Electronic Journal of Statistics
Vol. 6 (2012) 522–546
ISSN: 1935-7524
DOI: 10.1214/12-EJS681

High-dimensional additive hazards models and the Lasso

Stéphane Gaïffas[*]

Université Pierre et Marie Curie - Paris 6, Laboratoire de Statistique Théorique et Appliquée
e-mail: stephane.gaiffas@upmc.fr

and

Agathe Guilloux[*]

Université Pierre et Marie Curie - Paris 6, Laboratoire de Statistique Théorique et Appliquée
and Centre de Recherche Saint-Antoine (UMR S 893)
e-mail: agathe.guilloux@upmc.fr

Abstract: We consider a general high-dimensional additive hazards model in a non-asymptotic setting, including regression for censored-data. In this context, we consider a Lasso estimator with a fully data-driven ℓ_1 penalization, which is tuned for the estimation problem at hand. We prove sharp oracle inequalities for this estimator. Our analysis involves a new "data-driven" Bernstein's inequality, that is of independent interest, where the predictable variation is replaced by the optional variation.

AMS 2000 subject classifications: Primary 62N02; secondary 62H12.
Keywords and phrases: Survival analysis, counting processes, censored data, Aalen additive model, Lasso, high-dimensional covariates, data-driven Bernstein's inequality.

Received September 2011.

1. Introduction

Recent interests have grown on connecting gene expression profiles to survival patients' times, see e.g. [30, 34], where the aim is to assess the influence of gene expressions on the survival outcomes. The statistical analysis of such data faces two sorts of problems. First, the covariates are high-dimensional: the number of covariates is much larger than the number of observations. Second, the survival outcomes suffers from censoring, truncation, etc. The need of proper statistical methods to analyze such data, in particular high-dimensional right-censored data, led in the past years to numerous theoretical and computational contributions.

When the survival times suffer from right-censoring, the problem can be presented as follows. For an individual $i \in \{1, \ldots, n\}$, let T_i be the time of

[*] This work is supported by French Agence Nationale de la Recherche (ANR) ANR Grant "Prognostic" ANR-09-JCJC-0101-01.

interest (e.g. the patient survival time), let C_i be the censoring time and X_i be the vector of covariates in $\mathbb{R}^d$, assumed to be independent copies of T, C and $X = (X^1, \dots, X^d)$. We observe $Z_i = T_i \wedge C_i$, $\delta_i = \mathbf{1}(T_i \leq C_i)$ and X_i for $i = 1, \dots, n$.

The covariates vector X, where both genomic outcomes and clinical data may be recorded, is in high dimension $d \gg n$ and influences the distribution of T via its conditional hazard rate given $X = x$, defined by

$$\alpha_0(t, x) = \frac{f_{T|X}(t, x)}{1 - F_{T|X}(t, x)}$$

for $t > 0$, where $f_{T|X}$ and $F_{T|X}$ are respectively the conditional density and distribution functions of T given $X = x$. In the following, we assume that the conditional hazard fulfills the Aalen additive hazards model [1]:

$$\alpha_0(t, x) = \lambda_0(t) + x^\top \beta_0, \quad \forall t \geq 0,$$

where λ_0 is the baseline hazard function and β_0 measures the influence of the covariates on the conditional hazard function α_0. In [21], an additive hazards model is fitted to investigate the influence of the expression levels of 8810 genes on the (censored) survival times of 92 patients suffering from Mantel-Cell Lymphoma, see [30] for the data. The Aalen additive hazards model is indeed an useful alternative to the Cox model [10], in particular in situations where the proportional hazards assumption is violated. It can also "be seen as a first-order Taylor series expansion of a general intensity" (see [23], p. 103).

When the aim is then to understand the influence of X on the survival time T, one wants to estimate β_0 based on the observations. In small dimension $d \ll n$ and from the data $(Z_i, \delta_i, X_i)_{i=1,\dots,n}$, the least-squares estimator $\hat{\beta}$ of the unknown β_0 is the minimizer of the quadratic functional

$$R_n(\beta) = \beta^\top \mathbf{H}_n \beta - 2\beta^\top \boldsymbol{h}_n,$$

where $\mathbf{H}_n$ is the $d \times d$ symetrical positive semidefinite matrix with entries

$$(\mathbf{H})_{j,k} = \frac{1}{n} \sum_{i=1}^n \int_0^{Z_i} \Big(X_i^j - \frac{\sum_{l=1}^n X_l^j \mathbf{1}(Z_l \geq t)}{\sum_{l=1}^n \mathbf{1}(Z_l \geq t)} \Big) \Big(X_i^k - \frac{\sum_{l=1}^n X_l^k \mathbf{1}(Z_l \geq t)}{\sum_{l=1}^n \mathbf{1}(Z_l \geq t)} \Big) dt,$$

and where $\boldsymbol{h}_n \in \mathbb{R}^d$ has coordinates

$$(\boldsymbol{h}_n)_j = \frac{1}{n} \sum_{i=1}^n \delta_i \Big(X_i^j - \frac{\sum_{k=1}^n X_k^j \mathbf{1}(Z_k \geq Z_i)}{\sum_{k=1}^n \mathbf{1}(Z_k \geq Z_i)} \Big).$$

When $d \leq n$ and if $\mathbf{H}_n$ is full rank, we can write

$$\hat{\beta} = (\mathbf{H}_n)^{-1} \boldsymbol{h}_n,$$

see also [19] or [25]. The estimator $\hat{\beta}$ is $\sqrt{n}$-consistent and asymptotically Gaussian, see e.g. [2].

When X contains genomic outcomes, one typically has $d \gg n$, and the matrix $\mathbf{H}_n$ is no longer of full rank. A sparsity assumption is then natural in this setting: we expect only a few genes to have an influence on the survival times, so we expect β_0 to be sparse, which means that it has only a few non-zero coordinates. Several papers use sparsity inducing penalization in the context of survival analysis, mainly for the Cox multiplicative risks model or the Aalen additive risks model, we refer to [35] for a review. Most procedures are based on ℓ_1-penalization, where one considers

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \Big\{ R_n(\beta) + \lambda \sum_{j=1}^d w_j |\beta_j| \Big\}. \tag{1}$$

The smoothing parameter $\lambda > 0$ makes the balance between goodness-of-fit and sparsity, and the $w_j \geq 0$, $j = 1, \ldots, d$ are weights allowing for a precise tuning of the penalization. The Lasso penalization corresponds to the simple choice $w_j = 1$, while in the adaptive Lasso [38] one chooses $w_j = |\tilde{\beta}_j|^{-\gamma}$ where $\tilde{\beta}_j$ is a preliminary estimator and $\gamma > 0$ a constant. The idea behind this is to correct the bias of the Lasso in terms of variable selection accuracy, see [38] and [37] for regression analysis. The weights w_j can also be used to scale each variable at the same level, which is suitable when some variable has a strong variance compared to the others. As a by-product of the theoretical analysis given in this paper, we introduce a new way of scaling the variables using data-driven weights $\hat{w}_j$ in the ℓ_1 penalization, see (14) below.

In the Cox proportional hazards model, $R_n(\beta)$ is the partial likelihood (see e.g. [10] or [2]), for which the Lasso, adaptive Lasso, smooth clipped absolute deviation penalizations and the Dantzig selector are considered, respectively, in [31, 39, 36, 12] and [3].

For the additive risks, [22] considers principal component regression, [21] considers a Lasso with a least-squares criterion that differs from the one considered here, [18, 25] considers the ridge, Lasso and adaptive Lasso penalizations and [24] considers the partial least-squares and ridge regression estimators.

A serious advantage, from the computational point of view, in using additive risks over multiplicative risks has to be highlighted. Indeed, for the additive risks, the estimating Equation (1) has a least-squares form, so that one can apply in this case the fast Lars algorithm [11] in order to obtain the whole path of solutions of the Lasso, as explained in [18] for instance. This point is particularly relevant in practice, since one typically uses splitting techniques, such as cross-validation, to select the smoothing parameter, or ensemble feature methods, such as stability selection [27], to select covariates. The motivations and main contributions of this work are enumerated in the following.

First motivation. Among the papers that propose some mathematical analysis of the statistical properties of estimators of the form (1) (upper bounds, support recovery, etc.), the results are asymptotic in the number of observations. This can be a problem since, in practice, one can not, in general, consider that the asymptotic regime has been reached: in [30], for example, the expression levels of 8810 genes and survival information are measured for only 92 patients. Considering only the references that are the closest to the work proposed here, the oracle property for the adaptive Lasso is given in [18], which is an asymptotic property about the support and the asymptotic distribution of the estimator, and asymptotic normality and consistency in variable selection for the adaptive Lasso is proved in [25], where results about the Dantzig selector are also derived using the restricted isometry property and the uniform uncertainty principle from [8]. While non-asymptotic results, like sparse oracle inequalities for instance, are now well-known for regression or density estimation (see for instance [7, 5, 4], among many others), such results are not yet available for survival data. In this paper, we establish the first results of this kind for survival analysis.

Second motivation. We give sharp oracle inequalities (with leading constant 1) for the prediction error associated to the survival problem. The results are stated for general counting processes, including the censoring case, while most papers consider censored data only. Our results are stated without the assumption that the intensity is linear in the covariates. In fact, our Lasso estimator can be computed using an arbitrary dictionary of functions, so that one can expect a better approximation of the true underlying intensity.

Third motivation. In order to prove our results, we need a new version of Bernstein's inequality for martingales with jumps, where the predictable variation, which is not observable in this problem, is replaced by the optional variation, which is observable. This concentration inequality is of independent interest, and could be useful for other statistal problems as well.

Fourth motivation. Finally, and more importantly, our non-asymptotic analysis leads to an adaptive data-driven weighting of the ℓ_1-norm, that involves the optional variation of each element of the dictionary (or of each covariate in the linear case). More precisely, our sharp control of the noise term exhibits the fact that the ℓ_1-penalization (see (1)) should be scaled using data-driven weights of order (writing only the dominating terms, see Section 3 for details)

$$\hat{w}_j \approx \sqrt{\frac{x + \log d}{n} \hat{V}_j},$$

where

$$\hat{V}_j = \frac{1}{n} \sum_{i=1}^{n} \delta_i \Big(X_i^j - \frac{\sum_{k=1}^n X_k^j \mathbf{1}(Z_k \geq Z_i)}{\sum_{k=1}^n \mathbf{1}(Z_k \geq Z_i)} \Big)^2$$

corresponds, roughly, to an estimate of the variance of variable j. Hence, our theoretical analysis exhibits a new way of tuning the ℓ_1 penalization, by multiplying each coordinate by this empirical variance term, in order to make less apparent eventual differences between the variability of each X^j for $j = 1, \ldots, d$. This particular form of weighting, or scaling of the variables, was not previsouly noticed in literature.

The paper is organized as follows. Section 2 describes the model. The Lasso estimator is constructed in Section 3. Oracle inequalities for the Lasso are given in Section 4, see Theorems 1 and 2. Some details about the construction of the least-squares criterion are given in Section 6.1. The data-driven Bernstein's inequality is stated in Section 5, see Theorem 3, and the proofs of our results are given in Section 6.

2. High dimensional Aalen model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ a filtration satisfying the usual conditions: increasing, right-continuous and complete (see [14]). Let N be a marked counting process with compensator Λ with respect to $(\mathcal{F}_t)_{t\geq 0}$, so that $M = N - \Lambda$ is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. We assume that N is a marked point process satisfying the *Aalen multiplicative intensity model*. This means that Λ writes

$$\Lambda(t) = \int_0^t \alpha_0(s, X) Y_s ds \tag{2}$$

for all $t \geq 0$, where:

- the intensity α_0 is an unknown deterministic and nonnegative function called *intensity*
- $X \in \mathbb{R}^d$ is a $\mathcal{F}_0$-measurable random vector called *covariates* or *marks*;
- Y is a predictable random process in $[0, 1]$.

With differential notations, this model can be written has

$$dN_t = \alpha_0(t, X) Y_t \, dt + dM_t \tag{3}$$

for all $t \geq 0$ with the same notations as before, and taking $N_0 = 0$. Now, assume that we observe n i.i.d. copies

$$D_n = \{(X_i, N_t^i, Y_t^i) : t \in [0, \tau], 1 \leq i \leq n\} \tag{4}$$

of $\{(X, N_t, Y_t) : t \in [0, \tau]\}$, where τ is the end-point of the study. Without loss of generality, we set $\tau = 1$. We can write

$$dN_t^i = \alpha_0(t, X_i) Y_t^i dt + dM_t^i$$

for any $i = 1, \ldots, n$ where M^i are independent $(\mathcal{F}_t)_{t\geq 0}$-martingales. In this setting, the random variable N_t^i is the number of observed failures during the time interval $[0, t]$ of the individual i. This model encompasses several particular examples: censored data, marked Poisson processes and Markov processes, see e.g. [2] for a precise exposition. In the censored case, described in the Introduction, the random processes in D_n, see Equation (4), are given by

$$N^i(t) = \mathbf{1}(Z_i \leq t, \delta_i = 1) \text{ and } Y^i(t) = \mathbf{1}(Z_i \geq t)$$

for $i = 1, \ldots, n$ and $0 \leq t \leq 1$.

In this paper, we assume that the intensity function satisfies the Aalen additive model in the sense that it writes

$$\alpha_0(t, x) = \lambda_0(t) + h_0(x), \tag{5}$$

where $\lambda_0 : \mathbb{R}_+ \to \mathbb{R}_+$ is a nonparametric *baseline* intensity and $h_0 : \mathbb{R}^d \to \mathbb{R}_+$. Note that in the "usual" Aalen additive model, see [19, 26, 24, 25], the function h_0 is linear:

$$h_0(x) = x^\top \beta_0,$$

where β_0 is an unknown vector in $\mathbb{R}^d$. The aim of the paper is to recover the function h_0 based on the observation of the sample D_n.

3. Construction of an ℓ_1-penalization procedure

3.1. A least-squares type functional

The problem considered here is a regression problem: we want to explain the influence of the covariates X_i on the survival data N^i and Y^i. Namely, we want to infer on h_0, while the baseline function λ_0 is considered as a nuisance parameter. Thanks to the additive structure (5), we can construct an estimator of h_0 without any estimation of λ_0, so that the influence of the covariates on the survival data can be infered without any knowledge on λ_0. This classical principle leads to the construction of the partial likelihood in the Cox model (multiplicative risks, see [10]) and to the construction of the "partial" least-squares (in reference to the partial likelihood) for the additive risks, see [19], which is the one considered here. The "partial least-squares" criterion for a "covariate" function $h : \mathbb{R}^d \to \mathbb{R}^+$ is defined as:

$$h \mapsto \frac{1}{n}\sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))^2 Y_t^i dt - \frac{2}{n}\sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t)) dN_t^i, \qquad (6)$$

where

$$\bar{h}_Y(t) = \frac{\sum_{i=1}^n h(X_i) Y_t^i}{\sum_{i=1}^n Y_t^i}.$$

It has been first introduced in [19]. The main steps leading to (6) are described in Section 6.1 below, where we explain why it is indeed suitable for the estimation of h_0 (see in particular Equation (20)).

Now, we consider a set

$$\mathcal{H} = \{h_1, \dots, h_M\}$$

of functions $h_j : \mathbb{R}^M \to \mathbb{R}^+$, called *dictionary*, where M is large ($M \gg n$). The set $\mathcal{H}$ can be a collection of basis functions, that can approximate the unknown h, like wavelets, splines, kernels, etc. They can be also estimators computed using an independent training sample, like several estimators computed using different tuning parameters, leading to the so-called aggregation problem, see [6] for instance. Implicitely, it is assumed that the unknown h_0 is well-approximated by a linear combination

$$h_\beta(x) = \sum_{i=1}^M \beta_j h_j(x), \qquad (7)$$

where $\beta \in \mathbb{R}^M$ is to be estimated. However, note that we won't assume, for the statements of our results, that the unknown h_0 is equal to h_{β_0} for some unknown $\beta_0 \in \mathbb{R}^M$, hence allowing for a model bias. Note that the setting considered here includes the linear case: if $h_j(x) = x_j$ with $d = M$, then the estimator has the form $\hat{h}(x) = x^\top \hat{\beta}$. Introducing

$$\bar{h}_{j,Y}(t) = \frac{\sum_{i=1}^n h_j(X_i) Y_t^i}{\sum_{i=1}^n Y_t^i} \quad \text{and} \quad \bar{h}_{\beta,Y}(t) = \sum_{j=1}^M \beta_j \bar{h}_{j,Y}(t), \tag{8}$$

we define the least-squares risk of $\beta \in \mathbb{R}^M$ as

$$R_n(\beta) = \frac{1}{n} \sum_{i=1}^n \int_0^1 (h_\beta(X_i) - \bar{h}_{\beta,Y}(t))^2 Y_t^i dt - \frac{2}{n} \sum_{i=1}^n \int_0^1 (h_\beta(X_i) - \bar{h}_{\beta,Y}(t)) dN_t^i, \tag{9}$$

which is equal to the functional (6) where we applied (7). Note that (9) is a least-squares criterion, since

$$R_n(\beta) = \beta^\top \mathbf{H}_n \beta - 2\beta^\top \boldsymbol{h}_n, \tag{10}$$

where $\mathbf{H}_n$ is the $M \times M$ matrix with entries

$$(\mathbf{H})_{j,k} = \frac{1}{n} \sum_{i=1}^n \int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t))(h_k(X_i) - \bar{h}_{k,Y}(t)) Y_t^i dt, \tag{11}$$

and where $\boldsymbol{h}_n \in \mathbb{R}^M$ has coordinates

$$(\boldsymbol{h}_n)_j = \frac{1}{n} \sum_{i=1}^n \int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t)) dN_t^i.$$

Since $\mathbf{H}_n$ is a symetrical positive semidefinite matrix, we can take

$$\mathbf{G}_n = \mathbf{H}_n^{1/2},$$

so that

$$R_n(\beta) = |\mathbf{G}_n \beta|_2^2 - 2\beta^\top \boldsymbol{h}_n,$$

where $|x|_2$ stands for the ℓ_2-norm of $x \in \mathbb{R}^n$. Note that we will denote by $|x|_p$ the ℓ_p norm of x.

3.2. ℓ_1-penalization for the Aalen model

For the problem considered here, we have seen that the empirical risk R_n has to be chosen with care. This is also the case for the ℓ_1 penalization to be used for this problem. Namely, for a well-chosen sequence of positive data-driven weights $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_M)$, we consider the weighted ℓ_1-norm

$$\text{pen}(b) = |b|_{1,\hat{w}} = \sum_{j=1}^M \hat{w}_j |b_j|, \tag{12}$$

and choose $\hat{\beta}$ according to the following penalized criterion:

$$\hat{\beta}_n \in \operatorname*{argmin}_{b \in B} \Big\{ R_n(b) + \text{pen}(b) \Big\} \tag{13}$$

where B is an arbitrary convex set (typically $B = \mathbb{R}^M$ or $B = \mathbb{R}^M_+$, the latter making $h_{\hat{\beta}_n}$ non-negative). The weights considered in (13) are given by $\hat{w}_j = \hat{w}(h_j)$ (where we recall that $h_j \in \mathcal{H}$) and where for any function h, we take

$$\hat{w}(h) = c_1 \sqrt{\frac{x + \log M + \hat{\ell}_{n,x}(h)}{n} \hat{V}(h)} + c_2 \frac{x + 1 + \log M + \hat{\ell}_{n,x}(h)}{n} \|h\|_{n,\infty}, \tag{14}$$

where:

- $x > 0$ and $c_1 = 2\sqrt{2}$, $c_2 = 4\sqrt{14/3} + 2/3$,
- $\|h\|_{n,\infty} = \max_{i=1,\dots,n} |h(X_i)|$,
- $\hat{V}(h)$ is a term corresponding to the "observable empirical variance" of h (see below for details), given by

$$\hat{V}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))^2 dN_t^i,$$

- $\hat{\ell}_{n,x}(h)$ is a small technical term coming out of our analysis:

$$\hat{\ell}_{n,x}(h) = 2 \log\log \Big(\frac{6en\hat{V}(h) + 56x\|h\|^2_{n,\infty}}{24x\|h\|^2_{n,\infty}} \vee e \Big).$$

Note that the weights $\hat{w}_j$ are fully data-driven. The shape of these weights comes from a new empirical Bernstein's inequality involving the optional variation of the noise process of the model, see Theorem 3 in Section 5 below.

The penalization (12) is tuned for the estimation problem at hand. It uses the estimator $\hat{V}(h)$ of the (unobservable) predictable quadratic variation

$$V(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))^2 \alpha_0(t, X_i) Y_t^i dt,$$

and it does not depend on an uniform upper bound for h. As a consequence, it can give, from a practical point of view, some insight into the tuning of the ℓ_1-penalization. In particular, our analysis prove that the j-th coordinate of β in the ℓ_1 penalization should be rescaled by $\hat{V}(h_j)^{1/2}$. Note that this was not previously noticed in literature, in part because most results are stated using an asymptotic point of view, see the references mentioned in Introduction.

4. Oracle inequalities

If $\beta \in \mathbb{R}^M$, we denote its support by $J(\beta) = \{j \in \{1, \dots, M\} : \beta_j \neq 0\}$ and its *sparsity* is $|\beta|_0 = |J(\beta)| = \sum_{j=1}^M \mathbf{1}(\beta_j \neq 0)$, where $\mathbf{1}(A)$ is the indicator of A

and $|B|$ is the cardinality of a finite set B. If $J \subset \{1, \dots, M\}$, we also introduce the vector β_J such that $(\beta_J)_j = \beta_j$ if $j \in J$ and $(\beta_J)_j = 0$ if $j \in J^{\complement}$, where $J^{\complement} = \{1, \dots, M\} - J$. We define the empirical norm of a function h by

$$\|h\|_n^2 = \frac{1}{n} \sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))^2 Y_t^i \, dt, \tag{15}$$

and remark that $\|h_\beta\|_n^2 = |\mathbf{G}_n \beta|_2^2 / n$.

Below are two oracle inequalities for $h_{\hat{\beta}}$. The first one (Theorem 1) is a "slow" oracle inequality, with a rate of order $(\log M/n)^{1/2}$, which holds without any assymption on the Gram matrix $\mathbf{G}_n$. The second one (Theorem 2) is an oracle inequality with a fast rate of order $\log M/n$, that holds under an assumption on the restricted eigenvalues of $\mathbf{G}_n$.

Theorem 1. *Let $x > 0$ be fixed, and let $\hat{h} = h_{\hat{\beta}}$, where*

$$\hat{\beta}_n \in \operatorname*{argmin}_{b \in B} \Big\{ R_n(b) + \operatorname{pen}(b) \Big\},$$

with $\operatorname{pen}(b)$ *given by* (12). *Then we have, with a probability larger than* $1 - 29e^{-x}$:

$$\|\hat{h} - h_0\|_n^2 \le \inf_{\beta \in B} \Big(\|h_\beta - h_0\|_n^2 + 2 \operatorname{pen}(\beta) \Big).$$

Note that

$$\operatorname{pen}(\beta) \le |\beta|_1 \max_{j=1,\dots,M} \Bigg[c_1 \sqrt{\frac{x + \log M + \hat{\ell}_{n,x}(h_j)}{n} \hat{V}(h_j)} + c_2 \frac{x + 1 + \log M + \hat{\ell}_{n,x}(h_j)}{n} \|h_j\|_{n,\infty} \Bigg]$$

for any $\beta \in \mathbb{R}$, so the dominant term in $\operatorname{pen}(\beta)$ is, up to the slow $\log \log$ term, of order $|\beta|_1 \sqrt{\log M/n}$, which is the expected slow rate for $\hat{h}$ involving the ℓ_1-norm (see [5] for the regression model and [7, 4] for density estimation).

For the proof of oracle inequalities with a fast $\log M/n$ rate, the *restricted eigenvalue* (RE) condition introduced in [5] and [15, 16] is of importance. Restricted eigenvalue conditions are implied by, and in general weaker than, the so-called *incoherence* or RIP assumptions, which excludes strong correlations between covariates. This condition is acknowledged to be one of the weakest to derive fast rates for the Lasso. One can find in [33] an exhaustive survey and comparison of the assumptions used to prove fast oracle inequalities for the Lasso, where the so-called "compatibility condition", which is slightly more general than RE, is described.

The restricted eigenvalue condition is defined below. Note that our presentation (and arguments used in the proof of Theorem 2) is close to [17], where oracle inequalities for the matrix Lasso are given. Let us introduce, for any $\beta \in \mathbb{R}^M$ and $c_0 > 0$, the cone

$$\mathbb{C}_{\beta,c_0} = \big\{ b \in \mathbb{R}^M : |b_{J(\beta)^\complement}|_{1,\hat{w}} \le c_0 |b_{J(\beta)}|_{1,\hat{w}} \big\}. \tag{16}$$

The cone $\mathbb{C}_{\beta,c_0}$ consists of vectors that have a support close to the support of β. Then, introduce

$$\mu_{c_0}(\beta) = \inf \Big\{ \mu > 0 : |b_{J(\beta)}|_2 \le \frac{\mu}{\sqrt{n}} |\mathbf{G}_n b|_2 \quad \forall b \in \mathbb{C}_{\beta,c_0} \Big\}. \tag{17}$$

The number $1/\mu_{c_0}(\beta)$ is an uniform lower bound for $|\mathbf{G}_n b|_2 / |b_{J(\beta)}|_2$ over $b \in \mathbb{C}_{\beta,c_0}$. Hence, it is a lower bound for "eigenvalues" restricted over vectors with a support close to the support of β. Also, note that $c \mapsto \mu_c(\beta)$ is non-increasing.

Theorem 2. *Let $x > 0$ be fixed and let $\hat{h} = h_{\hat{\beta}}$, where*

$$\hat{\beta}_n \in \operatorname*{argmin}_{b \in B} \Big\{ R_n(b) + 2 \operatorname{pen}(b) \Big\},$$

with $\operatorname{pen}(b)$ *given by* (12)*. Then we have, with a probability larger than* $1 - 29e^{-x}$:

$$\|h_{\hat{\beta}} - h_0\|_n^2 \le \inf_{\beta \in B} \Big(\|h_\beta - h_0\|_n^2 + \frac{9}{4} \mu_3(\beta)^2 |\hat{w}_{J(\beta)}|_2^2 \Big),$$

where

$$|\hat{w}_{J(\beta)}|_2^2 = \sum_{j \in J(\beta)} \hat{w}_j^2.$$

Note that

$$|\hat{w}_{J(\beta)}|_2^2 \le 2 |\beta|_0 \max_{j \in J(\beta)} \Bigg[c_1^2 \frac{x + \log M + \hat{\ell}_{n,x}(h_j)}{n} \hat{V}(h_j) + c_2^2 \Big(\frac{x + 1 + \log M + \hat{\ell}_{n,x}(h_j)}{n} \|h_j\|_{n,\infty} \Big)^2 \Bigg],$$

so the dominant term is (up to the $\log\log$ term) of order $|\beta|_0 \log M / n$. This is the fast rate to be found in sparse oracle inequalities [5, 15, 8]. Moreover, note that the (sparse) oracle inequality in Theorem 2 is sharp, in the sense that there is a constant 1 in front of the oracle term $\inf_{\beta \in B} \|h_\beta - h_0\|_n^2$, see Remark 2 below.

Now, let us state Theorem 2 under the restricted eigenvalue condition.

Assumption 1 (RE(s, c_0) [5])**.** *For some integer $s \in \{1, \dots, M\}$ and a constant $c_0 > 0$, we assume that $\mathbf{G}_n$ satisfies:*

$$0 < \kappa(s, c_0) = \min_{\substack{J \subset \{1,\dots,M\}, \\ |J| \le s}} \; \min_{\substack{b \in \mathbb{R}^M \setminus \{0\}, \\ |b_{J^\complement}|_{1,\hat{w}} \le c_0 |b_J|_{1,\hat{w}}}} \frac{|\mathbf{G}_n b|_2}{\sqrt{n} |b_J|_2}$$

Note that using the previous notations, we have

$$\kappa(s, c_0) = \min_{\substack{b \in \mathbb{R}^M \setminus \{0\} \\ |b|_0 \leq s}} \frac{1}{\mu_{c_0}(b)}.$$

Corollary 1. *Let $x > 0$, $s \in \{1, \ldots, M\}$ be fixed and let $\hat{h}$ be the same as in Theorem 2. Then, under Assumption* $\mathrm{RE}(s, 3)$*, we have, with a probability larger than $1 - 29e^{-x}$:*

$$\|h_{\hat{\beta}} - h_0\|_n^2 \leq \inf_{\substack{\beta \in B \\ |\beta|_0 \leq s}} \Big(\|h_\beta - h_0\|_n^2 + \frac{9}{4\kappa(s,3)^2} |\hat{w}_{J(\beta)}|_2^2 \Big).$$

Remark 1. Note that the constant $c_0 = 3$ (for $\mu_{c_0}(\beta)$) is used in Theorem 2. This is because with a large probability, $\hat{\beta} - \beta$ belongs to the cone $\mathbb{C}_{\beta,3}$. Such an argument of cone constraint is at the core of the convex analysis underlying the proof of fast oracle inequalities for the Lasso, see for instance [8, 5, 17].

Remark 2. We were able to prove a sharp sparse oracle inequality (with leading constant 1), because we adapted in our context a recent argument from [17], that uses some tools from convex analysis (such as the fact that the subdifferential mapping is monotone, see [29]) in the study of $\hat{\beta}$ as the minimum of the convex functional $R_n + \mathrm{pen}$.

5. An empirical Bernstein's inequality

The proofs of Theorems 1 and 2 require a sharp control of the "noise term" arising from model (3). For a fixed function h, this noise term is the stochastic process

$$Z_t(h) = \frac{1}{n} \sum_{i=1}^n \int_0^t (h(X_i) - \bar{h}_Y(s)) dM_s^i,$$

where we recall that $M_t^i = N_t^i - \Lambda_t^i$ are i.i.d. martingales with jumps with jumps of size $+1$, as we assume the existence of the intensity function α_0, see (2). In order to give an upper bound on $|Z_t|$ that holds with a large probability, one can use Bernstein's inequality for martingales with jumps, see [20], and note that a proof of this fact is implicit in the proof of Theorem 3, see Section 6 below. Applied to the process $Z_t(h)$, this writes

$$\mathbb{P}\Big[|Z_t(h)| \geq \sqrt{\frac{2vx}{n}} + \frac{x}{3n}, V_t(h) \leq v \Big] \leq 2e^{-x}$$

for any $x, v > 0$, where

$$V_t(h) = n\langle Z(h) \rangle_t = \frac{1}{n} \sum_{i=1}^n \int_0^t (h(X_i) - \bar{h}_Y(s))^2 \alpha_0(s, X_i) Y_s^i \, ds$$

is the predictable variation of Z_t, which will also be referred to as variance term. But, since the term $V_t(h)$ depends explicitly on the unknown intensity α_0, one cannot use it in the penalizing term of the Lasso estimator. Morever, this result is stated on the event $\{V_t \leq v\}$ while we would like an inequality that holds in general. Hence, we need a new Bernstein's type inequality, that uses an observable empirical variance term instead of $V_t(h)$. We prove in Theorem 3 below that we can replace $V_t(h)$ by the optional variation of $Z_t(h)$, which can be also seen as an estimator of $V_t(h)$ and is defined as:

$$\hat{V}_t(h) = n[Z(h)]_t = \frac{1}{n}\sum_{i=1}^n \int_0^t (h(X_i) - \bar{h}_Y(s))^2 dN_s^i.$$

Moreover, our result holds in general, and not on $\{V_t(h) \leq v\}$. The counterpart for this is the presence of a small $\log\log$ term in the upper bound for $|Z_t(h)|$.

Theorem 3. *For any numerical constants $c_\ell > 1, \epsilon > 0$ and $c_0 > 0$ such that $ec_0 > 2(4/3+\epsilon)c_\ell$, the following holds for any $x > 0$:*

$$\mathbb{P}\left[|Z_t(h)| \geq c_1\sqrt{\frac{x+\hat{\ell}_{n,x}(h)}{n}\hat{V}_t(h)} + c_2\frac{x+1+\hat{\ell}_{n,x}(h)}{n}\|h\|_{n,\infty}\right] \leq c_3 e^{-x}, \quad (18)$$

where

$$\hat{\ell}_{n,x}(h) = c_\ell \log\log\Big(\frac{2en\hat{V}_t(h) + 8e(4/3+\epsilon)x\|h\|_{n,\infty}^2}{4(ec_0 - 2(4/3+\epsilon)c_\ell)\|h\|_{n,\infty}^2} \vee e\Big),$$
$$\|h\|_{n,\infty} = \max_{i=1,\dots,n} |h(X_i)|$$

and where

$$c_1 = 2\sqrt{1+\epsilon}, \quad c_2 = 2\sqrt{2\max(c_0, 2(1+\epsilon)(4/3+\epsilon))} + 2/3,$$
$$c_3 = 8 + 6(\log(1+\epsilon))^{-c_\ell}\sum_{j\geq 1} j^{-c_\ell}.$$

By choosing $c_\ell = 2$, $\epsilon = 1$ and $c_0 = 4(4/3+\epsilon)c_\ell/e = 56/(3e)$, Inequality (18) holds with the following numerical values:

$$c_1 = 2\sqrt{2}, \quad c_2 = 4\sqrt{14/3} + 2/3 \leq 9.31$$
$$c_3 = 8 + (\log 2)^{-2}\pi^2 + 4 \leq 28.55,$$
$$\hat{\ell}_{n,x}(h) = 2\log\log\Big(\frac{2en\hat{V}_t(h) + 56ex\|h\|_{n,\infty}^2/3}{8\|h\|_{n,\infty}^2} \vee e\Big).$$

The concentration inequality (18) is fully data-driven, since the random variable that upper bounds $|Z_t(h)|$ with a large probability is observable. Note that the numerical values given in Theorem 3 are the one used in the construction of the ℓ_1-penalization (12). These are chosen for the sake of simplicity, but another combination of numerical values can be considered as well.

The idea of using Bernstein's deviation inequality with an estimated variance is of importance for statistical problems. In [4] for instance, a Bernstein's inequality with empirical variance is derived in order to study the Dantzig selector for density estimation. Note that, however, we are not aware of a previous result such as Theorem 3 for continuous time martingales with jumps, excepted for a work in progress [13], which uses a similar concentration inequalities in the context of point processes.

Acknowledgements

We wish to thank the authors of [13] for giving us a preliminary version of their work, that helped in simplifying the proof of Theorem 3. We thank also Sylvain Delattre and Sarah Lemler for helpful discussions.

6. Proofs

6.1. Decomposition of the least-squares

In this section, we give the details of the construction of the partial least-squares (6). It is based on the decomposition, using the additive structure (5), of the least-squares risk for counting processes depending on covariates, see for instance [28] and [9]. In model (3), on the basis of the observations (4), the least-squares functional to be considered for the estimation of α_0 is given by

$$L_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \int_0^1 \alpha^2(t, X_i) Y_t^i \, dt - \frac{2}{n}\sum_{i=1}^n \int_0^1 \alpha(t, X_i) dN_t^i,$$

where $\alpha : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}_+$. Now, if $\alpha(t,x) = \lambda(t) + h(x)$, we can decompose L_n in the following way:

$$L_n(\alpha) = L_{n,1}(\lambda) + L_{n,2}(h) + L_{n,3}(\lambda, h), \tag{19}$$

where

$$\begin{aligned} L_{n,1}(\lambda) &= \frac{1}{n}\sum_{i=1}^n \int_0^1 (\lambda(t) + \bar{h}_Y(t))^2 Y_t^i \, dt - \frac{2}{n}\sum_{i=1}^n \int_0^1 (\lambda(t) + \bar{h}_Y(t)) dN_t^i \\ L_{n,2}(h) &= \frac{1}{n}\sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))^2 Y_t^i \, dt - \frac{2}{n}\sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t)) dN_t^i \\ L_{n,3}(\lambda, h) &= \frac{2}{n}\sum_{i=1}^n \int_0^1 (\lambda(t) + \bar{h}_Y(t))(h(X_i) - \bar{h}_Y(t)) Y_t^i \, dt, \end{aligned}$$

where, as introduced in Section 3:

$$\bar{h}_Y(t) = \frac{\sum_{i=1}^n h(X_i) Y_t^i}{\sum_{i=1}^n Y_t^i}.$$

Now, the point is that, according to Lemma 1 below, the term $L_{n,3}$ is zero.

Lemma 1. *For any function $h : \mathbb{R}^d \to \mathbb{R}^+$ and any function $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$, we have*

$$\sum_{i=1}^n \int_0^1 \varphi(t)(h(X_i) - \bar{h}_Y(t))Y_t^i \, dt = 0.$$

Lemma 1 follows from an easy computation which is omitted. The term $L_{n,2}$ in (19) is the partial least-squares criterion considered in Section 3, see Equation (6). We now explain why it is suitable for the estimation of h_0. If the Aalen additive model holds, we have $dN_t^i = (\lambda_0(t) + h_0(X_i))Y_t^i \, dt + dM_t^i$ for all $i = 1, \dots, n$, so we can write, using again Lemma 1:

$$\begin{aligned} L_{n,2}(h) = \frac{1}{n} \sum_{i=1}^n \int_0^1 & (h(X_i) - \bar{h}_Y(t))^2 Y_t^i \, dt \\ & - \frac{2}{n} \sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))(h_0(X_i) - \bar{h}_{0,Y}(t))Y_t^i \, dt \\ & - \frac{2}{n} \sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t)) dM_t^i, \end{aligned}$$

where

$$\bar{h}_{0,Y}(t) = \frac{\sum_{i=1}^n h_0(X_i)Y_t^i}{\sum_{i=1}^n Y_t^i}.$$

Now, using the empirical norm $\| \cdot \|_n^2$ defined in Equation (15), see Section 3 above, we can finally write

$$L_{n,2}(h) = \|h - h_0\|_n^2 - \|h_0\|_n^2 - \frac{2}{n} \sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t)) dM_t^i. \tag{20}$$

The last term in the right hand side of (20) is a noise term, with tails controlled in Section 5 above. It is now understood that finding a minimizer of $L_{n,2}$, or a penalized version of it, is a natural way of estimating h_0. We refer the reader to [25] for an other justification of the "partial least-squares" criterion in the linear case $h_0(x) = x^\top \beta_0$.

6.2. Proof of Theorem 3

For $i = 1, \dots, n$, the processes N^i are i.i.d. counting processes satisfying the Doob-Meyer decomposition $N_t^i - \int_0^t \alpha_0(s, X_i)Y_s^i \, ds = M_t^i$, see Equation (3). This implies that the processes M^i are i.i.d. centered martingales, with predictable variation $\langle M^i \rangle_t = \int_0^t \alpha_0(s, X_i)Y_s^i \, dt$ and optional variation $[M^i]_t = N_t^i$, see e.g. [2] for details. Moreover, the jumps of each M^i, denoted by $\Delta M_t^i = M_t^i - M_{t_-}^i$, are in $\{0, 1\}$. Introduce the process

$$U_t = \frac{1}{n} \sum_{i=1}^n \int_0^t H_s^i dM_s^i$$

where

$$H_t^i = \frac{h(X_i) - \bar{h}_Y(t)}{2 \max_{i=1,\ldots,n} |h(X_i)|}.$$

Note that $|H_t^i| \leq 1$. Since H^i is predictable and bounded, the process U is a square integrable martingale, as a sum of square integrable martingales. Its predictable variation $\langle U \rangle$ is given by:

$$\vartheta_t = n\langle U \rangle_t = \frac{1}{n} \sum_{i=1}^n \int_0^t (H_s^i)^2 \alpha_0(s, X_i) Y_s^i \, ds,$$

while its optional variation $[U]$ is given by

$$\hat{\vartheta}_t = n[U]_t = \frac{1}{n} \sum_{i=1}^n \int_0^t (H_s^i)^2 dN_s^i.$$

From [32], we know that

$$\exp(\lambda U_t - S_\lambda(t)) \tag{21}$$

is a supermartingale if S_λ is the compensator of

$$E_t = \sum_{0 \leq s \leq t} \big\{ \exp(\lambda \Delta U_s) - 1 - \lambda \Delta U_s \big\}.$$

We now derive the expression of S_λ. The process E can also be written as

$$\begin{aligned} E_t &= \sum_{s \leq t} \sum_{k \geq 2} \frac{\lambda^k}{k!} (\Delta U(s))^k = \sum_{s \leq t} \sum_{k \geq 2} \frac{\lambda^k}{k! n^k} \Big(\Delta (\sum_{i=1}^n \int_0^s H_u^i dM_u^i) \Big)^k \\ &= \sum_{s \leq t} \sum_{k \geq 2} \frac{\lambda^k}{k! n^k} \sum_{i=1}^n \Big(\Delta \int_0^s H_u^i dM_u^i \Big)^k, \end{aligned}$$

where the last inequality holds almost surely, since the M^i are independent, hence do not jump at the same time (with probability 1). Now, note that

$$\Big(\Delta \int_0^s H_u^i dM_u^i \Big)^k = (H_s^i)^k \Delta M^i(s) = (H_s^i)^k \Delta N^i(s),$$

so that we have

$$S_\lambda(t) = \sum_{i=1}^n \int_0^t \phi\Big(\frac{\lambda}{n} H_s^i \Big) \alpha_0(s, X_i) Y_s^i \, ds$$

with $\phi(x) = e^x - x - 1$. The fact that (21) is a supermartingale entails

$$\mathbb{P}\Big[U_t \geq \frac{S_\lambda(t)}{\lambda} + \frac{x}{\lambda} \Big] \leq e^{-x} \tag{22}$$

for any $\lambda, x > 0$. The following facts hold true:

- $\phi(xh) \le h^2\phi(x)$ for any $0 \le h \le 1$ and $x > 0$;
- $\phi(\lambda) \le \frac{\lambda^2}{2(1-\lambda/3)}$ for any $\lambda \in (0,3)$;
- $\min_{\lambda\in(0,1/b)} \big(\frac{a\lambda}{1-b\lambda} + \frac{x}{\lambda}\big) = 2\sqrt{ax} + bx$, for any $a, b, x > 0$.

For any $w > 0$, they entail the following embeddings:

$$\begin{aligned}\Big\{U_t \ge \sqrt{\frac{2wx}{n}} + \frac{x}{3n}, \vartheta_t \le w\Big\} &= \Big\{U_t \ge \frac{\lambda_w}{2(n-\lambda_w/3)} w + \frac{x}{\lambda_w}, \vartheta_t \le w\Big\} \\ &\subset \Big\{U_t \ge \frac{\phi(\lambda_w/n)}{\lambda_w} n\vartheta_t + \frac{x}{\lambda_w}, \vartheta_t \le w\Big\} \\ &\subset \Big\{U_t \ge \frac{S_{\lambda_w}(t)}{\lambda_w} + \frac{x}{\lambda_w}, \vartheta_t \le w\Big\}, \end{aligned} \tag{23}$$

where λ_w achieves the infimum. This leads to the standard Bernstein's inequality:

$$\mathbb{P}\Big[U_t \ge \sqrt{\frac{2wx}{n}} + \frac{x}{3n}, \vartheta_t \le w\Big] \le e^{-x}.$$

By choosing $w = c_0(x+1)/n$ for some constant $c_0 > 0$, this gives the following inequality, which says that when the variance term ϑ_t is small, the sub-exponential term is dominating in Bernstein's inequality:

$$\mathbb{P}\Big[U_t \ge \Big(\sqrt{2c_0} + \frac{1}{3}\Big)\frac{x+1}{n}, \vartheta_t \le \frac{c_0(x+1)}{n}\Big] \le e^{-x}. \tag{24}$$

For any $0 < v < w < +\infty$, we have

$$\Big\{U_t \ge \sqrt{\frac{2w\vartheta_t x}{vn}} + \frac{x}{3n}\Big\} \cap \{v < \vartheta_t \le w\} \subset \Big\{U_t \ge \sqrt{\frac{2wx}{n}} + \frac{x}{3n}\Big\} \cap \{v < \vartheta_t \le w\},$$

so, together with (22) and (23), we obtain

$$\mathbb{P}\Big[U_t \ge \sqrt{\frac{2w\vartheta_t x}{vn}} + \frac{x}{3n}, v < \vartheta_t \le w\Big] \le e^{-x}. \tag{25}$$

Now, we want to replace ϑ_t by the observable $\hat\vartheta_t$ in the deviation (25). Note that the process $\tilde U_t$ given by

$$\begin{aligned}\tilde U_t = \hat\vartheta_t - \vartheta_t &= \frac{1}{n}\sum_{i=1}^n \int_0^t (H_s^i)^2 \big(dN_s^i - \alpha_0(s,X_i)Y_s^i\,ds\big) \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^t (H_s^i)^2 dM_s^i\end{aligned}$$

is a martingale, so following the same steps as for U_t, we obtain that $\exp(\tilde U_t - \lambda \tilde S_\lambda(t))$ is a supermartingale, with

$$\tilde S_\lambda(t) = \sum_{i=1}^n \int_0^t \phi\Big(\frac{\lambda}{n}(H_s^i)^2\Big)\alpha_0(s,X_i)Y_s^i\,ds.$$

Now, writing again (23) for $\tilde{U}_t$ with the fact that $|H_t^i| \leq 1$ and using the same arguments as before, we arrive at

$$\mathbb{P}\Big[|\hat{\vartheta}_t - \vartheta_t| \geq \frac{\phi(\lambda/n)}{\lambda} n\vartheta_t + \frac{x}{\lambda}\Big] \leq 2e^{-x}$$

and

$$\mathbb{P}\Big[|\hat{\vartheta}_t - \vartheta_t| \geq \sqrt{\frac{2w\vartheta_t x}{vn}} + \frac{x}{3n}, v < \vartheta_t \leq w\Big] \leq 2e^{-x}. \tag{26}$$

But, if ϑ_t satisfies

$$|\hat{\vartheta}_t - \vartheta_t| \leq \sqrt{\frac{2w\vartheta_t x}{vn}} + \frac{x}{3n},$$

then it satisfies

$$\vartheta_t \leq 2\hat{\vartheta}_t + 2\Big(\frac{w}{v} + \frac{1}{3}\Big)\frac{x}{n}$$

and

$$\hat{\vartheta}_t \leq 2\vartheta_t + 2\Big(\frac{1}{3} + \sqrt{\frac{w}{v}\Big(\frac{1}{3} + \frac{w}{v}\Big)} + \frac{2w}{v}\Big)\frac{x}{n},$$

simply by using the fact that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ for any $a, A, b > 0$. This proves that

$$\begin{aligned}\Big\{U_t \leq & \sqrt{\frac{2w\vartheta_t x}{vn}} + \frac{x}{3n}\Big\} \cap \Big\{|\hat{\vartheta}_t - \vartheta_t| \leq \sqrt{\frac{2w\vartheta_t x}{vn}} + \frac{x}{3n}\Big\} \\ & \subset \Big\{U_t \leq 2\sqrt{\frac{wx}{vn}\hat{\vartheta}_t} + \Big(2\sqrt{\frac{w}{v}\Big(\frac{w}{v} + \frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}\Big\},\end{aligned} \tag{27}$$

so using (25) and (26), we obtain

$$\mathbb{P}\Big[U_t \geq 2\sqrt{\frac{wx}{vn}\hat{\vartheta}_t} + \Big(2\sqrt{\frac{w}{v}\Big(\frac{w}{v} + \frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}, v \leq \vartheta_t < w\Big] \leq 3e^{-x}.$$

This inequality is similar to (25), where we replaced ϑ_t by the observable $\hat{\vartheta}_t$ in the sub-Gaussian term. It remains to remove the event $\{v \leq \vartheta_t < w\}$ from this inequality. First, recall that (24) holds, so we can work on the event $\{\vartheta_t > c_0(x+1)/n\}$ from now on. We use a peeling argument: define, for $j \geq 0$:

$$v_j = c_0 \frac{x+1}{n}(1+\epsilon)^j,$$

and use the following decomposition into disjoint sets:

$$\{\vartheta_t > v_0\} = \bigcup_{j \geq 0} \{v_j < \vartheta_t \leq v_{j+1}\}.$$

We have

$$\mathbb{P}\Big[U_t \geq c_{1,\epsilon}\sqrt{\frac{x}{n}\hat{\vartheta}_t} + c_{2,\epsilon}\frac{x}{n}, v_j < \vartheta_t \leq v_{j+1}\Big] \leq 3e^{-x},$$

where we introduced the constants

$$c_{1,\epsilon} = 2\sqrt{1+\epsilon} \text{ and } c_{2,\epsilon} = 2\sqrt{(1+\epsilon)(4/3+\epsilon)} + 1/3.$$

Let us introduce

$$\ell = c_\ell \log\log\Big(\frac{\vartheta_t}{v_0} \vee e\Big),$$

where $c_\ell > 1$. On the event

$$\Big\{|\hat\vartheta_t - \vartheta_t| \le \sqrt{\frac{2(1+\epsilon)\vartheta_t(x+\ell)}{n}} + \frac{x+\ell}{3n}\Big\}$$

we have

$$\vartheta_t \le 2\hat\vartheta_t + 2(4/3+\epsilon)\frac{x}{n} + \frac{2(4/3+\epsilon)c_\ell}{n}\log\log\Big(\frac{\vartheta_t}{v_0} \vee e\Big),$$

which entails, assuming that $ec_0 > 2(4/3+\epsilon)c_\ell$:

$$\vartheta_t \le \frac{ec_0}{ec_0 - 2(4/3+\epsilon)c_\ell}\Big(2\hat\vartheta_t + 2(4/3+\epsilon)\frac{x}{n}\Big),$$

where we used the fact that $\log\log(x) \le x/e - 1$ for any $x \ge e$. This entails, together with (27), the following embedding:

$$\begin{aligned}\Big\{U_t \le & \sqrt{\frac{2(1+\epsilon)\vartheta_t(x+\ell)}{n}} + \frac{x+\ell}{3n}\Big\} \cap \Big\{|\hat\vartheta_t - \vartheta_t| \le \sqrt{\frac{2(1+\epsilon)\vartheta_t(x+\ell)}{n}} + \frac{x+\ell}{3n}\Big\} \\ & \subset \Big\{U_t \le c_{1,\epsilon}\sqrt{\frac{\hat\vartheta_t(x+\hat\ell)}{n}} + c_{2,\epsilon}\frac{x+\hat\ell}{n}\Big\},\end{aligned}$$

where

$$\hat\ell = c_\ell \log\log\Big(\frac{2en\hat\vartheta_t + 2e(4/3+\epsilon)x}{ec_0 - 2(4/3+\epsilon)c_\ell} \vee e\Big).$$

Now, using the previous embeddings together with (25) and (26), we obtain

$$\begin{aligned}&\mathbb{P}\Big[U_t \ge c_{1,\epsilon}\sqrt{\frac{\hat\vartheta_t(x+\hat\ell)}{n}} + c_{2,\epsilon}\frac{x+\hat\ell}{n}, \vartheta_t > v_0\Big] \\ &\le \sum_{j\ge0}\mathbb{P}\Big[U_t \ge \sqrt{\frac{2(1+\epsilon)\vartheta_t(x+\ell)}{vn}} + \frac{x+\ell}{3n}, v_j < \vartheta_t \le v_{j+1}\Big] \\ &\quad + \sum_{j\ge0}\mathbb{P}\Big[|\hat\vartheta_t - \vartheta_t| \ge \sqrt{\frac{2(1+\epsilon)\vartheta_t(x+\ell)}{n}} + \frac{x+\ell}{3n}, v_j < \vartheta_t \le v_{j+1}\Big] \\ &\le 3\Big(e^{-x} + \sum_{j\ge1} e^{-(x+c_\ell\log\log(v_j/v_0))}\Big) \\ &= 3\Big(1 + (\log(1+\epsilon))^{-c_\ell}\sum_{j\ge1} j^{-c_\ell}\Big)e^{-x}.\end{aligned}$$

Together with (24), this gives

$$\mathbb{P}\Big[U_t \geq c_{1,\epsilon}\sqrt{\frac{\hat{\vartheta}_t(x+\hat{\ell})}{n}} + c_{3,\epsilon}\frac{x+1+\hat{\ell}}{n}\Big] \leq \Big(4 + 3(\log(1+\epsilon))^{-c_\ell}\sum_{j\geq 1} j^{-c_\ell}\Big)e^{-x},$$

where $c_{3,\epsilon} = \sqrt{2\max(c_0, 2(1+\epsilon)(4/3+\epsilon))} + 1/3$. Now, it suffices to multiply both sides of the inequality

$$U_t \geq c_{1,\epsilon}\sqrt{\frac{x+\hat{\ell}}{n}\hat{\vartheta}_t} + c_{3,\epsilon}\frac{x+1+\hat{\ell}}{n}$$

by $2\|h\|_{n,\infty}$ to recover the statement of Theorem 3. □

6.3. Some notations and preliminary results for the proof of the oracle inequalities

Let us introduce the following notations. Let $\boldsymbol{h}(\cdot) = (h_1(\cdot), \ldots, h_M(\cdot))^\top$ and $\bar{\boldsymbol{h}}_Y(\cdot) = (\bar{h}_{1,Y}(\cdot), \ldots, \bar{h}_{M,Y}(\cdot))^\top$, so that $h_\beta = \boldsymbol{h}^\top\beta$ and $\bar{h}_{\beta,Y} = \bar{\boldsymbol{h}}_Y^\top\beta$. We will use the notation $\langle\cdot,\cdot\rangle_n$ for the following "empirical" inner-product between to functions $h, h' : \mathbb{R}^d \to \mathbb{R}^+$ (two "covariates" functions):

$$\langle h, h'\rangle_n = \frac{1}{n}\sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))(h'(X_i) - \bar{h}'_Y(t))Y_t^i\, dt,$$

and the corresponding empirical norm:

$$\|h\|_n^2 = \frac{1}{n}\sum_{i=1}^n \int_0^1 (h(X_i) - \bar{h}_Y(t))^2 Y_t^i\, dt.$$

Note that with these notations, we have:

$$\beta^\top \mathbf{H}_n \beta' = \langle h_\beta, h_{\beta'}\rangle_n.$$

To avoid any possible confusion, we will always write $\beta^\top\beta'$ for the Euclidean inner product between two vectors β and β' in $\mathbb{R}^M$.

In view of (11), we can write

$$\mathbf{H}_n = \frac{1}{n}\sum_{i=1}^n \int_0^1 (\boldsymbol{h}(X_i) - \bar{\boldsymbol{h}}_Y(t))(\boldsymbol{h}(X_i) - \bar{\boldsymbol{h}}_Y(t))^\top Y_t^i\, dt,$$

and

$$\boldsymbol{h}_n = \frac{1}{n}\sum_{i=1}^n \int_0^1 (\boldsymbol{h}(X_i) - \bar{\boldsymbol{h}}_Y(t))dN_t^i.$$

Now, in view of (5) and (3), the following holds:

$$\boldsymbol{h}_n = \boldsymbol{h}'_n + \boldsymbol{Z}_n, \tag{28}$$

where:

$$(\boldsymbol{h}'_n)_j = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t))(\lambda_0(t) + h_0(X_i))Y_t^i\,dt,$$

$$(\boldsymbol{Z}_n)_j = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t))dM_t^i.$$

Using Lemma 1 two times, we obtain:

$$\begin{aligned}(\boldsymbol{h}'_n)_j &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t))(\lambda_0(t) + h_0(X_i))Y_t^i dt\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t))h_0(X_i)Y_t^i dt\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 (h_j(X_i) - \bar{h}_{j,Y}(t))(h_0(X_i) - \bar{h}_{0,Y}(t))Y_t^i dt,\end{aligned}$$

namely

$$(\boldsymbol{h}'_n)_j = \langle h_j, h_0\rangle_n. \tag{29}$$

6.4. Proof of Theorem 1

Recall that the empirical risk R_n is given by (10). As a consequence of (28) and (29), we obtain the following decomposition of the empirical risk:

$$R_n(\beta) = \beta^\top \mathbf{H}_n\beta - 2\beta^\top \boldsymbol{h}_n = \|h_\beta\|_n^2 - 2\langle h_\beta, h_0\rangle_n - 2\beta^\top \boldsymbol{Z}_n,$$

so, for any $\beta \in \mathbb{R}^M$, the following holds:

$$R_n(\hat{\beta}) - R_n(\beta) = \|h_{\hat{\beta}} - h_0\|_n^2 - \|h_\beta - h_0\|_n^2 + 2(\beta - \hat{\beta})^\top \boldsymbol{Z}_n.$$

By definition of $\hat{\beta}$, we have

$$R_n(\hat{\beta}) + \mathrm{pen}(\hat{\beta}) \le R_n(\beta) + \mathrm{pen}(\beta)$$

for any $\beta \in \mathbb{R}^M$, so:

$$\|h_{\hat{\beta}} - h_0\|_n^2 \le \|h_\beta - h_0\|_n^2 + 2(\hat{\beta} - \beta)^\top \boldsymbol{Z}_n + \mathrm{pen}(\beta) - \mathrm{pen}(\hat{\beta}).$$

Let us introduce the event

$$A = \bigcap_{j=1}^{M}\Big\{2|(\boldsymbol{Z}_n)_j| \le \hat{w}_j\Big\}, \tag{30}$$

where the weights $\hat{w}_j$ are given by (14). Using Theorem 3 together with an union bound, we have that

$$\mathbb{P}(A) \ge 1 - c_3 e^{-x},$$

where c_3 is a purely numerical positive constant from Theorem 3. On A, we have

$$|2(\hat{\beta}-\beta)^\top \boldsymbol{Z}_n| \le \sum_{j=1}^M \hat{w}_j|\hat{\beta}_j-\beta_j| = |\hat{\beta}-\beta|_{1,\hat{w}},$$

so recalling that $\mathrm{pen}(\beta)=\sum_{j=1}^M \hat{w}_j|\beta_j|$, we obtain

$$\|h_{\hat{\beta}}-h_0\|_n^2 \le \|h_\beta-h_0\|_n^2 + 2\,\mathrm{pen}(\beta)$$

for any $\beta\in\mathbb{R}^M$, which is the statement of Theorem 1. □

6.5. Proof of Theorem 2

Recall the following notation: for any $J\subset\{1,\dots,M\}$ and $x\in\mathbb{R}^M$, we define the vector $x_J\in\mathbb{R}^M$ with coordinates by $(x_J)_j=x_j$ when $j\in J$ and $(x_J)_j=0$ if $j\in J^{\complement}$, where $J^{\complement}=\{1,\dots,M\}-J$. Recall that

$$\hat{\beta}\in\operatorname*{argmin}_{b\in B}\Big\{R_n(b)+2\,\mathrm{pen}(b)\Big\}, \tag{31}$$

where B is a convex set. This proof uses arguments from [17]. We denote by $\partial\phi$ the subdifferential mapping of a convex function ϕ. The function $b\mapsto R_n(b)$ is differentiable, so the subdifferential of $R_n(\cdot)+2\,\mathrm{pen}(\cdot)$ at a point $b\in\mathbb{R}^M$ is given by

$$\partial(R_n+2\,\mathrm{pen})(b)=\{\nabla R_n(b)\}+2\partial\,\mathrm{pen}(b)=\{2\mathbf{H}_n b-2\boldsymbol{h}_n\}+2\partial\,\mathrm{pen}(b).$$

So, Equation (31) means that there is $\hat{\beta}_\partial\in\partial\,\mathrm{pen}(\hat{\beta})$ such that $\nabla R_n(\hat{\beta})+2\hat{\beta}_\partial$ belongs to the normal cone of B at $\hat{\beta}$:

$$(2\mathbf{H}_n\hat{\beta}-2\boldsymbol{h}_n+2\hat{\beta}_\partial)^\top(\hat{\beta}-\beta)\le 0 \quad \forall\beta\in B. \tag{32}$$

Inequality (32) can be written, using (28) and (29), in the following way:

$$2\langle h_{\hat{\beta}}-h_\beta, h_{\hat{\beta}}-h_0\rangle_n + 2(\hat{\beta}_\partial-\beta_\partial)^\top(\hat{\beta}-\beta)\le -2\beta_\partial^\top(\hat{\beta}-\beta)+2\boldsymbol{Z}_n^\top(\hat{\beta}-\beta),$$

where chose any $\beta_\partial\in\partial\,\mathrm{pen}(\beta)$. Now, we use the fact that the subdifferential mapping is monotone (this is an immediate consequence of its definition, see [29], Chapter 24, p. 240) to say that $(\hat{\beta}_\partial-\beta_\partial)^\top(\hat{\beta}-\beta)\ge 0$. Moreover, it is standard to see that

$$\partial|b|_{1,\hat{w}}=\Big\{e+f: e_j=\hat{w}_j\mathrm{sgn}\,(b_j) \text{ and } f_{J(b)}=0, |f_j|\le\hat{w}_j \text{ for any } j=1,\dots,M\Big\},$$

where $J(b)=\{j: b_j\ne 0\}$. Let $\beta\in B$ be fixed, and denote $J=J(\beta)=\{j:\beta_j\ne 0\}$. Consider e and f such that $\beta_\delta=e+f$, with $e_{J^{\complement}}=0$. We have $|e^\top(\hat{\beta}-\beta)|\le$

$|\hat\beta_J - \beta_J|_{1,\hat w}$ and we can take f such that $f^\top(\hat\beta - \beta) = f^\top \hat\beta_{J^\complement} = |\hat\beta_{J^\complement}|_{1,\hat w}$. This gives

$$2\langle h_{\hat\beta} - h_\beta, h_{\hat\beta} - h_0\rangle_n + 2|\hat\beta_{J^\complement}|_{1,\hat w} \le 2|\hat\beta_J - \beta_J|_{1,\hat w} + 2\boldsymbol{Z}_n^\top(\hat\beta - \beta).$$

Using Pythagora's Theorem, we have

$$2\langle h_{\hat\beta} - h_0, h_{\hat\beta} - h_\beta\rangle_n = \|h_{\hat\beta} - h_0\|_n^2 + \|h_{\hat\beta} - h_\beta\|_n^2 - \|h_\beta - h_0\|_n^2, \tag{33}$$

so

$$\begin{aligned}\|h_{\hat\beta} - h_0\|_n^2 + &\|h_{\hat\beta} - h_\beta\|_n^2 + 2|\hat\beta_{J^\complement}|_{1,\hat w} \\ &\le \|h_{\hat\beta} - h_0\|_n^2 + 2|\hat\beta_J - \beta_J|_{1,\hat w} + 2\boldsymbol{Z}_n^\top(\hat\beta - \beta).\end{aligned}$$

If $\langle h_{\hat\beta} - h_0, h_{\hat\beta} - h_\beta\rangle_n < 0$, we have $\|h_{\hat\beta} - h_0\|_n < \|h_\beta - h_0\|_n$, which entails the Theorem, so we assume that $\langle h_{\hat\beta} - h_0, h_{\hat\beta} - h_\beta\rangle_n \ge 0$. In this case

$$2|\hat\beta_{J^\complement}|_{1,\hat w} \le 2\langle h_{\hat\beta} - h_0, h_{\hat\beta} - h_\beta\rangle_n + 2|\hat\beta_{J^\complement}|_{1,\hat w} \le 2|\hat\beta_J - \beta_J|_{1,\hat w} + 2\boldsymbol{Z}_n^\top(\hat\beta - \beta),$$

which entails, together with the fact that, on A (see (30)), we have

$$2|\boldsymbol{Z}_n^\top(\hat\beta - \beta)| = 2|(\boldsymbol{Z}_n)_J^\top(\hat\beta_J - \beta_J)| + 2|(\boldsymbol{Z}_n)_{J^\complement}^\top \hat\beta_{J^\complement}| \le |\hat\beta_J - \beta_J|_{1,\hat w} + |\hat\beta_{J^\complement}|_{1,\hat w},$$

that

$$|\hat\beta_{J^\complement}|_{1,\hat w} \le 3|\hat\beta_J - \beta_J|_{1,\hat w}.$$

This means that $\hat\beta - \beta \in \mathbb{C}_{\beta,3}$ (see (16)). So, using (17), we have

$$|\hat\beta_J - \beta_J|_2 \le \mu_3(\beta)|\mathbf{G}_n(\hat\beta - \beta)|_2. \tag{34}$$

Note that, on A, we have:

$$\|h_{\hat\beta} - h_0\|_n^2 + \|h_{\hat\beta} - h_\beta\|_n^2 + |\hat\beta_{J^\complement}|_{1,\hat w} \le \|h_\beta - h_0\|_n^2 + 3|\hat\beta_J - \beta_J|_{1,\hat w}.$$

A consequence of (34) is

$$|\hat\beta_J - \beta_J|_{1,\hat w} \le |\hat w_J|_2 |\hat\beta_J - \beta_J|_2 \le \mu_3(\beta)|\hat w_J|_2 |\mathbf{G}_n(\hat\beta - \beta)|_2,$$

so we arrive at

$$\|h_{\hat\beta} - h_0\|_n^2 \le \|h_\beta - h_0\|_n^2 + 3\mu_3(\beta)|\hat w_J|_2 \|h_{\hat\beta} - h_\beta\|_n - \|h_{\hat\beta} - h_\beta\|_n^2,$$

and finally

$$\|h_{\hat\beta} - h_0\|_n^2 \le \|h_\beta - h_0\|_n^2 + \frac{9}{4}\mu_3(\beta)^2 |\hat w_J|_2^2,$$

using the fact that $ax - x^2 \le a^2/4$ for any $a, x > 0$. □

References

[1] Odd Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisła, 1978)*, volume 2 of *Lecture Notes in Statist.*, pages 1–25. Springer, New York, 1980. MR0577267

[2] Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. MR1198884

[3] Anestis Antoniadis, Piotr Fryzlewicz, and Frédérique Letué. The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.*, 37(4):531–552, 2010. MR2779635

[4] Karine Bertin, Erwan Le Pennec, and Vincent Rivoirard. Adaptive dantzig density estimation. *Annales de l'IHP, Probabilités et Statistiques*, 47(1):43–74, 2011. MR2779396

[5] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. MR2533469

[6] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007. MR2351101

[7] Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp, and Adrian Barbu. Spades and mixture models. *Ann. Statist.*, 38(4):2525–2558, 2010. MR2676897

[8] Emmanuel Candès and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007. MR2382644

[9] Fabienne Comte, Stéphane Gaïffas, and Agathe Guilloux. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 47(4):1171–1196, 2011.

[10] David R. Cox. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220, 1972. MR0341758

[11] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. MR2060166

[12] Jianqing Fan and Runze Li. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, 30(1):74–99, 2002. MR1892656

[13] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. Work in progress, personnal communication.

[14] Jean Jacod and Albert N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1987. MR0959133

[15] Vladimir Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009. MR2555200

[16] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009. MR2500227

[17] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, 39(39):2302–2329, 2011.

[18] Chenlei Leng and Shuangge Ma. Path consistent model selection in additive risk model via Lasso. *Stat. Med.*, 26(20):3753–3770, 2007. MR2395831

[19] Danyu Lin and Zhiliang Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994. MR1279656

[20] Robert Sh. Liptser and Albert N. Shiryayev. *Theory of martingales*, volume 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989. Translated from the Russian by K. Dzjaparidze [Kacha Dzhaparidze]. MR1022664

[21] Shuangge Ma and J. Huang. Additive risk survival model with microarray data. *BMC bioinformatics*, 8(1):192, 2007.

[22] Shuangge Ma, Mickael R. Kosorok, and Jason P. Fine. Additive risk models for survival data with high-dimensional covariates. *Biometrics*, 62(1):202–210, 2006. MR2226574

[23] Torben Martinussen and Thomas H. Scheike. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York, 2006. MR2214443

[24] Torben Martinussen and Thomas H. Scheike. The additive hazards model with high-dimensional regressors. *Lifetime Data Anal.*, 15(3):330–342, 2009. MR2519717

[25] Torben Martinussen and Thomas H. Scheike. Covariate selection for the semiparametric additive risk model. *Scand. J. Stat.*, 36(4):602–619, 2009. MR2572578

[26] Ian W. McKeague and Peter D. Sasieni. A partly parametric additive risk model. *Biometrika*, 81(3):501–514, 1994. MR1311093

[27] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473, 2010. MR2758523

[28] Patricia Reynaud-Bouret. Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006. MR2248231

[29] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970. MR0274683

[30] Andreas Rosenwald, George Wright, Adrian Wiestner, Wing C. Chan, Joseph M. Connors, Elias Campo, Randy D. Gascoyne, Thomas M. Grogan, H. Konrad Muller-Hermelink, Erlend B. Smeland, Michael Chiorazzi, Jena M. Giltnane, Elaine M. Hurt, Hong Zhao, Lauren Averett, Sarah Henrickson, Liming M. Yang, John Powell, Wyndham H. Wilson, Elaine S.

Jaffe, Richard Simon, Richard D. Klausner, Emilio Montserrat, Francesc Bosch, Timohy C. Greiner, Dennis D. Weisenburger, Warren G. Sanger, Bhavana J. Dave, James C. Lynch, Julie Vose, James O. Armitage, Richard I. Fisher, Thomas P. Miller, Michael LeBlanc, German Ott, Stein Kvaloy, Harald Holte, Jan Delabie, and Louis M. Staudt. The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *CANCER CELL*, 3(2):185–197, 2003.

[31] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statist. in Med.*, 16:385–395, 1997.

[32] Sara A. van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, 23(5):1779–1801, 1995. MR1370307

[33] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. MR2576316

[34] Laura. J. van 't Veer, Hongyue Dai, Marc J. van de Vijver, and et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):484–5, 2002.

[35] Daniela M. Witten and Robert Tibshirani. Survival analysis with high-dimensional covariates. *Statistical methods in medical research*, 19(1):29, 2010. MR2744491

[36] Hao H. Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691, 2007. MR2410017

[37] Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, 11:1081–1107, 2010. MR2629825

[38] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006. MR2279469

[39] Hui Zou. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, 95(1):241–247, 2008. MR2409726