Automated effects selection via regularization in Cox frailty models

Andreas Groll* Trevor Hastie Thomas Kneib Gerhard Tutz

*Department of Statistics,
TU Dortmund University

Research Seminar Summer Term 2024
WU Vienna, Institute for Statistics and Mathematics
June 14^{th} 2024

technische universität dortmund

Motivation: PAIRFAM study

Data basis: Germany's current panel analysis of intimate relationships and family dynamics (**PAIRFAM**), release 4.0 (Nauck et al., 2013; Huinink et al., 2011).

Data has to be suitably prepared and structured for time-to-event data analysis:

Motivation: PAIRFAM study

Data basis: Germany's current panel analysis of intimate relationships and family dynamics (**PAIRFAM**), release 4.0 (Nauck et al., 2013; Huinink et al., 2011).

Data has to be suitably prepared and structured for time-to-event data analysis:

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Motivation: PAIRFAM study

Data basis: Germany's current panel analysis of intimate relationships and family dynamics (**PAIRFAM**), release 4.0 (Nauck et al., 2013; Huinink et al., 2011).

Data has to be suitably prepared and structured for time-to-event data analysis:

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Time-varying covariates $\Longrightarrow$ the 2,501 observations of the regarded women have to be split when time-varying covariates change $\Longrightarrow$ new data set: 20,550 lines

Outline

1. The Cox frailty model with time-varying effects

Outline

1. The Cox frailty model with time-varying effects
2. Penalization in Cox frailty models

Outline

1. The Cox frailty model with time-varying effects
2. Penalization in Cox frailty models
3. An application on the PAIRFAM data

Outline

1. The Cox frailty model with time-varying effects
2. Penalization in Cox frailty models
3. An application on the PAIRFAM data
4. Boosting for Cox frailty models

Outline

Introduction: The Cox Model

Cox model with semi-parametric hazard:

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

- $\lambda(t|\boldsymbol{x_i})$: hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x_i} = (x_{i1}, \dots, x_{ip})^T$
- $\lambda_0(t)$: shared baseline hazard
- β: fixed effects vector

Introduction: The Cox Model

Cox model with semi-parametric hazard:

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

- $\lambda(t|\boldsymbol{x_i})$: hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$
- $\lambda_0(t)$: shared baseline hazard
- β: fixed effects vector
- $\lambda(t|\boldsymbol{x}_i) := \lim\limits_{\Delta t \to 0} P(t \leq T < t + \Delta t | T \geq t, \boldsymbol{x}_i)/\Delta t,$

Introduction: The Cox Model

Cox model with semi-parametric hazard:

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

- $\lambda(t|\boldsymbol{x_i})$: hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x_i} = (x_{i1}, \dots, x_{ip})^T$
- $\lambda_0(t)$: shared baseline hazard
- β: fixed effects vector
- $\lambda(t|\boldsymbol{x}_i) := \lim_{\Delta t \to 0} P(t \leq T < t + \Delta t | T \geq t, \boldsymbol{x}_i)/\Delta t,$
- Inference: (usually) maximization of the corresponding partial likelihood

Introduction: The Cox Model

Cox model with semi-parametric hazard:

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

- $\lambda(t|\boldsymbol{x_i})$: hazard for observation i at time t, conditionally on the covariates $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^T$
- $\lambda_0(t)$: shared baseline hazard
- β: fixed effects vector
- $\lambda(t|\boldsymbol{x}_i) := \lim\limits_{\Delta t \to 0} P(t \leq T < t + \Delta t | T \geq t, \boldsymbol{x}_i)/\Delta t,$
- Inference: (usually) maximization of the corresponding partial likelihood
- $p > n$: LASSO (Tibshirani, 1997) extends the likelihood by the penalty term

$$\xi\, J(\beta) = \xi \sum_{j=1}^{p} |\beta_j|$$

Introduction: The Cox Frailty Model

Dependencies within clusters of observations or heterogeneity between clusters:

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

with frailties $b_i, i = 1, \ldots, n, j = 1, \ldots, N_i$

Introduction: The Cox Frailty Model

Dependencies within clusters of observations or heterogeneity between clusters:

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

with frailties $b_i, i = 1, \ldots, n, j = 1, \ldots, N_i$

- for mathematical convenience: frequently assumed $b_i \sim \Gamma(\cdot)$

Introduction: The Cox Frailty Model

Dependencies within clusters of observations or heterogeneity between clusters:

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

with frailties $b_i, i = 1, \ldots, n, j = 1, \ldots, N_i$

- for mathematical convenience: frequently assumed $b_i \sim \Gamma(\cdot)$
- R-packages: `frailtypack` (Rondeau et al., 2012), `survival` (Therneau, 2013), `frailtyHL` (Do Ha et al., 2012)

Introduction: The Cox Frailty Model

Dependencies within clusters of observations or heterogeneity between clusters:

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

with frailties $b_i, i = 1, \ldots, n, j = 1, \ldots, N_i$

- for mathematical convenience: frequently assumed $b_i \sim \Gamma(\cdot)$
- R-packages: `frailtypack` (Rondeau et al., 2012), `survival` (Therneau, 2013), `frailtyHL` (Do Ha et al., 2012)

With log-normal frailties

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

- $\boldsymbol{u_{ij}} = (u_{ij1}, \ldots, u_{ijq})^T$ covariate vector associated with random effects
- $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta}))$

Cox Frailty Model with Time-Varying Coefficients

Incorporate time-varying effects $\gamma_k(t)$:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^T \beta + \sum_{k=1}^{r} z_{ijk} \gamma_k(t) + \boldsymbol{u}_{ij}^T \boldsymbol{b}_i, \right)$$

with covariates $z_{ij1}, \ldots, z_{ijr}$ being associated with time-varying effects.

Cox Frailty Model with Time-Varying Coefficients

Incorporate time-varying effects $\gamma_k(t)$:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp\left(\boldsymbol{x}_{ij}^T \beta + \sum_{k=1}^{r} z_{ijk}\gamma_k(t) + \boldsymbol{u}_{ij}^T \boldsymbol{b}_i,\right)$$

with covariates $z_{ij1}, \ldots, z_{ijr}$ being associated with time-varying effects.

Estimation: expand time-varying effects $\gamma_k(t)$ in B-splines:

$$\gamma_k(t) = \sum_{m=1}^{M} \alpha_{k,m} B_m(t; d)$$

- $\alpha_{k,m}, m = 1, \ldots, M$: unknown spline coefficients
- $B_m(t; d)$: m-th B-spline basis function of degree d (see e.g. Eilers & Marx, 1996; Wood, 2017)

Cox Frailty Model with Time-Varying Coefficients

With $\gamma_0(t) := \log(\lambda_0(t))$ and $z_{ij0} = 1 \; \forall i, j$:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \exp\left(\overbrace{\boldsymbol{x}_{ij}^T \beta + \sum_{k=0}^{r} z_{ijk} \left(\sum_{m=1}^{M} \alpha_{k,m} B_m(t; d)\right) + \boldsymbol{u}_{ij}^T \boldsymbol{b}_i}^{\eta_{ij}(t)}\right), \qquad (1)$$

Now, $\boldsymbol{z_{ij}} = (1, z_{ij1}, \ldots, z_{ijr})^T$ is associated with both baseline hazard and time-varying effects.

Cox Frailty Model with Time-Varying Coefficients

With $\gamma_0(t) := \log(\lambda_0(t))$ and $z_{ij0} = 1 \ \forall i, j$:

$$\lambda(t|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_{ij}, \boldsymbol{b}_i) = \exp\left(\overbrace{\boldsymbol{x}_{ij}^T\beta + \sum_{k=0}^{r} z_{ijk}\left(\sum_{m=1}^{M}\alpha_{k,m}B_m(t;d)\right) + \boldsymbol{u}_{ij}^T\boldsymbol{b}_i}^{\eta_{ij}(t)}\right), \qquad (1)$$

Now, $\boldsymbol{z_{ij}} = (1, z_{ij1}, \ldots, z_{ijr})^T$ is associated with both baseline hazard and time-varying effects.

Estimation of parameters in (1) can be based on **Cox's full log-likelihood**:

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

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

Cox Frailty Model with Time-Varying Coefficients

A possible strategy to maximize the full log-likelihood (2) is based on PQL.

With $\boldsymbol{\delta}^T := (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{b}^T)$, the corresponding **marginal** log-likelihood yields

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

with random effects density $p(\boldsymbol{b}_i|\boldsymbol{\theta})$.

Cox Frailty Model with Time-Varying Coefficients

A possible strategy to maximize the full log-likelihood (2) is based on PQL.

With $\boldsymbol{\delta}^T := (\beta^T, \boldsymbol{\alpha}^T, \boldsymbol{b}^T)$, the corresponding **marginal** log-likelihood yields

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

with random effects density $p(\boldsymbol{b}_i|\boldsymbol{\theta})$.

Laplace approximation along the lines of Breslow & Clayton (1993) yields

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

Regularization via penalization

Major questions of model selection:

- which covariates should be included in the model?

Major questions of model selection:

- which covariates should be included in the model?
- which of those covariates included have a time-varying effect?

Regularization via penalization

Major questions of model selection:

- which covariates should be included in the model?
- which of those covariates included have a time-varying effect?

Two strategies:

Regularization via penalization

Major questions of model selection:

- which covariates should be included in the model?
- which of those covariates included have a time-varying effect?

Two strategies:

$\Longrightarrow$ Penalization

Regularization via penalization

Major questions of model selection:

- which covariates should be included in the model?
- which of those covariates included have a time-varying effect?

Two strategies:

$\Longrightarrow$ Penalization

$\Longrightarrow$ Boosting

Outline

Penalization (Groll, Hastie and Tutz, 2017)

$\Longrightarrow$ incorporate the following penalty into the Cox frailty log-likelihood:

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi \left(\zeta \sum_{k=1}^{r} \psi_k w_{\Delta,k} ||(\vartheta_{k,2}, \ldots, \vartheta_{k,M})||_2 + (1-\zeta) \sum_{k=1}^{r} \phi_k w_k ||\boldsymbol{\alpha}_k||_2 \right),$$

where $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$.

Penalization (Groll, Hastie and Tutz, 2017)

$\Longrightarrow$ incorporate the following penalty into the Cox frailty log-likelihood:

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi \left(\zeta \sum_{k=1}^{r} \psi_k w_{\Delta,k} ||(\vartheta_{k,2}, \ldots, \vartheta_{k,M})||_2 + (1-\zeta) \sum_{k=1}^{r} \phi_k w_k ||\boldsymbol{\alpha}_k||_2 \right),$$

where $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$.

The weights $\psi_k := \sqrt{M-1}$ and $\phi_k := \sqrt{M}$ assign different amounts of penalization to different parameter groups, relative to the respective group size.

Penalization (Groll, Hastie and Tutz, 2017)

$\Longrightarrow$ incorporate the following penalty into the Cox frailty log-likelihood:

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi \left(\zeta \sum_{k=1}^{r} \psi_k w_{\Delta,k} ||(\vartheta_{k,2}, \ldots, \vartheta_{k,M})||_2 + (1-\zeta) \sum_{k=1}^{r} \phi_k w_k ||\boldsymbol{\alpha}_k||_2 \right),$$

where $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$.

The weights $\psi_k := \sqrt{M-1}$ and $\phi_k := \sqrt{M}$ assign different amounts of penalization to different parameter groups, relative to the respective group size.

The adaptive weights $w_{\Delta,k} := 1/||\hat{\boldsymbol{\vartheta}}_k^{(ML)}||_2$ and $w_k := 1/||\hat{\boldsymbol{\alpha}}_k^{(ML)}||_2$ are based on the (slightly ridge-penalized) ML-estimator.

Penalization (Groll, Hastie and Tutz, 2017)

$\Longrightarrow$ incorporate the following penalty into the Cox frailty log-likelihood:

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi \left(\zeta \sum_{k=1}^{r} \psi_k w_{\Delta,k} \|(\vartheta_{k,2}, \dots, \vartheta_{k,M})\|_2 + (1-\zeta) \sum_{k=1}^{r} \phi_k w_k \|\boldsymbol{\alpha}_k\|_2 \right),$$

where $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$.

The weights $\psi_k := \sqrt{M-1}$ and $\phi_k := \sqrt{M}$ assign different amounts of penalization to different parameter groups, relative to the respective group size.

The adaptive weights $w_{\Delta,k} := 1/\|\hat{\boldsymbol{\vartheta}}_k^{(ML)}\|_2$ and $w_k := 1/\|\hat{\boldsymbol{\alpha}}_k^{(ML)}\|_2$ are based on the (slightly ridge-penalized) ML-estimator.

Tuning parameters ξ and ζ are chosen by appropriate technique, e.g. K-fold CV.

Penalization (Groll, Hastie and Tutz, 2017)

$\Longrightarrow$ incorporate the following penalty into the Cox frailty log-likelihood:

$$\xi \cdot J_\zeta(\boldsymbol{\alpha}) = \xi\left(\zeta \sum_{k=1}^{r} \psi_k w_{\Delta,k} \|(\vartheta_{k,2},\ldots,\vartheta_{k,M})\|_2 + (1-\zeta)\sum_{k=1}^{r} \phi_k w_k \|\boldsymbol{\alpha}_k\|_2\right),$$

where $\xi \geq 0$ and $\zeta \in (0,1)$ are tuning parameters and $\vartheta_{k,l} = \alpha_{k,l} - \alpha_{k,l-1}$.

The weights $\psi_k := \sqrt{M-1}$ and $\phi_k := \sqrt{M}$ assign different amounts of penalization to different parameter groups, relative to the respective group size.

The adaptive weights $w_{\Delta,k} := 1/\|\hat{\boldsymbol{\vartheta}}_k^{(ML)}\|_2$ and $w_k := 1/\|\hat{\boldsymbol{\alpha}}_k^{(ML)}\|_2$ are based on the (slightly ridge-penalized) ML-estimator.

Tuning parameters ξ and ζ are chosen by appropriate technique, e.g. K-fold CV.

Penalization of baseline hazard:

$$\xi_0 \cdot J_0(\boldsymbol{\alpha}_0) = \xi_0\left(\sum_{l=2}^{M}(\alpha_{0,l} - \alpha_{0,l-1})^2\right).$$

Estimation

- maximization of the penalized log-likelihood:

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

Estimation

- maximization of the penalized log-likelihood:

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

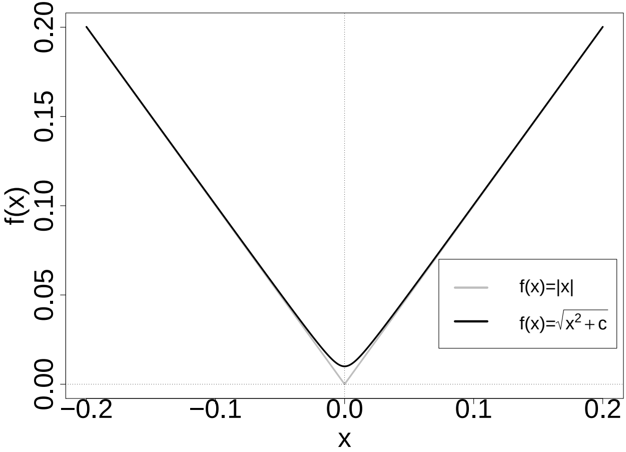
- local quadratic approximations of the penalty terms (Oelker & Tutz, 2017).

Estimation

- maximization of the penalized log-likelihood:

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

- local quadratic approximations of the penalty terms (Oelker & Tutz, 2017).



Estimation

- maximization of the penalized log-likelihood:

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

- local quadratic approximations of the penalty terms (Oelker & Tutz, 2017).
- estimation based on conventional Newton-Raphson

Estimation

- maximization of the penalized log-likelihood:

$$l^{pen}(\boldsymbol{\delta},\boldsymbol{\theta}) = l^{app}(\boldsymbol{\delta},\boldsymbol{\theta}) - \xi_0 \cdot J_0(\boldsymbol{\alpha}_{\mathbf{0}}) - \xi \cdot J_\zeta(\boldsymbol{\alpha}).$$

- local quadratic approximations of the penalty terms (Oelker & Tutz, 2017).
- estimation based on conventional Newton-Raphson

Algorithm `PenCoxFrail`

---

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

Estimation

- maximization of the penalized log-likelihood:

$$l^{pen}(\boldsymbol{\delta},\boldsymbol{\theta}) = l^{app}(\boldsymbol{\delta},\boldsymbol{\theta}) - \xi_0 \cdot J_0(\boldsymbol{\alpha}_{\mathbf{0}}) - \xi \cdot J_\zeta(\boldsymbol{\alpha}).$$

- local quadratic approximations of the penalty terms (Oelker & Tutz, 2017).
- estimation based on conventional Newton-Raphson

Algorithm `PenCoxFrail`

1. *Initialization* Choose starting values $\hat{\beta}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$
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 $\mathbf{s}^{pen}(\boldsymbol{\delta}) = \partial l^{pen}/\partial\boldsymbol{\delta}$ and information matrix $\mathbf{F}^{pen}(\boldsymbol{\delta})$ the general form of a single Newton-Raphson step is:

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

   As the fit is within an iterative procedure it is sufficient to use a single step.

Estimation

- maximization of the penalized log-likelihood:

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

- local quadratic approximations of the penalty terms (Oelker & Tutz, 2017).
- estimation based on conventional Newton-Raphson

Algorithm `PenCoxFrail`

1. *Initialization* Choose starting values $\hat{\beta}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$
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 $\mathbf{s}^{pen}(\boldsymbol{\delta}) = \partial l^{pen}/\partial\boldsymbol{\delta}$ and information matrix $\mathbf{F}^{pen}(\boldsymbol{\delta})$ the general form of a single Newton-Raphson step is:

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

   As the fit is within an iterative procedure it is sufficient to use a single step.

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

Outline

Data basis: Germany's current panel analysis of intimate relationships and family dynamics (**PAIRFAM**), release 4.0 (Nauck et al., 2013; Huinink et al., 2011).

The data has to be suitably prepared and structured for the event data analysis:

Data basis: Germany's current panel analysis of intimate relationships and family dynamics (**PAIRFAM**), release 4.0 (Nauck et al., 2013; Huinink et al., 2011).

The data has to be suitably prepared and structured for the event data analysis:

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Application: PAIRFAM

Data

Data basis: Germany's current panel analysis of intimate relationships and family dynamics (**PAIRFAM**), release 4.0 (Nauck et al., 2013; Huinink et al., 2011).

The data has to be suitably prepared and structured for the event data analysis:

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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Time-varying covariates $\Longrightarrow$ the 2,501 observations of the regarded women have to be split when time-varying covariates change $\Longrightarrow$ new data set: 20,550 lines

Application: PAIRFAM

Distribution of time-constant (left) and time-varying (right) covariates in the sample

	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

	# 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	

- regional fertility differences $\Longrightarrow$ **random intercept** for the German federal state where the women are born.

- regional fertility differences $\Longrightarrow$ **random intercept** for the German federal state where the women are born.
- PenCoxFrail: $n > 20\,000 \Longrightarrow$ ad-hoc strategy to determine optimal ξ (Chouldechova & Hastie, 2015; Liu et al., 2007):
  - $\zeta = 0.5$

- regional fertility differences $\Longrightarrow$ **random intercept** for the German federal state where the women are born.
- PenCoxFrail: $n > 20\,000 \Longrightarrow$ ad-hoc strategy to determine optimal ξ (Chouldechova & Hastie, 2015; Liu et al., 2007):
  - $\zeta = 0.5$
  - include 10 additional simulated noise variables

Application: PAIRFAM

Model Fit

- regional fertility differences $\Longrightarrow$ **random intercept** for the German federal state where the women are born.
- PenCoxFrail: $n > 20\,000 \Longrightarrow$ ad-hoc strategy to determine optimal ξ (Chouldechova & Hastie, 2015; Liu et al., 2007):
  - $\zeta = 0.5$
  - include 10 additional simulated noise variables
  - stop right before the first of them enters the model

Implementation

Call in R using the package PenCoxFrail:

```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
                rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

Implementation

Call in R using the package PenCoxFrail:

```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
                rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

Implementation

Call in R using the package PenCoxFrail:

```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
                rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

Implementation

Call in R using the package PenCoxFrail:

```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
                rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

Implementation

Call in R using the package PenCoxFrail:

```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
                rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

Implementation

Call in R using the package PenCoxFrail:

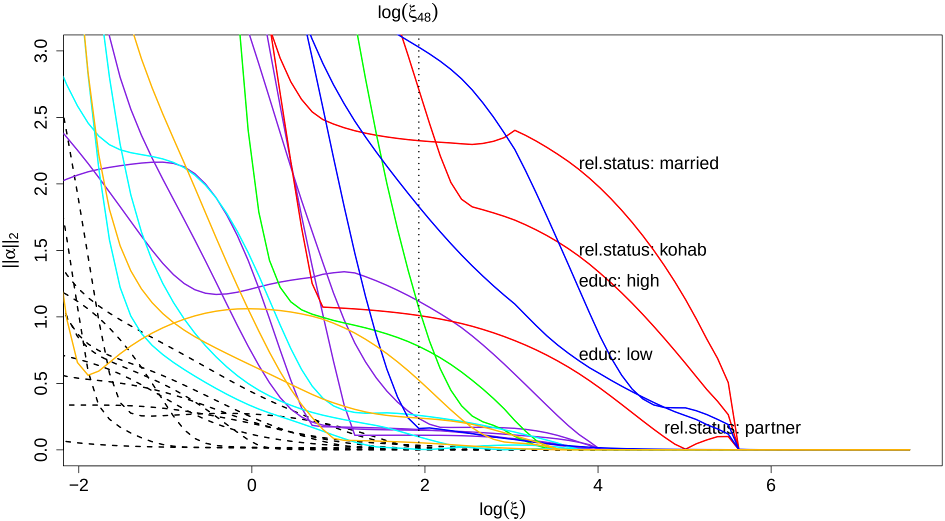
```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
                rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

Implementation

Call in R using the package PenCoxFrail:

```
>pencox.obj <- pencoxfrail(Surv(time,event) ~ 1, vary.coef = ~ relat.status + ...,
               rnd = list(fed.state = ~ 1), data = pairfam, xi = 100, control = list(...))
```

original 6 variables (colored solid lines) and simulated noise variables (black dashed lines); horizontal dotted line: chosen tuning parameter $\xi_{48} = 6.09$

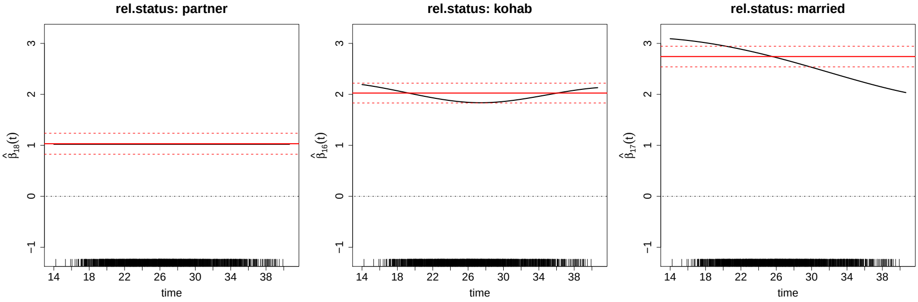


Application: PAIRFAM

Estimated Time-Varying Effects

Estimated effect of the categorical covariate "relation ship status" (black solid line) vs. time (women's age in years) at chosen tuning parameter $\xi_{48} = 6.09$.

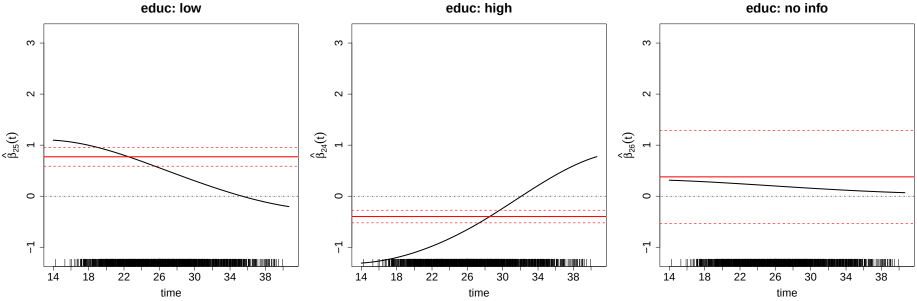
For comparison, time-constant effects of a conventional Cox model are shown (red solid line) together with 95% confidence interval.



Reference level: "single"

Estimated effect of the categorical covariate "education level" (black solid line) vs. 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 (red solid line) together with 95% confidence interval.



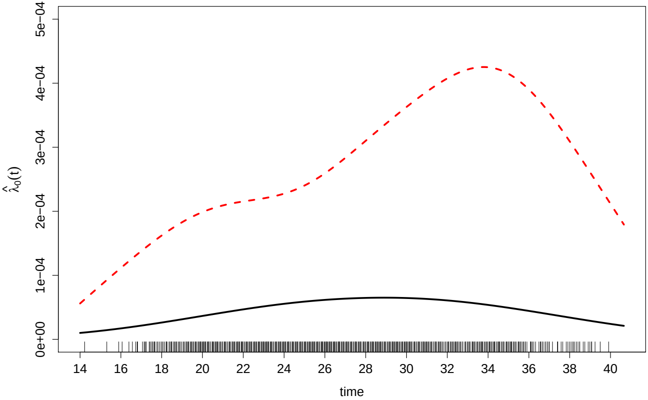
Reference level: "medium"

Application: PAIRFAM

Estimated Baseline Hazard

Estimated baseline hazard (black solid line) 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)



Heterogeneity between German federal states: $\hat{\sigma}_b = 0.179$ (0.179 for simple Cox)

Outline

Boosting

Basic idea:

Fahrmeir et al. (2004): re-parametrization of P-splines $\Longrightarrow$ split potentially time-varying effect $\gamma(t)$ of a covariate z into

$$\gamma(t)\cdot z = \underbrace{\alpha_0 \cdot z + \alpha_1 t \cdot z + \ldots \alpha_{d-1} t^{d-1} \cdot z}_{\text{unpenalized part}} \quad + \quad \underbrace{\gamma_{\text{centered}}(t)\cdot z}_{\text{smooth penalized part}} \; .$$

Boosting

Basic idea:

Fahrmeir et al. (2004): re-parametrization of P-splines $\Longrightarrow$ split potentially time-varying effect $\gamma(t)$ of a covariate z into

$$\gamma(t) \cdot z = \underbrace{\alpha_0 \cdot z + \alpha_1 t \cdot z + \ldots \alpha_{d-1} t^{d-1} \cdot z}_{\textbf{unpenalized part}} \quad + \quad \underbrace{\gamma_{\textbf{centered}}(t) \cdot z}_{\textbf{smooth penalized part}} \; .$$

(the vector of regression coefficients $\boldsymbol{\alpha}$ is decomposed into $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_{\textbf{unpen}}^T, \boldsymbol{\alpha}_{\textbf{pen}}^T)$ using spectral decomposition of the penalty matrix.)

Boosting

Basic idea:

Fahrmeir et al. (2004): re-parametrization of P-splines $\Longrightarrow$ split potentially time-varying effect $\gamma(t)$ of a covariate z into

$$\gamma(t) \cdot z = \underbrace{\alpha_0 \cdot z + \alpha_1 t \cdot z + \ldots \alpha_{d-1} t^{d-1} \cdot z}_{\textbf{unpenalized part}} \quad + \quad \underbrace{\gamma_{\textbf{centered}}(t) \cdot z}_{\textbf{smooth penalized part}} \ .$$

(the vector of regression coefficients $\boldsymbol{\alpha}$ is decomposed into $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_{\textbf{unpen}}^T, \boldsymbol{\alpha}_{\textbf{pen}}^T)$ using spectral decomposition of the penalty matrix.)

We use first order differences (with cubic B-splines):

$$\gamma(t) \cdot z = \alpha_0 \cdot z + \gamma_{\textbf{centered}}(t) \cdot z \,, \tag{3}$$

which simply decomposes the time-varying effect into a linear (time-constant) effect and a smooth time-varying part.

Effects selection:

We specify two base-learners for each (potentially) time-varying effect:

- a **linear** base learner, i.e. $\alpha_0 \cdot z$,
- a **smooth** deviation from linearity, i.e. $\gamma_{\text{centered}}(t) \cdot z$.

Boosting

Effects selection:

We specify two base-learners for each (potentially) time-varying effect:

- a **linear** base learner, i.e. $\alpha_0 \cdot z$,
- a **smooth** deviation from linearity, i.e. $\gamma_{\text{centered}}(t) \cdot z$.

$\Longrightarrow$ a covariate can be included with time-varying or time-constant effects, or can be excluded completely from the model!

Boosting

Effects selection:

We specify two base-learners for each (potentially) time-varying effect:

- a **linear** base learner, i.e. $\alpha_0 \cdot z$,
- a **smooth** deviation from linearity, i.e. $\gamma_{\textsf{centered}}(t) \cdot z$.

$\Longrightarrow$ a covariate can be included with time-varying or time-constant effects, or can be excluded completely from the model!

For fair comparison: force smooth base-learner $\gamma_{\textsf{centered}}(t) \cdot z$ to exactly one degree of freedom

Boosting

Effects selection:

We specify two base-learners for each (potentially) time-varying effect:

- a **linear** base learner, i.e. $\alpha_0 \cdot z$,
- a **smooth** deviation from linearity, i.e. $\gamma_{\text{centered}}(t) \cdot z$.

$\Longrightarrow$ a covariate can be included with time-varying or time-constant effects, or can be excluded completely from the model!

For fair comparison: force smooth base-learner $\gamma_{\text{centered}}(t) \cdot z$ to exactly one degree of freedom

DFs can be derived based on the penalized and unpenalized Fisher information:

$$\text{df} = \text{trace}\left(\mathbf{F} \cdot (\mathbf{F} + \xi \cdot \text{diag}(1, \ldots, 1))^{-1}\right),$$

see, e.g., Hofner et al. (2011).

Iterative component-wise boosting procedure

Algorithm `CoxFrailBoost`

---

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

Iterative component-wise boosting procedure

Algorithm `CoxFrailBoost`

---

1. *Initialization* Choose starting values $\hat{\beta}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$
2. *Iteration* For $l = 1, 2, \ldots, l_{stop}$:

   (a) *Computation of parameters*:

   (i) For $\tilde{\boldsymbol{\delta}} := (\hat{\beta}, \hat{\boldsymbol{\alpha}}_{(0)}, \hat{\mathbf{b}})$, calculate $\hat{\hat{\boldsymbol{\delta}}}^{(l)} = \hat{\hat{\boldsymbol{\delta}}}^{(l-1)} + (\tilde{\mathbf{F}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\, \tilde{\mathbf{s}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)})$;

Iterative component-wise boosting procedure

Algorithm `CoxFrailBoost`

1. *Initialization* Choose starting values $\hat{\beta}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$
2. *Iteration* For $l = 1, 2, \ldots, l_{stop}$:
   (a) *Computation of parameters*:
      (i) For $\tilde{\boldsymbol{\delta}} := (\hat{\beta}, \hat{\boldsymbol{\alpha}}_{(0)}, \hat{\mathbf{b}})$, calculate $\hat{\tilde{\boldsymbol{\delta}}}^{(l)} = \hat{\tilde{\boldsymbol{\delta}}}^{(l-1)} + (\tilde{\mathbf{F}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\, \tilde{\mathbf{s}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)})$;
      (ii) For $k \in \{1, \ldots, r\}$ derive score component $s_k^{lin}(\boldsymbol{\delta}) = \partial l^{app} / \partial \alpha_{1,k}$ and information matrix component $F_k^{lin}(\boldsymbol{\delta})$;

$$\Longrightarrow \hat{\alpha}_{1,k}^{(l)} = \hat{\alpha}_{1,k}^{(l-1)} + s_k^{lin}(\hat{\boldsymbol{\delta}}^{(l-1)}) / F_k^{lin}(\hat{\boldsymbol{\delta}}^{(l-1)})$$

Iterative component-wise boosting procedure

Algorithm `CoxFrailBoost`

1. *Initialization* Choose starting values $\hat{\beta}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$
2. *Iteration* For $l = 1, 2, \ldots, l_{stop}$:

   (a) *Computation of parameters*:

   (i) For $\tilde{\boldsymbol{\delta}} := (\hat{\beta}, \hat{\boldsymbol{\alpha}}_{(0)}, \hat{\mathbf{b}})$, calculate $\hat{\tilde{\boldsymbol{\delta}}}^{(l)} = \hat{\tilde{\boldsymbol{\delta}}}^{(l-1)} + (\tilde{\mathbf{F}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1} \, \tilde{\mathbf{s}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)})$;

   (ii) For $k \in \{1, \ldots, r\}$ derive score component $s_k^{lin}(\boldsymbol{\delta}) = \partial l^{app} / \partial \alpha_{1,k}$ and information matrix component $F_k^{lin}(\boldsymbol{\delta})$;

   $$\Longrightarrow \hat{\alpha}_{1,k}^{(l)} = \hat{\alpha}_{1,k}^{(l-1)} + s_k^{lin}(\hat{\boldsymbol{\delta}}^{(l-1)}) / F_k^{lin}(\hat{\boldsymbol{\delta}}^{(l-1)})$$

   (iii) For $k \in \{1, \ldots, r\}$ derive score function $\mathbf{s}_k^{smo}(\boldsymbol{\delta}) = \partial l^{pen} / \partial \boldsymbol{\alpha}_{[-1],k}$ and information matrix $\mathbf{F}_k^{smo}(\boldsymbol{\delta})$;

   $$\Longrightarrow \hat{\boldsymbol{\alpha}}_{[-1],k}^{(l)} = \hat{\boldsymbol{\alpha}}_{[-1],k}^{(l-1)} + (\mathbf{F}_k^{smo}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1} \mathbf{s}_k^{smo}(\hat{\boldsymbol{\delta}}^{(l-1)})$$

Iterative component-wise boosting procedure

Algorithm `CoxFrailBoost`

1. *Initialization* Choose starting values $\hat{\beta}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\mathbf{b}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}$
2. *Iteration* For $l = 1, 2, \dots, l_{stop}$:

   (a) *Computation of parameters*:

   (i) For $\tilde{\boldsymbol{\delta}} := (\hat{\beta}, \hat{\boldsymbol{\alpha}}_{(0)}, \hat{\mathbf{b}})$, calculate $\hat{\tilde{\boldsymbol{\delta}}}^{(l)} = \hat{\tilde{\boldsymbol{\delta}}}^{(l-1)} + (\tilde{\mathbf{F}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1}\, \tilde{\mathbf{s}}^{app}(\hat{\boldsymbol{\delta}}^{(l-1)})$;

   (ii) For $k \in \{1, \dots, r\}$ derive score component $s_k^{lin}(\boldsymbol{\delta}) = \partial l^{app} / \partial \alpha_{1,k}$ and information matrix component $F_k^{lin}(\boldsymbol{\delta})$;

   $$\Longrightarrow \hat{\alpha}_{1,k}^{(l)} = \hat{\alpha}_{1,k}^{(l-1)} + s_k^{lin}(\hat{\boldsymbol{\delta}}^{(l-1)}) / F_k^{lin}(\hat{\boldsymbol{\delta}}^{(l-1)})$$

   (iii) For $k \in \{1, \dots, r\}$ derive score function $\mathbf{s}_k^{smo}(\boldsymbol{\delta}) = \partial l^{pen} / \partial \boldsymbol{\alpha}_{[-1],k}$ and information matrix $\mathbf{F}_k^{smo}(\boldsymbol{\delta})$;

   $$\Longrightarrow \hat{\boldsymbol{\alpha}}_{[-1],k}^{(l)} = \hat{\boldsymbol{\alpha}}_{[-1],k}^{(l-1)} + (\mathbf{F}_k^{smo}(\hat{\boldsymbol{\delta}}^{(l-1)}))^{-1} \mathbf{s}_k^{smo}(\hat{\boldsymbol{\delta}}^{(l-1)})$$

   (b) *Selection step*:

   Select from (ii) and (iii) the component j that leads to the best improvement of the likelihood and denote it by $\hat{\alpha}_1^*$ or $\hat{\boldsymbol{\alpha}}_{[-1]}^*$, respectively.

Iterative component-wise boosting procedure

(c) ***Weak** update of best predictor*:

For $k \in \{1, \ldots, r\}$ and $0 < \nu \leq 1$ set

$$\hat{\alpha}_{1,k}^{(l)} = \begin{cases} \hat{\alpha}_{1,k}^{(l-1)} & \text{if } k \neq j, \\ \hat{\alpha}_{1,k}^{(l-1)} + \nu \cdot \hat{\alpha}_1^* & \text{if } k = j, \end{cases}$$

and

$$\hat{\boldsymbol{\alpha}}_{[-1],k}^{(l)} = \begin{cases} \hat{\boldsymbol{\alpha}}_{[-1],k}^{(l-1)} & \text{if } k \neq j, \\ \hat{\boldsymbol{\alpha}}_{[-1],k}^{(l-1)} + \nu \cdot \hat{\boldsymbol{\alpha}}_{[-1]}^* & \text{if } k = j. \end{cases}$$

Iterative component-wise boosting procedure

(c) ***Weak*** *update of best predictor*:

For $k \in \{1, \ldots, r\}$ and $0 < \nu \leq 1$ set

$$\hat{\alpha}^{(l)}_{\mathbf{1},k} = \begin{cases} \hat{\alpha}^{(l-1)}_{\mathbf{1},k} & \text{if } k \neq j, \\ \hat{\alpha}^{(l-1)}_{\mathbf{1},k} + \nu \cdot \hat{\alpha}^{*}_{\mathbf{1}} & \text{if } k = j, \end{cases}$$

and

$$\hat{\boldsymbol{\alpha}}^{(l)}_{[-1],k} = \begin{cases} \hat{\boldsymbol{\alpha}}^{(l-1)}_{[-1],k} & \text{if } k \neq j, \\ \hat{\boldsymbol{\alpha}}^{(l-1)}_{[-1],k} + \nu \cdot \hat{\boldsymbol{\alpha}}^{*}_{[-1]} & \text{if } k = j. \end{cases}$$

(d) *Computation of variance-covariance components*:
Estimates $\hat{\mathbf{Q}}^{(l)}$ are obtained as approximate EM-type estimates, yielding $\hat{\boldsymbol{\theta}}^{(l)}$.

Summary

Conclusions:

- 2 regularization approaches for Cox frailty models with time-varying coefficients and log-normal frailties: **penalization** and **boosting**

Summary

Conclusions:

- 2 regularization approaches for Cox frailty models with time-varying coefficients and log-normal frailties: **penalization** and **boosting**
- the methods yield flexible and sparse hazard rate models for modeling time-to-event data

Summary

Conclusions:

- 2 regularization approaches for Cox frailty models with time-varying coefficients and log-normal frailties: **penalization** and **boosting**
- the methods yield flexible and sparse hazard rate models for modeling time-to-event data
- (good performance in simulations)

Summary

Conclusions:

- 2 regularization approaches for Cox frailty models with time-varying coefficients and log-normal frailties: **penalization** and **boosting**
- the methods yield flexible and sparse hazard rate models for modeling time-to-event data
- (good performance in simulations)
- reasonable estimates in application (at least for the penalty approach)

Summary

Conclusions:

- 2 regularization approaches for Cox frailty models with time-varying coefficients and log-normal frailties: **penalization** and **boosting**
- the methods yield flexible and sparse hazard rate models for modeling time-to-event data
- (good performance in simulations)
- reasonable estimates in application (at least for the penalty approach)
- boosting looks even more promising and will be faster, because
  - component-wise parts of the algorithm can be parallelized
  - we brute-force the EDFs of each boosting update

    $\Rightarrow$ avoid K-fold CV and use AIC / BIC to determine optimal # of boosting steps

References & Software

Penalization:

- Groll, A., T. Hastie & G. Tutz (2017). Selection of Effects in Cox Frailty Models by Regularization Methods, *Biometrics*, **73(3)**, 846 – 856.

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

References & Software

Penalization:

Groll, A., T. Hastie & G. Tutz (2017). Selection of Effects in Cox Frailty Models by Regularization Methods, *Biometrics*, **73(3)**, 846 – 856.

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

Boosting:

Groll, A., T. Hastie, T. Kneib & G. Tutz (2018). Boosting Methods for Effects Selection in Cox Frailty Models, *Proceedings of the 33rd International Workshop on Statistical Modelling*, **(1)**, 122-127.

Groll, A. (2018). *CoxFrailBoost: Boosting in Cox Frailty Models.* R package version 0.0, (to appear soon).

The 1st package is available on CRAN (see `http://www.r-project.org`).

Further References

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

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

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

Eilers, P. H. C. & Marx, B. D. (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**, 731 – 761.

Hofner, B., T. Kneib, W. Hartl, and H. Küchenhoff (2011b). Building Cox-type structured hazard regression models with time-varying effects. *Statistical Modelling* **11(1)**, 3 – 24.

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

Further References II

- Liu, H., Wasserman, L., Lafferty, J. D., & Ravikumar, P. K. (2007). SpAM: Sparse additive models. In *NIPS*, 1201 – 1208.
- Nauck, B., J. Brüderl, J. Huinink, & S.Walper (2013). The german family panel (pairfam). *GESIS Data Archive, Cologne*. ZA5678 Data file Version 4.0.0.
- Oelker, M.-R. & Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, **11(1)**, 97 – 120.
- Rondeau, V., Mazroui, Y. & Gonzalez, J. R. (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.
- Therneau, T. M. (2013). *A package for survival analysis in S*. R package version 2.37-4.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 267–288.
- Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine*, **16**, 385 – 395.
- Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd edition). London: Chapman & Hall.

Determination of Optimal Tuning Parameters

- ξ_0, controlling the smoothness of the log-baseline hazard $\gamma_0(t) = \log(\lambda_0(t))$; in general, no complex selection procedure necessary
  $\Longrightarrow$ estimation procedure is already stabilized for a moderate choice of ξ_0.

Determination of Optimal Tuning Parameters

- ξ_0, controlling the smoothness of the log-baseline hazard $\gamma_0(t) = \log(\lambda_0(t))$; in general, no complex selection procedure necessary
$\Longrightarrow$ estimation procedure is already stabilized for a moderate choice of ξ_0.

ζ and ξ are determined via K-fold CV:

- ξ: controls overall amount of penalization, and hence, both smoothness and variable selection, it is of particular importance $\Longrightarrow$ use a fine grid
- ζ: controls apportionment between smoothness and shrinkage $\Longrightarrow$ rougher grid is sufficient.

Determination of Optimal Tuning Parameters

- ξ_0, controlling the smoothness of the log-baseline hazard $\gamma_0(t) = \log(\lambda_0(t))$; in general, no complex selection procedure necessary
  $\Longrightarrow$ estimation procedure is already stabilized for a moderate choice of ξ_0.

ζ and ξ are determined via K-fold CV:

- ξ: controls overall amount of penalization, and hence, both smoothness and variable selection, it is of particular importance $\Longrightarrow$ use a fine grid
- ζ: controls apportionment between smoothness and shrinkage $\Longrightarrow$ rougher grid is sufficient.
- **CV error measure**: evaluate log-likelihood (2) on the test data, i.e.

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

  where n^{test} denotes the number of clusters in the test data and N_i^{test} the corresponding cluster sizes.

Score function

Let $\mathbf{B}^T(t) := (B_1(t; d), \ldots, B_M(t; d))$ represent the vector-valued evaluations of the M basis functions in t and define $\mathbf{\Phi}^T(t) := (z_{ij0} \cdot \mathbf{B}^T(t), z_{ij1} \cdot \mathbf{B}^T(t), \ldots, z_{ijr} \cdot \mathbf{B}^T(t))$. Then, $\mathbf{s}^{pen}(\boldsymbol{\delta}) = \partial l^{pen}(\boldsymbol{\delta})/\partial \boldsymbol{\delta}$ has vector components

$$
\begin{aligned}
\mathbf{s}_{\boldsymbol{\beta}}^{pen}(\boldsymbol{\delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{N_i} \mathbf{x}_{ij}\left(d_{ij} - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right), \\
\mathbf{s}_{\boldsymbol{\alpha}}^{pen}(\boldsymbol{\delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{N_i} \left(d_{ij}\mathbf{\Phi}(t_{ij}) - \int_0^{t_{ij}} \exp(\eta_{ij}(s))\mathbf{\Phi}(s)ds\right) - \mathbf{A}_{\xi_{\mathbf{0}},\xi,\zeta}\,\boldsymbol{\alpha}, \\
\mathbf{s}_{i}^{pen}(\boldsymbol{\delta}) &= \sum_{j=1}^{N_i} \mathbf{u}_{ij}\left(d_{ij} - \int_0^{t_{ij}} \exp(\eta_{ij}(s))ds\right) - \mathbf{Q}^{-1}(\boldsymbol{\theta})\mathbf{b}_i, \quad i = 1, \ldots, n.
\end{aligned}
$$

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

Penalty matrix

The penalty matrix $\mathbf{A}_{\xi_0,\xi,\zeta}$ is block-diagonal: $\mathbf{A}_{\xi_0,\xi,\zeta} = diag(\mathbf{A}_{\xi_0}, \mathbf{A}_{\xi,\zeta})$. The first matrix $\mathbf{A}_{\xi_0} = \xi_0 \boldsymbol{\Delta}_M^T \boldsymbol{\Delta}_M$ corresponds to penalization of the squared differences between adjacent spline coefficients $\boldsymbol{\alpha}_0$ of the baseline hazard. $\boldsymbol{\Delta}_M$ denotes the $((M-1) \times M)$-dimensional difference operator matrix of degree one, defined as

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

The second penalty matrix $\mathbf{A}_{\xi,\zeta}$ results from local quadratic approximation of penalty $\xi \cdot J_\zeta(\boldsymbol{\alpha})$ (Oelker & Tutz, 2016). It is block-diagonal, i.e. $\mathbf{A}_{\xi,\zeta} = diag(\mathbf{A}_{1,\xi,\zeta}, \dots, \mathbf{A}_{r,\xi,\zeta})$, for $k = 1, \dots, r$ the single blocks have the form

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

where c is a small positive number (e.g. $c \approx 10^{-5}$), $\boldsymbol{\alpha}_k^T = (\alpha_{k,1}, \dots, \alpha_{k,M})$ contains all spline coefficients corresponding to the k-th time-varying effect and the matrix $\tilde{\boldsymbol{\Delta}}_M$ is equal to $\boldsymbol{\Delta}_M$, except that its first row consist of zeros only.

Information matrix

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

$$\mathbf{F}_{\beta\beta} = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\mathbf{x}_{ij}\mathbf{x}_{ij}^T\int_0^{t_{ij}}\exp(\eta_{ij}(s))ds,$$

$$\mathbf{F}_{\beta\alpha} = \mathbf{F}_{\alpha\beta}^T = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha}^T} = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\mathbf{x}_{ij}\int_0^{t_{ij}}\exp(\eta_{ij}(s))\boldsymbol{\Phi}^T(s)ds,$$

$$\mathbf{F}_{\alpha\alpha} = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\int_0^{t_{ij}}\exp(\eta_{ij}(s))\boldsymbol{\Phi}(s)\boldsymbol{\Phi}^T(s)ds + \mathbf{A}_{\xi_{\mathbf{0}},\xi,\zeta},$$

$$\mathbf{F}_{\beta i} = \mathbf{F}_{i\beta}^T = -\frac{\partial^2 l^{pen}(\boldsymbol{\delta})}{\partial\boldsymbol{\beta}\partial\mathbf{b}_i^T} = -\sum_{j=1}^{N_i}\mathbf{x}_{ij}\mathbf{u}_{ij}^T\int_0^{t_{ij}}\exp(\eta_{ij}(s))ds,$$

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

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

Variance-Covariance Components

With $\tilde{\beta}^T := (\beta^T, \boldsymbol{\alpha}^T)$, we get the simpler block structure

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

If the cluster sizes N_i are large enough: $\hat{\boldsymbol{\delta}} \overset{a}{\sim} N(\boldsymbol{\delta}, \mathbf{F}^{pen}(\hat{\boldsymbol{\delta}})^{-1})$

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

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

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

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