Variable Selection and Model Choice in Survival Models with Time-Varying Effects

Boosting Survival Models

Benjamin Hofner[1]
Thomas Kneib Torsten Hothorn

Department of Statistics
Ludwig-Maximilians-Universität München

14.05.2008

[1] benjamin.hofner@stat.uni-muenchen.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

Data Example

- **response:** 90-day survival
- **predictors:** 14 categorical predictors (sex, peritonitis (y/n), ...)
  6 continous predictors (age, Apache II Score, ...)
- **origin:** local database
  (Department of Surgery, Campus Großhadern, LMU Munich)
- **period of observation:** March 1993 – February 2005
- **N:** 462 septic patients (180 observations right-censored)

Question:

Do surgical patients with severe sepsis have a treatment benefit in terms of 90-day survival from an activity-guided AT 3 therapy?

- **Previous study showed:**
  Need for (three) time-varying and (two) smooth effects

Additive Hazard Regression

Model from previous study:

(Hofner, Kneib, Hartl, & Küchenhoff, 2008)

$$\begin{aligned}\lambda(t) = \exp(\ & \mathbf{g_0(t)} + \\ & g_1(\text{t}) \cdot \text{fungal infection} + g_2(\text{t}) \cdot \text{peritonitis} + \\ & f_1(\text{Apache II}) + f_2(\text{Horowitz ratio}) + f_3(\text{Haemoglobin conc.}) + \\ & \beta_1 \cdot \text{palliative operation} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{treatment period} + \\ & \beta_4 \cdot \text{malignant prim. diseases} + \beta_5 \cdot \text{sex} + \\ & \beta_6 \cdot \text{creatinin conc.} + \beta_7 \cdot \text{catecholamine therapy} + \\ & \beta_8 \cdot \text{surgery (toracic disease)} + \beta_9 \cdot \text{renal replacement therapy} \\ )\end{aligned}$$

log baseline hazard ($log(\lambda_0(t))$)

- Next slide: Generalized notation

Additive Hazard Regression

Model from previous study:

(Hofner et al., 2008)

$$\begin{aligned}\lambda(t) = \exp(\; & g_0(t) + \\ & \mathbf{g_1(t)} \cdot \text{fungal infection} + \mathbf{g_2(t)} \cdot \text{peritonitis} + \\ & f_1(\text{Apache II}) + f_2(\text{Horowitz ratio}) + f_3(\text{Haemoglobin conc.}) + \\ & \beta_1 \cdot \text{palliative operation} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{treatment period} + \\ & \beta_4 \cdot \text{malignant prim. diseases} + \beta_5 \cdot \text{sex} + \\ & \beta_6 \cdot \text{creatinin conc.} + \beta_7 \cdot \text{catecholamine therapy} + \\ & \beta_8 \cdot \text{surgery (toracic disease)} + \beta_9 \cdot \text{renal replacement therapy} \\ ) \end{aligned}$$

time-varying effects

- Next slide: Generalized notation

Additive Hazard Regression

Model from previous study:
(Hofner et al., 2008)

$$\begin{aligned}\lambda(t) = \exp(\; & g_0(\text{t}) + \\ & g_1(\text{t}) \cdot \text{fungal infection} + g_2(\text{t}) \cdot \text{peritonitis} + \\ & \mathbf{f_1(\text{Apache II}) + f_2(\text{Horowitz ratio}) + f_3(\text{Haemoglobin conc.})} + \\ & \beta_1 \cdot \text{palliative operation} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{treatment period} + \\ & \beta_4 \cdot \text{malignant prim. diseases} + \beta_5 \cdot \text{sex} + \\ & \beta_6 \cdot \text{creatinin conc.} + \beta_7 \cdot \text{catecholamine therapy} + \\ & \beta_8 \cdot \text{surgery (toracic disease)} + \beta_9 \cdot \text{renal replacement therapy} \\ )\end{aligned}$$

smooth effects

- Next slide: Generalized notation

Additive Hazard Regression

Model from previous study:
(Hofner et al., 2008)

$$\begin{aligned}\lambda(t) = \exp(\; & g_0(\text{t}) + \\ & g_1(\text{t}) \cdot \text{fungal infection} + g_2(\text{t}) \cdot \text{peritonitis} + \\ & f_1(\text{Apache II}) + f_2(\text{Horowitz ratio}) + f_3(\text{Haemoglobin conc.}) + \\ & \beta_1 \cdot \text{palliative operation} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{treatment period} + \\ & \beta_4 \cdot \text{malignant prim. diseases} + \beta_5 \cdot \text{sex} + \\ & \beta_6 \cdot \text{creatinin conc.} + \beta_7 \cdot \text{catecholamine therapy} + \\ & \beta_8 \cdot \text{surgery (toracic disease)} + \beta_9 \cdot \text{renal replacement therapy} \\ )\end{aligned}$$

linear effects

- Next slide: Generalized notation

Additive Hazard Regression

Generalisation: Additive Hazard Regression

(Kneib & Fahrmeir, 2007)

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

with

$$\eta_i(t) = \sum_{j=0}^{J} f_j(t, x_{ij})$$

where $f_j(t, x_{ij})$ can be

a) $g_o(t) = \log(\lambda_0(t))$ log-baseline ($\Rightarrow$ full likelihood available)

b) $g_j(t) \cdot x_{ij}$ time-varying effect of covariate x_{ij}

c) $g_j(x_{ij})$ smooth effect of covariate x_{ij}

d) $x_{ij}\beta_j$ linear effect

P-Splines

flexible terms can be represented using P-splines
(Eilers & Marx, 1996)

- model term (x can be either x_{ij} or t):

$$g_j(x) = \sum_{m=1}^{M} \beta_{jm} B_{jm}(x) \quad (j = 1, \ldots, J)$$

- penalty:

$$\text{pen}_j(\boldsymbol{\beta_j}) = \begin{cases} \kappa_j \, \boldsymbol{\beta_j}' \mathbf{K} \boldsymbol{\beta_j} & \text{cases a) to c)} \\ 0 & \text{case d)} \end{cases}$$

with

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

Inference

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

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

- 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

Estimation

component-wise boosting as a means of estimation and variable selection combined with model choice.

Cox$_{flex}$Boost

Likelihood-based Boosting

- Iterative algorithm
- Aim: Maximization of the high dimensional log-likelihood

Cox$_{flex}$Boost:

(i) **Initialization:** Iteration index $m := 0$. Function estimates $\hat{f}_j^{[0]}(\cdot) \equiv 0$. Offset (ML-estimate for constant log hazard):

$$\hat{\eta}^{[0]}(t) \equiv \arg\max_c \sum_{i=1}^{n} \delta_i \cdot c - \exp(c) \cdot t_i$$

$\text{Cox}_{flex}\text{Boost}$

Likelihood-based Boosting

- Iterative algorithm
- Aim: Maximization of the high dimensional log-likelihood

$\text{Cox}_{flex}\text{Boost}$:

(i) **Initialization:** Iteration index $m := 0$. Function estimates $\hat{f}_j^{[0]}(\cdot) \equiv 0$. Offset (ML-estimate for constant log hazard):

$$\hat{\eta}^{[0]}(t) \equiv \arg\max_c \sum_{i=1}^{n} \delta_i \cdot c - \exp(c) \cdot t_i$$

(ii) **Estimation:** $m := m + 1$. Fit (P-spline) base-learners

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

determined by

$$\hat{\boldsymbol{\beta}}_j = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_j^{[m]}(\boldsymbol{\beta})$$

with the penalized log-likelihood[2]

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

[2] (can also be interpreted as negative loss function)

(iii) **Selection:** Choose base-learner f_{j^*} with

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

(iv) **Update:**

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

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

- additive predictor (= fit):

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

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

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

Differences to classical LH-Based Boosting

	LH-based Boosting (Tutz & Binder, 2006)	Cox$_{flex}$Boost
Estimation	one-step Fisher scoring	step-length $\nu \times$ full maximum
Selection	based on deviance	based on penalized log-LH $\mathcal{L}_j^{[m]}$
Definition of Base-Learners	specified by smoothing parameter κ	specified by (initial) degrees of freedom df_{start}

For the ideas in Cox$_{flex}$Boost see, e.g., model based boosting (Bühlmann & Hothorn, 2007)

Early Stopping

1. Run the algorithm m_{stop}-times (previously defined).
2. Define new $m_{stop,new} \leq m_{stop}$:
   - ... based on out-of-bag sample (with simulations easy to use)
   - ... based on information criterion, e.g., AIC

$\Rightarrow$ Prevents algorithm to stop in a local maximum (of the log-likelihood)

$\Rightarrow$ Early stopping prevents overfitting

Variable Selection

Achieved by

- Selection of base-learner (step (iii)) (i.e., component-wise boosting)
  and
- Early stopping

Early Stopping

1. Run the algorithm m_{stop}-times (previously defined).
2. Define new $m_{stop,new} \leq m_{stop}$:
   - ... based on out-of-bag sample (with simulations easy to use)
   - ... based on information criterion, e.g., AIC

$\Rightarrow$ Prevents algorithm to stop in a local maximum (of the log-likelihood)

$\Rightarrow$ Early stopping prevents overfitting

Variable Selection

Achieved by

- Selection of base-learner (step (iii)) (i.e., component-wise boosting) **and**
- Early stopping

Model Choice

Recall:

$f_j(t, x_{ij})$ can be

b) $g_j(t)x_{ij}$ time-varying effect of covariate x_{ij}

c) $g_j(x_{ij})$ smooth effect of covariate x_{ij}

d) $x_{ij}\beta_j$ linear effect

(a) $g_o(t) = \log(\lambda_0(t))$ can be seen as special case of any of the above)

- $\Rightarrow$ We see: x_{ij} can enter the model in 3 different ways
- But how?
- Add all possibilities as base-learners to the model.
  Boosting can chose between the possibilities
- But the *df* must be comparable!
  Otherwise: more flexible base-learners are preferred

Model Choice

Recall:

$f_j(t, x_{ij})$ can be

b) $g_j(t)x_{ij}$ time-varying effect of covariate x_{ij}

c) $g_j(x_{ij})$ smooth effect of covariate x_{ij}

d) $x_{ij}\beta_j$ linear effect

(a) $g_o(t) = \log(\lambda_0(t))$ can be seen as special case of any of the above)

- $\Rightarrow$ We see: x_{ij} can enter the model in 3 different ways
- But how?
- Add all possibilities as base-learners to the model.
  Boosting can chose between the possibilities
- But the *df* must be comparable!
  Otherwise: more flexible base-learners are preferred

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

Decomposition
(based on Kneib, Hothorn, and Tutz (2007))

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

- 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
- Again early stopping
  $\Rightarrow$ selection of base-learners
  $\Rightarrow$ model choice
- Base-learner selected / not-selected:
  $\Rightarrow$ model choice

Example: age

possible base-learners: linear base-learner $\tilde{f}_1(age)$ and centered P-Spline base-learner $\tilde{f}_2(age)$

selected base-learners: only linear base-learner $\tilde{f}_1(age)$ selected until m_{stop}
$\Rightarrow$ no flexible term $\tilde{f}_2(age)$ needed

- Technical realization:
  decomposing the vector of regression coefficients $\boldsymbol{\beta}$ into $(\widetilde{\boldsymbol{\beta}}_{unpen}, \widetilde{\boldsymbol{\beta}}_{pen})$
  See Fahrmeir, Kneib, and Lang (2004) for more details

Degrees of Freedom

Definition of *df* in Survival Models (Gray, 1992)

- F (observed) Fisher matrix
- F_{pen} (observed) penalized Fisher matrix

$$df := trace\left[F \cdot F_{pen}^{-1}\right]$$

Compare: *df* in Linear Models

$$df = trace[X(X'X + \kappa \cdot K)^{-1}X'] = trace[\overbrace{X'X}^{F}(\overbrace{X'X + \kappa \cdot K}^{F_{pen}})^{-1}]$$

- Specifying *df* more intuitive than specifying smoothing parameter κ
- Comparable to other modeling components, e.g., linear effects
- $\Rightarrow$ Needed for model choice
- Problem: Not constant over the iterations
  - Preliminary result: No big deviation from the initial df_{start}

Degrees of Freedom

Definition of *df* in Survival Models (Gray, 1992)

- F (observed) Fisher matrix
- F_{pen} (observed) penalized Fisher matrix

$$df := trace\left[F \cdot F_{pen}^{-1}\right]$$

Compare: *df* in Linear Models

$$df = trace[X(X'X + \kappa \cdot K)^{-1}X'] = trace[\overbrace{X'X}^{F}(\overbrace{X'X + \kappa \cdot K}^{F_{pen}})^{-1}]$$

- Specifying *df* more intuitive than specifying smoothing parameter κ
- Comparable to other modeling components, e.g., linear effects
- $\Rightarrow$ Needed for model choice
- Problem: Not constant over the iterations
  - Preliminary result: No big deviation from the initial df_{start}

Computational Aspects

Cox$_{flex}$Boost is implemented using R (www.r-project.org)

- Crucial computation: Integral in $\mathcal{L}_j^{[m]}(\beta)$:

$$\int_0^{t_i} \exp\left\{\hat{\eta}_i^{[m-1]}(\tilde{\mathbf{t}}) + f_j(\tilde{\mathbf{t}}, x_{ij}; \beta)\right\} d\tilde{\mathbf{t}}$$

  - time consuming
  - very often evaluated (maximization of $\mathcal{L}_j^{[m]}(\beta)$)
- R-function `integrate()` slow in this context
  $\Rightarrow$ (specialized) vectorized trapezoid integration implemented
  $\Rightarrow \approx$ 100 times quicker
- Efficient storage of matrices can reduce computational burden
  $\Rightarrow$ recycling of results

Computational Aspects

Cox$_{flex}$Boost is implemented using R (www.r-project.org)

- Crucial computation: Integral in $\mathcal{L}_j^{[m]}(\beta)$:

$$\int_0^{t_i} \exp\left\{\hat{\eta}_i^{[m-1]}(\tilde{\mathbf{t}}) + f_j(\tilde{\mathbf{t}}, x_{ij}; \beta)\right\} d\tilde{\mathbf{t}}$$

  - time consuming
  - very often evaluated (maximization of $\mathcal{L}_j^{[m]}(\beta)$)
- R-function `integrate()` slow in this context
  $\Rightarrow$ (specialized) vectorized trapezoid integration implemented
  $\Rightarrow \approx$ 100 times quicker
- Efficient storage of matrices can reduce computational burden
  $\Rightarrow$ recycling of results

Computational Aspects

Cox$_{flex}$Boost is implemented using R (www.r-project.org)

- Crucial computation: Integral in $\mathcal{L}_j^{[m]}(\beta)$:

$$\int_0^{t_i} \exp\left\{\hat{\eta}_i^{[m-1]}(\tilde{\mathbf{t}}) + f_j(\tilde{\mathbf{t}}, x_{ij}; \beta)\right\} d\tilde{\mathbf{t}}$$

  - time consuming
  - very often evaluated (maximization of $\mathcal{L}_j^{[m]}(\beta)$)
- R-function `integrate()` slow in this context
  $\Rightarrow$ (specialized) vectorized trapezoid integration implemented
  $\Rightarrow \approx$ 100 times quicker
- Efficient storage of matrices can reduce computational burden
  $\Rightarrow$ recycling of results

Summary & Outlook

Cox_{flex}Boost ...

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

To be continued ...

- Simulation Studies
  - for deviation from df_{start}
  - for model choice capabilities
  - general predictive power
- Formula for AIC (for Boosting in Survival Models)
- Include mandatory covariates (update in each step)
- Measure for variable importance: e.g., $\int |\hat{f}_j^{[m_{stop}]}(\cdot)|$

Summary & Outlook

Cox_{flex}Boost ...

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

To be continued ...

- Simulation Studies
  - for deviation from df_{start}
  - for model choice capabilities
  - general predictive power
- Formula for AIC (for Boosting in Survival Models)
- Include mandatory covariates (update in each step)
- Measure for variable importance: e.g., $\int |\hat{f}_j^{[m_{stop}]}(\cdot)|$

Literature

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*(4), 477-505.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 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. *JASA*, *87*, 942–951.

Hofner, B., Kneib, T., Hartl, W., & Küchenhoff, H. (2008). *Model choice in Cox-type additive hazard regression models with time-varying effects.* (*Submitted*)

Kneib, T., & Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scand. J. Statist.*, *34*, 207–228.

Kneib, T., Hothorn, T., & Tutz, G. (2007). *Variable selection and model choice in geoadditive regression.* (Submitted to Biometrics)

Tutz, G., & Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood-based boosting. *Biometrics*, *62*, 961–971.