Variable Selection and Model Choice in Structured Survival Models

Boosting Survival Models

Benjamin Hofner [1]
Institut für Medizininformatik, Biometrie und Epidemiologie (IMBE)
Friedrich-Alexander-Universität Erlangen-Nürnberg

joint work with Torsten Hothorn and Thomas Kneib
Institut für Statistik
Ludwig-Maximilians-Universität München

55. Biometrisches Kolloquium
Hannover - 2009

[1] benjamin.hofner@imbe.med.uni-erlangen.de

Introduction

Cox PH model:

$$\lambda_i(t) = \lambda(t, \mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta})$$

with

- $\lambda_i(t)$ hazard rate of observation i $[i = 1, \ldots, n]$
- $\lambda_0(t)$ baseline hazard rate
- $\mathbf{x}_i$ vector of covariates for observation i $[i = 1, \ldots, n]$
- $\boldsymbol{\beta}$ vector of regression coefficients

Problem: restrictive model, not allowing for

- non-proportional hazards (e.g., time-varying effects)
- non-linear effects

Structured Survival Models

Generalization: Structured Survival Models

(Kneib & Fahrmeir, 2007)

$$\lambda_i(t) = \exp(\eta_i(t))$$

with additive predictor

$$\eta_i(t) = \sum_{l=1}^{L} f_l(\mathbf{x_i}(t)),$$

Generic representation of covariate effects $f_l(\mathbf{x}_i)$

a) linear effects: $f_l(\mathbf{x}_i(t)) = f_{l,\text{linear}}(\tilde{x}_i) = \tilde{x}_i\beta$

b) smooth effects: $f_l(\mathbf{x}_i(t)) = f_{l,\text{smooth}}(\tilde{x}_i)$

c) time-varying effects: $f_l(\mathbf{x}_i(t)) = f_{l,\text{smooth}}(t) \cdot \tilde{x}_i$

where $\tilde{x}_i$ is a covariate from $\mathbf{x}_i(t)$.

Note:

c) includes log-baseline ($\tilde{x}_i \equiv 1$)

P-Splines

flexible terms can be represented using P-splines

(Eilers & Marx, 1996)

- model term (x can be either $\tilde{x}_i$ or t):

$$f_{l,\text{smooth}}(x) = \sum_{m=1}^{M} \beta_{lm} B_{lm}(x)$$

- penalty: $\text{pen}_l(\boldsymbol{\beta_l}) = \kappa_l \, \boldsymbol{\beta_l}' \mathbf{K} \boldsymbol{\beta_l}$ cases b), c)
  ($\text{pen}_l(\boldsymbol{\beta_l}) = 0$ in case a))

with

- $\mathbf{K} = \mathbf{D}'\mathbf{D}$ (i.e., cross product of difference matrix $\mathbf{D}$)
  $\mathbf{D} \overset{e.g.}{=} \begin{pmatrix} 1 & -2 & 1 & \dots & \\ 0 & 1 & -2 & 1 & \dots \end{pmatrix}$
- κ_l smoothing parameter
  (larger $\kappa_l \Rightarrow$ more penalization $\Rightarrow$ smoother fit)

Estimation

Penalized Likelihood Criterion: (NB: this is the **full** log-likelihood)

$$\mathcal{L}_{\text{pen}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\delta_i \eta_i(t_i) - \int_0^{t_i} \exp(\eta_i(t))\, dt \right] - \sum_{l=0}^{L} \text{pen}_l(\boldsymbol{\beta}_l)$$

- T_i true survival time
- C_i censoring time
- $t_i = \min(T_i, C_i)$ observed survival time (right censoring)
- $\delta_i = \mathbb{1}(T_i \leq C_i)$ indicator for non-censoring

Problem:

Estimation and in particular model choice

CoxFlexBoost

Aim:

Maximization of a (potentially) high-dimensional log-likelihood with different modeling alternatives

Thus, we use:

- Iterative algorithm
- Likelihood-based boosting algorithm
- Component-wise base-learners

Therefore:

- Use one base-learner $g_j(\cdot)$ for each covariate (or each model component) $\quad [\, j \in \{1, \ldots, J\} \,]$

Component-Wise Boosting

as a means of estimation and variable selection combined with model choice.

CoxFlexBoost

Aim:

Maximization of a (potentially) high-dimensional log-likelihood with different modeling alternatives

Thus, we use:

- Iterative algorithm
- Likelihood-based boosting algorithm
- Component-wise base-learners

Therefore:

- Use one base-learner $g_j(\cdot)$ for each covariate (or each model component) $\quad [\, j \in \{1, \ldots, J\} \,]$

Component-Wise Boosting

as a means of estimation and variable selection combined with model choice.

CoxFlexBoost

Aim:

Maximization of a (potentially) high-dimensional log-likelihood with different modeling alternatives

Thus, we use:

- Iterative algorithm
- Likelihood-based boosting algorithm
- Component-wise base-learners

Therefore:

- Use one base-learner $g_j(\cdot)$ for each covariate (or each model component) $\quad [\, j \in \{1, \ldots, J\} \,]$

Component-Wise Boosting

as a means of estimation and variable selection combined with model choice.

CoxFlexBoost Algorithm

(i) **Initialization:** Iteration index $m := 0$.

- Function estimates (for all $j \in \{1, \ldots, J\}$):

$$\hat{f}_j^{[0]}(\cdot) \equiv 0$$

- Offset (MLE for constant log hazard):

$$\hat{\eta}^{[0]}(\cdot) \equiv \log\left(\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}\right)$$

(ii) **Estimation:** $m := m + 1$.

Fit all (linear/P-spline) base-learners separately

$$\hat{g}_j = g_j(\cdot\,; \hat{\boldsymbol{\beta}}_j),\ \forall j \in \{1, \ldots, J\},$$

by penalized MLE.

Details on pMLE

$$\hat{\boldsymbol{\beta}}_j = \arg\max_{\boldsymbol{\beta}} \mathcal{L}^{[m]}_{j,\text{pen}}(\boldsymbol{\beta})$$

with the penalized log-likelihood (analogously as above)

$$\begin{aligned}\mathcal{L}^{[m]}_{j,\text{pen}}(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \Bigg[& \delta_i \cdot (\hat{\eta}_i^{[m-1]} + g_j(x_i(t_i); \boldsymbol{\beta})) \\ & - \int_0^{t_i} \exp\left\{ \hat{\eta}_i^{[m-1]}(\tilde{\mathbf{t}}) + g_j(x_i(\tilde{\mathbf{t}}); \boldsymbol{\beta}) \right\} d\tilde{\mathbf{t}} \Bigg] - \text{pen}_j(\boldsymbol{\beta}),\end{aligned}$$

with the additive predictor η_i split

- into the estimate from previous iteration $\hat{\eta}_i^{[m-1]}$
- and the current base-learner $g_j(\cdot; \boldsymbol{\beta})$

(iii) **Selection:** Choose base-learner $\hat{g}_{j^*}$ with

$$j^* = \arg \max_{j \in \{1,\ldots,J\}} \mathcal{L}^{[m]}_{j,\text{unpen}}(\hat{\beta}_j)$$

(iv) **Update:**

- Function estimates (for all $j \in \{1, \ldots, J\}$):

$$\hat{f}_j^{[m]} = \begin{cases} \hat{f}_j^{[m-1]} + \nu \cdot \hat{g}_j & j = j^* \\ \hat{f}_j^{[m-1]} & j \neq j^* \end{cases}$$

- Additive predictor (= fit):

$$\hat{\eta}^{[m]} = \hat{\eta}^{[m-1]} + \nu \cdot \hat{g}_{j^*}$$

with step-length $\nu \in (0, 1]$ (here: $\nu = 0.1$)

(v) **Stopping rule:** Continue iterating steps (ii) to (iv) until $m = m_{\text{stop}}$

We stated that we use

Component-Wise Boosting

as a means of estimation and variable selection combined with model choice.

How?

Variable Selection and Model Choice

. . . is achieved by

- selection of base-learner (in step (iii) of CoxFlexBoost),
  i.e., component-wise boosting

and

- early stopping,
  i.e., choose $\widehat{m}_{\text{stop,opt}}$ via cross validation, out-of-bag sample, . . .

- Variable selection (without model choice):
  Define one base-learner per covariate
  e.g. flexible base-learner with 4 df
- Variable selection and model choice:
  Define one base-learner per modelling possibility
  But the df must be comparable!
  Otherwise: more flexible base-learners are preferred

Degrees of Freedom to Specify Smoothness

- Specifying df more intuitive than specifying smoothing parameter κ
- Smooth effects comparable to other modeling components, e.g., linear effects

Use initial $\widetilde{df}_j$ ($\overset{e.g.}{=} 4$) and solve

$$\text{df}(\kappa_j) - \widetilde{df}_j \overset{!}{=} 0$$

for κ_j, where

$$\text{df}(\kappa_j) = \text{trace}\Big[\overbrace{\mathbf{F}_j^{[0]}}^{\text{Fisher matrix}} \; (\underbrace{\mathbf{F}_j^{[0]} + \kappa_j \mathbf{K}_j}_{\text{penalized Fisher matrix}})^{-1} \Big] \qquad \text{(Gray, 1992).}$$

- Problem 1: Not constant over the (boosting) iterations

  But simulation studies showed: No big deviation from the initial $\widetilde{df}_j$

Degrees of Freedom to Specify Smoothness

- Specifying df more intuitive than specifying smoothing parameter κ
- Smooth effects comparable to other modeling components, e.g., linear effects

Use initial $\widetilde{df}_j$ ($\overset{e.g.}{=} 4$) and solve

$$\text{df}(\kappa_j) - \widetilde{df}_j \overset{!}{=} 0$$

for κ_j, where

$$\text{df}(\kappa_j) = \text{trace}\Big[\overbrace{\mathbf{F}_j^{[0]}}^{\text{Fisher matrix}} \; (\underbrace{\mathbf{F}_j^{[0]} + \kappa_j \mathbf{K}_j}_{\text{penalized Fisher matrix}})^{-1} \Big] \qquad \text{(Gray, 1992).}$$

- Problem 1: Not constant over the (boosting) iterations

  But simulation studies showed: No big deviation from the initial $\widetilde{df}_j$

Problem 2

- For higher order differences ($d \geq 2$): $\text{df} > 1\,(\kappa \to \infty)$
- Polynomial of order $d-1$ remains unpenalized
- **Solution:**

Decomposition (based on Kneib, Hothorn, & Tutz, 2008)

$$f_{\text{smooth}}(x) = \underbrace{\beta_0 + \beta_1 x + \ldots + \beta_{d-1} x^{d-1}}_{\text{unpenalized, parametric part}} + \underbrace{f_{\text{smooth,centered}}(x)}_{\text{deviation from polynomial}}$$

- Add unpenalized part as separate, parametric base-learners
- Assign df $= 1$ to the centered effect (and add as P-spline base-learner)
- Analogously for time-varying effects

Technical realization (see Fahrmeir, Kneib, & Lang, 2004):

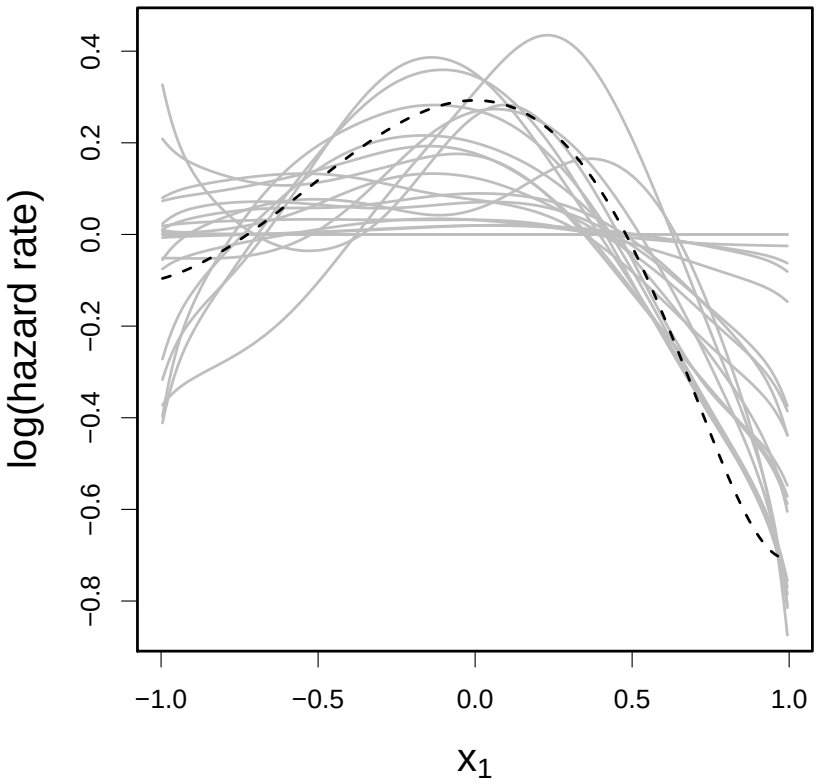
decomposing the vector of regression coefficients β into $(\widetilde{\beta}_{\text{unpen}}, \widetilde{\beta}_{\text{pen}})$ utilizing a spectral decomposition of the penalty matrix

Results

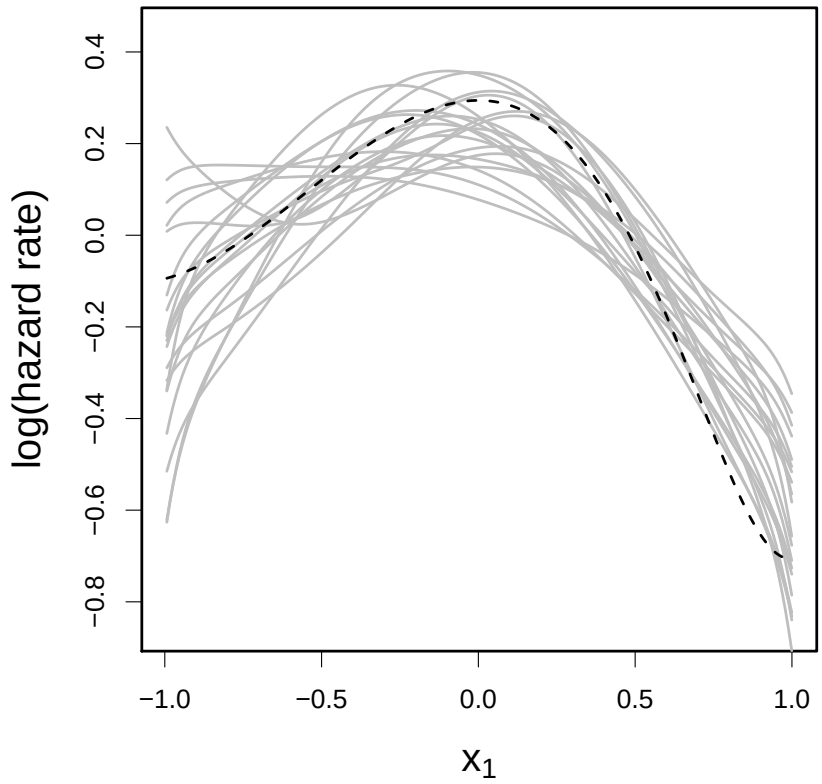
Simulation-Results (in short)

- Good variable selection strategy
- Good model choice strategy if only linear and smooth effects are used
- Selection bias in favor of time-varying base-learners (if present)
  $\Rightarrow$ standardizing time could be a solution
- Estimates are better if decomposition for model choice is used (compared to one flexible base-learner with 4 df)

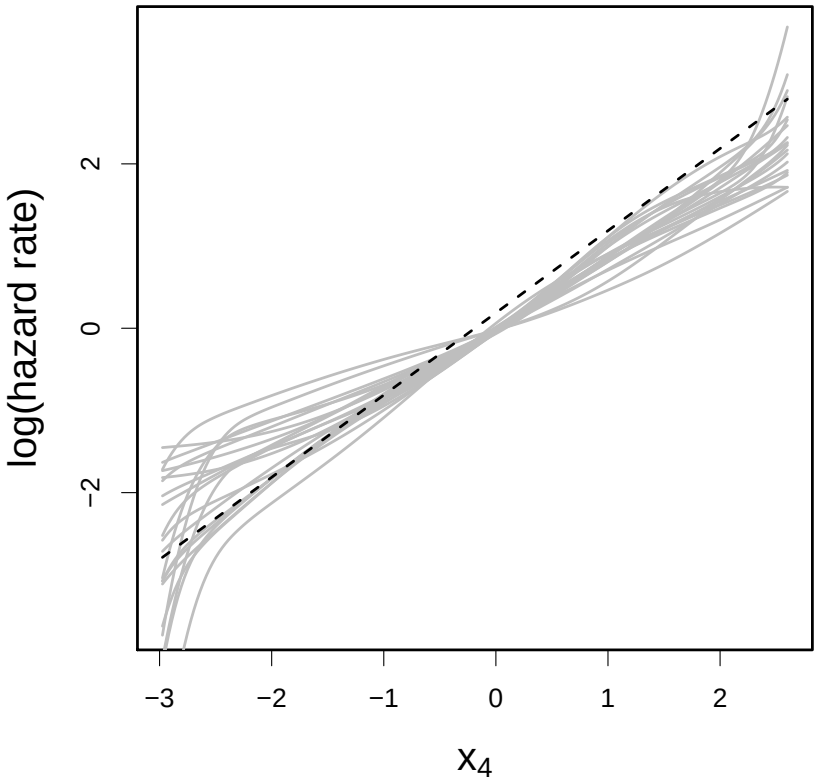
without model choice
($f_{\text{smooth}}(x_1)$)



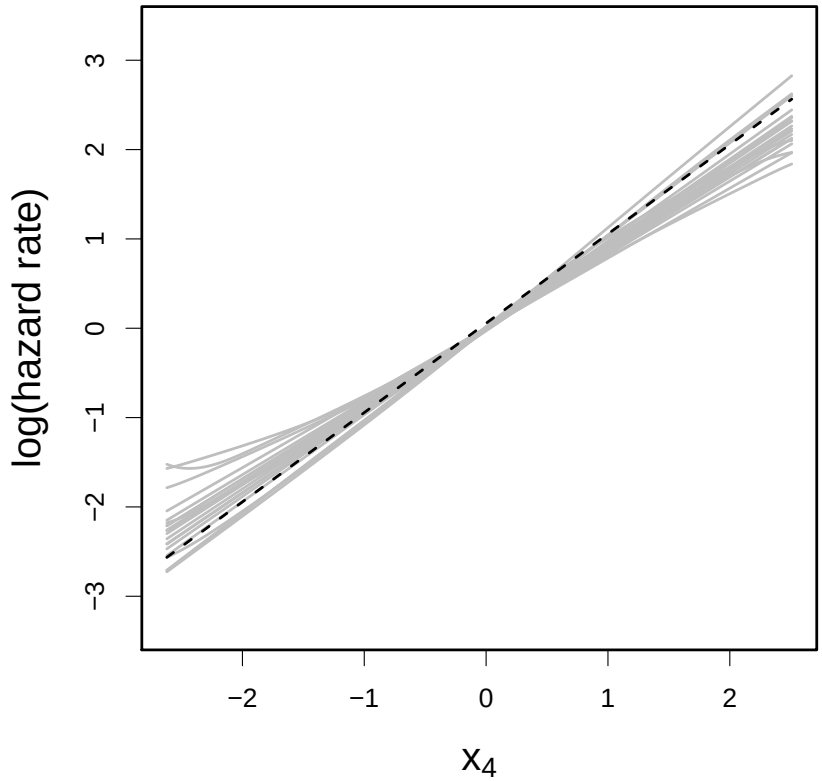
with model choice
($f_{\text{linear}}(x_1) + f_{\text{smooth,centered}}(x_1)$)



without model choice
($f_{\text{smooth}}(x_4)$)



with model choice
($f_{\text{linear}}(x_4) + f_{\text{smooth,centered}}(x_4)$)



Summary & Outlook

R-package **CoxFlexBoost** available (Hofner, 2008)

CoxFlexBoost . . .

- . . . allows for variable selection and model choice.
- . . . allows for flexible modeling
  - flexible, non-linear effects
  - time-varying effects (i.e., non-proportional hazards)
- . . . provides convenient functions to manipulate and show results (`summary()`, `plot()`, `subset()`, . . .)

Literature

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

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, *14*, 731–761.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association*, *87*, 942–951.

Hofner, B. (2008). *CoxFlexBoost: Boosting Flexible Cox Models (with Time-Varying Effects).* (R package version 0.6-0)

Hofner, B., Hothorn, T., & Kneib, T. (2008). *Variable selection and model choice in structured survival models* (Tech. Rep. No. 43). Department of Statistics, Ludwig-Maximilans-Universität München.

Kneib, T., & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, *34*, 207–228.

Kneib, T., Hothorn, T., & Tutz, G. (2008). Variable selection and model choice in geoadditive regression models. *Biometrics*. (accepted)

Find out more: `http://benjaminhofner.de/`