Learning high-dimensional time-varying Aalen and Cox models

Mokhtar Z. Alaya

LSTA Laboratoire de Statistique Théorique et Appliquée

UPMC SORBONNE UNIVERSITÉS

48^{e} Journées de Statistique - Montpellier, 2 juin 2016

Joint work with ...

Thibault Allart
LSTA – UPMC, CNAM & Ubisoft

Agathe Guilloux
LSTA – UPMC & CMAP – École Polytechnique

Sarah Lemler
MICS – École Centrale Supélec

Welcome !

☹ No graph concepts

Welcome !

☹ No graph concepts

☺ Fused Lasso concepts (Tibshirani et al. (2005))

Plan

1. Models
2. Estimation procedure
3. Theoretical guaranties
4. Algorithm + Numerical experiments

Framework and models: counting process

For individual $i, i = 1, \ldots, n$:

- $N_i(t)$ is a marked counting process over a fixed time interval $[0, \tau]$, with marker $Y_i(t)$.
- N_i has intensity, namely

$$\mathbb{P}[N_i \text{ has a jump in } [t, t + dt)|\mathcal{F}_t] = Y_i(t)\lambda_\star(t, X_i(t))dt,$$

  where $\mathcal{F}_t$ is some filtration.
- $X_i(t) = (X_i^1(t), \ldots, X_i^p(t))$ is a process of covariables.
- High-dimensional setting: p is large.

Framework and models: example of right censoring

- T_i survival time, C_i censoring time, and $\delta_i = \mathbb{1}(T_i \leq C_i)$ censoring indicator.
- Observed data : $Z_i = T_i \wedge C_i, \delta_i, X_i = (X_i^1(t), \ldots, X_i^p(t))$.
- Counting process:
  - $Y_i(t) = \mathbb{1}(Z_i \geq t)$ the at-risk process,
  - $N_i(t) = \mathbb{1}(\{Z_i \leq t, \delta_i = 1\})$ counting process.

Conditional hazard rate function of the survival time T_i

$$\lambda_\star(t, X_i(t)) = \lim_{dt \downarrow 0} \frac{\mathbb{P}[t < T_i \leq t + dt | T_i > t, X_i(t)]}{dt}$$

which characterizes the conditional distribution of T_i.

Framework and models: dynamic Aalen and Cox models

- We consider two dynamic models for the function $\lambda_\star$:
  - a time-varying Aalen model

$$\lambda_\star^{\mathsf{A}}(t, X(t)) = X(t)\beta^\star(t),$$

  - a time-varying Cox model

$$\lambda_\star^{\mathsf{M}}(t, X(t)) = \exp\left(X(t)\beta^\star(t)\right),$$

- $\beta^\star$ is an unknown p-dimensional function from $[0, \tau]$ to $\mathbb{R}^p$.
- Estimating the parameter $\beta^\star$ on the basis of data from n independent individuals:

$$\mathcal{D}_n = \left\{(X_i(t), Y_i(t), N_i(t)) : t \in [0, \tau], i = 1, \ldots, n\right\}.$$

Learning dictionary: piecewise constant functions

- We consider sieves (or histogram) based estimators of the p-dimensional unknown function $\beta^\star$ (Murphy and Sen (1991)).
- We hence consider a L-partition of the time interval $[0, \tau]$, where $L \in \mathbb{N}^*$

$$\varphi_l = \sqrt{\frac{L}{\tau}} \mathbb{1}(I_l) \text{ and } I_l = \Big(\frac{l-1}{L}\tau, \frac{l}{L}\tau \Big].$$

- Let the set of univariate piecewise constant functions

$$\mathcal{H}_L = \Big\{ \alpha(\cdot) = \sum_{l=1}^{L} \alpha_l \varphi_l(\cdot) : (\alpha_l)_{1 \leq l \leq L} \in \mathbb{R}_+^L \Big\}.$$

Learning dictionary: piecewise constant functions

- We define the sets of candidates for estimation as

$$\Lambda^{\mathsf{A}} = \{x, t \in [0, \tau] \mapsto \lambda^{\mathsf{A}}_{\beta}(t, x(t)) = x(t)\beta(t) \mid \forall j \ \beta_j \in \mathcal{H}_L\}.$$

  for the Aalen model and

$$\Lambda^{\mathsf{M}} = \{x, t \in [0, \tau] \mapsto \lambda^{\mathsf{M}}_{\beta}(t, x(t)) = \exp\left(x(t)\beta(t)\right) \mid \forall j \ \beta_j \in \mathcal{H}_L\}.$$

  for the Cox model.

- We consider

$$\beta = (\beta_{1,.}^{\top}, \ldots, \beta_{p,.}^{\top})^{\top} = (\beta_{1,1}, \ldots, \beta_{1,L}, \ldots, \beta_{p,1}, \ldots, \beta_{p,L})^{\top},$$

$$\forall j = 1 \ldots, p, \ \forall l = 1, \ldots, L. \text{ and } \forall t \in I_l, \beta_j(t) = \sqrt{\frac{L}{\tau}}\beta_{j,l}.$$

Estimation procedure: goodness-of-fit

- Least squares functional: time-varying Aalen model

$$\ell_n^{\mathrm{A}}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big\{ \int_0^\tau \left(\lambda_\beta^{\mathrm{A}}(t, X_i(t))\right)^2 Y_i(t)dt - 2\int_0^\tau \lambda_\beta^{\mathrm{A}}(t, X_i(t))dN_i(t)\Big\}.$$

  Reynaud-Bouret (2003, 2006), Martinussen and Scheike (2007), Gaiffas and Guilloux (2012).

- Full likelihood functional: time-varying Cox model

$$\ell_n^{\mathrm{M}}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{ \int_0^\tau \log\left(\lambda_\beta^{\mathrm{M}}(t, X_i(t))\right) dN_i(t) - \int_0^\tau Y_i(t)\lambda_\beta^{\mathrm{M}}(t, X_i(t))dt\Big\}.$$

  Martinussen and Scheike (2007), Lemler (2013).

Estimation procedure: weighted $(\ell_1 + \ell_1)$-total-variation penalization

- We introduce a well-chosen vector of data-driven weights $\hat{\gamma} = (\hat{\gamma}_{1,\cdot}{}^\top, \ldots, \hat{\gamma}_{p,\cdot}{}^\top)^\top$

$$\hat{\gamma}_{j,l} \asymp \sqrt{\frac{L \log pL}{n} \hat{V}_{j,l}},$$

where

$$\hat{V}_{j,l} = \frac{1}{n} \sum_{i=1}^{n} \int_{(l-1)\tau/L}^{\tau} (X_i^j(t))^2 dN_i(t),$$

- Our specific covariate weighted $(\ell_1 + \ell_1)$-total-variation penalty is given by

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \sum_{j=1}^{p} \left(\hat{\gamma}_{j,1} |\beta_{j,1}| + \sum_{l=2}^{L} \hat{\gamma}_{j,l} |\beta_{j,l} - \beta_{j,l-1}| \right)$$

for $\beta \in \mathbb{R}^{p \times L}$.

Estimation procedure: definition of estimators

- Our estimators are then respectively defined as

$$\hat{\lambda}^{\mathsf{A}} = \lambda^{\mathsf{A}}_{\hat{\beta}^{\mathsf{A}}}, \text{ and } \hat{\lambda}^{\mathsf{M}} = \lambda^{\mathsf{M}}_{\hat{\beta}^{\mathsf{M}}}.$$

- $$\hat{\beta}^{\mathsf{A}} = \underset{\beta \in \mathbb{R}^{p \times L}}{\operatorname{argmin}} \left\{ \ell_n^{\mathsf{A}}(\beta) + \|\beta\|_{\mathsf{gTV},\hat{\gamma}} \right\},$$

  in the Aalen model.

- $$\hat{\beta}^{\mathsf{M}} = \underset{\beta \in \mathbb{R}^{p \times L}}{\operatorname{argmin}} \left\{ \ell_n^{\mathsf{M}}(\beta) + \|\beta\|_{\mathsf{gTV},\hat{\gamma}} \right\}$$

  in the Cox model.

Theoretical guaranties: oracles inqualities

- Theoretical properties: proving classical non-asymptotic oracle approaches.
- Weighted empirical quadratic norm $\|\lambda_\beta^{\mathsf{A}}\|_n$ defined for any $\lambda_\beta^{\mathsf{A}} \in \Lambda^{\mathsf{A}}$ by

$$\|\lambda_\star^{\mathsf{A}} - \lambda_\beta^{\mathsf{A}}\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\lambda_\star^{\mathsf{A}}(t, X_i(t)) - \lambda_\beta^{\mathsf{A}}(t, X_i(t))\right)^2 Y_i(t)dt},$$

- Empirical Kullback divergence $K_n(\lambda_\star^{\mathsf{M}}, \lambda_\beta^{\mathsf{M}})$ defined for $\lambda_\beta^{\mathsf{M}} \in \Lambda^{\mathsf{M}}$ by

$$K_n(\lambda_\star^{\mathsf{M}}, \lambda_\beta^{\mathsf{M}}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\lambda_\star^{\mathsf{M}}(t, X_i(t))}{\lambda_\beta^{\mathsf{M}}(t, X_i(t))}\right)\lambda_\star^{\mathsf{M}}(t, X_i(t))Y_i(t)dt$$
$$- \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\lambda_\star^{\mathsf{M}}(t, X_i(t)) - \lambda_\beta^{\mathsf{M}}(t, X_i(t))\right) Y_i(t)dt.$$

Slow oracles inqualities in prediction

Time-varying Aalen model

Theorem 1

For $x > 0$ fixed, the estimator $\hat{\lambda}^{\mathrm{A}}$ verifies with a probability larger than $1 - 29e^{-x}$,

$$\|\lambda_\star^{\mathrm{A}} - \hat{\lambda}^{\mathrm{A}}\|_n^2 \leq \inf_{\beta \in \mathbb{R}^{p \times L}} \left(\|\lambda_\star^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2 + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right).$$

Time-varying Cox model

Theorem 2

For $x > 0$ fixed, the estimator $\hat{\lambda}^{\mathrm{M}}$ verifies with a probability larger than $1 - 29e^{-x}$,

$$K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) \leq \inf_{\beta \in \mathbb{R}^{p \times L}} \left(K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \right).$$

Slow oracles inqualities in prediction: remarks

- Two terms are involved on the right hand side of above inequalities.
- The first one measures how far are the true functions of interest $\lambda_\star^{\mathsf{A}}$ and $\lambda_\star^{\mathsf{M}}$ from their approximations on Λ^{A} and Λ^{M}.
- The second one can be viewed as a variance term that satisfies

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \asymp \|\beta\|_{\mathrm{gTV}} \max_{j=1,\ldots,p} \max_{l=1,\ldots,L} \sqrt{\frac{L \log pL}{n}} \hat{V}_{j,\ell}.$$

- To obtain fast oracle inequalities: we suppose that Restricted eigenvalues (RE) conditions are fulfilled (Bickel et al. (2009)).

Proximal operator + algorithm

- We are interested in computing a solution

$$x^\star = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \{g(x) + h(x)\},$$

  where g is smooth and h is simple (prox-calculable).
- prox is the proximal operator defined by,

$$\operatorname{prox}_h(x) = \underset{v \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|x - v\|_2^2 + h(v).$$

- Proximal gradient descent (PGD) algorithm is based on

$$x^{(k+1)} = \operatorname{prox}_{\varepsilon_k h}\left(x^{(k)} - \varepsilon_k \nabla g(x^{(k)})\right).$$

Daubechies et al. (2004) (ISTA) , Beck and Teboulle (2009) (FISTA)

Stochastic proximal gradient descent (SPGD) algorithm

- PGD uses all of the training instances, by evaluation n gradients, to update the model parameters in each iteration.
  $\rightarrow$ too long for n and p really large.
- Go back to "old" algorithms (when computers were really slow),
  $\rightarrow$ Stochastic gradient Descent (SGD) Robbins and Monro (1951)
- At each iteration $k = 1, 2, \ldots,$ we pick i_k randomly (uniformly) from $\{1, 2, \ldots, n\}$, and take the following update:

$$x^{(k+1)} = \text{prox}_{\varepsilon_k h}\left(x^{(k)} - \varepsilon_k \nabla g_{i_k}(x^{(k)})\right)$$

- The computational cost of SPGD per iteration is $1/n$.
- Learning rate ε_k: Bottou (2012), AdaGrad (Duchi et al. (2011)), AdaDelta (Zeiler (2012)), vSGD (Schaul et al. (2012)), ...

Proximal operator of the weighted $(\ell_1 + \ell_1)$-total-variation penalization

- $\theta = \text{prox}_{\|\cdot\|_{\text{gTV}},\hat{\gamma}}(\beta)$

$$\theta = \underset{x \in \mathbb{R}^{p \times L}}{\text{argmin}} \left\{ \frac{1}{2}\|\beta - x\|_2^2 + \sum_{j=1}^{p} \left(\hat{\gamma}_{j,1} x_{j,1} + \sum_{l=2}^{L} \hat{\gamma}_{j,l} |x_{j,l} - x_{j,l-1}| \right) \right\}.$$

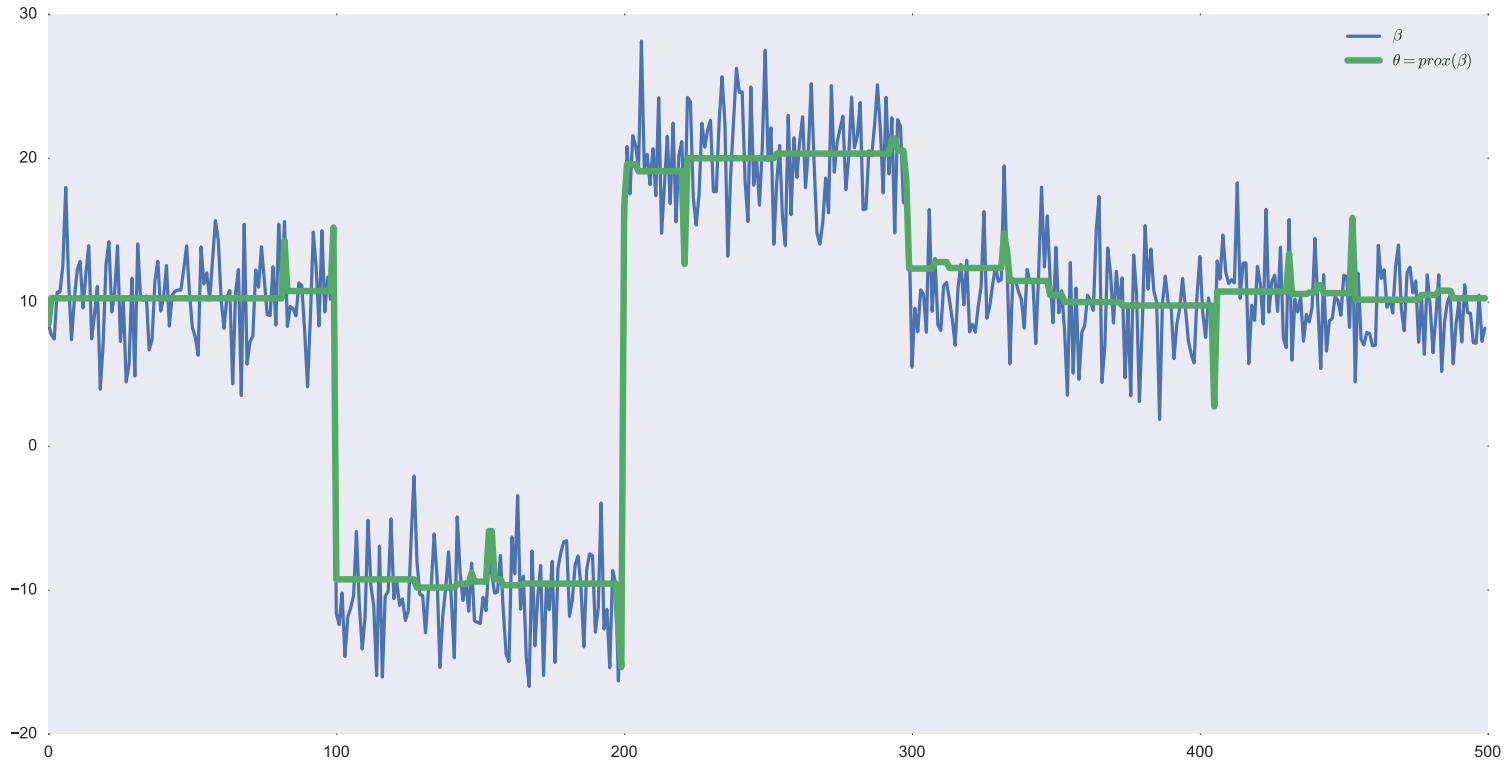
Algorithm 1:

for $j = 1, \ldots, p$ **do**
  set $\mu \leftarrow \beta_{j,\cdot}$; $w \leftarrow \hat{\gamma}_{j,\cdot} \backslash \{\hat{\gamma}_{j,1}\}$;
  $\eta \leftarrow \text{prox}_{\|\cdot\|_{\text{TV}},w}(\mu)$;
  $\theta_{j,\cdot} \leftarrow \eta - \left(\eta_1 - \text{sgn}(\eta_1) \max\left(0, |\eta_1| - \frac{w_1}{L}\right) \right) \mathbf{1}_L$;
  return $\theta_{j,\cdot}$

- $\text{prox}_{\|\cdot\|_{\text{TV}},w}$ (Alaya et al. (2015)).

.

Proximal operator of the weighted $(\ell_1 + \ell_1)$-total-variation penalization



Proximal operator of the weighted $(\ell_1 + \ell_1)$-total-variation penalization

SPGD for time-varying Cox model = CoxSGD

Algorithm 2: CoxSGD

1. Parameters: Integer $K > 0$;
2. Initialization: $(\hat{\beta}^{\mathrm{M}})^{(1)} = 0 \in \mathbb{R}^{p \times L}$, and $r^{(1)} \in [0, 1]$;
3. **for** $k = 1, \ldots, K$ **do**

   Choose randomly $i_k \in \{1, \ldots, n\}$ and compute $\nabla^{\mathrm{M}}_{i_k} = \nabla \ell^{\mathrm{M}}_{i_k}((\hat{\beta}^{\mathrm{M}})^{(k)})$;

   Update moving averages

   $a^{(k)} \leftarrow (1 - (r^{(k)})^{-1}) a^{(k)} + (r^{(k)})^{-1} \nabla^{\mathrm{M}}_{i_k}$;

   $b^{(k)}_j \leftarrow (1 - (r^{(k)})^{-1}) b^{(k)}_j + (r^{(k)})^{-1} \| \nabla_{j,\cdot} \|^2$;

   $c^{(k)} \leftarrow (1 - (r^{(k)})^{-1}) c^{(k)} + (r^{(k)})^{-1} \operatorname{diag}(H_{i_k})$ where $H_{i_k} = \left(\frac{\partial^2 \left(\ell^{\mathrm{M}}_{i_k}((\hat{\beta}^{\mathrm{M}})^{(k)}) \right)}{\partial^2 \beta} \right)$;

   Estimate learning rate

   $\varepsilon^{(k)}_j \leftarrow \frac{1}{c^+_j} \frac{\sum_{l=1}^{L} (a^{(k)}_{j,l})^2}{b^{(k)}_j}$; where $c^+_j = \max_{1 \le l \le L} c_{j,l}$

   $\eta_j \leftarrow \varepsilon^{(k)}_j$;

   $\varepsilon^{(k)} \leftarrow (\varepsilon^{(k)}_1 \mathbf{1}_L, \ldots, \varepsilon^{(k)}_p \mathbf{1}_L)^\top$;

   Update memory size

   $r^{(k)} \leftarrow \left(1 - \frac{\sum_{l=1}^{L} (a^{(k)}_{j,l})^2}{b^{(k)}_j}\right) \odot r^{(k)} + 1$;

   $(\hat{\theta}^{\mathrm{M}})^{(k)} \leftarrow (\hat{\beta}^{\mathrm{M}})^{(k)} - \varepsilon^{(k)} \odot \nabla^{\mathrm{M}}_{i_k}$;

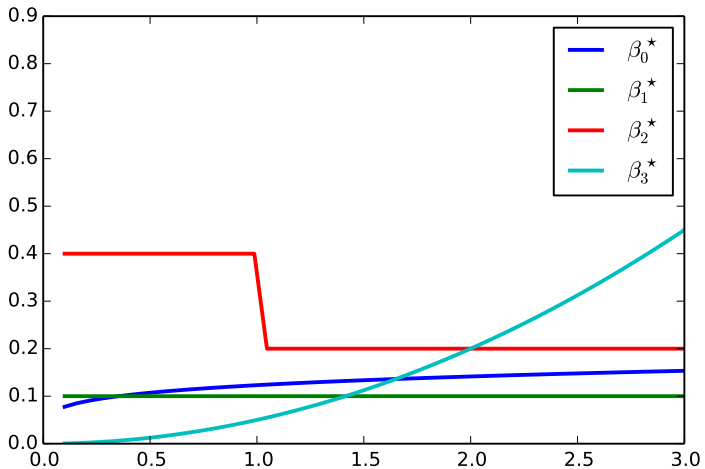
   $(\hat{\beta}^{\mathrm{M}})^{(k+1)} \leftarrow \left(\operatorname{prox}_{\eta_1 \| \cdot \|_{\mathrm{gTV}, \hat{\gamma}}} ((\hat{\theta}^{\mathrm{M}})^{(k)}_{1,\cdot}), \ldots, \operatorname{prox}_{\eta_p \| \cdot \|_{\mathrm{gTV}, \hat{\gamma}}} ((\hat{\theta}^{\mathrm{M}})^{(k)}_{p,\cdot}) \right)^\top$;

4. **return** $(\hat{\beta}^{\mathrm{M}})^{(K)}$

Schaul et al. (2012)

Simulated data in the time-varying Cox model

- Right censoring: $n = 1000$, and T has hazard rate $\lambda_\star(t, X) = \beta_0^\star(t) \exp(X(t)\beta^\star(t))$.
- $p = 10$ covariates processes $X_i(t)_{i=1,\dots,n}$ which are $\mathcal{N}(0, 0.5)$ $i.i.d$ piecewise constant over a 50-partition of the time interval $[0, 3]$.
- The baseline $\beta_0^\star$ is defined through a Weibull $\mathcal{W}(1.2, 0.15)$.
- We draw the true regression functions $\beta_1^\star$, $\beta_2^\star$, and $\beta_3^\star$. We set $\beta_j^\star \equiv 0$, for $j = 4, \dots, 10$.



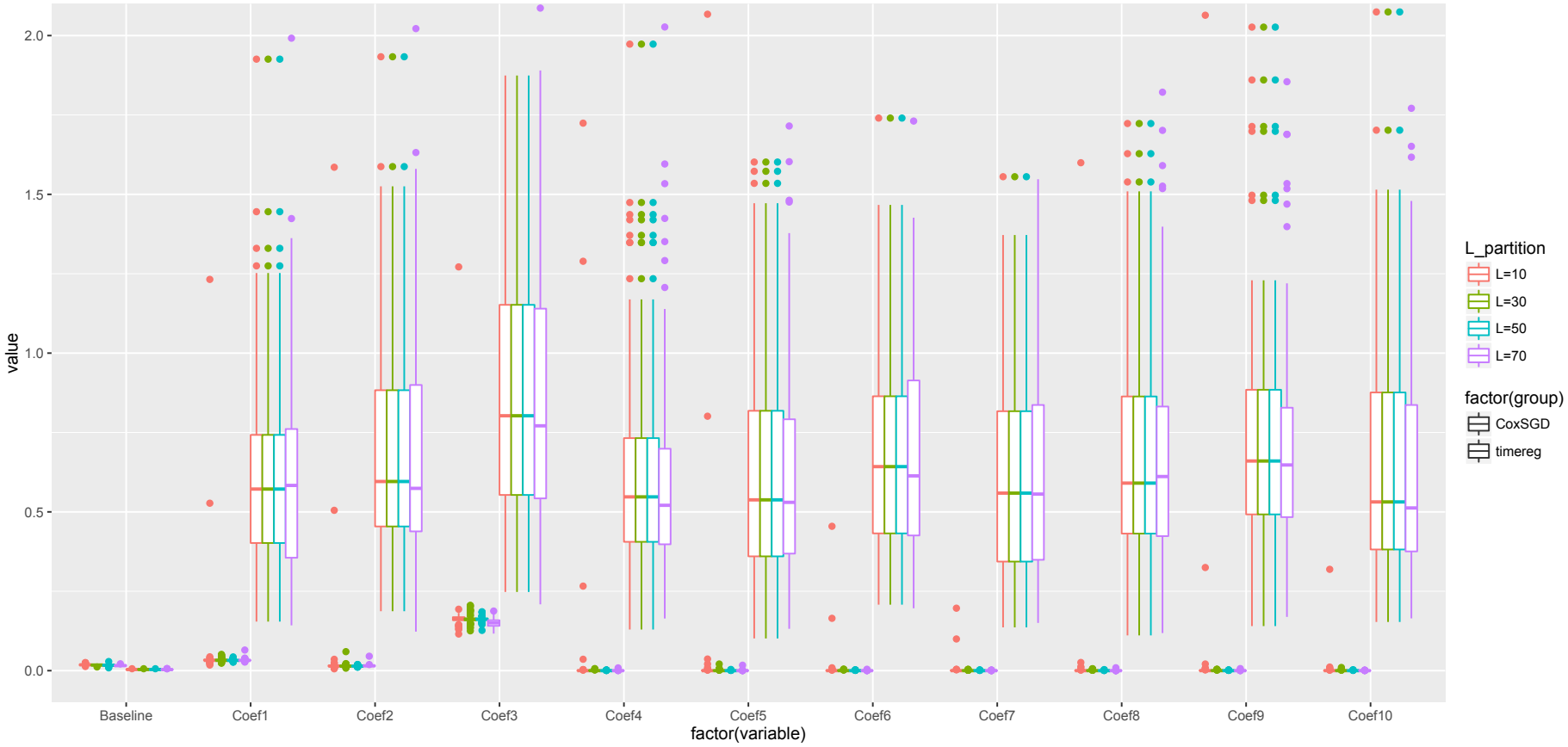
Simulated data in the time-varying Cox model

- We run 100 Monte-Carlo experiments of training data as described above.
- The estimation accuracy is investigated via a mean squared error defined as

$$\mathrm{MSE}_j = \frac{1}{100} \sum_{m=1}^{100} \|(\hat{\beta}_j^{\mathsf{M}})_m - \beta_j^{\star}\|_2^2,$$

where $(\hat{\beta}_j^{\mathsf{M}})_m$ is the estimation of $\beta_j^{\star}$ in the sample m, for all $j = 1, \ldots, p$.

Simulated data in the time-varying Cox model

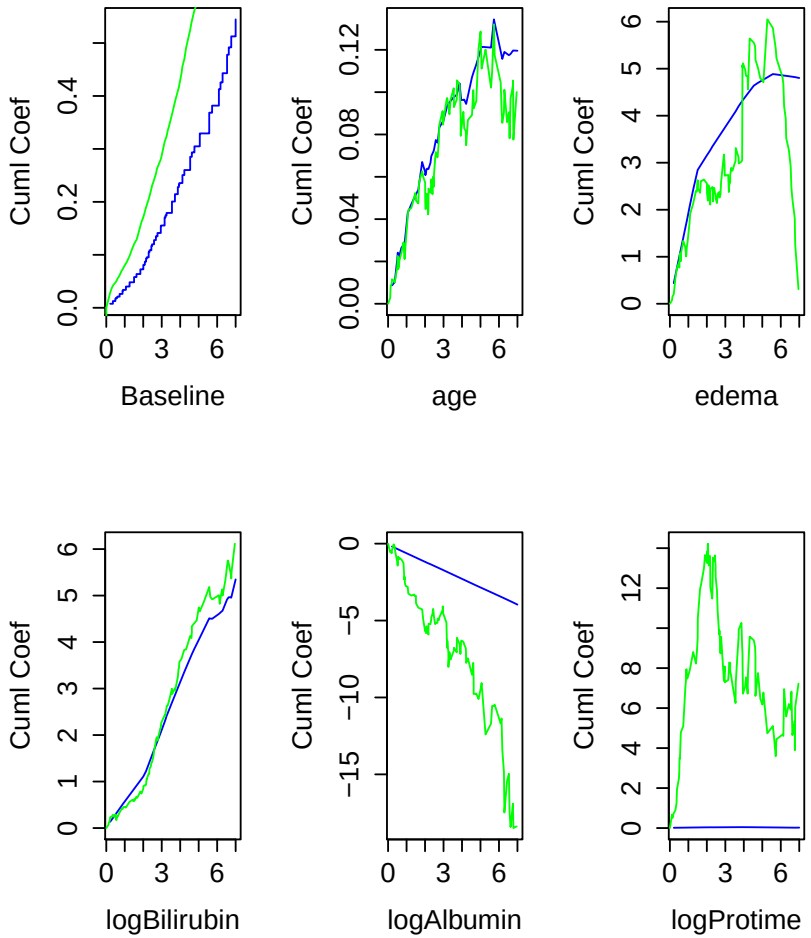


A boxplots of the MSE_j of estimated regression coefficients over L-partition ($L \in \{10, 30, 50, 70\}$) with CoxSGD and `timereg` R-package (Martinussen and Scheike (2007)).

PBC data: time-varying Cox model

- Primary Biliary Cirrhosis (PBC) of the liver and was conducted between 1974 and 1984 (Feleming 1991).
- A total of 418 patients are included in the dataset and were followed until death or censoring.
- We restrict attention to the first 8 years days of the study.
- We consider the covariates: age, edema, log(bilirubin), log(albumin) and log(protime).

PBC data: time-varying Cox model



Estimated cumulative regression coefficients $\hat{B}_j^{\mathrm{M}}(t) = \int_0^t \hat{\beta}_j(s)ds$ on PBC data with: CoxSGD (blue) and `timereg` R-package (green).

Take-home message

- Introducing the weighted $(\ell_1 + \ell_1)$-total-variation penalization procedure for high-dimensional time-varying Aalen and Cox models.
- Theoretical and algorithmic guaranties for this approach.

References

Alaya, M. Z., S. Gaïffas, and A. Guilloux (2015). Learning the intensity of time events with change-points. *Information Theory, IEEE Transactions on 61*(9), 5148–5171.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences 2*(1), 183–202.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009, 08). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist. 37*(4), 1705–1732.

Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, and K. R. Muller (Eds.), *Neural Networks: Tricks of the Trade (2nd ed.)*, Volume 7700 of *Lecture Notes in Computer Science*, pp. 421–436. Springer.

Daubechies, I., M. Defrise, and C. De Mol (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics 57*(11), 1413–1457.

Duchi, J., E. Hazan, and Y. Singer (2011, July). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res. 12*, 2121–2159.

Gaiffas, S. and A. Guilloux (2012). High-dimensional additive hazards models and the lasso. *Electron. J. Statist. 6*, 522–546.

Lemler, S. (2013). Oracle inequalities for the lasso in the high-dimensional multiplicative aalen intensity model. *Les Annales de l'Institut Henri Poincaré, arXiv preprint*.

Martinussen, T. and T. H. Scheike (2007). *Dynamic regression models for survival data*. Springer Science & Business Media.

Murphy, S. A. and P. K. Sen (1991). Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications 39*(1), 153–180.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields 126*(1), 103–153.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli 12*(4), 633–661.

Schaul, T., S. Zhang, and Y. LeCun (2012). No more pesky learning rates. *arXiv preprint arXiv:1206.1106*.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Zeiler, M. D. (2012). Adadelta: An adaptive learning rate method. *CoRR abs/1212.5701*.


Thank you.