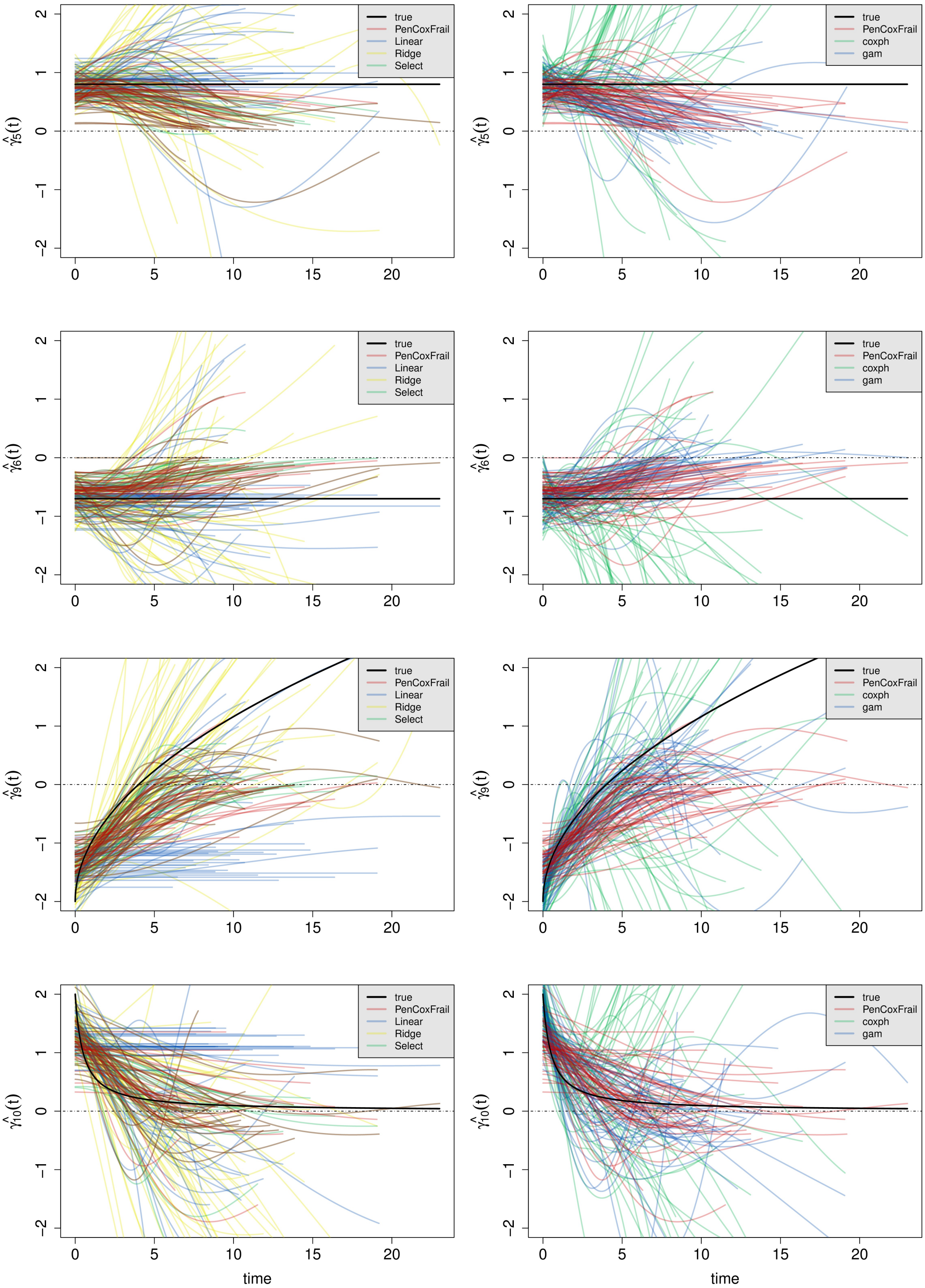
Web Table 4: *Proportion of non-convergent simulation runs.*



Web Figure 1: *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*

Scenario	σ_b		`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	0	null	0	0	0.44	0.30	0.54	0
	0	constant	0	0.92	0.01	0.77	0.01	0
	0	smooth	1	0.12	1	0.24	0.99	1
	0	exact	0	0	0	0	0	0
	0.5	null	0	0	0.55	0.46	0.53	0
	0.5	constant	0	0.96	0.04	0.77	0.02	0
	0.5	smooth	0.99	0.02	1	0.24	1	1
	0.5	exact	0	0	0	0	0	0
	1	null	0	0.01	0.48	0.44	0.55	0
	1	constant	0	0.96	0.06	0.66	0.04	0
	1	smooth	0.98	0.05	0.97	0.32	0.98	1
	1	exact	0	0	0	0	0	0
B	0	null	0	0	0.34	0.28	0.54	0
	0	constant	0	0.65	0.01	0.17	0	0
	0	smooth	1	0.55	1	0.92	1	1
	0	exact	0	0	0	0	0	0
	0.5	null	0	0.02	0.46	0.46	0.49	0
	0.5	constant	0	0.60	0.01	0.08	0	0
	0.5	smooth	1	0.50	0.98	0.95	1	1
	0.5	exact	0	0	0	0	0	0
	1	null	0	0.02	0.50	0.53	0.58	0
	1	constant	0	0.73	0.08	0.12	0.01	0
	1	smooth	1	0.38	0.98	0.94	1	1
	1	exact	0	0	0	0	0	0
C	0	null	0	0.02	0.36	0.38	0.42	0
	0	constant	0	0.90	0	0.77	0	0
	0	smooth	-	-	-	-	-	-
	0	exact	0	0.02	0	0.26	0	0
	0.5	null	0	0	0.44	0.34	0.46	0
	0.5	constant	0	0.95	0.01	0.78	0	0
	0.5	smooth	-	-	-	-	-	-
	0.5	exact	0	0	0	0.16	0	0
	1	null	0	0	0.44	0.36	0.50	0
	1	constant	0	0.92	0.02	0.63	0	0
	1	smooth	-	-	-	-	-	-
	1	exact	0	0	0	0.14	0	0

Web Table 5: *Proportions of correctly identified null (null), constant (constant) and time-varying effects (smooth) as well as proportions of correctly identified exact true model structure (exact).*

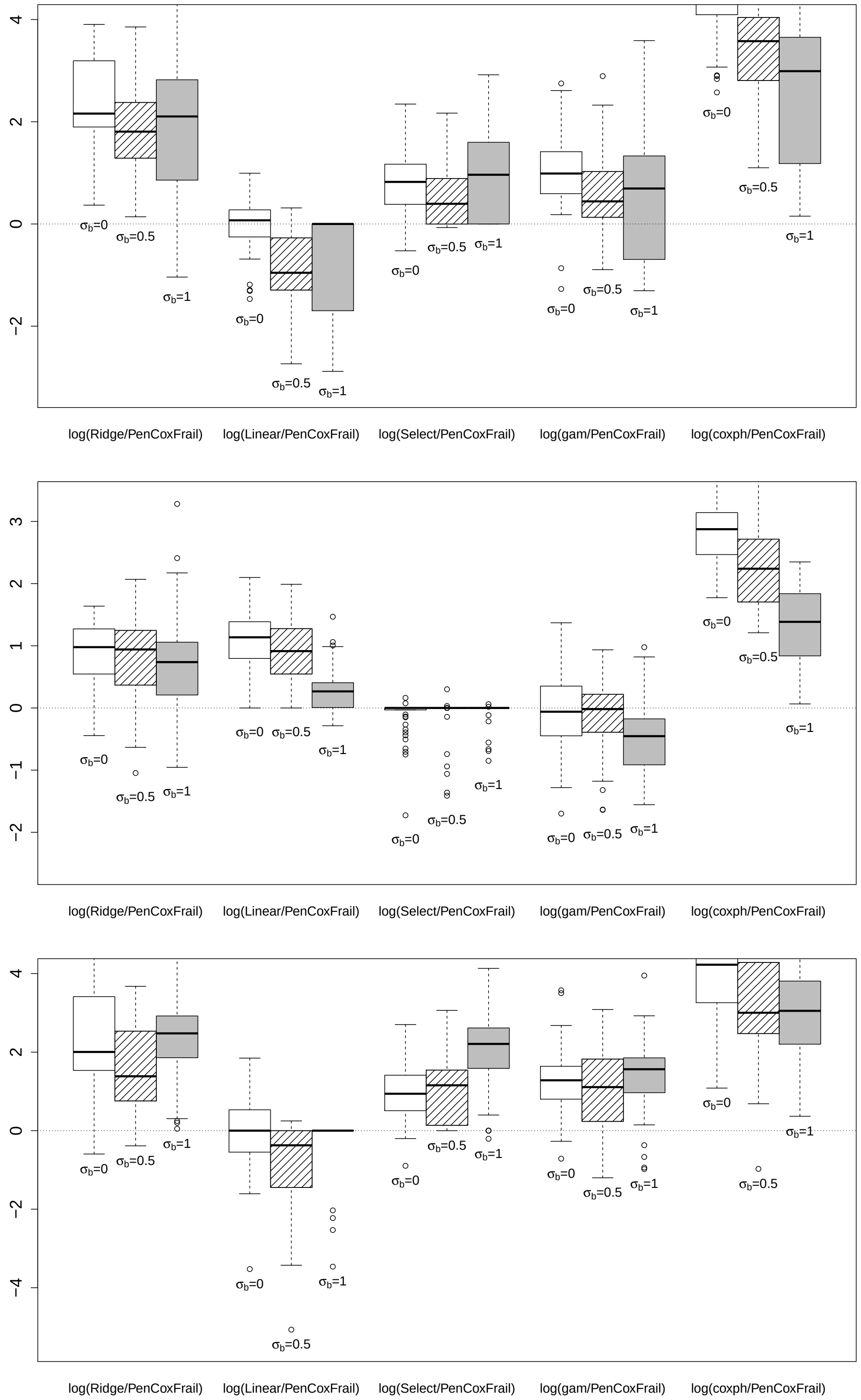


Web Figure 2: *Estimated (partly time-varying) effects* $\widehat{\gamma_5}(t), \widehat{\gamma_6}(t), \widehat{\gamma_9}(t), \widehat{\gamma_{10}}(t)$*, exemplarily for Scenario B and* $\sigma_b = 1$*; left:* `Ridge` *(yellow),* `Linear` *(blue),* `Select` *(green) and* `PenCoxFrail` *(red); right:* `gam` *(blue),* `coxph` *(green) and* `PenCoxFrail` *(red); true effect in black*

B.2 Simulation Study II

Simulation Study II ($n = 500, N_i = 1$)
Similar to Simulation Study I, 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 Web Figure 3, 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 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.



Web Figure 3: *Simulation Study II: boxplots of* $\log(mse_\gamma(\cdot)/mse_\gamma(\texttt{PenCoxFrail}))$ *for Scenario A (top), B (middle) and C (bottom)*

B.3 Simulation Study III

Simulation Study III ($n = 100, N_i = 5$)
In order to investigate the methods' robustness with regard to violations of the normal distribution assumption of the random effects, we consider a third simulation scenario. We use exactly the same simulation setting as in Simulation Study I, with the only difference that now the random effects are specified by $b_i \sim \Gamma(\zeta, \theta)$ with $\zeta = \theta = 1$. With respect to the quantities mse_0 and mse_γ very similar result are obtained as in Simulation Study I, compare Web Table 6 and 7. However, while again most methods also yield good results in terms of mse_{σ_b}, even though the normal distribution assumption of the random effects is violated, the `gam` method yields rather defective results, see Web Table 8.

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	1	226 (349)	116 (77)	170 (114)	147 (100)	167 (297)	841 (2235)
B	1	99 (97)	68 (45)	69 (40)	66 (36)	91 (55)	290 (403)
C	1	150 (137)	131 (84)	177 (115)	159 (107)	99 (75)	246 (193)

Web Table 6: *Simulation Study III: Results for* mse_0 *(standard errors in brackets).*

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	1	10613 (13337)	487 (630)	909 (400)	667 (554)	1564 (992)	9194 (4931)
B	1	6686 (9026)	3346 (4066)	1337 (598)	1440 (708)	1263 (606)	5690 (3944)
C	1	3555 (5924)	404 (1271)	609 (356)	373 (400)	836 (505)	3524 (3242)

Web Table 7: *Simulation Study III: Results for* mse_γ *(standard errors in brackets).*

Scenario	σ_b	`Ridge`	`Linear`	`Select`	`PenCoxFrail`	`gam`	`coxph`
A	1	.017 (.025)	.013 (.021)	.011 (.014)	.011 (.016)	580 (169)	.101 (.144)
B	1	.011 (.017)	.010 (.018)	.010 (.018)	.010 (.018)	486 (130)	.054 (.076)
C	1	.012 (.016)	.011 (.015)	.011 (.016)	.011 (.015)	564 (153)	.066 (.098)

Web Table 8: *Simulation Study III: Results for* mse_{σ_b} *(standard errors in brackets).*

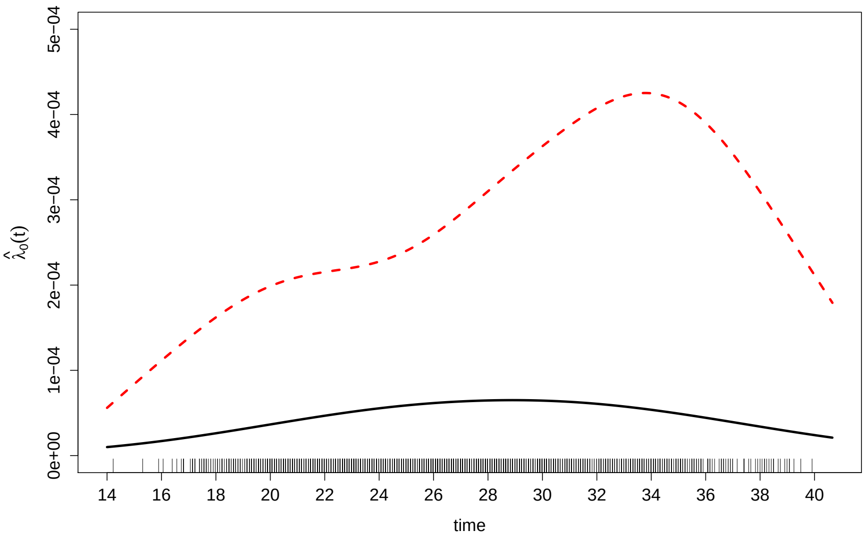
Web Appendix C: Application

	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

Web Table 9: *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	

Web Table 10: *Distribution of the time-varying covariates in the sample*



Web Figure 4: *pairfam data: estimated baseline hazard vs. time (women's age in years) at the chosen tuning parameter* $\xi_{48} = 6.09$*; for comparison, the estimated baseline hazard of a simple Cox model with time-constant effects is shown (red dashed line)*

References

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

Groll, A. (2016). *PenCoxFrail: Regularization in Cox Frailty Models*. R package version 1.0.1.

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

Oelker, M.-R. and G. Tutz (2016). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification, published online (DOI 10.1007/s11634-015-0205-y)*.

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