High-dimensional time-varying Aalen and Cox models

Mokhtar Z. Alaya, Sarah Lemler, Agathe Guilloux, Thibault Allart

High-dimensional time-varying Aalen and Cox models

Mokhtar Z. Alaya[1] Thibault Allart[2] Agathe Guilloux[3,5]
Sarah Lemler[4]

May 23, 2018

Abstract

We consider the problem of estimating the intensity of a counting process in high-dimensional time-varying Aalen and Cox models. We introduce a covariate-specific weighted total-variation penalization, using data-driven weights that correctly scale the penalization along the observation interval. We provide theoretical guaranties for the convergence of our estimators and present a proximal algorithm to solve the convex studied problems. The practical use and effectiveness of the proposed method are demonstrated by simulation studies and real data example.

Keywords. Dynamic regression, Time-dependent covariates; Time-varying coefficients; Total-variation; Oracle inequalities
2010 MSC: 62G08, 62N01

[1]Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France
[2]Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France, and Ubisoft
[3]LaMME, UEVE and UMR 8071, Université Paris Saclay, Évry, France
[4]École CentraleSupélec, Laboratoire de Mathématiques et Informatique pour la Complexité des Systèmes, France
[5]Corresponding author, Email adress:agathe.guilloux@math.cnrs.fr

Introduction

Longitudinal events data arise in medicine, insurance, economics, game analytics, etc. For example, electronically health records (EHR) databases and patient registries collect, for large numbers of patients and for several decades now, numerous longitudinal clinical markers, together with times to (adverse) events, see e.g. Häyrinen, Saranto, and Nykänen 2008; Baker et al. 2004; Coloma et al. 2011. For these longitudinal data, dynamical regression models, such as Cox Cai and Sun 2003; Tian, D. Zucker, and Wei 2005 or Aalen Torben Martinussen and Thomas H Scheike 1999 models with time-varying coefficients, are flexible and popular models for assessing the (time-varying) influence of each covariates on the risk, see examples of applications in Kalantar-Zadeh et al. 2006; Bellera et al. 2010.

When a large number of covariates are recorded for each individuals or patients, such dynamical models face three major difficulties that have to be addressed by statistical algorithms: the number of covariates, the complexity of the model, a function has to be estimated for each covariates, and the size of the data, when the number of individuals and/or covariates and the number of records per individual grow.

In the existing literature, the first two difficulties have been addressed via model selection. The most recent contributions for the Cox model with time-varying coefficients include Yan and Jian Huang 2012; T. Honda and Härdle 2014; T. Honda and Yabe 2017. They both propose to perform model selection via sparsity inducing penalties (LASSO and SCAD) with spline proposals. We also refer the reader to Cheng et al. 2014 for similar methods in the classical longitudinal data model, where a response process is observed.

To the best of our knowledge, the third difficulty, due to the size of the data (in both direction) has only been addressed in Perperoglou, Cessie, and Houwelingen 2006; He et al. 2016, whereas neither of them considered the case of time-varying covariates. The existing algorithms, based either on kernels T. Martinussen and T. H. Scheike 2009b; Tian, D. Zucker, and Wei 2005, or splines Yan and Jian Huang 2012; T. Honda and Härdle 2014; T. Honda and Yabe 2017 are actually not scalable. For kernel estimators, model selection is entirely based on tests and, consequently, do not support high-dimensional covariates, see in particular the `timereg` package T. Scheike, Martinussen, and Silver 2016. For estimators based on splines, this is simply due to the fact that they involve dense predictor matrices. We refer the reader to R. J. Tibshirani 2014 for a discussion on this fact in the simple signal+noise model.

We introduce, in the present paper, a new estimator, based on sieves proposals, as introduced in Murphy and Sen 1991. These proposals are penalized via a data-driven and covariate specific total variation penalty, to which is added a lasso penalty of the first coefficient (see Equation (13) for a proper definition). This penalization induces covariates selection and temporal sparsity, and hence constant estimators. As the splines-based estimators of Yan and Jian Huang 2012; T. Honda and Härdle 2014; T. Honda and Yabe 2017, our estimator self-adapts to the three kinds of covariates: it selects covariates that are relevant in the model, seek for constant estimators, for covariates with constant coefficients, and gives a sieves estimate for covariates with time-varying coefficients. Hence it addresses the first two aforementioned difficulties.

Our estimator addresses the third difficulty in two ways. First, it is based on sieves proposals, hence only involves sparse predictor matrices, as opposed to splines methods. In addition, the optimization problems at hand are solved via stochastic proximal gradient descent (SPGD) algorithms (see Bottou 2010; Bottou 2012; Rosasco, Villa, and Vu n.d.; Atchade, Fort, and Moulines 2014), and are, as such, scalable.

On the theoretical part, asymptotic rates of convergence have been established in T. Honda and Härdle 2014; T. Honda and Yabe 2017 for the Cox model with time-varying coefficients but only in the context of censored data. To the best of our knowledge, there is no non-asymptotic results neither for Cox nor Aalen models with time-varying coefficients. In the present paper, we establish oracle inequalities for our penalized estimators of the complete intensity function in the general counting processes setting for both Cox and Aalen models with time-varying coefficients. We ameliorate the rates of convergence established in T. Honda and Härdle 2014

The paper is organized as follows. Sections and are devoted to the framework and the definition of our estimators. In Section we define our estimation procedure and in Sectionsec:theory, we state the theoretical properties of the estimators. In Section , we describe our algorithms. This section also include an algorithm for simulating in time-varying Aalen and Cox model. A quantitative comparison of the speeds of the different algorithms is proposed. Simulation results and illustration on a real dataset are presented in Section .

Framework and models

Consider the usual counting process framework where a process $\tilde{N}$ counts the number of occurring events of interest over a fixed time interval, say $[0, \tau]$ with $0 < \tau < \infty$, and the convention $\tilde{N}(0) = 0$ (see Andersen et al. 1993; T. Martinussen and T. H Scheike 2007). Let $\lambda_\star$ denote the intensity of the process $\tilde{N}$ depending on both time and a p-dimensional predictable process of covariates denoted by X (possibly including an intercept).

We consider that the process $\tilde{N}$ may be independently filtered (see Andersen et al. 1993) by a censoring predictable process Y and the resulting observed process is denoted by N. The intensity of N is then given for all $t \in [0, \tau]$ by

$$Y(t)\lambda_\star(t, X(t)).$$

Assumption 1. *We assume that* $\mathbb{P}[Y(\tau) > 0] > 0$.

This is a classical hypothesis in survival analysis (see for instance Andersen et al. 1993).

In this framework, we consider two dynamic models for the function $\lambda_\star$:

- a time-varying Aalen model

$$\lambda_\star^{\mathrm{A}}(t, X(t)) = X(t)\beta^\star(t), \tag{1}$$

- a time-varying Cox model

$$\lambda_\star^{\mathrm{M}}(t, X(t)) = \exp\big(X(t)\beta^\star(t)\big), \tag{2}$$

where, in both cases, $\beta^\star$ is an unknown function from $[0, \tau]$ to $\mathbb{R}^p$ to be estimated. We consider the problem of estimating the parameter $\beta^\star$ in dynamic models (1) and (2) on the basis of data from n independent individuals:

$$\mathcal{D}_n = \big\{(X_i(t), Y_i(t), N_i(t)) : t \in [0, \tau], i = 1, \ldots, n\big\}. \tag{3}$$

Estimation in models (1) and (2) are received a lot of attention in the past four decades. References for the additive Aalen model include Aalen 1980; Aalen 1989; Aalen 1993; McKeague 1988; Huffer and McKeague 1991, for the time-varying Cox models include D. M. Zucker and Karr 1990; Murphy and Sen 1991; Grambsch and Therneau 1994; T. Martinussen, T. H. Scheike, and Skovgaard 2002; Cai and Sun 2003; Winnett and Sasieni 2003 and very recently Yan and Jian Huang 2012; T. Honda and Härdle 2014; T. Honda and Yabe 2017. In T. Martinussen and T. H Scheike 2007 may be found a complete presentation of the models, estimation methods and results. The R package `timereg`, see Appendix C in T. Martinussen and T. H Scheike 2007, implements these procedures.

These models are known extensions of the classical Aalen Aalen 1980 and Cox D. R. Cox 1972 models with constant regression parameters. Dynamic models are obviously more flexible than their constant counterparts, but they suffer from their complexities: p unknown functions are to be estimated from the data. We propose in the present paper a penalized procedure that reaches a compromise between these two extreme situations, and in addition performs variable selection.

Penalized piecewise constant estimators

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

$$\varphi_l = \sqrt{\frac{L}{\tau}} \mathbf{1}(I_l) \text{ and } I_l = \Big(\frac{l-1}{L}\tau, \frac{l}{L}\tau \Big]. \tag{4}$$

For all $j = 1, \dots, p$, candidates for the estimation of the j-th coordinate $\beta_j^\star$ of $\beta^\star$ belongs to the set of univariate piecewise constant functions

$$\mathcal{H}_L = \Big\{ \alpha(\cdot) = \sum_{l=1}^{L} \alpha_l \varphi_l(\cdot) : (\alpha_l)_{1 \le l \le L} \in \mathbb{R}_+^L \Big\}. \tag{5}$$

For moderate sample size n and/or high-dimensional covariates and/or a fine partition, the resulting estimators would suffer from over-parametrization, in the sense that $\sqrt{n}$ could be much lesser than $p \times L$. On the other hand, simpler forms of Cox and Aalen models, when the functions $\beta_j^\star$ are constant over $[0, \tau]$, are often too poor to accurate (see the discussions on page 205 and following in T. Martinussen and T. H Scheike 2007 and in Paragraph).

We here seek to reach a compromise between these two extreme situations by introducing a covariate specific weighted $\ell_1 + \ell_1$-total-variation penalty (defined in (13)). The total-variation part in the penalty induces simple, interpretable estimators, which do not vary much over the time. The ℓ_1 part allows our procedure to support high-dimensional (with a large p) covariates.

Our algorithms bears similarities with the class of fused Lasso algorithms. The latter have been introduced and studied, for noised piecewise constant signals, by R. Tibshirani et al. 2005, Rinaldo 2009, Harchaoui and Lévy-Leduc 2010, or Dalalyan, M.H., and Lederer 2014. A total-variation penalized estimator has been investigated in Alaya, Gaïffas, and Guilloux 2015 for estimating the intensity of a counting process, while Bouaziz and Guilloux 2015; Alaya, Gaïffas, and Guilloux 2017 proposed related estimators in a other context.

Lasso estimators in the context of survival analysis with high-dimensional covariates have been introduced and studied in T. Martinussen and T. H. Scheike 2009a; Gaïffas and Guilloux 2012 in the Aalen model and R. Tibshirani 1997; J. Huang et al. 2013; Lemler 2013 in the Cox model, among others.

The main contributions of the paper are the following.

- Theoretical: we propose new estimators, see the definitions in (14) and (15) for the problem at hand. We investigate their theoretical properties by proving oracle inequalities, stated in Section , that assure their convergences.
- Practical: we propose an algorithm, see Section , for computing our estimators in the dynamic models of Equations (1) and (2). We demonstrate in Section that they outperform existing algorithms in terms of estimation precision, variable selection and timings.

Estimation procedures

We describe in this section our novel estimation procedures, which involve a $\ell_1 + \ell_1$-total-variation penalization of criteria, specific to either the multiplicative or additive models. We first give more details on models (1) and (2).

Estimation

Estimation in traditional models with constant coefficients is based on a minimization of a (partial) least-square criterion in the usual Aalen Aalen 1980 model and a (partial) log-likelihood maximization in the usual Cox D. R. Cox 1972 model. We refer the reader to T. Martinussen and T. H. Scheike 2009a; Gaïffas and Guilloux 2012 and D. R. Cox 1975 for the details.

We now introduce some notations. For each individual i with a p-dimensional process of covariates X_i, we denote by X_i^j the process associated to its j-th covariate. Accordingly, for any p-dimensional function β, candidate for the estimation of $\beta^\star$, the univariate function β_j is its j-th coordinate. We define the sets of candidates for estimation as

$$\Lambda^{\mathrm{A}} = \{x, t \in [0,\tau] \mapsto \lambda^{\mathrm{M}}_\beta(t, x(t)) = x(t)\beta(t) \mid \forall j,\ \beta_j \in \mathcal{H}_L\}. \tag{6}$$

for the Aalen model (1) and

$$\Lambda^{\mathrm{M}} = \{x, t \in [0,\tau] \mapsto \lambda^{\mathrm{M}}_\beta(t, x(t)) = \exp\left(x(t)\beta(t)\right) \mid \forall j,\ \beta_j \in \mathcal{H}_L\}. \tag{7}$$

for the Cox model (2).

As for a candidate in Λ^{A} or Λ^{M}, each time-varying coefficient β_j is piecewise constant, we will refer equivalently to β as a p-dimensional function or as the vector of dimension $p \times L$

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

where $\beta_{j,\cdot}$ is in $\mathbb{R}^L$ and $\beta_{j,l}$ is the value taken by the j-th coordinate on the l-th time interval in our L-partition $\{I_1, \ldots, I_L\}$:

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

Estimation in the time varying Aalen and Cox models

The existing estimators in the additive (1) and multiplicative (2) models with time varying coefficients are also defined via respectively the least-squares and log-likelihood (see pages 108 and following and 206 and following in T. Martinussen and T. H Scheike 2007).

The time-varying Aalen model. For the time-varying Aalen model, we consider the least square criterion for our data and a candidate $\lambda_\beta^{\mathrm{A}}$ defined by

$$\ell_n^{\mathrm{A}}(\beta) = \frac{1}{n}\sum_{i=1}^n \Big\{ \int_0^\tau \left(\lambda_\beta^{\mathrm{A}}(t, X_i(t))\right)^2 Y_i(t)dt - 2\int_0^\tau \lambda_\beta^{\mathrm{A}}(t, X_i(t))dN_i(t)\Big\}, \quad (8)$$

see Gaïffas and Guilloux 2012 for details on this criterion. When the candidate $\lambda_\beta^{\mathrm{A}}$ is in the class Λ^{A}, Equation (8) simplifies to

$$\begin{aligned}\ell_n^{\mathrm{A}}(\beta) &= \frac{1}{n}\sum_{i=1}^n \Big\{ \int_0^\tau \Big(\sum_{j=1}^p X_i^j(t)\beta_j(t)\Big)^2 Y_i(t)dt - 2\int_0^\tau \sum_{j=1}^p X_i^j(t)\beta_j(t)dN_i(t)\Big\} \\ &= \frac{1}{n}\sum_{i=1}^n\sum_{l=1}^L \Big\{\frac{L}{\tau}\int_{I_l}\Big(\sum_{j=1}^p X_i^j(t)\beta_{j,l}\Big)^2 Y_i(t)dt - 2\sqrt{\frac{L}{\tau}}\sum_{j=1}^p \Big(\int_{I_l} X_i^j(t)dN_i(t)\Big)\beta_{j,l}\Big\}. \end{aligned} \quad (9)$$

The time-varying Cox model. We consider, in this paragraph, estimation in the time-varying Cox model. Minus the log-likelihood for our data and a candidate $\lambda_\beta^{\mathrm{M}}$ is given by

$$\ell_n^{\mathrm{M}}(\beta) = -\frac{1}{n}\sum_{i=1}^n \Big\{ \int_0^\tau \log\left(\lambda_\beta^{\mathrm{M}}(t, X_i(t))\right)dN_i(t) - \int_0^\tau Y_i(t)\lambda_\beta^{\mathrm{M}}(t, X_i(t))dt\Big\}, \quad (10)$$

see Andersen et al. 1993 for details. When $\lambda_\beta^{\mathrm{M}}$ is in the class Λ^{M}, the last expression reduces to

$$\begin{aligned}\ell_n^{\mathrm{M}}(\beta) &= -\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \sum_{j=1}^{p} X_i^j(t)\beta_j(t)dN_i(t) - \int_0^\tau Y_i(t)\exp\Big(\sum_{j=1}^{p} X_i^j(t)\beta_j(t)\Big)dt\Big\} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\Big\{\sqrt{\frac{L}{\tau}}\sum_{j=1}^{p}\Big(\int_{I_l} X_i^j(t)dN_i(t)\Big)\beta_{j,l} \\ &\qquad - \int_{I_l} Y_i(t)\exp\Big(\sqrt{\frac{L}{\tau}}\sum_{j=1}^{p} X_i^j(t)\beta_{j,l}\Big)dt\Big\}. \qquad (11)\end{aligned}$$

Estimation procedure. We introduce a well-chosen vector of data-driven weights $\hat{\gamma} = (\hat{\gamma}_{1,\cdot}^\top, \ldots, \hat{\gamma}_{p,\cdot}^\top)^\top$. The weights $\hat{\gamma}_{j,l} > 0$ are fully data-driven, and their shape is giving by

$$\hat{\gamma}_{j,l} = \mathcal{O}\Big(\sqrt{\frac{\log pL}{n}\hat{V}_{j,l}}\Big), \text{ with } \hat{V}_{j,l} = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=l}^{L}\frac{L}{\tau}\int_{I_u}(X_i^j(t))^2 dN_i(t). \quad (12)$$

We write here only the dominating terms, see Definition 1 in Supplementary Material for its explicit form. Our covariate specific weighted $\ell_1 + \ell_1$-total-variation penalty is defined by

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \sum_{j=1}^{p}\Big(\hat{\gamma}_{j,1}|\beta_{j,1}| + \sum_{l=2}^{L}\hat{\gamma}_{j,l}|\beta_{j,l} - \beta_{j,l-1}|\Big) \qquad (13)$$

for any $\beta \in \mathbb{R}^{pL}$. Our estimators are then respectively defined as $\hat{\lambda}^{\mathrm{A}} = \lambda^{\mathrm{A}}_{\hat{\beta}^{\mathrm{A}}}$ and $\hat{\lambda}^{\mathrm{M}} = \lambda^{\mathrm{M}}_{\hat{\beta}^{\mathrm{M}}}$, where

$$\hat{\beta}^{\mathrm{A}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{pL}}\Big\{\ell_n^{\mathrm{A}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}}\Big\}, \qquad (14)$$

in the Aalen model, and

$$\hat{\beta}^{\mathrm{M}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{pL}}\Big\{\ell_n^{\mathrm{M}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}}\Big\} \qquad (15)$$

in the Cox model.

Theoretical guaranties

In this section we address the statistical properties of the weighted $\ell_1 + \ell_1$-total-variation estimation procedure presented in the previous section. Our first results establish theoretical properties of our estimators by using the classical non-asymptotic oracle approaches.

Towards this end, we first introduce the weighted empirical quadratic norm $\|\lambda^{\mathrm{A}}\|_n$ defined for any $\lambda^{\mathrm{A}} \in \Lambda^{\mathrm{A}}$ by

$$\|\lambda^{\mathrm{A}}\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big(\lambda^{\mathrm{A}}(t, X_i(t))\big)^2 Y_i(t)dt},$$

and the empirical Kullback divergence $K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}})$ defined for $\lambda_\beta^{\mathrm{M}} \in \Lambda^{\mathrm{M}}$ by

$$\begin{aligned} K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big(\log \lambda_\star^{\mathrm{M}}(t, X_i(t)) - \log \lambda_\beta^{\mathrm{M}}(t, X_i(t))\Big)\lambda_\star^{\mathrm{M}}(t, X_i(t))Y_i(t)dt \\ - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big(\lambda_\star^{\mathrm{M}}(t, X_i(t)) - \lambda_\beta^{\mathrm{M}}(t, X_i(t))\Big)Y_i(t)dt. \end{aligned}$$

Theorem 1. *For $x > 0$ fixed, the estimator $\hat{\lambda}^{\mathrm{A}}$ defined in (14), verifies with a probability larger than $1 - C_{\mathrm{A}}e^{-x}$ for a some constant $C_{\mathrm{A}} > 0$,*

$$\|\lambda_\star^{\mathrm{A}} - \hat{\lambda}^{\mathrm{A}}\|_n^2 \leq \inf_{\beta \in \mathbb{R}^{pL}} \Big(\|\lambda_\star^{\mathrm{A}} - \lambda_\beta^{\mathrm{A}}\|_n^2 + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}}\Big). \tag{16}$$

Theorem 2. *For $x > 0$ fixed, the estimator $\hat{\lambda}^{\mathrm{M}}$ defined in (15), verifies with a probability larger than $1 - C_{\mathrm{M}}\mathrm{e}^{-x}$ for a some constant $C_{\mathrm{M}} > 0$,*

$$K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) \leq \inf_{\beta \in \mathbb{R}^{pL}} \Big(K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 2||\beta||_{\mathrm{gTV},\hat{\gamma}}\Big). \tag{17}$$

The proofs of Theorems 1 and 2 are presented respectively in Section 1 of Supplementary Material. Two terms are involved on the right hand side of (16) and (17). The first one measures how far are the true functions of interest $\lambda_\star^{\mathrm{A}}$ and $\lambda_\star^{\mathrm{M}}$ from their approximations on Λ^{A} and Λ^{M}. The second one can be viewed as a variance term that satisfies

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \leq \|\beta\|_{\mathrm{gTV}} \max_{j=1,\ldots,p} \max_{l=1,\ldots,L} \sqrt{\frac{\log pL}{n}\hat{V}_{j,\ell}} = \mathcal{O}\Big(\sqrt{\frac{\log(pL)}{n}}\Big). \tag{18}$$

for any $\beta \in \mathbb{R}^{pL}$. Here, $\|\cdot\|_{\mathrm{gTV}}$ stands for the unweighted $\ell_1 + \ell_1$-total-variation ($\hat{\gamma}_{j,l} = 1$). The dominant term in (18) is of order $\|\beta\|_{\mathrm{gTV}}\big(\log(pL)/n\big)^{1/2}$, which is the expected slow rate for $\hat{\lambda}^{\mathrm{A}}$ and $\hat{\lambda}^{\mathrm{M}}$ involving the total-variation penalization. Such oracle inequality is now classical in the huge literature of the sparsity procedures see for instance Bickel, Ritov, and A. B. Tsybakov 2009; Gaïffas and Guilloux 2012; Bunea, A. Tsybakov, and Wegkamp 2007; Van de Geer and Bühlmann 2009; Alaya, Gaïffas, and Guilloux 2015; Hansen, Reynaud-Bouret, and Rivoirard 2015. Most of these papers aim at establishing oracle inequalities under weak assumptions on the design matrix.

We establish in addition fast oracle inequalities for $\lambda_\star^{\mathrm{A}}$ and $\lambda_\star^{\mathrm{M}}$ respectively (see Theorems S1 and S2 in Supplementary Material), leading to variance terms of order with a variance term of order

$$\frac{\log(pL)}{n}.$$

In that sense, we improve the rate of convergence obtained in T. Honda and Härdle 2014, which is of order $\sqrt{pL/n}$ in L_2-norm in a basis of size L. Moreover, we established non-asymptotic oracle inequalities on the whole intensity, whereas the results in T. Honda and Härdle 2014; T. Honda and Yabe 2017 are only on the regression function $\beta(t)$. To our best knowledge, this is the only works that deal also with variable selection.

These fast oracle inequalities require additional assumptions of the *restricted eigenvalue* (RE) assumption type. In classical setting, the RE assumption excludes strong correlations between covariates and it was introduced in Bickel, Ritov, and A. B. Tsybakov 2009. In Van de Geer and Bühlmann 2009, one can find an exhaustive survey and comparison of the assumptions used to prove fast oracle inequalities.

Implementation and numerical experiments

Algorithms

For computing solutions (see Algorithm 2) of the regularized problems (14) and (15) respectively,we implemented the proximal gradient descent (PGD) algorithm (see Daubechies, Defrise, and De Mol 2004; Beck and Teboulle 2009; Bach et al. 2012; Parikh and Boyd 2014) via the fast iterative shrinkage-thresholding procedure Beck and Teboulle 2009. In Algorithm 2 below, we need the proximal operator of the weighted $\ell_1 + \ell_1$-total-variation, namely

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

Since $\|\cdot\|_{\mathrm{gTV},\hat{\gamma}}$ is separable by blocks, we have $\left(\mathrm{prox}_{\|\cdot\|_{\mathrm{gTV},\hat{\gamma}}}(\beta)\right)_{j,\cdot} = \mathrm{prox}_{\|\cdot\|_{\mathrm{gTV},\hat{\gamma}}}(\beta_{j,\cdot})$ for all $j = 1, \ldots, p$ (see Bach et al. 2012). Thus, we can focus on a single j-th block. Algorithm 1 expresses $\mathrm{prox}_{\|\cdot\|_{\mathrm{gTV},\hat{\gamma}}}(\beta)$ from the proximal operator of the weighted total-variation penalization Alaya, Gaïffas, and Guilloux 2015, namely $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{\gamma}}}$.

Proposition 1. *Algorithm 1 computes the proximal operator of the weighted $\ell_1 + \ell_1$-total-variation given by* (13).

Algorithm 1: Proximal operator of the weighted $\ell_1+\ell_1$-total-variation (see ())

input : vector $\beta \in \mathbb{R}^{p\times d}$ and weights $\hat{\gamma}_{j,l}$ for $j = 1, \dots, p$ and $l = 1, \dots, L$.

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

for $j = 1, \dots, p$ **do**

  $\eta_{j,\cdot} \leftarrow \text{prox}_{\|\cdot\|_{\text{TV},\hat{\omega}_{j,\cdot}}}(\beta_{j,\cdot})$, where $\hat{\omega}_{j,\cdot} = \hat{\gamma}_{j,\cdot}\backslash\{\hat{\gamma}_{j,1}\}$

  $\vartheta_{j,\cdot} \leftarrow \eta_{j,\cdot} - \eta_{j,1}\mathbf{1}_L$

  $\theta_{j,\cdot} \leftarrow \vartheta_{j,\cdot} + \eta_{j,1} \max\Big(1 - \frac{\hat{\gamma}_{j,1}}{L|\eta_{j,1}|}, 0\Big)\mathbf{1}_L$

return θ

Algorithm 2: PGD for the time-varying Aalen and Cox model, see (14) and (15)

1. Parameters: Integer $K > 0$;, function ℓ_n ($= \ell_n^{\text{A}}$ or $= \ell_n^{\text{M}}$)
2. Calculus of the Lipschitz constant L of the operator $\nabla\ell_n$;
3. Initialization: $(\hat{\beta})^{(0)} = 0 \in \mathbb{R}^{pL}$; $(\hat{\mu})^{(0)} = (\hat{\beta})^{(0)}$; and $t_1 = 1$;
4. **for** $k = 1, \dots, K$ **do**

   $\hat{\theta}^{(k)} \leftarrow \hat{\mu}^{(k)} - \frac{1}{L}\nabla\ell_n(\hat{\mu}^{(k)})$;

   $\hat{\beta}^{(k)} \leftarrow \text{prox}_{\frac{1}{L}\|\cdot\|_{\text{gTV},\hat{\gamma}}}(\hat{\theta}^{(k)})$;

   $t_{k+1} \leftarrow \frac{1+\sqrt{1+4t_k^2}}{2}$;

   $\mu^{(k+1)} \leftarrow \hat{\beta}^{(k)} + \big(\frac{t_k-1}{t_{k+1}}\big)\big(\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\big)$;
5. **return** $\hat{\beta}^{(K)}$

Details on the implementation

The data-driven weights of our algorithm are given in a compact form in Equation (12) (and in exact form in Definition S1 in Supplementary material. Following Equation (12), the weights used in practice are set to

$$\hat{\gamma}_{j,l} = \gamma\sqrt{\frac{\log pL}{n}}\frac{1}{n}\sum_{i=1}^{n}\sum_{u=l}^{L}\frac{L}{\tau}\int_{I_u}(X_i^j(t))^2 dN_i(t) \tag{19}$$

in our algorithm, where $\gamma > 0$ is a tunning parameter, which allows to modulate the strength of the penalty. We select the best value using grid search and a K-fold cross validation, based on the criterion at hand, e.g. the log-likelihood in the multiplicative Cox model.

Once a penalized estimator (refereed to as "CoxTV") has been computed, we compute each coefficient support and re-run an unpenalized estimation on these supports. The estimator associated to this re-estimation is refereed to as "On support" in what follows.

Simulated data in the time-varying Cox model

We carried out simulations in the time-varying Cox model for survival time with different sample sizes ($n = 100$, $n = 1000$) to compare our estimator to the splines-based group-lasso estimator of T. Honda and Härdle 2014 (refereed to as "H&H estimator") and to the kernels-based estimator of the `timereg` library T. Scheike, Martinussen, and Silver 2016 (refereed to as "`timereg` estimator"). The estimator of T. Honda and Härdle 2014 (whose code was kindly provided to us by T. Honda) is only implemented in the case where individuals experience only one event and have constant covariates, hence we restricted our simulations to this case. The implementation of our estimator however allows time-varying covariates and/or with repeated events (see Section 5 for an example).

The intensity of the process N_i for individual i ($i = 1, \dots, n$) is in this case given by

$$t \mapsto \lambda^\star(t, X_i(t))\mathbf{1}N_i(t) \leq 1,$$

and we set

$$\lambda^\star(t, X) = \exp(\beta_0^\star(t) + X\beta^\star(t)).$$

The covariates X_i, $i = 1, \dots, n$, are realizations of i.i.d. centered Gaussian random variables of dimension p, with variance proportional to the identity matrix. The true function $\beta^\star(t)$ is defined as:

- $\beta_0^\star(t) = \log\left(b^a a t^{a-1}\right)$ with $a = 1.2$ and $b = 0.25$, this corresponds to a Weibull baseline,
- $\beta_1^\star(t) = 0.18$,
- $\beta_2^\star(t) = 0.2\mathbf{1}(t \leq 1 + 0.05\mathbf{1}(t > 1)$,
- $\beta_3^\star(t) = 0.1t^2$,

and, for $j > 3$, we set $\beta_j^\star(t) = 0$. Given $t \mapsto \lambda_\star(t, X_i)$ and the covariates X_i, $i = 1, \dots, n$, the times T_i were simulated as the first event of a nonhomogeneous Poisson process with intensity $\lambda_\star(t, X_i)$ via thinning (see Lewis and Shedler 1979).

Our estimator was computed with the weights of Equation (19) and the tunning parameters was chosen via 3-fold cross-validation as described earlier. In addition, we computed a second estimation with no penalty, as described in Subsection 5.

To evaluate the performance of the three estimators, we run $M = 500$ Monte-Carlo experiments in the model described above. The estimation accuracy of each estimator is investigated via a mean squared error defined as

$$\mathrm{MSE}(\hat{\beta}^{\mathrm{M}}_{j,m}) = \frac{1}{M}\sum_{m=1}^{M}\int_0^\tau \left(\hat{\beta}^{\mathrm{M}}_{j,m}(t) - \beta^\star_j(t)\right)^2,$$

where $\hat{\beta}^{\mathrm{M}}_{j,m}$ is the estimation of $\beta^\star_j$ in the sample m, for $j = 1, \ldots, p$. The integrals are approximated on a grid of length 50 equispaced on $[0, \tau]$. We then considered a cumulative MSE defined as

$$\mathrm{MSE}(\hat{\beta}^{\mathrm{M}}_{m}) = \sum_{j=1}^{p}\mathrm{MSE}(\hat{\beta}^{\mathrm{M}}_{j,m}).$$

In addition, we evaluate the capacity of the estimators to detect no active covariates. To this end, we define the numbers of true/false positives, and true negatives as

$$\begin{aligned}
\mathrm{TP}(\hat{\beta}^{\mathrm{M}}_{m}) &= \#\{j = 1, \ldots, p : \hat{\beta}^{\mathrm{M}}_{j,m} \neq 0 \text{ and } \beta^\star_j \neq 0\}\\
\mathrm{FP}(\hat{\beta}^{\mathrm{M}}_{m}) &= \#\{j = 1, \ldots, p : \hat{\beta}^{\mathrm{M}}_{j,m} \neq 0 \text{ and } \beta^\star_j = 0\}\\
\mathrm{TN}(\hat{\beta}^{\mathrm{M}}_{m}) &= \#\{j = 1, \ldots, p : \hat{\beta}^{\mathrm{M}}_{j,m} = 0 \text{ and } \beta^\star_j = 0\}.
\end{aligned}$$

	CoxTV			H&H		
	TP	FP	TN	TP	FP	TN
$p = 5$	4.27	0.00	0.00	3.30	0.00	0.00
$p = 10$	2.38	0.98	4.02	4.20	3.00	2.00
$p = 50$	1.00	3.00	42.00	NA	NA	NA

Table 1: True/false positives, true negatives for $n = 100$, "NA" values indicate that the algorithm did not converge.

	CoxTV		
	TP	FP	TN
$p = 10$	4.47	1.92	3.08
$p = 50$	3.76	4.15	40.85
$p = 100$	3.61	6.78	88.22

Table 2: True/false positives, true negatives for $n = 1000$, "NA" values indicate that the algorithm did not converge.

	CoxTV	On support	H&H	`timereg`
$p = 5$	4.32	2.35	1.56	2.33
$p = 10$	4.70	2.53	1.45e+5	10.88
$p = 50$	4.93	4.39	NA	NA

Table 3: MSE for $n = 100$, "NA" values indicate that the algorithm did not converge.

	CoxTV	On support	H&H	`timereg`
$p = 10$	1.95	0.54	NA	0.62
$p = 50$	3.55	0.78	NA	4.56
$p = 100$	4.39	1.20	NA	23.81

Table 4: MSE for $n = 1000$, "NA" values indicate that the algorithm did not converge.

The results show that our penalized estimator performs well in selecting active variables and setting coefficients of non active variables to zero, see Tables 1 for a comparison with the estimator in T. Honda and Härdle 2014 in the situation where $n = 100$. For larger n, the selection performances are still very satisfactory, see Table 2. As a consequence, in situations where p reaches (or exceeds) $\sqrt{n}$, the "On support" version of our estimator outperforms both the estimator of T. Honda and Härdle 2014 and the one of T. Scheike, Martinussen, and Silver 2016 in terms of mean square error, see Tables 3 and 4.

We insist in addition that our method outperforms existing estimators, the ones of T. Honda and Härdle 2014 and T. Scheike, Martinussen, and Silver 2016 in terms of time of computation, as soon as n is larger than 100. In Table 5, we report timings for the three methods, notice that for our method (we consider the timings of the first fit and the re-fit as one estimation step)

	CoxTV + On support	H&H	Timereg
$n = 100$ and $p = 5$	11.05	19.69	0.71
$n = 1000$ and $p = 10$	1.63	NA	49.36

Table 5: Timings in minutes

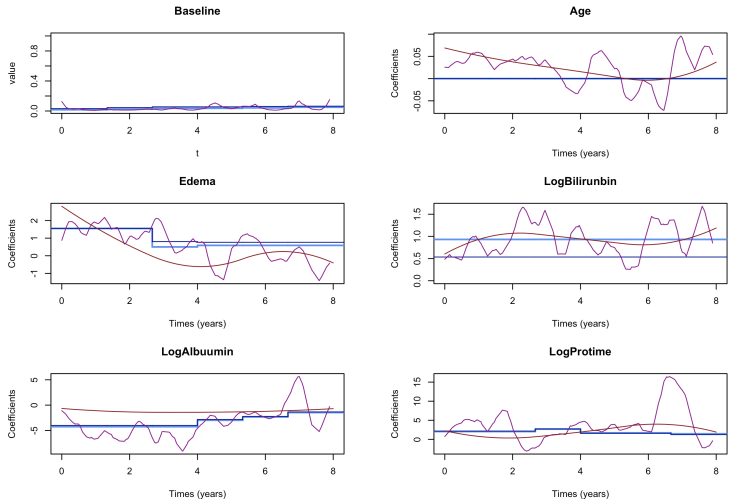

Figure 1: Estimated regression coefficients on PBC data: with "CoxTV" (dark blue), "On support" (light blue), "H&H"(brown) and "timereg" estimation (magenta).

Real data

Our method is illustrated on PBC dataset described in Fleming and Harrington 1991, and originates from a Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver and was conducted between 1974 and 1984. 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, and we consider the covariates: age, edema, log(bilirubin), log(albumin) and log(protime), as in the example 6.0.2 of T. Martinussen and T. H Scheike 2007. All covariates are centered around their averages. We estimated the regression coefficients on PBC data with the four methods: our methods ("CoxTV", and "On support"), the estimator of T. Honda and Härdle 2014 ("'H&H") and the `timereg` estimator(in R package timereg T. Scheike, Martinussen, and Silver 2016). The good performances of our estimators are illustrated in Figure 1. Our estimators and the one of T. Honda and Härdle 2014 ("'H&H") share the nice property of being smoother than the one obtained via R package timereg T. Scheike, Martinussen, and Silver 2016. The first methods hence give easily interpretable estimated coefficient. In this example, our method also performs variable selection (the coefficient of "age" is set as zero), which is not the case for the competing methods. These results give an illustration of the nice behavior of our estimators seen in Section .

References

[Aal80] O. O. Aalen. "A model for nonparametric regression analysis of counting processes". In: *Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisła, 1978)*. Vol. 2. Lecture Notes in Statist. Springer, New York-Berlin, 1980, pp. 1–25.

[Aal89] O. O. Aalen. "A linear regression model for the analysis of life times". In: *Statistics in medicine* 8.8 (1989), pp. 907–925.

[Aal93] O. O. Aalen. "Further results on the non-parametric linear regression model in survival analysis". In: *Statistics in medicine* 12.17 (1993), pp. 1569–1588.

[AFM14] Y. F. Atchade, G. Fort, and E. Moulines. "On stochastic proximal gradient algorithms". In: *arXiv preprint arXiv:1402.2365* (2014).

[AGG15] M. Z. Alaya, S. Gaïffas, and A. Guilloux. "Learning the Intensity of Time Events With Change-Points". In: *Information Theory, IEEE Transactions on* 61.9 (2015), pp. 5148–5171. ISSN: 0018-9448. DOI: 10.1109/TIT.2015.2448087.

[AGG17] M. Z. Alaya, S. Gaïffas, and A. Guilloux. "Binarsity: a Penalization for One-Hot Encoded Features". In: *preprint* (2017).

[And+93] P. K. Andersen et al. *Statistical models based on counting processes*. Springer Series in Statistics. New York: Springer-Verlag, 1993, pp. xii+767. ISBN: 0-387-97872-0. DOI: 10.1007/978-1-4612-4348-9. URL: http://dx.doi.org/10.1007/978-1-4612-4348-9.

[Bac+12] F. Bach et al. "Optimization with sparsity-inducing penalties". In: *Foundations and Trends® in Machine Learning* 4.1 (2012), pp. 1–106.

[Bak+04] G Ross Baker et al. "The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada". In: *Canadian medical association journal* 170.11 (2004), pp. 1678–1686.

[Bel+10] Carine A Bellera et al. "Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer". In: *BMC medical research methodology* 10.1 (2010), p. 20.

[BG15] O. Bouaziz and A. Guilloux. "A penalized algorithm for event-specific rate models for recurrent events". In: *Biostatistics* 16.2 (2015), pp. 281–294. DOI: 10.1093/biostatistics/kxu046.

[Bot10] L. Bottou. “Large-scale machine learning with stochastic gradient descent”. In: *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[Bot12] L. Bottou. “Stochastic Gradient Descent Tricks.” In: *Neural Networks: Tricks of the Trade (2nd ed.)* Ed. by Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller. Vol. 7700. Lecture Notes in Computer Science. Springer, 2012, pp. 421–436. ISBN: 978-3-642-35288-1.

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. “Simultaneous analysis of lasso and Dantzig selector”. In: *Ann. Statist.* 37.4 (2009), pp. 1705–1732. ISSN: 0090-5364. DOI: 10.1214/08-AOS620. URL: http://dx.doi.org/10.1214/08-AOS620.

[BT09] A. Beck and M. Teboulle. “A fast iterative shrinkage-thresholding algorithm for linear inverse problems”. In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202.

[BTW07] F. Bunea, A. Tsybakov, and M. Wegkamp. “Sparsity oracle inequalities for the Lasso”. In: *Electron. J. Statist.* 1 (2007), pp. 169–194. DOI: 10.1214/07-EJS008. URL: http://dx.doi.org/10.1214/07-EJS008.

[Che+14] Ming-Yen Cheng et al. “Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data”. In: *The Annals of Statistics* 42.5 (2014), pp. 1819–1849.

[Col+11] Preciosa M Coloma et al. “Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project”. In: *Pharmacoepidemiology and drug safety* 20.1 (2011), pp. 1–11.

[Cox72] D. R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246. URL: http://www.jstor.org/stable/2985181.

[Cox75] D. R. Cox. “Partial likelihood”. In: *Biometrika* 62.2 (1975), pp. 269–276. DOI: 10.1093/biomet/62.2.269. eprint: http://biomet.oxfordjournals.org/content/62/2/269.full.pdf+html. URL: http://biomet.oxfordjournals.org/content/62/2/269.abstract.

[CS03] Z. Cai and Y. Sun. “Local Linear Estimation for Time-Dependent Coefficients in Cox's Regression Models”. In: *Scandinavian Journal of Statistics* 30.1 (2003), pp. 93–111.

[DDD04] I. Daubechies, M. Defrise, and C. De Mol. "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint". In: *Communications on Pure and Applied Mathematics* 57.11 (2004), pp. 1413–1457. ISSN: 1097-0312. DOI: 10.1002/cpa.20042. URL: http://dx.doi.org/10.1002/cpa.20042.

[DML14] A. S. Dalalyan, M.H., and J. Lederer. *On the Prediction Performance of the Lasso*. Bernoulli 1402.1700. arXiv, 2014, pp. 1–30. URL: http://arxiv.org/pdf/1402.1700v1.pdf.

[FH91] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991, pp. xiv+429. ISBN: 0-471-52218-X.

[GG12] S. Gaïffas and A. Guilloux. "High-dimensional additive hazards models and the Lasso". In: *Electron. J. Stat.* 6 (2012), pp. 522–546. ISSN: 1935-7524. DOI: 10.1214/12-EJS681. URL: http://dx.doi.org/10.1214/12-EJS681.

[GT94] P. M. Grambsch and T. M. Therneau. "Proportional hazards tests and diagnostics based on weighted residuals". In: *Biometrika* 81.3 (1994), pp. 515–526.

[He+16] Kevin He et al. "Modeling Time-varying Effects with Large-scale Survival Data: An Efficient Quasi-Newton Approach". In: *Journal of Computational and Graphical Statistics* just-accepted (2016).

[HH14] T. Honda and W. K. Härdle. "Variable selection in Cox regression models with varying coefficients". In: *Journal of Statistical Planning and Inference* 148 (2014), pp. 67–81.

[HL10] Z. Harchaoui and C. Lévy-Leduc. "Multiple change-point estimation with a total variation penalty". In: *J. Amer. Statist. Assoc.* 105.492 (2010), pp. 1480–1493. ISSN: 0162-1459. DOI: 10.1198/jasa.2010.tm09181. URL: http://dx.doi.org/10.1198/jasa.2010.tm09181.

[HM91] F. W. Huffer and I. W. McKeague. "Weighted least squares estimation for Aalen's additive risk model". In: *Journal of the American Statistical Association* 86.413 (1991), pp. 114–129.

[HRR15] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. "Lasso and probabilistic inequalities for multivariate point processes". In: *Bernoulli* 21.1 (2015), pp. 83–143. DOI: 10.3150/13-BEJ562. URL: http://dx.doi.org/10.3150/13-BEJ562.

[HSN08] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. "Definition, structure, content, use and impacts of electronic health records: a review of the research literature". In: *International journal of medical informatics* 77.5 (2008), pp. 291–304.

[Hua+13] J. Huang et al. "Oracle inequalities for the lasso in the Cox model". In: *Ann. Statist.* 41.3 (2013), pp. 1142–1165.

[HY17] T. Honda and R. Yabe. "Variable selection and structure identification for varying coefficient Cox models". In: *Journal of Multivariate Analysis* 161 (2017), pp. 103–122.

[Kal+06] K Kalantar-Zadeh et al. "Survival predictability of time-varying indicators of bone disease in maintenance hemodialysis patients". In: *Kidney international* 70.4 (2006), pp. 771–780.

[Lem13] S. Lemler. "Oracle inequalities for the lasso in the high-dimensional multiplicative Aalen intensity model". In: *Les Annales de l'Institut Henri Poincaré, arXiv preprint* (2013).

[LS79] P. A. W Lewis and G. S. Shedler. "Simulation of nonhomogeneous poisson processes by thinning". In: *Naval Research Logistics Quarterly* 26.3 (1979), pp. 403–413. ISSN: 1931-9193. DOI: 10.1002/nav.3800260304. URL: http://dx.doi.org/10.1002/nav.3800260304.

[McK88] I. W. McKeague. "Asymptotic Theory for Weighted Least Squares Estimators in". In: *Statistical Inference from Stochastic Processes: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference Held August 9-15, 1987, with Support from the National Science Foundation and the Army Research Office*. Vol. 80. American Mathematical Soc. 1988, p. 139.

[MS07] T. Martinussen and T. H Scheike. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007.

[MS09a] T. Martinussen and T. H. Scheike. "Covariate selection for the semiparametric additive risk model". In: *Scand. J. Statist.* 36.4 (2009), pp. 602–619. ISSN: 0303-6898. DOI: 10.1111/j.1467-9469.2009.00650.x. URL: https://acces-distant.upmc.fr:443/http/dx.doi.org/10.1111/j.1467-9469.2009.00650.x.

[MS09b] T. Martinussen and T. H. Scheike. "The additive hazards model with high-dimensional regressors". In: *Lifetime Data Anal.* 15.3 (2009), pp. 330–342. ISSN: 1380-7870. DOI: 10.1007/s10985-009-9111-y. URL: http://dx.doi.org/10.1007/s10985-009-9111-y.

[MS91] Susan Allbritton Murphy and Pranab Kumar Sen. "Time-dependent coefficients in a Cox-type regression model". In: *Stochastic Processes and their Applications* 39.1 (1991), pp. 153–180.

[MS99] Torben Martinussen and Thomas H Scheike. "A semiparametric additive regression model for longitudinal data". In: *Biometrika* (1999), pp. 691–702.

[MSS02] T. Martinussen, T. H. Scheike, and I. M. Skovgaard. "Efficient Estimation of Fixed and Time-varying Covariate Effects in Multiplicative Intensity Models". In: *Scandinavian Journal of Statistics* 29.1 (2002), pp. 57–74.

[PB14] N. Parikh and S. P. Boyd. "Proximal Algorithms." In: *Foundations and Trends in optimization* 1.3 (2014), pp. 127–239.

[PCH06] Aris Perperoglou, Saskia le Cessie, and Hans C van Houwelingen. "A fast routine for fitting Cox models with time varying effects of the covariates". In: *Computer methods and programs in biomedicine* 81.2 (2006), pp. 154–161.

[Rin09] A. Rinaldo. "Properties and refinements of the fused lasso". In: *Ann. Statist.* 37.5B (2009), pp. 2922–2952. ISSN: 0090-5364. DOI: 10.1214/08-AOS665. URL: http://dx.doi.org/10.1214/08-AOS665.

[RVV] L. Rosasco, S. Villa, and B. C. Vu. "Learning with stochastic proximal gradient". In: ().

[SMS16] T Scheike, T Martinussen, and J Silver. "Timereg: timereg package for flexible regression models for survival data". In: *R package version* (2016), pp. 1–9.

[Tib+05] R. Tibshirani et al. "Sparsity and smoothness via the fused lasso". In: *J. R. Statist. Soc. Ser. B Statist. Methodol.* 67.1 (2005), pp. 91–108. ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2005.00490.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2005.00490.x.

[Tib14] Ryan J Tibshirani. "ATTENTION". In: *The Annals of Statistics* 42.1 (2014), pp. 285–323.

[Tib97] R. Tibshirani. "The lasso method for variable selection in the Cox model". In: *Statist. Med.* 16 (1997), pp. 385–395.

[TZW05] Lu Tian, David Zucker, and LJ Wei. "On the Cox model with time-varying regression coefficients". In: *Journal of the American statistical Association* 100.469 (2005), pp. 172–183.

[VB09] S. A. Van de Geer and P. Bühlmann. “On the conditions used to prove oracle results for the Lasso”. In: *Electron. J. Statist.* 3 (2009), pp. 1360–1392. DOI: 10.1214/09-EJS506. URL: http://dx.doi.org/10.1214/09-EJS506.

[WS03] A. Winnett and P. Sasieni. “Iterated residuals and time-varying covariate effects in Cox regression”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 473–488.

[YH12] Jun Yan and Jian Huang. “Model Selection for Cox Models with Time-Varying Coefficients”. In: *Biometrics* 68.2 (2012), pp. 419–428.

[ZK90] D. M. Zucker and A. F. Karr. “Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach”. In: *The Annals of Statistics* (1990), pp. 329–353.