Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates

Agathe Guilloux, Sarah Lemler, Marie-Luce Taupin

Adaptive kernel estimation of the baseline function in the Cox model with high-dimensional covariates

Agathe Guilloux
Laboratoire de Statistique Théorique et Appliquée,
Université Pierre et Marie Curie - Paris 6
Centre de Mathématiques Appliquées, Ecole Polytechnique,
CNRS UMR 7641, 91128 Palaiseau, France
e-mail : agathe.guilloux@upmc.fr

Sarah Lemler
Laboratoire Mathématiques et Informatique pour la Complexité et les Systèmes,
École CentraleSupélec, France
e-mail : sarah.lemler@centralesupelec.fr

Marie-Luce Taupin
Laboratoire de Mathématiques et Modélisation d'Évry, UMR CNRS 8071- USC INRA,
Université d'Évry Val d'Essonne, France
Unité MaIAGE, INRA Jouy-En-Josas, France
e-mail : marie-luce.taupin@genopole.cnrs.fr

Abstract

We propose a novel kernel estimator of the baseline function in a general high-dimensional Cox model, for which we derive non-asymptotic rates of convergence. To construct our estimator, we first estimate the regression parameter in the Cox model via a LASSO procedure. We then plug this estimator into the classical kernel estimator of the baseline function, obtained by smoothing the so-called Breslow estimator of the cumulative baseline function. We propose and study an adaptive procedure for selecting the bandwidth, in the spirit of Goldenshluger and Lepski [14]. We state non-asymptotic oracle inequalities for the final estimator, which leads to a reduction in the rate of convergence when the dimension of the covariates grows.

Keywords: Conditional hazard rate function; Semi-parametric model; Counting process; Kernel estimation; Goldenshluger and Lepski method, Non-asymptotic oracle inequality; Survival analysis

1 Introduction

The Cox model, introduced by Cox [9], is a regression model often considered in survival analysis, which relates the distribution of a time T to the values of covariates. The hazard function of T is then defined by

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \exp({\boldsymbol{\beta_0}}^\top \boldsymbol{Z}), \tag{1}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_p)^\top$ is a p-dimensional vector of covariates, $\boldsymbol{\beta_0} = (\beta_{0_1}, \ldots, \beta_{0_p})^\top$ the vector of regression coefficients and α_0 the baseline hazard function.

The regression parameter $\boldsymbol{\beta_0}$ and the baseline function α_0 are the two unknown parameters in this model. In previous works, more attention has been paid to the estimation of the regression parameter than to the estimation of the baseline function.

There are good reasons for this. First, the Cox partial log-likelihood, introduced by Cox [9], allows us to estimate $\boldsymbol{\beta_0}$ without knowledge of α_0. Second, the regression parameter is directly related to the covariates. Therefore, in order to select the relevant covariates that best explain the survival time, we need to estimate the regression parameter. Many papers deal with the problem of the estimation of $\boldsymbol{\beta_0}$, the number of covariates p being large (or not) compared with the number of individuals n. When p is smaller than n, the usual estimator of $\boldsymbol{\beta_0}$ is obtained by maximizing the Cox partial log-likelihood (see Andersen et al. [2] as a good reference). When the number of covariates grows, the LASSO procedure is often considered. This consists of a minimization of the negative ℓ_1-penalized Cox partial log-likelihood. Asymptotic results are stated in Bradic et al. [4], Kong and Nan [18] and Bradic and Song [5]. Lastly, the non-asymptotic rate of convergence of the LASSO is now known to be of order $\sqrt{\ln p/n}$, see Huang et al. [17].

The estimation of the baseline function α_0 has been less studied. One known estimator of the baseline function is a kernel estimator, introduced by Ramlau-Hansen [23, 24]. We present here its form in the special case of right-censoring. Let us consider, for the moment, that we observe for $i = 1, \ldots, n$, $(X_i, \delta_i, \boldsymbol{Z_i})$, where $X_i = \min(T_i, C_i)$, $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$, T_i is the time of interest, and C_i the censoring time. The usual kernel estimator is then obtained from an estimator of the cumulative baseline function A_0 defined by $A_0(t) = \int_0^t \alpha_0(s)\mathrm{d}s$. This estimator is called the Breslow estimator and is defined, for $t > 0$, by

$$\hat{A}_0(t, \hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \frac{\delta_i}{S_n(X_i, \hat{\boldsymbol{\beta}})}, \qquad \text{with } S_n(t, \hat{\boldsymbol{\beta}}) = \sum_{i:T_i \geq t} \exp(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z_i}), \tag{2}$$

see Ramlau-Hansen [24] and Andersen et al. [2] for details. From $\hat{A}_0(., \hat{\boldsymbol{\beta}})$, the kernel function estimator for α_0 is derived by smoothing the increments of the Breslow estimator. It is defined by

$$\hat{\alpha}_h^{\hat{\beta}}(t) = \frac{1}{h} \int_0^\tau K\Big(\frac{t-u}{h}\Big) \mathrm{d}\hat{A}_0(u, \hat{\boldsymbol{\beta}}), \quad \tau \geq 0, \tag{3}$$

with $K : \mathbb{R} \mapsto \mathbb{R}$ a kernel with integral 1, and h a positive parameter called the bandwidth. This estimator was introduced and studied by Ramlau-Hansen [23, 24] within the

framework of the multiplicative intensity model for counting processes, thereby extending its use to censored survival data. Consistency and asymptotic normality are proven in Ramlau-Hansen [24] with fixed bandwidth.

The choice of the bandwidth in kernel estimation is crucial, in particular when one is interested in establishing non-asymptotic adaptive inequalities. State-of-the-art methods are based on cross-validation. Ramlau-Hansen [22] has suggested the cross-validation method to select the bandwidth but without any theoretical guarantees. For randomly censored survival data, Marron and Padgett [21] have shown that the cross-validation method gives the optimal bandwidth for estimating the density: the ratio between the integrated squared error for the cross-validation bandwidth and the infimum of the integrated squared error for any bandwidth almost surely converges to 1. Grégoire [15] has considered the cross-validated method suggested by Ramlau-Hansen [22] for adaptive estimation of the intensity of a counting process and has proved some consistency and asymptotic normality results for the cross-validated kernel estimator.

However, all the results for the adaptive kernel estimator with a cross-validated bandwidth are asymptotic. No non-asymptotic oracle inequalities have to date been stated for the kernel estimator of the baseline function. In addition, to our knowledge, the construction of $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ has not yet been considered for high-dimensional covariates. The goal of the present paper is thus twofold: whatever the dimension, we aim to propose an estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}$ of the baseline function, for which we can establish a non-asymptotic oracle inequality to measure its performance. The loss of prediction quality of $|\hat{\alpha}^{\hat{\boldsymbol{\beta}}} - \alpha_0|$ when p increases will be quantified.

To fulfill these purposes, the idea is to first estimate the regression parameter $\boldsymbol{\beta_0}$ via a LASSO procedure applied to the Cox partial log-likelihood, then to plug this estimator in the usual kernel estimator (3) of the baseline hazard function; then, lastly, to select a data-driven bandwidth, following a procedure adapted from Goldenshluger and Lepski [14]. In the latter, the problem of bandwidth selection in kernel density estimation is addressed and an adaptive estimator is derived, which satisfies non-asymptotic minimax bounds. This method was then considered by Doumic et al. [11] for estimating the division rate of a size-structured population in a non-parametric setting, by Bouaziz et al. [3] to estimate the intensity function of a recurrent event process, and by Chagny [8] for the estimation of a real function via a warped kernel strategy. In the present paper, we consider this method in order to obtain an adaptive kernel estimator of the baseline function with a data-driven bandwidth. We establish the first adaptive and non-asymptotic oracle inequality, which guarantees the theoretical performance of this kernel estimator. The oracle inequality depends on non-asymptotic control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ deduced from an estimation inequality in Huang et al. [17] and extended to the case of unbounded counting processes (see Guilloux et al. [16] for details).

The paper is organized as follows. In Section 3, we describe the two-step procedure to estimate the baseline function: first, we describe the estimation of $\boldsymbol{\beta_0}$ as a preliminary step and give the bound for $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$, and then we focus on the kernel estimation of α_0 and describe the adaptive estimation procedure of Goldenshluger and Lepski to select a data-driven bandwidth. In Section 4, we establish a non-asymptotic oracle inequality for

the adaptive kernel estimator. Proofs are gathered in Section 6. Lastly, supplementary materials provide some technical results needed for the proofs.

2 Notation and preliminaries

2.1 Framework for counting processes

Consider the general setting of counting processes, which embraces the classical case of right censoring. We follow here the now classical setting of Andersen et al. [2] or Fleming and Harrington [13]. For n independant individuals, we observe for $i = 1, \dots, n$ a counting process N_i, a random process Y_i with values in $[0, 1]$, and a vector of covariates $\boldsymbol{Z_i} = (Z_{i,1}, \dots, Z_{i,p})^\top \in \mathbb{R}^p$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ the filtration defined by

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), 0 \leq s \leq t, \boldsymbol{Z_i}, i = 1, \dots, n\}.$$

From the Doob-Meyer decomposition, we know that each N_i admits a compensator denoted by Λ_i, such that $M_i = N_i - \Lambda_i$ is a $(\mathcal{F}_t)_{t\geq 0}$ local square-integrable martingale (see Andersen et al. [2] for details). We assume in the following that N_i satisfies an Aalen multiplicative intensity model.

Assumption 2.1. *For each $i = 1, \dots, n$ and all $t \geq 0$,*

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i}) Y_i(s) \mathrm{d}s, \tag{4}$$

where $\lambda_0(t, \boldsymbol{z}) = \alpha_0(t) \mathrm{e}^{\beta^\top \boldsymbol{z}}$, for $\boldsymbol{z} \in \mathbb{R}^p$.

This general setting, introduced by Aalen [1], includes the particular examples of censored data, marked Poisson processes and Markov processes (see Andersen et al. [2] for further details). This framework generalizes the case considered in Ramlau-Hansen [24] to unbounded counting processes and hence widens the scope of applications: we can consider the jumps of the counting to happen at times of relapse from a disease in biomedical research, monetization times in marketing, blogging times in social network studies, etc.

2.2 Notation

For a real number $q \geq 1$ and a function $f : \mathbb{R} \longmapsto \mathbb{R}$ such that $|f|^q$ is integrable and bounded, we consider

$$||f||_{\mathbb{L}^q(\mathbb{R})} = \left(\int_{\mathbb{R}} |f(x)|^q \mathrm{d}x \right)^{1/q} \text{ and } ||f||_\infty = \sup_{x\in\mathbb{R}} |f(x)|.$$

The integrals and the supremum are restricted to the support of f, and for τ a positive real number, we set $||f||_{\infty,\tau} = \sup_{x\in[0,\tau]} |f(x)|$ and simply denote by $||.||_2$ the $\mathbb{L}^2$-norm restricted to the interval $[0, \tau]$, so that

$$||f||_2^2 = \int_0^\tau f^2(x) \mathrm{d}x.$$

For h a positive real number, we define $f_h(.) = f(./h)/h$. For square-integrable functions f and g from $\mathbb{R}$ to $\mathbb{R}$, we denote the convolution of f and g by $f * g$. For a vector $\boldsymbol{b} \in \mathbb{R}^p$ and a real $q \geq 1$, we denote $|\boldsymbol{b}|_q = (\sum_{j=1}^p |b_j|^q)^{1/q}$.

For quantities $\gamma(n)$ and $\eta(n)$, the notation $\gamma(n) = \mathcal{O}(\eta(n))$ means that there exists a positive constant c such that $\gamma(n) \leq c\eta(n)$.

Lastly, let $\boldsymbol{Z} \in \mathbb{R}^p$ denote the generic vector of covariates with the same distribution as the vectors of covariates $\boldsymbol{Z_i}$ of each individual i, and Z_j its j-th component, namely the j-th covariate of the vector $\boldsymbol{Z}$.

3 Estimation procedure

In this section, we describe the two-step procedure for estimating the baseline function. We begin by recalling the usual estimation of the regression parameter $\boldsymbol{\beta_0}$ in the high-dimensional setting. We then focus our study on the second step, which consists of the adaptive kernel estimation of the baseline function α_0.

3.1 Preliminary estimation of $\boldsymbol{\beta_0}$

The regression parameter $\boldsymbol{\beta_0}$ is estimated via a LASSO procedure applied to the so-called Cox partial log-likelihood introduced by Cox [9], and defined, for all $\boldsymbol{\beta} \in \mathbb{R}^p$, by

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \ln \frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})} \mathrm{d}N_i(t), \quad \text{where } S_n(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^n \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z_i}} Y_i(t). \tag{5}$$

The estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta_0}$ is then defined by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{B}(0,R)}{\arg\min}\{-l_n^*(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \quad \text{with} \quad \mathrm{pen}(\boldsymbol{\beta}) = \Gamma_n |\boldsymbol{\beta}|_1, \tag{6}$$

where Γ_n is a well-chosen positive regularization parameter, and $\mathcal{B}(0,R)$ the ball defined by

$$\mathcal{B}(0,R) = \{b \in \mathbb{R}^p : |b|_1 \leq R\}, \quad \text{with } R > 0.$$

The ball constraint has already been considered by van de Geer [27] and Kong and Nan [18]. Roughly speaking, it means that we have restricted our attention to a (possibly very large) ball around $\boldsymbol{\beta_0}$, for which the following (very mild) assumption is needed, ensuring control of the kernel estimator of the baseline function $\boldsymbol{\beta_0}$.

Assumption 3.1. *We assume that* $|\boldsymbol{\beta_0}|_1 < +\infty$.

Concerning the covariates, we introduce the following assumption.

Assumption 3.2. *There exists a positive constant* B *such that for all* $j \in \{1, \ldots, p\}$,

$$|Z_j| \leq B.$$

Assumption 3.2 is a classical assumption in the Cox model to obtain oracle inequalities (see Huang et al. [17] and Bradic and Song [5]) and seems reasonable since in practice the covariates are bounded.

We define by $\dot{\boldsymbol{l}}_{\boldsymbol{n}}^{*}(\boldsymbol{\beta}) = (\dot{l}_{n,1}^{*}(\boldsymbol{\beta}), \ldots, \dot{l}_{n,p}^{*}(\boldsymbol{\beta}))^{\top} = \partial l_n^{*}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ the gradient of the Cox partial log-likelihood $l_n^{*}(\boldsymbol{\beta})$ defined by (5), and $\ddot{\boldsymbol{l}}_{\boldsymbol{n}}^{*}(\boldsymbol{\beta}) = \partial^2 l_n^{*}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\top}$ the Hessian matrix.

Let us also denote by $\mathcal{O} = \{j : (\beta_0)_j \neq 0\}$ the sparsity set of $\boldsymbol{\beta_0}$, where $(\beta_0)_j$ is the j-th coordinate of vector $\boldsymbol{\beta_0}$, and $s = |\mathcal{O}|$ the cardinality of $\mathcal{O}$, called the sparsity index. For any $\xi > 1$, we define the cone

$$\mathcal{C}(\xi, \mathcal{O}) = \{\boldsymbol{b} \in \mathbb{R}^p : |\boldsymbol{b}_{\mathcal{O}^c}|_1 \leq \xi |\boldsymbol{b}_{\mathcal{O}}|_1\}.$$

For this cone, we define the following condition:

Assumption 3.3. *For any $\xi > 1$, let us assume that*

$$0 < \kappa(\xi, \mathcal{O}) = \inf_{0 \neq \boldsymbol{b} \in \mathcal{C}(\xi, \mathcal{O})} \frac{s^{1/2}(\boldsymbol{b}\ddot{\boldsymbol{l}}_{\boldsymbol{n}}^{*}(\boldsymbol{\beta_0})\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_{\mathcal{O}}|_1}.$$

This term corresponds to the compatibility factor introduced by van de Geer [26]. It is one of the classical conditions used to obtain non-asymptotic oracle inequalities for the LASSO estimator (6). See also Bühlmann and van de Geer [6] for more details about this compatibility factor, and a comparison of this criterion with other assumptions, such as the restricted eigenvalue condition among others.

We now give a general version of the estimation inequality of Theorem 3.1 of Huang et al. [17]. We refer to Guilloux et al. [16] for a proof of Proposition 3.4 in the general case.

Proposition 3.4. *Let $k > 0$, $c > 0$, and let s be the sparsity index of $\boldsymbol{\beta_0}$. Assume that $||\alpha_0||_{\infty,\tau} < \infty$ and*

$$\Gamma_n = C_0 B \frac{\xi + 1}{\xi - 1} \sqrt{\frac{2\ln(pn^k)}{n}},$$

where $\xi > 1$ and $C_0 > \sqrt{\tau ||\alpha_0||_{\infty,\tau} \mathbb{E}\{\mathrm{e}^{\beta_0^{\top} \boldsymbol{Z}}\}}$. Let $\nu = B(\xi+1)s\Gamma_n/\{2\kappa^2(\xi, \mathcal{O})\}$, and assume that $\nu \leq 1/\mathrm{e}$. Then, under Assumptions 3.1, 3.2 and 3.3, with probability larger than $1 - cn^{-k}$, we have

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq \frac{\mathrm{e}^{\eta}(\xi+1)s}{2\kappa^2(\xi, \mathcal{O})}\Gamma_n =: C(s)\sqrt{\frac{\ln(pn^k)}{n}},$$

where $\eta \leq 1$ is the smaller solution of $\eta \mathrm{e}^{-\eta} = \nu$.

For the sake of simplicity, we will present the bound of the estimation error as follows:

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\ln(pn^k)}{n}}, \tag{7}$$

where $C(s) > 0$ is a constant that depends on the sparsity index s.

In the rest of the paper, the conditions of Proposition 3.4 will be fulfilled, so $\hat{\boldsymbol{\beta}}$ satisfies inequality (7). The assumption $||\alpha_0||_{\infty,\tau} < \infty$ is part of Assumption 3.5.(iii) described in Section 3.2.1.

3.2 Estimation of α_0

In this section, we define the kernel estimator of the baseline hazard function α_0 on which our procedure relies. We state several functional and kernel assumptions, and describe the Goldenshluger and Lepski procedure for selecting a data-driven bandwidth.

3.2.1 Kernel estimator

We first recall the definition of the kernel estimator introduced by Ramlau-Hansen [24] by using kernel functions to smooth the increments of the non-parametric Breslow estimator (2) of the cumulative intensity.

Let $K : \mathbb{R} \to \mathbb{R}$ be a function such that $\int_{\mathbb{R}} K(x)\mathrm{d}x = 1$, wich will play the role of a kernel. The usual functional estimation of α_0 is defined by

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}N_i(u), \tag{8}$$

with

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad S_n(u,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z_i}} Y_i(u), \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p.$$

The parameter $h > 0$ is called the bandwidth. The bandwidth of kernel estimators has to be chosen by the user. Grégoire [15] has defined a cross-validation procedure for selecting the bandwidth for a smooth estimate of the intensity in the Aalen counting process. To our knowledge, all theoretical results for the kernel function estimator (8) with a bandwidth selected by cross-validation are asymptotic. The cross-validation provides no theoretical adaptive guarantees when the size of the sample n is fixed, and not as large as is the case for medical surveys where only a few patients can be observed. This explains our interest in providing a data-driven method to automatically select the bandwidth and obtain a kernel function estimator, for which we can guarantee certain non-asymptotic properties.

In the following, we denote the estimator under study by $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ into which the LASSO estimator (6) has been plugged.

3.2.2 Functional and kernel assumptions

Classical conditions are required on the intensity function and the kernel K.

Assumption 3.5.

(i) *For all $i \in \{1,\dots,n\}$, the random process Y_i takes its values in $\{0,1\}$.*

(ii) *For* $S(t, \boldsymbol{\beta_0}) = \mathbb{E}\{e^{\boldsymbol{\beta_0}^\top \boldsymbol{Z}_i} Y_i(t)\}$, *there exists a positive constant* c_S *such that,*

$$S(t, \boldsymbol{\beta_0}) \geq c_S, \quad \forall t \in [0, \tau].$$

(iii) $||\alpha_0||_{\infty,\tau} := \sup_{t\in[0,\tau]} \alpha_0(t) < \infty$.

Assumption 3.5.(i) is satisfied for all the examples quoted in the introduction. In fact, this assumption is needed to ensure that the random process Y_i has a lower bound when it is nonzero. We could also have considered a modified estimator of $S_n(u, \boldsymbol{\beta})$, defined by (5), as it is usually done in the censoring case without covariates. Assumption 3.5.(ii) is common in the context of estimation with censored observations (see Andersen et al. [2]). Assumption 3.5.(iii) is required to obtain Lemma 6.1 and Theorem 4.1 below. Nevertheless, the value $||\alpha_0||_{\infty,\tau}$ is not needed to compute the estimator (see Section 5).

The following assumptions are fulfilled by many standard kernel functions and are standard for kernel estimators.

Assumption 3.6.

(i) $||K||_\infty = \sup_{u\in\mathbb{R}} |K(u)| < \infty$ *and* $||K||_2^2 = \int_{\mathbb{R}} K^2(u)\mathrm{d}u < \infty$.

(ii) $nh \geq 1$ *and* $0 < h < 1$.

(iii) *The kernel* K *is of order* 1, *i.e., for* $j \in \{0, 1, 2\}$ *the function* $x \mapsto x^j K(x)$ *is integrable and*

$$\int_{\mathbb{R}} xK(x)\mathrm{d}x = 0 \quad \textit{and} \quad \int_{\mathbb{R}} x^2 K(x)\mathrm{d}x < \infty.$$

Assumptions 3.6.(i) and 3.6.(ii) are rather standard in kernel density estimation (see Goldenshluger and Lepski [14]) and have also been considered in kernel intensity estimation by Bouaziz et al. [3]. Assumption 3.6.(iii) is only required to ensure that $K_h * \alpha_0(t) \underset{h\to 0}{\longrightarrow} \alpha_0(t)$ for all $t \in [0, \tau]$.

Remark 3.7. *In this paper, we do not assume that the kernel* K *has a compact support, unlike* Bouaziz et al. [3]. *The Breslow estimator (8) and the pseudo-estimator (11) are then well-defined for all* $t \in [0, \tau]$.

3.2.3 Collections of estimators

Let $\mathcal{H}_n$ be a grid of bandwidths $h > 0$, satisfying the following assumptions:

Assumption 3.8.

(i) $\mathrm{Card}(\mathcal{H}_n) \leq n$.

(ii) *For some* $a \geq 0$, $\sum_{h\in\mathcal{H}_n} \frac{1}{nh} = \mathcal{O}(\ln^a(n))$.

(iii) *For all* $b > 0$, $\sum_{h \in \mathcal{H}_n} \exp(-b/h) < +\infty$.

Assumptions 3.8.(i)-(iii) mean that the bandwidth collection should not be too large. Let us give an example of a grid $\mathcal{H}_n$ that satisfies the three previous assumptions.

Example 3.9 (Example of $\mathcal{H}_n$)**.** *The following grid is considered in the simulations in Section 5:*

$$\mathcal{H}_n = \left\{ h_j = \frac{1}{2^j}, j = 1, \ldots, \lfloor \ln(n)/\ln(2) \rfloor \right\}.$$

For this grid, all of the assumptions on the bandwidths are satisfied. Indeed, $\text{Card}(\mathcal{H}_n) \leq \ln(n)/\ln(2) \leq n$ *and* $\forall k = 1, \ldots, \lfloor \ln(n)/\ln(2) \rfloor$*, we have* $h_j \in [n^{-1}, 1]$*. Moreover, Assumption 3.8.(ii) holds true since*

$$\sum_{j: h_j \in \mathcal{H}_n} \frac{1}{n h_j} = \frac{1}{n} \sum_{j=1}^{\lfloor \ln(n)/\ln(2) \rfloor} 2^j = O(1).$$

Lastly,

$$\sum_{j: h_j \in \mathcal{H}_n} \exp(-b/h_j) = \sum_{j=1}^{\lfloor \ln(n)/\ln(2) \rfloor} \mathrm{e}^{-b2^j} = O(1),$$

so Assumption 3.8.(iii) is satisfied.

On the grid $\mathcal{H}_n$, we obtain a set of kernel estimators $\mathcal{F}(\mathcal{H}_n) = \{\hat{\alpha}_h^{\hat{\beta}}, h \in \mathcal{H}_n\}$ of the baseline function α_0 from definition (8).

3.2.4 Adaptive selection of the bandwidth

We wish to automatically select a relevant bandwidth $h \in \mathcal{H}_n$, in such a way as to then be able to select a kernel estimator from the set $\mathcal{F}(\mathcal{H}_n)$. As usual, we must choose a bandwidth h which gives the best compromise between the squared-bias and variance terms. The choice should be data-driven. For this, we use a relatively recent method introduced by Goldenshluger and Lepski [14] for the problem of density estimation. The "Goldenshluger and Lepski method" has only been considered in two different settings: Bouaziz et al. [3] has applied it to provide an adaptive kernel function estimator of the intensity function of a recurrent event process, and Chagny [8] has used it to estimate a real-valued function from a sample of random pairs. Recently, Chagny [7] has also proposed a "mixed strategy", which consists in applying the "Goldenshluger and Lepski method" to select the relevant model in model selection methods for real-valued function in regression models. We use this method to obtain an adaptive kernel function estimator of the baseline function, for which we establish a non-asymptotic oracle inequality.

We start from the collection of estimators

$$\mathcal{F}(\mathcal{H}_n) = \{\hat{\alpha}_h^{\hat{\beta}}, h \in \mathcal{H}_n\}.$$

For $h \in \mathcal{H}_n$, the optimal bandwidth in this collection is the one that minimizes the excess risk:

$$\arg\min_{h\in\mathcal{H}_n}\Big\{\mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 \Big\}. \tag{9}$$

This optimal bandwidth, which can be seen as the oracle, is unavailable since it depends on the true unknown function α_0. The idea of the method is to estimate an upper bound of the excess risk and then to estimate the oracle. The first step consists in bounding the excess risk, by showing that it satisfies

$$\mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 \leq C\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E} \parallel \bar{\alpha}_h - K_h * \alpha_0 \parallel_2^2 + \mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h \parallel_2^2 \Big\}, \tag{10}$$

where $C > 0$, $K_h(.) = (1/h)K(./h)$ and $\mathbb{E}(\bar{\alpha}_h) = K_h * \alpha_0$, with $\bar{\alpha}_h$ a pseudo-estimator defined by

$$\bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{1}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u). \tag{11}$$

In fact $\bar{\alpha}_h$ corresponds to the kernel estimator of α_0 when $S(u,\boldsymbol{\beta_0}) = \mathbb{E}\{\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_i}} Y_i(u)\}$ is known (see Section 6.1 for more details about the pseudo-estimator). The second step consists in bounding $\mathbb{E} \parallel \bar{\alpha}_h - \hat{\alpha}_h^{\hat{\beta}} \parallel^2$ by a constant $C(\hat{\beta}, \beta_0)$ that does not depend on h (see Lemma 6.3). Hence we get that

$$\mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 \leq C(\hat{\beta}, \beta_0) + C\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E} \parallel \bar{\alpha}_h - K_h * \alpha_0 \parallel_2^2 \Big\}. \tag{12}$$

Now, let h^* be

$$h^* = \arg\min_{h\in\mathcal{H}_n}\Big\{ \parallel \alpha_0 - K_h * \alpha_0 \parallel_2^2 + \mathbb{E} \parallel \bar{\alpha}_h - K_h * \alpha_0 \parallel_2^2 \Big\}.$$

The idea is now to estimate h^* instead of the oracle defined in (9). The third step consists of estimating the term $B(h) = ||\alpha_0 - K_h * \alpha_0||_2^2$, which is unknown and can be seen as a bias term. We derive from the Goldenshluger and Lepski method an estimator of this bias term as follows:

$$A^{\hat{\beta}}(h) = \sup_{h'\in\mathcal{H}_n}\Big\{ \parallel \hat{\alpha}_{h,h'}^{\hat{\beta}} - \hat{\alpha}_{h'}^{\hat{\beta}} \parallel_2^2 - V(h')\Big\}_+, \tag{13}$$

where for any $t \geq 0$ and h, h' positive real numbers,

$$\hat{\alpha}_{h,h'}^{\hat{\beta}}(t) = K_{h'} * \hat{\alpha}_h^{\hat{\beta}}(t), \tag{14}$$

and where $V(h)$ is an upper bound of the variance term $\mathbb{E} \parallel \bar{\alpha}_h - K_h * \alpha_0 \parallel_2^2$. More precisely, $V(h)$ is obtained from the control of $\mathbb{E}\{A^{\hat{\beta}}(h)\}$ by applying a Talagrand inequality to show that

$$V(h) = \mathcal{O}\Big(\frac{1}{nh}\Big).$$

We refer to the proof of Lemma 6.5 for more details and the exact definition of $V(h)$.

The heuristic involved in this method, initially proposed by Goldenshluger and Lepski [14], can be found in Chagny [7]. In order to get closer to the bias term $\| \alpha_0 - K_h * \alpha_0 \|_2^2$, we replace α_0 with an estimator $\hat{\alpha}_{h'}^{\hat{\beta}}$ (with a fixed bandwidth h'), so that we obtain $\| \hat{\alpha}_{h'}^{\hat{\beta}} - K_h * \hat{\alpha}_{h'}^{\hat{\beta}} \|_2^2$. However, unlike the bias term, this quantity is random and thus contains some variability. We need to correct this variability by deducting the part of the variance $V(h')$. Lastly, since there is no reason to choose one bandwidth $h' \in \mathcal{H}_n$ over another, we consider the entire collection and take the maximum over it.

From these definitions, we deduce the following choice of bandwidth:

$$\hat{h}^{\hat{\beta}} = \underset{h \in \mathcal{H}_n}{\arg\min} \{ A^{\hat{\beta}}(h) + V(h) \}. \tag{15}$$

Our adaptive kernel estimator is then $\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}}$.

4 Non-asymptotic bounds for the kernel estimator

Let us now state the main theorems of the paper, which provide the first non-asymptotic oracle inequality for the adaptive kernel baseline estimator in the high-dimensional setting.

Theorem 4.1. *Under Assumptions 3.1, 3.2, 3.3, 3.5.(i)-(iii) and 3.6.(i)-(iii), if $\mathcal{H}_n$ is a finite discrete set of bandwidths such that 3.8.(i)-(iii) are satisfied, then there exists a constant κ such that $\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}}$ defined by (13) and (15) satisfies, for n large enough and $k \geq 12$,*

$$\mathbb{E} \| \hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \alpha_0 \|_2^2 \leq C \inf_{h \in \mathcal{H}_n} \left\{ \| \alpha_h - \alpha_0 \|_2^2 + V(h) \right\} + C'(s) \frac{\ln^a(n) \ln(pn^k)}{n}, \tag{16}$$

with

$$V(h) = \kappa \frac{||\alpha_0||_{\infty,\tau} \tau}{c_S^2} \left(||\alpha_0||_{\infty,\tau} \mathbb{E}\{ \mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}} \} \tau + \mathbb{E}\{ \mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}} \} \right) \frac{||K||_{\mathbb{L}^2(\mathbb{R})}^2}{nh}, \tag{17}$$

where C is a constant, $C'(s)$ a constant depending on τ, κ_b the bound of the Bürkholder Inequality, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_{\mathbb{L}^1(\mathbb{R})}$, $||K||_{\mathbb{L}^2(\mathbb{R})}$, and on the sparsity index s of $\boldsymbol{\beta_0}$.

Remark 4.2. *The condition $k \geq 12$ is simply used for technical convenience. It comes from the control of the kernel estimator (8) and the pseudo-estimator (11) introduced in Section 3.2.4. It is required to ensure convergence of order $1/n$ of the quadratic error between the kernel estimator and the pseudo-estimator. We refer the reader to the proof of Proposition 6.4 for more details.*

This inequality ensures that the adaptive kernel estimator $\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}}$ automatically makes the squared-bias/variance compromise. The selected bandwidth $\hat{h}^{\hat{\beta}}$ performs as well as the unknown oracle, up to the multiplicative constant C and up to a remaining term

of order $\ln^a(n)\ln(pn^k)/n$, which is negligible. In Inequality (16), the infimum term is a classical one in such oracle inequalities for kernel estimators: the bias term $||\alpha_h - \alpha_0||_2^2$ decreases when h decreases and the variance term $V(h)$ increases when h decreases. The remaining terms are of order

$$\frac{\ln^a(n)\ln(pn^k)}{n} = \frac{k\ln^{a+1}(n)}{n} + \frac{\ln^a(n)\ln(p)}{n}.$$

Chagny [8], in the context of an additive regression model, established an oracle inequality for the kernel estimator of the real-value regression function with a term remaining of order $1/n$. In the context of estimation of the intensity of a recurrent event process observed under a standard censoring scheme but without covariates, Bouaziz et al. [3] have a logarithm term which appears in their oracle inequality, with a term remaining of order $\ln^{1+a}(n)/n$ instead of the expected $1/n$. This logarithmic term comes from the control in $\ln(n)/n$ between the distribution function G and its modified Kaplan-Meier estimator $\hat{G}$, which appears in the kernel intensity estimator. The exponent a in the remaining term arises from Assumption 3.8.(ii), which is needed for control of the difference between the kernel intensity estimator involving $\hat{G}$ and a pseudo-estimator that does not depend on $\hat{G}$. As for Bouaziz et al. [3], our kernel estimator depends on another estimator, so we need Assumption 3.8.(ii) in order to control the difference between the kernel estimator (8) and the pseudo-estimator (11). If our kernel estimator did not involve another estimator, we would have considered condition $\sum 1/h \leq k_0 n^{a_0}$, as in Chagny [8], instead of Assumption 3.8.(ii). The term in $\ln(p)/n$ in the remaining term comes from the control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ given by Proposition 3.4. This term is typical of the estimation of the regression parameter $\boldsymbol{\beta_0}$ when the number of covariates is large. There is no hope of catching up to usual rates in this high-dimensional setting, but the loss in the variance term is only of order $\ln p/n$.

When we assume that the counting processes N_i are bounded for $i = 1, \ldots, n$, the variance term $V(h)$ is simpler and has the same form as the variance term in Bouaziz et al. [3]. In this particular case, Theorem 4.1 takes the following form.

Theorem 4.3. *Under the same assumptions as Theorem 4.1 and assuming also that there exists $c_\tau > 0$, such that $N_i(t) \leq c_\tau$ almost surely for every $t \in [0, \tau]$ and $i \in \{1, \ldots, n\}$, there exists a constant κ such that $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ defined by (13), (15) and*

$$V_b(h) = \kappa \frac{c_\tau \tau ||\alpha_0||_{\infty,\tau}}{c_S} \frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nh}, \tag{18}$$

satisfies for n large enough:

$$\mathbb{E} \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0 \parallel_2^2 \leq \tilde{C} \inf_{h \in \mathcal{H}_n} \left\{ \parallel \alpha_h - \alpha_0 \parallel_2^2 + V_b(h) \right\} + \tilde{C}'(s) \frac{\ln^a(n)\ln(np)}{n}, \tag{19}$$

where $\tilde{C}$ is a constant, and $\tilde{C}'(s)$ a constant depending on τ, c_τ, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_{\mathbb{L}^1(\mathbb{R})}$, $||K||_{\mathbb{L}^2(\mathbb{R})}$, and on the sparsity index s of $\boldsymbol{\beta_0}$.

The proof of Theorem 4.3 is similar to that of Theorem 4.1 and we refer to Lemler [19] for details.

5 Applications

5.1 Simulation study

The aim of this section is to illustrate the behavior of the kernel estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ of the baseline function in the case of right censoring, and to compare it with the usual kernel estimator with a bandwidth selected by cross-validation introduced by Ramlau-Hansen [24].

5.1.1 Simulated data: censored data.

We consider a cohort of size n with p covariates simulated according to the Cox model (1) with right censoring and with regression parameter $\boldsymbol{\beta_0}$ chosen as a vector of dimension p, defined by

$$\boldsymbol{\beta_0} = (0.1, 0.3, 0.5, 0, \dots, 0)^\top \in \mathbb{R}^p,$$

for various $p \geq 3$. Several choices of n and p have been considered, with sample size n taking values $n = 200$ and $n = 500$ and p varying between $p = \sqrt{n}$, being 15 and 22 respectively, and $p = n$, referred to as the high-dimensional case. For each n and p, the design matrix $\boldsymbol{Z} = (Z_{i,j})_{1\leq i\leq n, 1\leq j\leq p}$ is simulated independently from a uniform distribution on $[-1, 1]$, and survival times T_i, $i = 1, \dots, n$ are simulated according to Weibull distributions $\mathcal{W}(1.5, 1)$ and $\mathcal{W}(0.5, 2)$. Hence, the associated baseline function has the form $\alpha_0(t) = a\lambda^a t^{a-1}$, where a and λ stand for the parameters in $\mathcal{W}(a, \lambda)$. The censoring times C_i, for $i = 1, \dots, n$, are simulated independently from the survival times via an exponential distribution $\mathcal{E}\{1/\gamma\mathbb{E}(T_1)\}$, where γ is adjusted to the chosen rate of censorship: $\gamma = 4.5$ for 20% censorship and $\gamma = 1.2$ for 50%.

The time τ of the end of the study is taken as the quantile at 90% of $(T_i \wedge C_i)_{i=1,\dots,n}$. For $i = 1, \dots, n$, we compute the observed times $X_i = \min(T_i, \tilde{C}_i)$, where $\tilde{C}_i = C_i \wedge \tau$ and the censoring indicators are $\delta_i = \mathbb{1}_{T_i\leq C_i}$. The definition of $\tilde{C}_i$ ensures the existence of $i \in \{1, \dots, n\}$ such that $X_i \geq \tau$, so that all estimators are defined on the interval $[0, \tau]$, and prevents a certain edge effect. Each sample $(\boldsymbol{Z_i}, T_i, C_i, X_i, \delta_i, i = 1, \dots, n)$ is simulated $N_e = 100$ times.

The estimators of the baseline hazard function are both constructed with the Epanechnikov kernel, defined by

$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{\{|u|\leq 1\}}.$$

In both cases, we plug-in the LASSO regression parameter estimator $\hat{\boldsymbol{\beta}}$ defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\Big\{ - l_n^*(\boldsymbol{\beta}) + \mu|\boldsymbol{\beta}|_1\Big\},$$

implemented in the R-package *glmnet*. The regularization parameter μ has been chosen classically by cross-validation from the LASSO procedure, using the R-function *cv.glmnet*.

The goal of this simulation study is to compare two procedures for the data-driven choice of h: the Goldenshluger and Lepski method with the selected bandwidth denoted by $\hat{h}^{\hat{\boldsymbol{\beta}}}_{GL}$, and cross-validation with the selected bandwidth denoted by $\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}$.

5.1.2 The Goldenshluger and Lepski method

The adaptive bandwidth selection method we consider here is based on the grid of bandwidths $\mathcal{H}_n$ defined in Example 3.9 by

$$\mathcal{H}_n = \{1/2^k, k = 0, \dots, \lfloor \ln(n)/\ln(2) \rfloor\}.$$

In our procedure (13), the variance term $V(h)$ involves unknown quantities, so we consider a data-driven equivalent of it and use the fact that right-censoring implies that the counting processes are bounded. We thus implement the following procedure:

$$\hat{h}^{\hat{\boldsymbol{\beta}}}_{GL} = \arg\min_{h \in \mathbb{H}_n} \{A^{\hat{\boldsymbol{\beta}}}(h) + \hat{V}^{\hat{\boldsymbol{\beta}}}_b(h)\},$$

where, for any $t \geq 0$ and h, h' positive real numbers,

$$A^{\hat{\boldsymbol{\beta}}}(h) = \sup_{h' \in \mathcal{H}_n} \left\{ \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'} \parallel_2^2 - \hat{V}^{\hat{\boldsymbol{\beta}}}_b(h') \right\}_+, \qquad \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(t) = K_{h'} * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h(t),$$

and

$$\hat{V}^{\hat{\boldsymbol{\beta}}}_b(h) = \kappa' \frac{||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(h)}||_{\infty,\tau} ||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nh}.$$

In the variance term $\hat{V}^{\hat{\boldsymbol{\beta}}}_b(h)$, we have replaced the true unknown function α_0 by an estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(h)}$ computed for the largest bandwidth h in the grid $\mathcal{H}_n$ (see Bouaziz et al. [3]). The constant κ' is a universal constant that we tuned from the comparison of the MISEs for several candidate values in the range 10^{-4} to 1000, and for the two different distributions $\mathcal{W}(1.5, 1)$ and $\mathcal{W}(0.5, 2)$. We took $\kappa' = 1$.

5.1.3 Cross-validation method

The bandwidth $\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}$ selected by cross-validation is defined by:

$$\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV} = \arg\min_h \left\{ \mathbb{E} \int_0^\tau \{\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h(t)\}^2 \mathrm{d}t - 2 \sum_{i \neq j} \frac{1}{h} K\Big(\frac{X_i - X_j}{h}\Big) \frac{\delta_i}{\bar{Y}(X_i)} \frac{\delta_j}{\bar{Y}(X_j)} \right\},$$

where $\bar{Y} = \sum_{i=1}^n \mathbb{1}_{\{X_i \geq t\}}$.

5.1.4 Performance

The performance of these two estimators is evaluated via versions of the integrated squared error (ISE). For some function $\alpha \in \mathbb{L}^2([0, \tau])$ the standard ISE and the total ISE are respectively defined by

$$\mathrm{ISEstand}(\alpha) = \int_0^\tau \{\alpha(t) - \alpha_0(t)\}^2 \mathrm{d}t,$$

$$\mathrm{ISEtotal}(\alpha, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \{\alpha(t) \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{Z_i}} - \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_i}}\}^2 \mathrm{d}t.$$

The associated mean integrated squared errors are defined by $\mathrm{MISEg}(\alpha) = \mathbb{E}\{\mathrm{ISEg}(\alpha)\}$, for g=stand or total, where the expectation is taken on $(T_i, C_i, \boldsymbol{Z_i})$ (for the sake of simplicity, we write $\mathrm{MISEg}(\alpha)$ even if the MISE depends on $\boldsymbol{\beta}$). We obtain an estimation of the MISE by taking the empirical mean over the $N_e = 100$ replicates.

5.1.5 Results

We first study the case where we assume the regression parameter $\boldsymbol{\beta_0}$ to be known. Thus, we directly plug this parameter into the kernel estimators. Table 1 gives two empirical MISE of the kernel estimators, with a bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method, for survival times that are distributed from $\mathcal{W}(1.5, 1)$ in the case of a known regression parameter $\boldsymbol{\beta_0}$.

	MISEstand		MISEtotal	
	GL	CV	GL	CV
$n = 200$ and $p = 15$	0.01526653	0.01803263	0.01919891	0.02267852

Table 1 – MISE of the kernel estimators with bandwidth selected by the Goldenshluger and Lepski method (GL), and with bandwidth selected by cross-validation (CV) of the baseline function for a known regression parameter $\boldsymbol{\beta_0}$, with $n = 200$ and $p = 15$.

From Table 1, we see that the Goldenshluger and Lepski method gives smaller MISE than cross-validation; it appears that the Goldenshluger and Lepski method provides better results than cross-validation when the regression parameter $\boldsymbol{\beta_0}$ is assumed to be known.

Table 2 gives the two empirical MISE of the kernel estimators with bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method, for a LASSO estimator of the regression parameter, and survival times that are distributed from $\mathcal{W}(1.5, 1)$, in different censoring situations. We consider the results for two rates of censoring: a usual rate of 20% and a large rate of 50%.

As expected, with both procedures, the MISE degrade when the censoring rate increases. When we compare the standard and total MISE, the results are rather good for the standard MISE. This is understandable, since the total MISE measures the performance of the complete intensity estimators $\hat{\lambda}(t, \boldsymbol{Z}) = \hat{\alpha}^{\hat{\boldsymbol{\beta}}}(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}}$, including the error coming from $\hat{\boldsymbol{\beta}}$, whereas the standard MISE measures the performance of the estimators of the baseline function.

One can see that MISE resulting from the two procedures are very similar, with rather good results when using our procedure.

In Table 3, we give the standard MISE of the kernel estimators, with a bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method for different distributions of the survival times. We observe that the kernel estimator with a

Dimensions \ MISEs		20%				50%			
		MISEstand		MISEtot		MISEstand		MISEtot	
$n = 200$	$p = 15$	0.014	0.017	0.080	0.082	0.023	0.029	0.104	0.120
	$p = 500$	0.013	0.016	0.117	0.117	0.022	0.026	0.152	0.154
$n = 500$	$p = 22$	0.009	0.007	0.038	0.035	0.011	0.012	0.055	0.056
	$p = 1000$	0.008	0.008	0.068	0.064	0.011	0.013	0.094	0.096

Table 2 – MISE of the kernel estimators with bandwidth selected by the Goldenshluger and Lepski method (first column for each MISE), and with bandwidth selected by cross-validation of the baseline function with a LASSO estimator of the regression parameter, given two rates of censoring: 20% and 50%.

bandwidth selected by the Goldenshluger and Lepski method performs better than that with a bandwidth selected by cross-validation for the two Weibull distributions.

Dimensions \ Distributions		$\mathcal{W}(1.5, 1)$		$\mathcal{W}(0.5, 2)$	
$n = 200$	$p = 15$	0.056	0.088	1.02	1.561
	$p = 200$	0.06	0.085	0.923	1.556
$n = 500$	$p = 22$	0.025	0.037	1.006	1.521
	$p = 500$	0.027	0.033	1.098	1.515

Table 3 – MISEs for the kernel estimators with a bandwidth selected by the Goldenshluger and Lepski method (first column for each distribution) and with a bandwidth selected by cross-validation (second column for each distribution), with a LASSO estimator of the regression parameter for two different Weibull distributions of the survival times.

5.2 Application to a breast cancer dataset

In this section, we apply the proposed method to study the relapse free survival (RFS) from breast cancer in the high-dimensional covariates context for two groups of patients. We consider a Cox model (1) to link the RFS to covariates. The dataset is available on the website www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532.

The dataset consists of 414 patients in the cohort GSE6532 collected by Loi et al. [20] for the purpose of characterizing estrogen receptor (ER)-positive subtypes with gene expression profiles. Estrogen receptors are a group of proteins found inside cells, which are activated by the estrogen hormone. There are different forms of estrogen receptors, referred to as subtypes. When over-expressed, they are called ER-positive. The dataset has been studied from a survival analysis point of view by Tian et al. [25]. Following this analysis, we apply the two procedures to the same survival time of interest (the RFS). Excluding patients with incomplete information, as it is done by Loi et al. [20], there are

142 patients receiving Tamoxifen and 104 untreated patients. In this study, we have also excluded patients for which some data are missing. In addition to clinical information such as age, or size of tumor, we have 44 928 gene expression measurements for each of the 246 patients. Two different survival times are available in this study: the time of relapse free survival and the time of distant metastasis free survival. We are interested here in the time of relapse free survival, which is subject to right censoring due to incomplete follow-up. There is 60% of censorship in the group of untreated patients and 66% in that receiving Tamoxifen. Our goal is to compare the baseline functions in the two groups of patients.

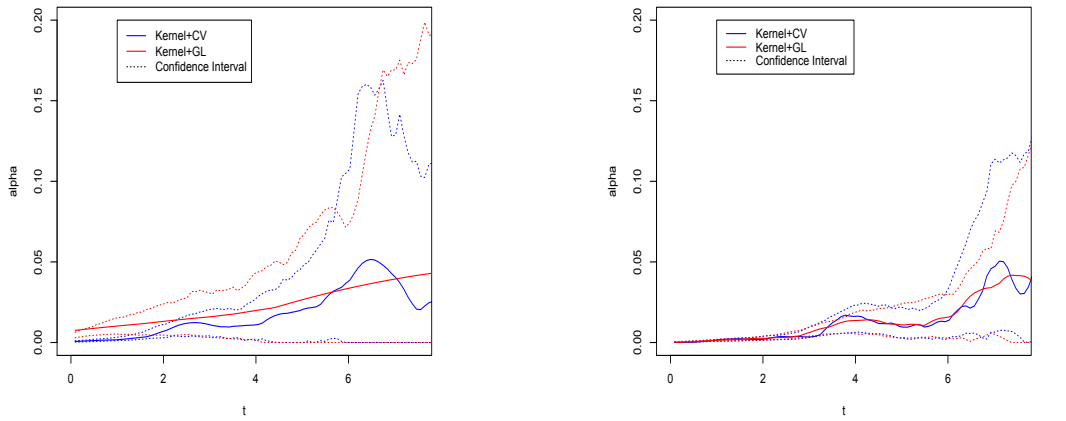
We start by preliminary variable selection among the 44928 gene expression measurements. This corresponds to a screening step (see Fan et al. [12]). This preliminary variable selection is based on the score statistics of each Cox model, considered for each variable separately. We only keep the variables whose score statistics are superior to a threshold. The difference from the procedure proposed by Fan et al. [12] is that we fix the number of covariates we want to keep and then tune a threshold to select this number. We define the threshold as the 95^{th} percentile of a chi-squared distribution with 1 degree of freedom, so that 996 probe-sets were selected, so along with the clinical covariates, we end up with $p = 1000$.

Figure 1 shows graphs of the kernel estimators of the baseline function, with a bandwidth selected by cross-validation or by the Goldenshluger and Lepski method, in the two groups of patients for $p = 1000$ and on the interval $[0, 7.5]$. Pointwise 90% confidence intervals are also represented. They were obtained via model based resampling, see Davison and Hinkley [10], for the survival times and from the estimated cumulative distribution function for the censoring times.

In Figure 1, we observe that the risk of relapse of breast cancer has slowed down with treatment, because the estimated baseline function is close to 0 until $t = 2.5$ for patients treated with tamoxifen, whereas it is already increasing at $t = 1.5$ for untreated patients. This leads us to believe that treatment has a positive influence on survival time.

6 Proofs

This section is organized as follows. First, we establish a lemma that allows to control the estimation error of the kernel estimator for a fixed bandwidth h, then we prove Theorem 4.1 from two fundamental lemmas that are also proved in this section. We add a supplementary material for all the other used technical lemmas, that are not essential for a first reading.



(a) Untreated patients (p=1000). (b) Tamoxifen patients (p=1000).

Figure 1 – Kernel estimator with bandwidth selected by cross-validation (blue) or with the Goldenshluger and Lepski method (red). The right-hand plot is associated with the group of untreated patients and the left-hand plot with the Tamoxifen patients, for $p = 1000$.

6.1 Intermediate lemma: bound for the kernel estimator of α_0 with a fixed bandwidth

We first establish a non-asymptotic global bound on Mean Integrated Squared Error (MISE) for the estimators $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$, with h fixed.

Lemma 6.1. *Under Assumptions 3.1, 3.2, 3.5.(ii)-(iii) and 3.6.(i)-(iii), for a fixed $h > 0$, n large enough and $k \geq 12$*

$$\mathbb{E} \parallel \hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0 \parallel_2^2 \leq 2 \parallel \alpha_h - \alpha_0 \parallel_2^2 + \frac{C_1}{nh} + C_2(s)\frac{\ln(pn^k)}{n} \tag{20}$$

where C_1 is a constant depending on τ, $||\alpha_0||_{\infty,\tau}$, c_S, $\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, τ and $||K||_{\mathbb{L}^2(\mathbb{R})}$ and $C_2(s)$ is a constant depending on B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_{\mathbb{L}^2(\mathbb{R})}$ and on the sparsity index s of $\boldsymbol{\beta_0}$.

To prove this lemma and link the kernel estimator to the true baseline function α_0, the trick is to introduce a pseudo-estimator, which does not depend on $\hat{\boldsymbol{\beta}}$. Consider for $h > 0$ the pseudo-estimator defined by (11). To justify the choice of the pseudo-estimator, let us calculate its expectation:

$$\begin{aligned}
\mathbb{E}\{\bar{\alpha}_h(t)\} &= \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{1}{S(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathbb{E}\{\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_i}}Y_i(u)\}\mathrm{d}u \\
&= \frac{1}{h}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\alpha_0(u)\mathrm{d}u \\
&= K_h * \alpha_0(t),
\end{aligned}$$

which is a unit approximation of α_0, so that $K_h * \alpha_0 \underset{h\to 0}{\longrightarrow} \alpha_0$ under mild conditions (see Bochner Lemma and Assumption 3.6.(iii)).
In the following, we define for all $t \in [0,\tau]$

$$\alpha_h(t) := \mathbb{E}\{\bar{\alpha}_h(t)\} = K_h * \alpha_0(t). \tag{21}$$

The proof is based on the following decomposition for $h > 0$

$$\mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 \leq \ 2\, \mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h \parallel_2^2 + 2\, \mathbb{E} \parallel \bar{\alpha}_h - \alpha_0 \parallel_2^2 . \tag{22}$$

The following lemma provides the classical bias/variance inequality for the pseudo-estimator (11).

Lemma 6.2. *Under Assumptions 3.5.(ii)-(iii), 3.6.(i)-(iii), for $h > 0$ fixed*

$$\mathbb{E} \parallel \bar{\alpha}_h - \alpha_0 \parallel_2^2 \leq \parallel \alpha_h - \alpha_0 \parallel_2^2 + \frac{2||\alpha_0||_{\infty,\tau}\tau}{c_S^2}\left\{\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}}) + ||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})\tau\right\}\frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nh}. \tag{23}$$

The next lemma controls the quadratic error between $\hat{\alpha}_h^{\hat{\beta}}$ and $\bar{\alpha}_h$. The term to be controlled in this difference is in fact the difference between the regression parameter $\boldsymbol{\beta_0}$ and its LASSO estimator $\hat{\boldsymbol{\beta}}$. The ℓ_1-norm of this difference is bounded from Proposition 3.4 by a term of order $\ln(np)/n$. This explains the obtained bound in the following lemma.

Lemma 6.3. *Under Assumptions 3.5.(ii)-(iii), 3.6.(i)-(iii), 3.1 and 3.2, for a fixed $h > 0$,*

$$\mathbb{E} \parallel \hat{\alpha}_h^{\hat{\beta}} - \bar{\alpha}_h \parallel_2^2 \leq c(s)\frac{\ln(n^k p)}{n},$$

where $c(s)$ is a constant depending on B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_{\mathbb{L}^2(\mathbb{R})}$ and s the sparsity index of $\boldsymbol{\beta_0}$.

From Equation (22), gathering Lemmas 6.2 and 6.3 provide directly Lemma 6.1.
Lemmas 6.2 and 6.3 are proved in the supplementary material.

6.2 Proof of the oracle inequality in Theorem 4.1

For all $h \in \mathcal{H}_n$, $A^{\hat{\beta}}(h)$ is defined by (13) and we can apply this definition for $h = \hat{h}^{\hat{\beta}}$. We deduce from this, using Definition (15) of $\hat{h}^{\hat{\beta}}$, that for all $h \in \mathcal{H}_n$

$$\begin{aligned}
\parallel \hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \alpha_0 \parallel_2^2 &\leq 3 \parallel \hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \hat{\alpha}_{h,\hat{h}^{\hat{\beta}}}^{\hat{\beta}} \parallel_2^2 + 3 \parallel \hat{\alpha}_{h,\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \hat{\alpha}_h^{\hat{\beta}} \parallel_2^2 + 3 \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 \\
&\leq 3(A^{\hat{\beta}}(h) + V(\hat{h}^{\hat{\beta}})) + 3(A^{\hat{\beta}}(\hat{h}^{\hat{\beta}}) + V(h)) + 3 \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 \\
&\leq 6(A^{\hat{\beta}}(h) + V(h)) + 3 \parallel \hat{\alpha}_h^{\hat{\beta}} - \alpha_0 \parallel_2^2 .
\end{aligned}$$

We obtain for $h \in \mathcal{H}_n$

$$\mathbb{E} \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\beta}}} - \alpha_0 \parallel_2^2 \leq 6\ \mathbb{E}\{A^{\hat{\boldsymbol{\beta}}}(h)\} + 6V(h) + 3\ \mathbb{E} \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h} - \alpha_0 \parallel_2^2 . \tag{24}$$

Lemma 6.1 gives a bound of $\mathbb{E} \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h} - \alpha_0 \parallel_2^2$, which reveals the bias term, the variance term of order $1/nh$ and a remaining term of order $\ln(np)/n$, and $V(h)$ is of the expected order $1/nh$. $\mathbb{E}\{A^{\hat{\boldsymbol{\beta}}}(h)\}$ is bounded in the following proposition.

Proposition 6.4. *Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 4.1, there exist constants C_1, $C_2(s)$, $C_3(s)$ such that,*

$$\mathbb{E}\{A^{\hat{\boldsymbol{\beta}}}(h)\} \leq C_1 \parallel \alpha_h - \alpha_0 \parallel_2^2 + C_2(s)\frac{\ln^a(n)\ln(n^k p)}{n} + C_3(s)\frac{\ln(n^k p)}{n}, \tag{25}$$

where the constant C_1 only depends on $||K||_1$.

We obtain Inequality (16) by applying Inequalities (20) and (25) in Equation (24) and by taking the infimum over $h \in \mathcal{H}_n$. This ends the proof of Theorem 4.1. □

6.3 Proof of Proposition 6.4

We introduce several additional notation $\bar{\alpha}_{h,h'} = K_{h'} * \bar{\alpha}_h$, $\alpha_h(t) = \mathbb{E}\{\bar{\alpha}_h(t)\}$, $\alpha_{h,h'}(t) = \mathbb{E}\{\bar{\alpha}_{h,h'}(t)\}$, and write

$$\begin{aligned}
A^{\hat{\boldsymbol{\beta}}}(h) &= \sup_{h' \in \mathcal{H}_n} \left\{ \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} \parallel_2^2 - V(h') \right\}_+ \\
&\leq 5 \sup_{h' \in \mathcal{H}_n} \left\{ \parallel \bar{\alpha}_{h'} - \alpha_{h'} \parallel_2^2 - V(h')/10 \right\}_+ + 5 \sup_{h' \in \mathcal{H}_n} \left\{ \parallel \bar{\alpha}_{h,h'} - \alpha_{h,h'} \parallel_2^2 - V(h')/10 \right\}_+ \\
&\quad + 5 \sup_{h' \in \mathcal{H}_n} \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'} - \bar{\alpha}_{h'} \parallel_2^2 + 5 \sup_{h' \in \mathcal{H}_n} \parallel \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} - \bar{\alpha}_{h,h'} \parallel_2^2 + 5 \sup_{h' \in \mathcal{H}_n} \parallel \alpha_{h'} - \alpha_{h,h'} \parallel_2^2 \\
&:= 5(T_1 + T_2 + T_3 + T_4 + T_5)
\end{aligned}$$

- Study of $\mathbb{E}[T_1]$: Recall that for all $h \in \mathcal{H}_n$

$$\parallel \bar{\alpha}_h - \alpha_h \parallel_2^2 = \sup_{f \in \mathbb{L}^2([0,\tau]), ||f||_2 = 1} \langle \bar{\alpha}_h - \alpha_h, f \rangle_2^2. \tag{26}$$

We introduce the centered empirical process $\nu_{n,h}(f) = \langle \bar{\alpha}_h - \alpha_h, f \rangle_2$, which is equal to

$$\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau f(t) \left[\int_0^\tau \frac{K_h(t-u)}{S(u, \boldsymbol{\beta_0})} \{\mathrm{d}N_i(u) - \alpha_0(u) S(u, \boldsymbol{\beta_0}) \mathrm{d}u\} \right] \mathrm{d}t.$$

As $f \longmapsto \nu_{n,h}(f)$ is continuous, the supremum in (26) can be taken over a countable dense subset of $\{f \in \mathbb{L}^2([0,\tau]), ||f||_2 = 1\}$, which we denote by $\mathcal{B}_\tau$. Therefore, we can write

$$\begin{aligned}\mathbb{E}(T_1) &\leq \mathbb{E}\Big[\Big\{\sup_{h' \in \mathcal{H}_n} \| \bar{\alpha}_{h'} - \alpha_{h'} \|_2^2 - V(h')/10\Big\}_+\Big] \\ &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E}\Big[\Big\{ \| \bar{\alpha}_{h'} - \alpha_{h'} \|_2^2 - V(h')/10\Big\}_+\Big] \\ &\leq \sum_{h' \in \mathcal{H}_n} \mathbb{E}\left[\left\{\sup_{f \in \mathcal{B}_\tau} \nu_n^2(f) - V(h')/10\right\}_+\right]. \end{aligned} \tag{27}$$

We introduce a key lemma which allows us to bound (27).

Lemma 6.5. *Let us introduce the centered process $\nu_{n,h}(f) = \langle \bar{\alpha}_h - \alpha_h, f\rangle_2$, for any $h \in \mathcal{H}_n$ and $f \in \mathbb{L}^2([0,\tau])$ and $\mathcal{B}_\tau = \{f \in \mathbb{L}^2([0,\tau]), ||f||_2 = 1\}$. Under the assumptions of Theorem 4.1, with $V(h')$ defined by (17) for any $h' \in \mathcal{H}_n$, there exists two constants c_6 and c_7 depending on the bound κ_b of the Bürkholder Inequality, τ, $||\alpha_0||_{\infty,\tau}$, the bound c_S of $S(t, \boldsymbol{\beta_0})$, $\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $||K||_{\mathbb{L}^1(\mathbb{R})}$ and $||K||_{\mathbb{L}^2(\mathbb{R})}$, such that*

$$\sum_{h \in \mathcal{H}_n} \mathbb{E}\left[\left\{\sup_{f \in \mathcal{B}_\tau(h)} \nu_{n,h}^2(f) - V(h)/10\right\}_+\right] \leq \frac{c_6}{n} + c_7 \frac{\ln^a(n)}{n}.$$

So, from Lemma 6.5, there exist two constants $c_6 > 0$ and $c_7 > 0$ such that

$$\mathbb{E}(T_1) \leq \frac{c_6}{n} + c_7 \frac{\ln^a(n)}{n}, \tag{28}$$

where c_6 and c_7 depend on τ, $||\alpha_0||_{\infty,\tau}$, c_S, $\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $||K||_{\mathbb{L}^1(\mathbb{R})}$ and $||K||_{\mathbb{L}^2(\mathbb{R})}$.

- Study of $\mathbb{E}[T_2]$: We study T_2 similarly as T_1 since

$$\mathbb{E}(T_2) \leq \sum_{h' \in \mathcal{H}_n} \mathbb{E}\left[\left\{\| \bar{\alpha}_{h,h'} - \alpha_{h,h'} \|_{2,h'}^2 - V(h')/10\right\}_+\right].$$

From Lemma 6.5 (see the remark at the end of the proof of Lemma 6.5), there exist two constants $c_8 > 0$ and $c_9 > 0$ such that

$$\mathbb{E}(T_2) \leq \frac{c_8}{n} + c_9 \frac{\ln^a(n)}{n}, \tag{29}$$

where c_8 and c_9 depend on τ, $||\alpha_0||_{\infty,\tau}$, c_S, $\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $||K||_{\mathbb{L}^1(\mathbb{R})}$ and $||K||_{\mathbb{L}^2(\mathbb{R})}$.

- Study of $\mathbb{E}(T_3)$: First, write for all $h \in \mathcal{H}_n$, that

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) - \bar{\alpha}_h(t) = \frac{1}{nh} \sum_{i=1}^n \int_0^\tau K\left(\frac{t-u}{h}\right) \frac{S(u, \boldsymbol{\beta_0}) \mathbb{1}_{\{\bar{Y}(u) > 0\}} - S_n(u, \hat{\boldsymbol{\beta}})}{S_n(u, \hat{\boldsymbol{\beta}}) S(u, \boldsymbol{\beta_0})} \mathrm{d}N_i(u)$$

For all $u \in [0,\tau]$, we have $S(u,\boldsymbol{\beta_0})\mathbb{1}_{\{\bar{Y}(u)>0\}} - S_n(u,\hat{\boldsymbol{\beta}}) = \{S(u,\boldsymbol{\beta_0}) - S_n(u,\hat{\boldsymbol{\beta}})\}\mathbb{1}_{\{\bar{Y}(u)>0\}}$. Indeed, for all $u \in [0,\tau]$, if $\mathbb{1}_{\{\bar{Y}(u)>0\}} = 0$, then for all $i \in \{1,\dots,n\}$, $Y_i(u) = 0$ and $S_n(u,\hat{\boldsymbol{\beta}}) = 0$. So, we can rewrite for all $h \in \mathcal{H}_n$ that

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) - \bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\left(\frac{t-u}{h}\right)\frac{S(u,\boldsymbol{\beta_0}) - S_n(u,\hat{\boldsymbol{\beta}})}{S_n(u,\hat{\boldsymbol{\beta}})S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{\bar{Y}(u)>0\}}\mathrm{d}N_i(u). \tag{30}$$

Consider the following sets:

$$\Omega_{H,k} = \left\{\omega : \forall u \in [0,\tau], |S_n(u,\hat{\boldsymbol{\beta}}) - S(u,\boldsymbol{\beta_0})| \leq 2C(s)B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}\sqrt{\frac{\ln(pn^k)}{n}}\right\}, \tag{31}$$

$$\Omega_{S_n} = \left\{\omega : \forall u \in [0,\tau], S_n(u,\hat{\boldsymbol{\beta}}) - S(u,\boldsymbol{\beta_0}) \geq -\frac{c_S}{2}\right\}, \tag{32}$$

$$\Omega_k = \Omega_{H,k} \cap \Omega_{S_n}. \tag{33}$$

We decompose T_3 on Ω_k and on its complement. On Ω_k^c, let introduce the following lemma:

Lemma 6.6. *Under Assumptions 3.5.(ii)-(iii), 3.6.(i)-(iii), 3.1 and 3.2, for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}\{\| \hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h \|_2^2 \mathbb{1}(\Omega_k^c)\} \leq c_3 n^{4-k/2},$$

where c_3 is a constant depending on B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_\infty$. Choosing $k \geq 10$ yields $\mathbb{E}\| \hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h \|_2^2 \leq c_3/n$.

From Lemma 6.6,

$$\begin{aligned}\mathbb{E}\left\{\sup_{h'\in\mathcal{H}_n}\| \hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_{h'} \|_2^2 \mathbb{1}(\Omega_k^c)\right\} &\leq \sum_{h'\in\mathcal{H}_n}\mathbb{E}\{\| \hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_{h'} \|_2^2 \mathbb{1}(\Omega_k^c)\} \\ &\leq \sum_{h'\in\mathcal{H}_n} c_3 n^{4-k/2},\end{aligned}$$

which is of order $1/n$ as long as $k \geq 12$. On the other hand, from (30) on Ω_k, we have

$$\begin{aligned}&\mathbb{E}\left\{\sup_{h'\in\mathcal{H}_n}\int_0^\tau (\hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_{h'})^2(t)\mathbb{1}(\Omega_k)\mathrm{d}t\right\} \\ &\quad\leq \frac{16C(s)^2B^2\mathrm{e}^{2BR}\mathrm{e}^{4B|\boldsymbol{\beta_0}|_1}}{c_S^2}\frac{\ln(pn^k)}{n}\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_0^\tau\left\{\int_0^\tau\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}N_i(u)\right)\right\}^2\right].\end{aligned}$$

Then, we decompose $N_i = (N_i - \Lambda_i) + \Lambda_i$ to obtain

$$\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_0^\tau\left\{\int_0^\tau\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{d}N_i(u)\Big)\right\}^2\mathrm{d}t\right]$$

$$\leq 2\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_0^\tau\left\{\int_0^\tau\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{d}N_i(u) - \alpha_0(u)S(u,\boldsymbol{\beta_0})\mathrm{d}u\Big)\right\}^2\mathrm{d}t\right] \tag{34}$$

$$+\,2\sup_{h'\in\mathcal{H}_n}\int_0^\tau\left\{\int_0^\tau|K_{h'}(t-u)|\alpha_0(u)\mathrm{d}u\right\}^2\mathrm{d}t. \tag{35}$$

The term (35) is bounded by $2\tau||\alpha_0||^2_{\infty,\tau}||K||^2_{\mathbb{L}^1(\mathbb{R})}$. Let us bound the term (34),

$$\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_0^\tau\left\{\int_0^\tau\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{d}N_i(u) - \alpha_0(u)S(u,\boldsymbol{\beta_0})\mathrm{d}u\Big)\right\}^2\mathrm{d}t\right]$$

$$\leq \sum_{h'\in\mathcal{H}_n}\int_0^\tau \mathrm{Var}\left[\int_0^\tau\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\frac{1}{n}\sum_{i=1}^n\mathrm{d}N_i(u)\right]$$

It remains to bound the variance term.

$$\mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau\frac{|K_h(t-u)|}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u)\right\} \leq \frac{1}{n}\mathbb{E}\left\{\left(\int_0^\tau\frac{|K_h(t-u)|}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right)^2\right\}.$$

We apply the Doob-Meyer decomposition to get

$$\mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^n\int_0^\tau\frac{|K_h(t-u)|}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u)\right\} \leq \frac{2}{n}\mathbb{E}\left\{\left(\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}M_1(u)\right)^2\right\} \tag{36}$$

$$+\,\frac{2}{n}\mathbb{E}\left\{\left(\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}}Y_1(u)\mathrm{d}u\right)^2\right\}. \tag{37}$$

The term (36) is bounded by

$$\frac{2}{n}\mathbb{E}\left\{\int_0^\tau\frac{K^2_{h'}(t-u)}{(S(u,\boldsymbol{\beta_0}))^2}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}}Y_1(u)\mathrm{d}u\right\} \leq \frac{2}{n}\frac{||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})}{c_S^2}\frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{h'}, \tag{38}$$

and from the Cauchy-Schwarz inequality, the term (37) is bounded by

$$\frac{2}{n}\frac{||\alpha_0||^2_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\tau}{c_S^2}\frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{h'}. \tag{39}$$

From (38) and (39), (34) is bounded by

$$\frac{4}{n}\frac{||\alpha_0||_{\infty,\tau}\tau}{c_S^2}\left\{\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}}) + ||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})\right\}||K||^2_{\mathbb{L}^2(\mathbb{R})}\sum_{h'\in\mathcal{H}_n}\frac{1}{nh'}. \tag{40}$$

From Condition 3.6.(ii) and bounds (37) and (40), we deduce that

$$\mathbb{E}\left\{\sup_{h'\in\mathcal{H}_n}\int_0^\tau(\hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}}-\bar{\alpha}_{h'})^2(t)\mathbb{1}(\Omega_k)\mathrm{d}t\right\}$$
$$\leq C(s,c_S,B,R,|\boldsymbol{\beta_0}|_1,||\alpha_0||_{\infty,\tau},\tau,||K||_{\mathbb{L}^2(\mathbb{R})},\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}}),\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}}))\frac{\ln^a(n)\ln(pn^k)}{n}.$$

Finally, there exists a constant $c_5>0$ such that

$$\mathbb{E}(T_3)\leq c_5\frac{\ln^a(n)\ln(n^kp)}{n}, \tag{41}$$

where c_5 depends on s, c_S, B, R, τ, $||\alpha_0||_{\infty,\tau}$, $|\boldsymbol{\beta_0}|_1$, $||K||_{\mathbb{L}^2(\mathbb{R})}$, $\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$ and $\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$.

- Study of $\mathbb{E}(T_4)$: Since

$$\hat{\alpha}_{h,h'}^{\hat{\boldsymbol{\beta}}}-\bar{\alpha}_{h,h'}=K_{h'}*(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}-\bar{\alpha}_h),$$

we have from Young Inequality (Lemma 2.2 in the supplementary material) with $p=1,q=2,r=2$,

$$\mathbb{E}(T_4)\leq||K||^2_{\mathbb{L}^1(\mathbb{R})}\mathbb{E}\parallel\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}-\bar{\alpha}_h\parallel_2^2\leq C(s)||K||^2_{\mathbb{L}^1(\mathbb{R})}\frac{\ln(n^kp)}{n}, \tag{42}$$

where the last inequality is obtained from Lemma 6.3.

- Study of $\mathbb{E}(T_5)$: From Young Inequality (Lemma 2.2 in the supplementary material) with $p=1,q=2,r=2$, we obtain that

$$||\alpha_{h'}-\alpha_{h,h'}||_2^2=||K_{h'}*(\alpha_0-K_h*\alpha_0)||_2^2\leq||K||^2_{\mathbb{L}^1(\mathbb{R})}||\alpha_0-K_h*\alpha_0||_2^2$$

Therefore, since $\alpha_h=K_h*\alpha_0$,

$$\mathbb{E}(T_5)\leq||K||^2_{\mathbb{L}^1(\mathbb{R})}||\alpha_0-\alpha_h||_2^2, \tag{43}$$

which corresponds to a bias term.

Finally, gathering the bounds of the five terms (28), (29), (41), (42) and (43), gives the result of Proposition 6.4. □

6.4 Proof of Lemma 6.5

We have to control the supremum of $\nu_{n,h}(f)$ defined by (44) over the ball $\mathcal{B}_\tau = \{g \in \mathbb{L}^2([0,\tau]), ||g||_2 = 1\}$. For all $h \in \mathcal{H}_n$ and $f \in \mathcal{B}_\tau$, we have

$$\nu_{n,h}(f) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f(t)\left\{\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\left(\mathrm{d}N_i(u) - \alpha_0(u)S(u,\boldsymbol{\beta_0})\mathrm{d}u\right)\right\}\mathrm{d}t. \tag{44}$$

Usually, to control such a process, we apply the Talagrand Inequality given in Theorem 2.3. However, since $\nu_{n,h}(f)$ is not bounded, we can not directly apply the Talagrand Inequality: we have to introduce a truncation (see Chagny [8] for a close approach). Let us define for a constant c,

$$\kappa_n = c\frac{\sqrt{n}}{\ln n},$$

and we decompose $\nu_{n,h}$ as

$$\nu_{n,h}(f) = \nu_{n,h}^{(1)}(f) + \nu_{n,h}^{(2)}(f),$$

where

$$\begin{aligned}\nu_{n,h}^{(1)}(f) = &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f(t)\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)\leq\kappa_n\}}\mathrm{d}N_i(u)\mathrm{d}t\\ &-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f(t)\int_0^\tau \mathbb{E}\left\{\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)\leq\kappa_n\}}\mathrm{d}N_i(u)\right\}\mathrm{d}t,\end{aligned}$$

and

$$\begin{aligned}\nu_{n,h}^{(2)}(f) = &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f(t)\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)>\kappa_n\}}\mathrm{d}N_i(u)\mathrm{d}t\\ &-\frac{1}{n}\sum_{i=1}^{n}\int_h^{\tau-h} f(t)\int_0^\tau \mathbb{E}\left\{\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)>\kappa_n\}}\mathrm{d}N_i(u)\right\}\mathrm{d}t.\end{aligned}$$

- Control of $\nu_{n,h}^{(1)}(f)$:

  We can apply a Talagrand Inequality to $\nu_{n,h}^{(1)}(f)$, which is bounded. To apply this concentration inequality, we need to determine the bounds H, M, W and the constant ε (see Theorem 2.3 in Appendix 2 for the notations).

  - Determination of the constant M:

Using the Cauchy-Schwarz Inequality, we have

$$\begin{aligned}
&\left|\int_0^\tau f(t)\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}N_1(u)\mathrm{d}t\right| \\
&\qquad\leq||f||_2\left|\int_0^\tau\Big(\int_0^\tau K_h^2(t-u)\mathrm{d}t\Big)^{1/2}\frac{\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right| \\
&\qquad\leq\frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{\sqrt{h}}\frac{|N_1(\tau)\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}|}{c_S} \\
&\qquad\leq\frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{c_S\sqrt{h}}\kappa_n := M = \mathcal{O}\Big(\frac{\sqrt{n}}{\ln n\sqrt{h}}\Big).
\end{aligned}$$

– Determination of the constant H:
Let us define

$$\psi_h(t) = \frac{1}{n}\sum_{i=1}^n\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)\leq\kappa_n\}}\mathrm{d}N_i(u).$$

We have $\sup_{f\in\mathcal{B}_\tau}(\nu^{(1)}_{n,h}(f))^2 = \sup_{f\in\mathcal{B}_\tau}\langle\psi_h-\mathbb{E}[\psi_h],f\rangle_2^2 = ||\psi_h-\mathbb{E}[\psi_h]||_2^2$. We deduce from the Doob-Meyer decomposition that

$$\begin{aligned}
\mathbb{E}\Big\{\sup_{f\in\mathcal{B}_\tau}(\nu^{(1)}_{n,h}(f))^2\Big\} &= \int_0^\tau \mathrm{Var}\{\psi_h(t)\}\mathrm{d}t \\
&\leq \frac{1}{n}\int_0^\tau\mathbb{E}\left[\left\{\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}N_1(u)\right\}^2\right]\mathrm{d}t \\
&\leq \frac{2||\alpha_0||_{\infty,\tau}\tau}{c_S^2}\Big\{||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\tau+\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\Big\}\frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nh} := H^2.
\end{aligned}$$

We have $H^2 = V(h)/\kappa$. Then, we set $\varepsilon^2 = 1/2$ and $\kappa = 80$ in order to have $2(1+2\varepsilon^2)H^2 = V(h)/20 = O(1/nh)$.

– Determination of the constant W:
Since $f\in\mathcal{B}_\tau$, we have

$$\begin{aligned}
&\mathrm{Var}\left\{\int_0^\tau f(t)\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}N_1(u)\mathrm{d}t\right\} \\
&\qquad\leq\mathbb{E}\left[\left\{\int_0^\tau f(t)\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}N_1(u)\mathrm{d}t\right\}^2\right] \\
&\qquad\leq\mathbb{E}\left[\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\left\{\int_0^\tau\frac{(K_h*f)(u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right\}^2\right].
\end{aligned}$$

So, from the Doob-Meyer decomposition and Young Lemma 2.2 in the supplementary material, we have

$$\begin{aligned}
&\mathrm{Var}\left\{\int_0^\tau f(t)\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}N_1(u)\mathrm{d}t\right\}\\
&\quad\leq\frac{2||\alpha_0||_{\infty,\tau}}{c_S^2}\left\{\tau||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})+\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\right\}||K_h*f||_2^2\\
&\quad\leq\frac{2||\alpha_0||_{\infty,\tau}}{c_S^2}\left\{\tau||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})+\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\right\}||K||_{\mathbb{L}^1(\mathbb{R})}^2:=W.
\end{aligned}$$

Then, from Assumptions 3.8.(ii) and (iii), we deduce that

$$\begin{aligned}
\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left\{\sup_{f\in\mathcal{B}_\tau}(\nu_{n,h}^{(1)}(f))^2-V(h)/20\right\}_+\right]&\leq\frac{\vartheta_1}{n}\sum_{h\in\mathcal{H}_n}\left\{\mathrm{e}^{-\frac{\vartheta_2}{h}}+\frac{1}{n\ln^2(n)h}\mathrm{e}^{-\vartheta_3\ln n}\right\}\\
&\leq\frac{\tilde{\vartheta}_1}{n}+\tilde{\vartheta}_2\frac{\ln^{a-2}n}{n^{\tilde{\vartheta}_3}}
\end{aligned}\tag{45}$$

with

$$V(h)=\kappa\frac{2||\alpha_0||_{\infty,\tau}\tau}{c_S^2}\left\{||\alpha_0||_{\infty,\tau}\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\tau+\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top\boldsymbol{Z_1}})\right\}\frac{||K||_{\mathbb{L}^2(\mathbb{R})}^2}{nh}.$$

- Control of $\nu_{n,h}^{(2)}(f)$:

  Now, let us focus on the second unbounded term $\nu_{n,h}^{(2)}(f)$. Let us consider the process $\Psi(t)$ defined as

$$\frac{1}{n}\sum_{i=1}^n\left[\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)>\kappa_n\}}\mathrm{d}N_i(u)-\mathbb{E}\left\{\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)>\kappa_n\}}\mathrm{d}N_i(u)\right\}\right],$$

  so that $\nu_{n,h}^{(2)}(f)=\int_0^\tau f(t)\Psi(t)\mathrm{d}t$. Using Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\mathbb{E}\left\{\sup_{f\in\mathcal{B}_\tau}(\nu_{n,h}^{(2)}(f))^2\right\}&\leq\mathbb{E}\left\{\int_0^\tau\Psi^2(t)\mathrm{d}t\right\}\\
&\leq\frac{1}{n}\int_0^\tau\mathrm{Var}\left\{\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}\mathrm{d}N_1(u)\right\}\mathrm{d}t\\
&\leq\frac{1}{n}\int_0^\tau\mathbb{E}\left[\left\{\int_0^\tau\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}\mathrm{d}N_1(u)\right\}^2\right]\mathrm{d}t.
\end{aligned}$$

Applying the Cauchy-Schwarz Inequality (see Lemma 2.1 in the supplementary material), we obtain that for all $k > 0$,

$$\begin{aligned}
\mathbb{E}\left\{\sup_{f\in\mathcal{B}_\tau}(\nu_{n,h}^{(2)}(f))^2\right\} &\leq \frac{1}{n}\int_0^\tau \mathbb{E}\left\{\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}N_1(\tau)\int_0^\tau \frac{K_h^2(t-u)}{S^2(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right\}\mathrm{d}t \\
&\leq \frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nhc_S^2}\mathbb{E}\{N_1^2(\tau)\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}\} \\
&\leq \frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nhc_S^2}\frac{\mathbb{E}\{N_1^{k+2}(\tau)\}}{\kappa_n^k} \\
&\leq \frac{||K||^2_{\mathbb{L}^2(\mathbb{R})}}{nhc_S^2}\frac{\mathbb{E}\{N_1^{k+2}(\tau)\}}{n}.
\end{aligned}$$

From Assumption 3.8.(ii), we deduce that for k large enough

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left\{\sup_{f\in\mathcal{B}_\tau}(\nu_{n,h}^{(2)}(f))^2\right\} \leq C\frac{\ln^a(n)\mathbb{E}\{N(\tau)^{k+1}\}}{n}.$$

It remains to verify that $\mathbb{E}\{N(\tau)^{k+1}\}$ is bounded. Using the fact that for all $a \geq 0$, $b \geq 0$ and $p \geq 1$, $(a+b)^k \leq 2^{k-1}(a^k + b^k)$ and from the Bürkholder Inequality, we can easily show by recurrence that for all $p \in \mathbb{N}^*$, $\mathbb{E}\{N(\tau)^k\} \leq C_k$. Thus, we conclude that for a good choice of p,

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left\{\sup_{f\in\mathcal{B}_\tau}(\nu_{n,h}^{(2)}(f))^2\right\} \leq \tilde{C}\frac{\ln^a(n)}{n}, \tag{46}$$

for a constant $\tilde{C} > 0$.

Combining (45) and (46), we finally get

$$\sum_{h\in\mathcal{H}_n}\mathbb{E}\left[\left\{\sup_{f\in\mathcal{B}_\tau}\nu_{n,h}^2(f) - V(h)/10\right\}_+\right] \leq \frac{c_6}{n} + c_7\frac{\ln^a(n)}{n},$$

where c_6 and c_7 depends on τ, $||\alpha_0||_{\infty,\tau}$, c_S, $\mathbb{E}(\mathrm{e}^{\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $\mathbb{E}(\mathrm{e}^{2\boldsymbol{\beta_0}^\top \boldsymbol{Z_1}})$, $||K||_{\mathbb{L}^1(\mathbb{R})}$ and $||K||_{\mathbb{L}^2(\mathbb{R})}$. □

Remark: A similar lemma can be obtained for the centered process $\langle\bar{\alpha}_{h,h'} - \alpha_{h,h'}, f\rangle_2$, where $\bar{\alpha}_{h,h'} = K_{h'} * \bar{\alpha}_h$ and $\alpha_{h,h'} = \mathbb{E}(\bar{\alpha}_{h,h'})$ for $h, h' \in \mathcal{H}_n$. Indeed, from Young Lemma 2.2 in the supplementary material, we have

$$\begin{aligned}
\langle\bar{\alpha}_{h,h'} - \alpha_{h,h'}, f\rangle_2 &= \int_0^\tau f(t)\Big[K_{h'} * \bar{\alpha}_h(t) - \mathbb{E}\{K_{h'} * \bar{\alpha}_h(t)\}\Big]\mathrm{d}t \\
&\leq ||f||_2||K_{h'} * (\bar{\alpha}_h - \mathbb{E}[\bar{\alpha}_h])||_2 \\
&\leq ||f||_2||K||_{\mathbb{L}^1(\mathbb{R})}||\bar{\alpha}_h - \mathbb{E}[\bar{\alpha}_h]||_2^2.
\end{aligned}$$

Just take the same constants M, H^2 and W than previously and multiply them by $||K||_{\mathbb{L}^1(\mathbb{R})}$.

The proofs of Lemmas 6.2, 6.3 and 6.6 are available in the supplementary material.

Acknowledgements

We gratefully acknowledge the editor and the referee for carefully reading the manuscript and for helpful advises and suggestions.

References

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

[2] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York, 1993.

[3] Bouaziz, O., Comte, F., and Guilloux, A. Nonparametric estimation of the intensity function of a recurrent event process. *Statistica Sinica 23*, 2 (2013), 635–665.

[4] Bradic, J., Fan, J., and Jiang, J. Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics 39*, 6 (12 2011), 3092–3120.

[5] Bradic, J., and Song, R. Structured estimation for the nonparametric cox model. *Electronic Journal of Statistics 9*, 1 (2015), 492–534.

[6] Bühlmann, P., and van de Geer, S. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics 3* (2009), pp. 1360–1392.

[7] Chagny, G. *Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie.* PhD thesis, Université René Descartes-Paris V, 2013.

[8] Chagny, G. Adaptive warped kernel estimators. *Scandinavian Journal of Statistics 42*, 2 (2015), 336–360.

[9] Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. (Methodological) 34*, 2 (1972), pp. 187–220.

[10] Davison, A., and Hinkley, D. *Bootstrap methods and their application*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 1997. With 1 IBM-PC floppy disk (3.5 inch; HD).

[11] Doumic, M., Hoffmann, M., Reynaud-Bouret, P., and Rivoirard, V. Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis 50*, 2 (2012), 925–950.

[12] Fan, J., Feng, Y., and Wu, Y. High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, vol. Volume 6. Institute of Mathematical Statistics, 2010, pp. 70–86.

[13] Fleming, T., and Harrington, D. *Counting processes and survival analysis*, vol. 169. John Wiley & Sons, 2011.

[14] Goldenshluger, A., and Lepski, O. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics 39*, 3 (2011), 1608–1632.

[15] Grégoire, G. Least squares cross-validation for counting process intensities. *Scandinavian Journal of Statistics 20*, 4 (1993), 343–360.

[16] Guilloux, A., Lemler, S., and Taupin, M. Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates. *Journal of Statistical Planning and Inference* (2015).

[17] Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C. Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics 41*, 3 (2013), 1142–1165.

[18] Kong, S., and Nan, B. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica 24*, 1 (2014), 25–42.

[19] Lemler, S. *Estimation for counting processes with high-dimensional covariates.* PhD thesis, Université d'Évry Val d'Essonne, Dec 2014.

[20] Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J., et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. *Journal of clinical oncology 25*, 10 (2007), 1239–1246.

[21] Marron, J., and Padgett, W. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics 15*, 4 (12 1987), 1520–1535.

[22] Ramlau-Hansen, H. *Udglatning med kernefunktioner i forbindelse med tælleprocesser: Del 1.* Forsikringsmatematisk Laboratorium, Københavns universitet, 1981.

[23] Ramlau-Hansen, H. The choice of a kernel function in the graduation of counting process intensities. *Scandinavian Actuarial Journal 1983*, 3 (1983), 165–182.

[24] Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics 11*, 2 (06 1983), 453–466.

[25] Tian, L., Alizadeh, A., Gentles, A., and Tibshirani, R. A simple method for detecting interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association 109*, 508 (10 2014), 1517–1532.

[26] van de Geer, S. The deterministic lasso. Tech. rep., ETH Zürich, Switzerland, 2007.

[27] van de Geer, S. High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*, 2 (2008), 614–645.