Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates

Agathe Guilloux, Sarah Lemler, Marie-Luce Taupin

Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates

Agathe Guilloux
Laboratoire de Statistique Théorique et Appliquée,
Université Pierre et Marie Curie - Paris 6
e-mail : agathe.guilloux@upmc.fr

Sarah Lemler
Laboratoire de Mathématiques et de Modélisation d'Evry, UMR CNRS 8071- USC INRA,
Université d'Évry Val d'Essonne, France
e-mail : sarah.lemler@genopole.cnrs.fr

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

Abstract

The purpose of this article is to provide an adaptive estimator of the baseline function in the Cox model with high-dimensional covariates. We consider a two-step procedure : first, we estimate the regression parameter of the Cox model via a Lasso procedure based on the partial log-likelihood, secondly, we plug this Lasso estimator into a least-squares type criterion and then perform a model selection procedure to obtain an adaptive penalized contrast estimator of the baseline function.

Using non-asymptotic estimation results stated for the Lasso estimator of the regression parameter, we establish a non-asymptotic oracle inequality for this penalized contrast estimator of the baseline function, which highlights the discrepancy of the rate of convergence when the dimension of the covariates increases.

Keywords: Survival analysis; Conditional hazard rate function; Cox's proportional hazards model; Right-censored data; Semi-parametric model; Nonparametric model; High-dimensional covariates; Model selection; Non-asymptotic oracle inequalities; Concentration inequalities

1 Introduction

Consider the following Cox model, introduced by Cox (1972) and defined, for a vector of covariates $\boldsymbol{Z} = (Z_1, ..., Z_p)^T$, by

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

where λ_0 denotes the hazard rate, $\boldsymbol{\beta_0} = (\beta_{0_1}, ..., \beta_{0_p})^T \in \mathbb{R}^p$ is the regression parameter and α_0 is the baseline hazard function. The Cox partial log-likelihood, introduced by Cox (1972), allows to estimate

$\boldsymbol{\beta_0}$ without the knowledge of α_0, considered as a functional nuisance parameter. For the estimation of α_0, one common way is to use a two step procedure, starting with the estimation of $\boldsymbol{\beta_0}$ alone and then to plug this estimator into a non parametric type estimator α_0, usually a kernel type estimator.

Let us be more specific.

When p is small compared to n, $\boldsymbol{\beta_0}$ is usually estimated by minimization of the opposite of the Cox partial log-likelihood. We refer to Andersen et al. (1993), as a reference book, for the proofs of the consistency and the asymptotic normality of $\hat{\boldsymbol{\beta}}$ when p is small compared to n. Thoses strategies only apply when $p < n$ and even more, they only apply when p is small compared to n. When p growths up, becoming of the same order as n and possibly larger than n, various well known problems appears. Among them, the minimization of the opposite of the Cox partial log-likelihood becomes difficult and even impossible if $p > n$.

In high-dimension, when p is large compared to n, the Lasso procedure is one of the classical considered strategies. The Lasso (Least Absolute Shrinkage and Selection Operator) has been first introduced by Tibshirani (1996) in the linear regression model. It has been largely considered in additive regression model (see for instance Knight and Fu (2000), Efron et al. (2004), Donoho et al. (2006), Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zhang and Huang (2008), Meinshausen and Yu (2009) and also Juditsky and Nemirovski (2000), Nemirovski (2000), Bunea et al. (2006; 2007a;b), Greenshtein and Ritov (2004) or Bickel et al. (2009)), and in density estimation (see Bunea et al. (2007c) and Bertin et al. (2011)). In the particular case of the semi-parametric Cox model, Tibshirani (1997) has proposed a Lasso procedure for the regression parameter. The Lasso estimator of the regression parameter $\hat{\boldsymbol{\beta}}$ is defined as the minimizer of the opposite of the Cox partial log-likelihood under an ℓ_1 type constraint, that is, suitably penalized with an ℓ_1-penalty function. Recent results exist on the estimation of $\boldsymbol{\beta_0}$ in high-dimension setting. Among them one can mention Bradic et al. (2012) who have proved asymptotic results for Lasso estimator. More recently, Bradic and Song (2012), Kong and Nan (2012) and Huang et al. (2013) establish the first non-asymptotic oracle inequalities (estimation and prediction bounds) for the Lasso estimator.

For the baseline hazard function and when p is small compared to n, the common estimator is a kernel estimator, which depends on $\hat{\boldsymbol{\beta}}$ obtained by minimization of the opposite of the Cox partial log-likelihood. This kernel estimator has been introduced by Ramlau-Hansen (1983a;b) from the Breslow estimator of the cumulative baseline function (see Ramlau-Hansen (1983b) and Andersen et al. (1993) for more details). In this context, Ramlau-Hansen (1983b) and Grégoire (1993) proved asymptotic results. No non-asymptotic results and no adaptive results have to date been established for the kernel estimator of the baseline function. Finally, when p is large compared to n, to our knowledge, the construction of an estimator of the baseline function has not been yet considered.

In this paper, we consider a two-step procedure to estimate $\boldsymbol{\beta_0}$ and α_0, the two parameters in the Cox model. But our contributions focus more on the estimation of α_0. In the Cox model we consider, it is noteworthy that the high-dimension only concerns the regression parameter, whereas the baseline function is a time function. Its estimation would not require a procedure specific to high-dimension, besides the first step concerning the estimation of $\boldsymbol{\beta_0}$. We propose a procedure for the construction of an estimator of the baseline hazard function α_0, p being either smaller than n or greater than n. It combines a Lasso procedure for $\boldsymbol{\beta_0}$ as a first step and a second step based on a model selection strategy for the estimation of the baseline function α_0. This model selection procedure takes its origins in the works of Akaike (1973) and Mallows (1973), more recently formalized by Birgé and Massart (1997) and Barron et al. (1999) for the estimation of densities and regression functions (see the book of Massart (2007) as a reference work on model selection). In survival analysis, the model selection has also been documented. Letué (2000) has adapted these methods to estimate the regression function of the non-parametric Cox model, when $p < n$. More recently, Brunel and Comte (2005), Brunel et al. (2009), Brunel et al. (2010) have obtained adaptive estimation of densities in a censoring setting. Model selection methods have also been used to estimate the intensity function of a counting process in the multiplicative Aalen intensity model (see Reynaud-Bouret (2006) and Comte et al. (2011)). However, the model selection procedure has never been considered, to our knowledge, for estimating the baseline hazard function in the Cox model.

Our contributions are at least threefold: Our procedure is the first that focus on the estimation of baseline function of the semi-parametric Cox model with high-dimentional covariates. This procedure provide an adaptive estimator of the baseline function that works as well for small p and large p compared to n (that is for possibly high-dimensional covariates). Furthermore, for this estimator, we state non-asymptotic oracle inequalities, that hold, once again, p being either smaller than n or greater than n. More precisely, we prove that the risk of this estimator achieves the best risk among estimators in a large collection. For each model, the risk of an estimator is bounded by the sum of three terms. The first term is a bias term involving to the approximation properties of the collection of models, through the distance evaluated in $\boldsymbol{\beta_0}$ between the true baseline and the orthogonal projection of α_0 on the best selected model. The second term is a penalty term of the same order than the variance on one model, that is of order the dimension of one model over n, as expected with ℓ_0-penalty. These two terms are the "usual" terms appearing in nonparametric estimation. It is noteworthy that these two terms do not involve any quantity related to the risk of the Lasso estimator of $\boldsymbol{\beta_0}$. The last term precisely comes from the properties of the Lasso estimator of $\boldsymbol{\beta_0}$. This last term is of order $\log(np)/n$, as expected for a Lasso estimator.

When p is small, the third last term is of order $\log(n)/n$ and, the rate is governed by the first two terms. In that case, the penalty term being of the same order than the variance over one model, we conclude that the model selection procedure achieves the "expected rate" of order $n^{-2\gamma/(2\gamma+1)}$ when the baseline function belongs to a Besov space with smoothness parameter γ. This continues to hold when p is of the same order than the sample size n. When p is larger than n, that is in the so-called ultra-high dimension (see Verzelen (2012)), the rate for estimating α_0 is changed, and more precisely degraded as a price to pay for being with high dimension covariates. This degradation follows accordingly to the order of p compared to n.

The main tools for stating our results are the theory of marked counting processes and martingales with jumps, the theory of penalized minimum contrast estimators and concentrations inequalities such as Talagrand inequality (see Talagrand (1996)) and a Bernstein inequality found in (see van de Geer (1995) and Comte et al. (2011)) for unbounded martingale process and combined with chaining methods (see Talagrand (2005) and Baraud (2010)).

The article is organized as follows. In Section 3, we describe the estimation procedure. Section 4 provides non-asymptotic oracle inequalities on the estimator of the baseline hazard function α_0, in a high-dimensional setting for $\boldsymbol{\beta_0}$. In section 5, we compare the performances of the resulting penalized contrast estimator to those of the usual kernel estimator on simulated data. Section 6 is devoted to the proofs: we state some technical results, then we establish the two main theorems and lastly we prove the technical results. Finally, Appendix A discusses the bound of the error estimation for the Lasso estimator of the regression parameter of the Cox model.

2 Notations and preliminaries

2.1 Framework with counting processes

Consider the general setting of counting processes, which embeds the classical case of right censoring. We follow here the now classical setting of Andersen et al. (1993) or Fleming and Harrington (2011). For n independant individuals, we observe for $i = 1, ..., n$ a counting process N_i, a random process Y_i with values in $[0, 1]$ and a vector of covariates $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ be the filtration defined by

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

From the Doob-Meyer decomposition, we know that each N_i admits a compensator denote 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. (1993) for details). We assume in the following that N_i has a satisfies an Aalen multiplicative intensity model.

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

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i}) Y_i(s) \mathrm{d}s, \tag{2}$$

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

We observe the independent and identically distributed (i.i.d.) data $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$, where $[0, \tau]$ is the time interval between the beginning and the end of the study.

This general setting, introduced by Aalen (1980), embeds several particular examples as censored data, marked Poisson processes and Markov processes (see Andersen et al. (1993) for further details). We give here details for the right censoring case. We observe for $i = 1, ..., n$, $(X_i, \delta_i, \boldsymbol{Z_i})$, 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. With these notations, the $(\mathcal{F}_t)_{t\geq 0}$-adapted processes Y_i and N_i are respectively defined as the at-risk process $Y_i(t) = \mathbb{1}_{\{X_i \geq t\}}$ and the counting process $N_i(t) = \mathbb{1}_{\{X_i \leq t, \delta_i = 1\}}$ which jumps when the ith individual dies.

2.2 Assumptions

Before describing the estimation procedure, we introduce few assumptions on the framework defined in Subsection 2.1.

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 by Z_j its j-th component, namely the j-th covariates of the vector $\boldsymbol{Z}$. Similarly, we denote by Y the generic version of the random process Y_i with values in $[0, 1]$.

We define the standard $\mathbb{L}^2$ and $\mathbb{L}^\infty$-norms, for $\alpha \in (\mathbb{L}^2 \cap \mathbb{L}^\infty)([0, \tau])$:

$$||\alpha||_2^2 = \int_0^\tau \alpha^2(t)\mathrm{d}t \quad \text{and} \quad ||\alpha||_{\infty,\tau} = \sup_{t\in[0,\tau]} |\alpha(t)|.$$

For a vector $\boldsymbol{b} \in \mathbb{R}^p$, we also introduce the ℓ_1-norm $|b|_1 = \sum_{j=1}^p |b_j|$.

Assumption 2.2.

(i) *There exists a positive constant* B *such that*

$$|Z_j| \leq B, \quad \forall j \in \{1, ..., p\}.$$

In the following, we denote $A = [-B, B]^p$.

(ii) *The vector of covariates* $\boldsymbol{Z}$ *admit a p.d.f.* $f_{\boldsymbol{Z}}$ *such that* $\sup_A |f_{\boldsymbol{Z}}| \leq f_1 < +\infty$.

(iii) *There exists* $f_0 > 0$*, such that* $\forall (t, \boldsymbol{z}) \in [0, \tau] \times A$,

$$\mathbb{E}[Y(t)|\boldsymbol{Z} = \boldsymbol{z}] f_{\boldsymbol{Z}}(\boldsymbol{z}) \geq f_0.$$

(iv) *For all* $t \in [0, \tau]$, $\alpha_0(t) \leq ||\alpha_0||_{\infty,\tau} < +\infty$.

Remark 2.3. *Let say a few word on these assumptions starting by noting that these four assumptions are quite classic and reasonnable. To be more specific, Assumption 2.2.(i), is very common to establish oracle inequalities of Lasso estimators in various frameworks. In particular, in the Cox model, see e.g. Huang et al. (2013) and Bradic and Song (2012) for the statement of non asymptotic oracle inequalities*

In the specific case of right censoring, Assumption 2.2.(iii) is automatically verified. Indeed, for T *the survival time and* C *the censoring time, we can write*

$$\mathbb{E}(Y(t)|\boldsymbol{Z} = \boldsymbol{z}) = \mathbb{E}(\mathbb{1}_{\{T \wedge C \leq t\}}|\boldsymbol{Z} = \boldsymbol{z}) = (1 - F_{T|\boldsymbol{Z}}(t))(1 - G_{C|\boldsymbol{Z}}(t-)),$$

where $F_{T|\boldsymbol{Z}}$ *and* $G_{C|\boldsymbol{Z}}$ *are the cumulative distribution functions of* $T|\boldsymbol{Z}$ *and* $C|\boldsymbol{Z}$ *respectively. It is known (see Andersen et al. (1993)) that the Kaplan-Meier estimator is consistent only on intervals of the form* $[0, \tau]$*, where* $\tau \leq \sup\{t \geq 0, (1 - F_{T|\boldsymbol{Z}}(t))(1 - G_{C|\boldsymbol{Z}}(t)) > 0\}$*. Hence when* $f_{\boldsymbol{Z}}$ *is bounded from below on* A*, there exists* $f_0 > 0$*, such that*

$$\forall (t, \boldsymbol{z}) \in [0, \tau] \times A, \quad \mathbb{E}[Y(t)|\boldsymbol{Z} = \boldsymbol{z}] f_{\boldsymbol{Z}}(\boldsymbol{z}) \geq f_0.$$

Assumption 2.2.(iii) is required in order to compare the natural norm of the baseline function induced by our contrast to the standard $\mathbb{L}^2$*-norm (see Proposition 6.1).*

3 Estimation procedure

We now describe our two-steps estimation procedure, starting by recalling the Lasso estimation of $\boldsymbol{\beta_0}$ and then giving a bound of its prediction risk. Then, we describe the contrast and the model selection procedure for the estimation of the baseline function.

3.1 Preliminary estimation of $\boldsymbol{\beta_0}$: procedure and results

The Lasso estimator $\hat{\boldsymbol{\beta}}$ of the regression parameter $\boldsymbol{\beta_0}$, introduced in Tibshirani (1997), is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{-l_n^*(\boldsymbol{\beta}) + \Gamma_n |\boldsymbol{\beta}|_1\}, \tag{3}$$

where Γ_n is a positive regularization parameter to be suitable chosen, $|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$ and l_n^* is the Cox partial log-likelihood defined by,

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} \log \frac{\mathrm{e}^{\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\mathrm{d}N_i(t), \quad \text{where } S_n(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{Z_i}}Y_i(t) \quad \forall t \geq 0. \tag{4}$$

The risk bounds for the estimator of α_0 will naturally involve the risk $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$, that have to be at least bounded. Thus, we rather consider the following procedure

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{B}(0,R_1)}{\arg\min}\{-l_n^*(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \quad \text{with} \quad \mathrm{pen}(\boldsymbol{\beta}) = \Gamma_n|\boldsymbol{\beta}|_1, \tag{5}$$

where $\mathcal{B}(0,R_1)$ is the ball defined by

$$\mathcal{B}(0,R_1) = \{b \in \mathbb{R}^p : |b|_1 \leq R_1\}, \quad \text{with } R_1 > 0.$$

Consider the following assumption:

Assumption 3.1. *We assume that* $|\boldsymbol{\beta_0}|_1 < R_2 < +\infty$.

We denote $R = \max(R_1, R_2)$, so that

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq 2R \quad \text{a.s.} \tag{6}$$

Such condition has already been considered by van de Geer (2008) or Kong and Nan (2012). Roughly speaking, it means that we can restrict our attention to a ball, possibly very large, in a neighborhood of $\boldsymbol{\beta_0}$ for finding a good estimator of $\boldsymbol{\beta_0}$.

As mentionned above, our risk bounds for the estimator of α_0 depend on the risk $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$. Such bounds on this risk already exist. In particular, in their Theorem 3.1, Huang et al. (2013) state a non asymptotic inequality for $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ in the specific case of bounded counting processes. We consider here more general processes, possibly unbounded. In the following proposition, we provide a generalization of the results established by Huang et al. (2013) to the case of unbounded counting processes. We refer to Appendix A for a proof of Proposition 3.2.

Proposition 3.2. *Let $k > 0$, $c > 0$ and $s := \mathrm{Card}\{j \in \{1,...,p\} : \beta_{0_j} \neq 0\}$ be the sparsity index of $\boldsymbol{\beta_0}$. Assume that $||\alpha_0||_{\infty,\tau} < \infty$. Then, under Assumptions 3.1 and (i), with probability larger than $1 - cn^{-k}$, we have*

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

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

As mentioned previously, this proposition is crucial to establish a non-asymptotic oracle inequality for the baseline function. In the rest of the paper, we consider that $\hat{\boldsymbol{\beta}}$ satisfies Inequality (7).

Assumption 3.3. *We assume that*

$$\lim_{n\to\infty} C(s)\frac{\log(np)}{n} = 0.$$

This assumption is clearly reasonable: when p is smaller than n or of the same order, this assumption is automatically fulfilled. It is not satisfied when p becomes too high compared to n. This case corresponds to the now well known case of ultra-high dimension framework. In this specific case, recent lower bounds in additive regression models typically say that the estimation of paramater is mostly impossible (see for example Verzelen (2012)).

3.2 Estimation of α_0

We now come to the estimation of the baseline function α_0 via a model selection procedure. As usual, such a procedure requires an empirical estimation criterion, a collection of models and a suitable penalty function, all being presented in the following.

3.2.1 Definition of the estimation criterion

We estimate the baseline function α_0 using a least-squares criterion. More precisely, based on the data $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$ and for a fixed $\boldsymbol{\beta}$, we consider the empirical least-squares type given for a function $\alpha \in (\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$ by

$$C_n(\alpha, \boldsymbol{\beta}) = -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \alpha(t)\mathrm{d}N_i(t) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha^2(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(t)\mathrm{d}t. \tag{8}$$

The use of such least-square empirical criterion in survival analysis is not so usual as for the additive regression model. Nevertheless, few recent studies have developped such very useful as strategies. Among them one can cite Reynaud-Bouret (2006) or Comte et al. (2011).

Let us define a deterministic scalar product and its associated deterministic norm for α_1, α_2 and α functions in $(\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$:

$$\begin{aligned} \langle \alpha_1, \alpha_2 \rangle_{det(\boldsymbol{\beta})} &= \int_0^\tau \alpha_1(t)\alpha_2(t)\mathbb{E}[e^{\boldsymbol{\beta^T Z}}Y(t)]\mathrm{d}t, \\ ||\alpha||^2_{det(\boldsymbol{\beta})} &= \int_0^\tau \alpha^2(t)\mathbb{E}[e^{\boldsymbol{\beta^T Z}}Y(t)]\mathrm{d}t. \end{aligned} \tag{9}$$

Using the Doob-Meyer decomposition $N_i = M_i + \Lambda_i$ and according to the multiplicative Aalen model (2), we get:

$$\mathbb{E}[C_n(\alpha, \boldsymbol{\beta_0})] = ||\alpha||^2_{det} - 2\langle \alpha, \alpha_0 \rangle_{det} = ||\alpha - \alpha_0||^2_{det} - ||\alpha_0||^2_{det},$$

which is minimum when $\alpha = \alpha_0$. Hence, minimizing $C_n(., \boldsymbol{\beta_0})$ is a relevant strategy to estimate α_0.

3.2.2 Model selection

We now describe the model selection procedure in our context, introducing first the collection of models.

Collections of models. Let $\mathcal{M}_n$ be a set of indices and $\{S_m, m \in \mathcal{M}_n\}$ be a collection of models:

$$S_m = \{\alpha : \alpha = \sum_{j \in J_m} a_j^m \varphi_j^m, a_j^m \in \mathbb{R}\},$$

where $(\varphi_j^m)_{j \in J_m}$ is an orthonormal basis of $(\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$ for the usual $\mathbb{L}_2(P)$- norm. We denote D_m the cardinality of S_m, i.e. $|J_m| = D_m$.

Sequence of estimators. Let us consider $\hat{\boldsymbol{\beta}}$ the Lasso estimator of $\boldsymbol{\beta_0}$ defined by (5). For each $m \in \mathcal{M}_n$, we define the estimator

$$\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}} = \arg\min_{\alpha \in S_m}\{C_n(\alpha, \hat{\boldsymbol{\beta}})\}. \tag{10}$$

Model selection. The relevant space is automatically selected by using following penalized criterion

$$\hat{m}^{\hat{\beta}} = \arg\min_{m \in \mathcal{M}_n} \{C_n(\hat{\alpha}_m^{\hat{\beta}}, \hat{\boldsymbol{\beta}}) + \mathrm{pen}(m)\}, \tag{11}$$

where $\mathrm{pen} : \mathcal{M}_n \to \mathbb{R}$ will be defined later.

Final estimator. The final estimator of α_0 is then $\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}}$.

Let us say few words on the optimisation problem. Denote by $\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\beta}}$ the random Gram matrix

$$\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\beta}} = \Big(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varphi_j(t)\varphi_k(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}Y_i(t)\mathrm{d}t\Big)_{(j,k)\in J_m^2}. \tag{12}$$

By definition, the estimator $\hat{\alpha}_m^{\hat{\beta}}$ is the solution of the equation $\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\beta}}\boldsymbol{A}_{\boldsymbol{m}}^{\hat{\beta}} = \boldsymbol{\Gamma}_{\boldsymbol{m}}$, where

$$\boldsymbol{A}_{\boldsymbol{m}}^{\hat{\beta}} = (\hat{a}_j^{\hat{\beta}})_{j\in J_m} \quad \text{and} \quad \boldsymbol{\Gamma}_{\boldsymbol{m}} = \Big(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varphi_j(t)\mathrm{d}N_i(t)\Big)_{j\in J_m}. \tag{13}$$

The Gram matrix $\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\beta}}$ may not be invertible in some cases. Hence we consider the set

$$\hat{\mathcal{H}}_m^{\hat{\beta}} = \left\{\min \mathrm{Sp}(\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\beta}}) \geq \max\left(\frac{\hat{f}_0\mathrm{e}^{-B|\beta_0|_1}\mathrm{e}^{-B|\beta_0-\hat{\beta}|_{\mathbf{1}}}}{6}, \frac{1}{\sqrt{n}}\right)\right\}, \tag{14}$$

where $\mathrm{Sp}(\boldsymbol{M})$ denotes the spectrum of matrix $\boldsymbol{M}$ and $\hat{f}_0$ satisfies the following assumption:

Assumption 3.4. *There exist a preliminary estimator $\hat{f}_0$ of f_0 and two positive constants $C_0 > 0$, $n_0 > 0$ such that*

$$\mathbb{P}(|\hat{f}_0 - f_0| > f_0/2) \leq C_0/n^6 \quad \textit{for any} \quad n \geq n_0.$$

From Assumptions 3.1, on the set $\hat{\mathcal{H}}_m^{\hat{\beta}}$, the matrix $\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\beta}}$ is invertible and $\hat{\alpha}_m^{\hat{\beta}}$ is thus uniquely defined as

$$\hat{\alpha}_m^{\hat{\beta}} = \begin{cases} \arg\min_{\alpha\in\mathcal{S}_m}\{C_n(\alpha, \hat{\boldsymbol{\beta}})\} & \text{on } \hat{\mathcal{H}}_m^{\hat{\beta}}, \\ 0 & \text{on } (\hat{\mathcal{H}}_m^{\hat{\beta}})^c. \end{cases}$$

3.2.3 Assumptions and examples of the models

The following assumptions on the models $\{S_m : m \in \mathcal{M}_n\}$ are usual in model selection procedures. They are verified by the spaces spanned by usual bases: trigonometric basis, regular piecewise polynomial basis, regular compactly supported wavelet basis and histogram basis. We refer to Barron et al. (1999) and Brunel and Comte (2005) for other examples and further discussions.

Assumption 3.5.

(i) *For all $m \in \mathcal{M}_n$, we assume that*

$$D_m \leq \frac{\sqrt{n}}{\log n}.$$

(ii) *For all $m \in \mathcal{M}_n$, there exists $\phi > 0$ such that for all α in S_m,*

$$\sup_{t\in[0,\tau]} |\alpha(t)|^2 \leq \phi D_m \int_0^\tau \alpha^2(t)\mathrm{d}t.$$

(iii) *The models are nested within each other: $D_{m_1} \leq D_{m_2} \Rightarrow S_{m_1} \subset S_{m_2}$. We denote by $\mathcal{S}_n$ the global nesting space in the collection and by $\mathcal{D}_n$ its dimension.*

Remark 3.6. *Assumption 3.5.(i) ensures that the sizes D_m of the models are not too large compared with the number of observations n. This assumption seems reasonable if we remember that D_m is the number of coefficients to be estimated: if this number is too large compared to the size of the panel, we cannot expect to obtain a relevant estimator. Assumption 3.5.(ii) implies a useful connection between the standard $\mathbb{L}^2$-norm and the infinite norm. Assumption 3.5.(iii) ensures that $\forall m, m' \in \mathcal{M}_n$, $S_m + S_{m'} \subset \mathcal{S}_n$. Thanks to this assumption, one does not have to browse through all models for the model selection, which reduces the algorithmic complexity of the procedure. In addition, we have from Assumption 3.5.(i) that $\mathcal{D}_n \leq \sqrt{n}/\log n$.*

4 Non-asymptotic oracle inequalities

We now are in a position to state our main theorem: a non-asymptotic oracle inequality for the estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ of the baseline function in the Cox model.

Theorem 4.1. *Let Assumptions 2.2.(i)-(iv), Assumptions 3.1, Assumption 3.3, Assumption 3.4 and Assumptions 3.5.(i)-(iii) hold. Let $\alpha_m^{\boldsymbol{\beta_0}}$ be the projection of α_0 on S_m with respect to the deterministic scalar product when $\boldsymbol{\beta_0}$ is known:*

$$\alpha_m^{\boldsymbol{\beta_0}} = \arg\min_{\alpha\in S_m} \mathbb{E}[C_n(\alpha, \boldsymbol{\beta}_0)] = \arg\min_{\alpha\in S_m} ||\alpha - \alpha_0||^2_{det}. \tag{15}$$

Let $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ be defined by (10) and (11) with

$$\mathrm{pen}(m) := K_0(1 + ||\alpha_0||_{\infty,\tau})\frac{D_m}{n}, \tag{16}$$

where K_0 is a numerical constant. Then, for any $n \geq n_0$, with n_0 a constant defined in Assumption 3.4,

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}] \leq \kappa_0 \inf_{m\in\mathcal{M}_n} \{||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} + 2\,\mathrm{pen}(m)\} + \frac{C_1}{n} + C_2 C(s)\frac{\log(np)}{n}, \tag{17}$$

where κ_0 is a numerical constant, C_1 and C_2 are constants depending on τ, ϕ, $||\alpha_0||_{\infty,\tau}$, f_0, $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$, $\mathbb{E}[\mathrm{e}^{2\boldsymbol{\beta_0^T Z}}]$, $\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]$, B, $|\boldsymbol{\beta_0}|_1$, the sparsity index s of $\boldsymbol{\beta_0}$ and κ_b a constant from the Bürkholder Inequality (see Theorem 6.9) and $C(s)$ the constant depending on the sparsity index of $\boldsymbol{\beta_0}$ in Proposition 3.2.

Inequality (17) provides the first non-asymptotic oracle inequality for an estimator of the baseline function. This inequality warrants the performances of our estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$. We refer to Subsection 6.2.1 for precisions about C_1 and C_2. In Inequality (17), the risk is bounded by the sum of four terms.

The third term of order $1/n$ is negligible compared to the others. The first two terms are respectively the bias and the variance terms. The bias term, $||\alpha_0 - \alpha_m^{\beta_0}||^2_{det}$, corresponds to the approximation error and decreases with the dimension D_m of the model S_m. It depends on the regularity of the true function, which is unknown: the more regular α_0 is, the smaller the bias is. The variance term $\mathrm{pen}(m)$ quantifies the estimation error and in contrary to the bias term, increases with D_m. It is of order D_m/n, which corresponds to the order of the variance term on one model. These three first terms do not involve quantities related to the estimation error of the Lasso estimator of $\boldsymbol{\beta_0}$.

The last term precisely comes from the non-asymptotic control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ given by Proposition 3.2. Indeed, we can rewrite Inequality (17) before using the bound of control (7):

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}] \leq \kappa_0 \inf_{m \in \mathcal{M}_n} \{||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} + 2\,\mathrm{pen}(m)\} + \frac{C_1}{n} + C_2 \mathbb{E}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^2].$$

This inequality makes clearer the role of the first step of the procedure in the control of the estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ of the baseline function. The bound obtained for this control is of order $\log(np)/n$, which explains the order of the fourth term. This term quantifies the influence of the high dimension on the estimation of the baseline hazard function. For small p, we obtain the expected rate of convergence in the case of a purely non-parametric estimation, but when is larger than n, the rate of convergence of the inequality is degraded. This is the price to pay for dealing with covariates in high dimension.

Corollary 4.2. *Assume that α_0 belongs to the Besov space $\mathcal{B}^{\gamma}_{2,\infty}([0,\tau])$, with smoothness γ. Then, under the assumptions of Theorem 4.1,*

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_2] \leq \tilde{C} n^{-\frac{2\gamma}{2\gamma+1}} + C_2 C(s) \frac{\log(np)}{n},$$

where $\tilde{C}$ and C_2 are constants depending on τ, ϕ, $||\alpha_0||_{\infty,\tau}$, f_0, $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$, $\mathbb{E}[\mathrm{e}^{2\boldsymbol{\beta_0^T Z}}]$, B, $|\boldsymbol{\beta_0}|_1$, the sparsity index s of $\boldsymbol{\beta_0}$ and κ_b a constant from the Bürkholder Inequality (see Theorem 6.9) and $C(s)$ the constant depending on the sparsity index of $\boldsymbol{\beta_0}$ from Proposition 3.2.

From Reynaud-Bouret (2006), we know that, for an intensity function without covariates in a Besov space with smoothness parameter γ, the minimax rate is $n^{-2\gamma/(2\gamma+1)}$. We infer that this would also be the optimal rate in our case when the term $\log(np)/n$ is negligible, namely when $p < n$. However, when the high-dimension $p \gg n$ is reached, the remaining term $\log(np)/n$ is not negligible anymore and there is a loss in the rate of convergence, which comes from the difficulty to estimate $\boldsymbol{\beta_0}$.

5 Applications: simulation study

The aim of this section is to illustrate the behavior of the penalized contrast estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ 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 (1983b).

5.1 Simulated data

Let consider the Cox model (1) in the case of right censoring. We consider a cohort of size n and p covariates. In the simulation study, several choices of n and p have been considered. The sample size n takes the values $n = 200$ and $n = 500$ and p varies between $p = \sqrt{n}$, being 15 and 22 respectively and $p = n$, referred to as the high-dimension case.

The true regression parameter $\boldsymbol{\beta_0}$ is chosen as a vector of dimension p, defined by

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

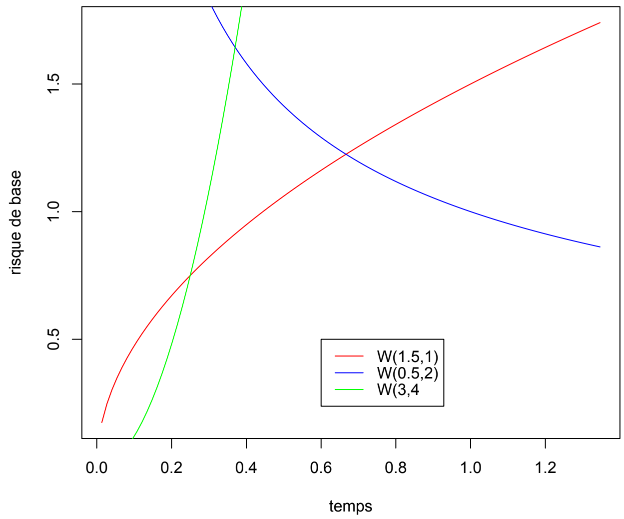

Figure 1 – Plots of the baseline hazard function for different parameters of a Weibull distribution $\mathcal{W}(a, \lambda)$

for various $p \geq 3$ and 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 $[-1, 1]$. We consider survival times T_i, $i = 1, ..., n$ that are distributed according to a Weibull distribution $\mathcal{W}(a, \lambda)$, namely the associated baseline function is of the form $\alpha_0(t) = a\lambda^a t^{a-1}$. We simulate three Weibull distribution $\mathcal{W}(0.5, 1)$, $\mathcal{W}(1, 1)$, $\mathcal{W}(3, 4)$ (see Figure 1). We consider a rate of censoring of 20% and the censoring times C_i, for $i = 1, ..., n$, are simulated independently from the survival times via an exponential distribution $\mathcal{E}(1/\gamma\mathbb{E}[T_1])$, where $\gamma = 4.5$ is adjusted to the rate of censorship. The time τ of the end of the study is taken as the quantile at 90% of $(T_i \wedge C_i)_{i=1,...,n}$. For $i = 1, ..., n$, 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 $\delta_i = \mathbb{1}_{T_i \leq C_i}$. The definition of $\tilde{C}_i$ ensures that there exist some $i \in \{1, ..., n\}$ for which $X_i \geq \tau$, so that all estimators are defined on the interval $[0, \tau]$ and it prevents from certain edge effect.

Each sample $(\boldsymbol{Z_i}, T_i, C_i, X_i, \delta_i, i = 1, ..., n)$ is repeated $N_e = 100$ times.

5.2 Estimation procedures

We implement $\hat{\alpha}_m^{\hat{\beta}}$ in a histogram basis defined, for $j = 1, ..., 2^m$, by

$$\varphi_j^m(t) = \frac{1}{\sqrt{\tau}} 2^{m/2} \mathbb{1}_{[(j-1)\tau/2^m, j\tau/2^m[}(t),$$

In this case, the cardinal of S_m is $D_m = 2^m$ and Assumption 3.5.(ii) is satisfied for $\phi = 1/\tau$. We take $m = 0, ..., \lfloor \log(n/\log(n))/\log(2) \rfloor$, so that Assumption 3.5.(i) is fulfilled. In this basis, the estimator is being written by

$$\hat{\alpha}_m^{\hat{\beta}}(t) = \sum_{j \in J_m} \hat{a}_j^{\hat{\beta}} \varphi_j^m(t), \quad \forall t \in [0, \tau], \tag{18}$$

where

$$\hat{a}_j^{\hat{\boldsymbol{\beta}}} = \frac{\tau}{2^m} \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}\Big(\Big(\min\Big(X_i, \dfrac{j\tau}{2^m}\Big) - \dfrac{(j-1)\tau}{2^m}\Big) \vee 0\Big)} \frac{1}{n}\sum_{i=1}^{n} \delta_i \frac{2^{m/2}}{\sqrt{\tau}} \mathbb{1}_{\left[\frac{(j-1)\tau}{2^m}, \frac{j\tau}{2^m}\right)}(X_i).$$

The final estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ is obtained from the implementation of the selection model procedure (10), replacing in the penalty term the unknown quantity $||\alpha_0||_{\infty,\tau}$ by $||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(m)}||_{\infty,\tau}$, an estimator of α_0 computed on the arbitrary larger space $S_{\max(m)}$.

We want to compare the performances of the estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ to those of the usual kernel estimator with a bandwidth selected by cross-validation introduced by Ramlau-Hansen (1983b), that we have also implemented. More precisely the usual kernel estimator is defined by

$$\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}}(t) = \frac{1}{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}} \sum_{i=1}^{n} \frac{\delta_i}{\sum_{j=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_j}} \mathbb{1}_{\{X_j \geq X_i\}}} K\left(\frac{t - X_i}{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}}\right), \tag{19}$$

where $K(u) = 0.75(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$ is the Epanechnikov kernel and the bandwidth $\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}$ has been selected by cross-validation:

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

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

Both estimators of the baseline hazard function are defined from the Lasso estimator $\hat{\boldsymbol{\beta}}$ of the regression parameter defined by (3).

The performances of these two estimators are evaluated via a random Mean Integrated Squared Error (MISErand) adapted to the Cox model and defined by $\mathrm{MISErand}(\alpha, \hat{\boldsymbol{\beta}}) = \mathbb{E}[\mathrm{ISErand}(\alpha, \hat{\boldsymbol{\beta}})]$, where the expectation is taken on $(T_i, C_i, \boldsymbol{Z_i})$ and

$$\mathrm{ISErand}(\alpha, \hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^{X_i} (\alpha(t) - \alpha_0(t))^2 \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}} \mathrm{d}t, \tag{20}$$

We obtain an estimation of the MISErand by taking the empirical mean for $N_e = 100$ replications.

In Table 1, we give the random MISE of the penalized contrast estimator and of the kernel estimator with a bandwidth selected by cross-validation for different distributions of the survival times.

First, as expected, the random MISEs are smaller for a large n and a small p. Then, we observe that the penalized contrast estimator performs better than the kernel estimator for the Weibull distributions $\mathcal{W}(0.5, 2)$ and $\mathcal{W}(3, 4)$. Note that the random MISEs are very high for this last distribution. This can easily be explained from the fact that the baseline hazard function associated to a $\mathcal{W}(3, 4)$ has the most complicated form since it increases steeply (see Figure 1). Lastly, for the distribution $\mathcal{W}(1.5, 1)$, the random MISEs are smaller in the case of the kernel estimator with a bandwidth selected by cross-validation than in the case of the penalized contrast estimator.

Dimensions \ Distributions		$\mathcal{W}(1.5,1)$		$\mathcal{W}(0.5,2)$		$\mathcal{W}(3,4)$	
$n=200$	$p=15$	0.072	0.021	0.626	1.09	5.26	8.48
	$p=200$	0.071	0.020	0.613	1.09	5.30	8.33
$n=500$	$p=22$	0.055	0.009	0.401	1.06	5.24	7.48
	$p=500$	0.059	0.008	0.402	1.06	5.25	8.10

Table 1 – Random empirical MISE for the penalized contrast estimator in a histogram basis (first column for each distribution) and for the kernel estimator with a bandwidth selected by cross-validation (second column for each distribution), with a Lasso estimator of the regression parameter, for three different Weibull distributions of the survival times.

6 Proofs

6.1 Technical results

In this section, we introduce some propositions and lemmas that are necessary to prove the theorems. Their proofs are postponed to Subsection 6.3.

Let us first introduce the random norm revealed from the contrast (8) and associated to the deterministic norm defined by (9), and its associated scalar product: for α, α_1 and α_2 functions in $(\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ fixed,

$$\begin{aligned} ||\alpha||^2_{rand(\boldsymbol{\beta})} &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \alpha^2(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(t)\mathrm{d}t, \\ \langle \alpha_1, \alpha_2\rangle_{rand(\boldsymbol{\beta})} &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \alpha_1(t)\alpha_2(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(t)\mathrm{d}t, \end{aligned} \tag{21}$$

Subsequently, to relieve the notations, we denote $||.||_{rand} := ||.||_{rand(\boldsymbol{\beta}_0)}$ and the same holds for the associated scalar product. We state a key relation between $\langle .,. \rangle_{rand(\boldsymbol{\beta})}$ and $C_n(.,\boldsymbol{\beta})$. By definition, for all $m \in \mathcal{M}_n$ and $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$C_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}, \boldsymbol{\beta}) + \mathrm{pen}(\hat{m}^{\boldsymbol{\beta}}) \le C_n(\hat{\alpha}^{\boldsymbol{\beta}}_m, \boldsymbol{\beta}) + \mathrm{pen}(m) \le C_n(\alpha^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}) + \mathrm{pen}(m), \tag{22}$$

where $\hat{m}^{\boldsymbol{\beta}} = \arg\min_{m\in\mathcal{M}_n}\{C_n(\hat{\alpha}^{\boldsymbol{\beta}}_m, \boldsymbol{\beta}) + \mathrm{pen}(m)\}$. Now, we write that

$$\begin{aligned} &C_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}, \boldsymbol{\beta}) - C_n(\alpha^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}) \\ &= -\frac{2}{n}\sum_{i=1}^n \int_0^\tau (\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m)(t)\mathrm{d}N_i(t) + \frac{1}{n}\sum_{i=1}^n \int_0^\tau (\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}(t)^2 - \alpha^{\boldsymbol{\beta_0}}_m(t)^2)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(t)\mathrm{d}t. \end{aligned}$$

Using the Doob-Meyer decomposition, we derive that

$$\begin{aligned} &C_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}, \boldsymbol{\beta}) - C_n(\alpha^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}) \\ &= -2\langle \hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha_0\rangle_{rand} + ||\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}||^2_{rand(\boldsymbol{\beta})} - ||\alpha^{\boldsymbol{\beta_0}}_m||^2_{rand(\boldsymbol{\beta})} - 2\nu_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m), \end{aligned}$$

where $\nu_n(\alpha)$ is defined by

$$\nu_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \alpha(t)\mathrm{d}M_i(t). \tag{23}$$

It follows that

$$C_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}, \boldsymbol{\beta}) - C_n(\alpha^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}) = ||\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand(\boldsymbol{\beta})} - 2\nu_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m)$$
$$+ 2\langle \hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m \rangle_{rand(\boldsymbol{\beta})} - 2\langle \hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha_0 \rangle_{rand}. \tag{24}$$

Let us now introduce the following events :

$$\Delta_1 = \left\{ \alpha \in \mathcal{S}_n : \left| \frac{||\alpha||^2_{rand}}{||\alpha||^2_{det}} - 1 \right| \le \frac{1}{2} \right\}, \quad \text{and} \quad \Omega = \left\{ \left| \frac{\hat{f}_0}{f_0} - 1 \right| \le \frac{1}{2} \right\} \tag{25}$$

$$\Delta_2 = \left\{ \alpha \in \mathcal{S}_n : \left| \frac{||\alpha||^2_{rand(\hat{\boldsymbol{\beta}})}}{||\alpha||^2_{rand}} - 1 \right| \le \frac{1}{2} \right\}. \tag{26}$$

On the sets Δ_1 and Δ_2 we have a relation between the random $||.||_{rand}$ and the deterministic $||.||_{det}$ norms and between the random norms $||.||_{rand}$ and $||.||_{rand(\hat{\boldsymbol{\beta}})}$ respectively. The following proposition state a relation between the deterministic norm (9) and the standard $\mathbb{L}^2$-norm:

Proposition 6.1 (Connections between the norms)**.** *From Assumptions 2.2.(i)-(iii), we deduce the following connection between the deterministic norm and the standard* $\mathbb{L}^2$*-norm:*

$$f_0 \mathrm{e}^{-B|\boldsymbol{\beta_0}|_1} ||\alpha||^2_2 \le ||\alpha||^2_{det} \le \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}] ||\alpha||^2_2 \le \mathrm{e}^{B|\boldsymbol{\beta_0}|_1} ||\alpha||^2_2.$$

The proof of this proposition is immediate using the fact that from Assumption 2.2.(ii), we can rewrite the deterministic norm as

$$||\alpha||^2_{det} = \int_0^\tau \int_A \alpha^2(t) e^{\boldsymbol{\beta_0^T z}} \mathbb{E}[Y(t)|\boldsymbol{Z} = \boldsymbol{z}] f_{\boldsymbol{Z}}(z) \mathrm{d}z \mathrm{d}t.$$

6.1.1 Results used in the proofs of Theorem 4.1

Recall that for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\hat{\mathcal{H}}^{\boldsymbol{\beta}}_m = \left\{ \min \mathrm{Sp}(\boldsymbol{G}^{\boldsymbol{\beta}}_{\boldsymbol{m}}) \ge \max\left(\frac{\hat{f}_0 \mathrm{e}^{-B|\beta_0|_1} \mathrm{e}^{-B|\beta_0 - \beta|_1}}{6}, \frac{1}{\sqrt{n}} \right) \right\}.$$

The following lemma ensures the existence of the estimators $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ on $\Delta_1 \cap \Delta_2 \cap \Omega$.

Lemma 6.2. *Under Assumptions 2.2.(i)-(iv), Assumptions 3.1 and Assumptions 3.5.(i)-(iii), for* $n \ge 16/(f_0 \mathrm{e}^{-3BR})^2$*, the following embedding holds:*

$$\Delta_1 \cap \Delta_2 \cap \Omega \subset \hat{\mathcal{H}}^{\hat{\boldsymbol{\beta}}} \cap \Omega, \quad \textit{where } \hat{\mathcal{H}}^{\hat{\boldsymbol{\beta}}} := \bigcap_{m \in \mathcal{M}_n} \hat{\mathcal{H}}^{\hat{\boldsymbol{\beta}}}_m.$$

From this lemma, for all $m \in \mathcal{M}_n$, the matrix $\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}$ is invertible on $\Delta_1 \cap \Delta_2 \cap \Omega$, and thus the estimator of α_0 is well defined. Proof 6.2 are available in Subsection 6.3.1.

The following proposition bounds the quadratic difference between $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\beta}}}$ and $\alpha^{\boldsymbol{\beta_0}}_m$ for $m \in \mathcal{M}_n$, on the complements of

$$\aleph_k = \Delta_1 \cap \Delta_2 \cap \Omega \cap \Omega_H^k,$$

where Ω_H^k, (the indice H is for "Huang", since the set has already been defined by Huang et al. (2013)), is defined for $k > 0$ by

$$\Omega_H^k = \left\{ |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}} \right\}, \tag{27}$$

for a constant $C(s)$ depending on the sparsity index of $\boldsymbol{\beta_0}$. From Proposition 3.2, $\mathbb{P}(\Omega_H^k) \geq 1 - cn^{-k}$ for a constant $c > 0$. Now, let us state the two following propositions.

Proposition 6.3. *Under Assumptions 2.2.(i)-(iv), Assumptions 3.1 and Assumptions 3.5.(i)-(iii),*

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} \mathbb{1}_{\aleph_k^c}] \leq \tilde{c}_1/n, \tag{28}$$

where $\tilde{c}_1$ is a constant depending on τ, ϕ, $||\alpha_0||_{\infty,\tau}$, f_0, $\mathbb{E}[e^{\boldsymbol{\beta_0^T Z}}]$, $\mathbb{E}[e^{2\boldsymbol{\beta_0^T Z}}]$, B, $|\boldsymbol{\beta_0}|_1$, the sparsity index s of $\boldsymbol{\beta_0}$ and κ_b a constant that comes from the Bürkholder Inequality (see Theorem 6.9).

We refer to Subsection 6.3.2 for the proof of Proposition 6.3. This propositions are directly used in the proof of Theorems 4.1 in Subsection 6.2.

Usually, in model selection (see for instance Massart (2007)), the penalty is obtained by using the so-called Talagrand's deviation inequality for the maximum of empirical processes. In the empirical process (23), the martingales M_i, $i = 1, ..., n$, are unbounded, Thus, we cannot directly use the Talagrand's inequality. We consider the following proposition proved in Comte et al. (2011). To obtain an uniform deviation of $\nu_n(.)$, Comte et al. (2011) have used tools from van de Geer (1995) to establish Bennett and Bernstein type inequalities and a $\mathbb{L}^2(det) - \mathbb{L}^\infty$ generic chaining type of technique (see Talagrand (2005) and Baraud (2010)).

Proposition 6.4. *Let $m, m' \in \mathcal{M}_n$. Define*

$$\mathcal{B}^{det}_{m,m'}(0,1) = \{\alpha \in S_m + S'_m : ||\alpha||_{\text{det}} \leq 1\}. \tag{29}$$

Under the assumptions of Theorem 4.1, there exists $\kappa > 0$ such that for

$$p(m, m') = \frac{\kappa}{K_0}(\text{pen}(m) + \text{pen}(m')), \tag{30}$$

where the constant K_0 and $\text{pen}(m)$ are defined in (16), then

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\Big(\Big(\sup_{\alpha \in \mathcal{B}^{det}_{m,m'}(0,1)} \nu_n^2(\alpha) - p(m, m')\Big)_+ \mathbb{1}_{\Delta_1}\Big) \leq \frac{C_3}{n}$$

for n large enough, where C_3 is a constant depending on f_0, $\mathbb{E}[e^{\boldsymbol{\beta_0^T Z}}]$, B, $|\boldsymbol{\beta_0}|_1$, $||\alpha_0||_{\infty,\tau}$ and the choice of the basis.

These propositions are applied to prove Theorem 4.1. We admit the proof of this proposition and refer to Comte et al. (2011) for a detailed proof of this result.

We need Proposition 6.5 to prove Theorem 4.1: the empirical centered process $\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})$, defined by

$$\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}}) = \frac{1}{n}\sum_{i=1}^{n}\Big(U_i(\alpha, \alpha_m^{\boldsymbol{\beta_0}}) - \mathbb{E}[U_i(\alpha, \alpha_m^{\boldsymbol{\beta_0}})]\Big),$$

where

$$U_i(\alpha, \alpha_m^{\boldsymbol{\beta_0}}) = \left(\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right)^2.$$

appears in the proof of Theorem 4.1, when we control the difference between the scalar products $\langle .,.\rangle_{rand} - \langle .,.\rangle_{rand(\hat{\beta})}$ (see Subsection 6.2.1). Proposition 6.5 allows to control this process.

Proposition 6.5. *Let introduce the ball $\mathcal{B}_n^{det}(0,1) \subset \mathcal{S}_n$ defined by*

$$\mathcal{B}_n^{det}(0,1) = \{\alpha \in \mathcal{S}_n : ||\alpha||_{det} \leq 1\}. \tag{31}$$

Under Assumptions 2.2.(i)-(iv) and Assumption 3.1, we have

$$\mathbb{E}\left[\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)} \eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})^2\right] \leq \frac{1}{n}\frac{\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]||\alpha_m^{\boldsymbol{\beta_0}}||_2^4}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}f_0)^2}.$$

Proposition 6.5 is proved in Subsection 6.3.3.

6.1.2 Technical lemmas for the proofs of Proposition ?? and 6.3

In order to prove Proposition 6.3, we need three lemmas:

Lemma 6.6. *Under Assumptions 2.2.(i)-(iv), Assumptions 3.1 and Assumptions 3.5.(i)-(iii), we have*

$$\mathbb{E}[||\hat{\alpha}_{\hat{m}^{\hat{\beta}}}^{\hat{\beta}}||_2^4] \leq C_b n^4,$$

where C_b is constant depending on $||\alpha_0||_{\infty,\tau}$, τ, $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$ and $\mathbb{E}[\mathrm{e}^{2\boldsymbol{\beta_0^T Z}}]$, κ_b, the constant of the Bürkholder Inequality (see Theorem 6.9) and on the choice of the basis.

Lemma 6.7. *Under Assumptions 2.2.(i)-(iv) and Assumptions 3.5.(i)-(iii), we have*

$$\mathbb{P}(\Delta_1^c) \leq \frac{C_k^{(\Delta_1)}}{n^k}, \quad \forall k \geq 1,$$

where $C_k^{(\Delta_1)}$ is a constant depending on f_0, B and $|\boldsymbol{\beta_0}|_1$.

Lemma 6.8. *Under Assumptions 2.2.(i)-(iv), Assumptions 3.1 and Assumption 3.3, we have for n large enough,*

$$\mathbb{P}(\Delta_2^c) \leq \frac{C_k^{(\Delta_2)}}{n^k}, \quad \forall k \geq 1,$$

where the constant $C_k^{(\Delta_2)}$ depends on τ, $||\alpha_0||_{\infty,\tau}$ and $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$.

These three lemmas are required to prove Proposition 6.3. There are proved in Subsection 6.3.

6.1.3 A classical inequality: the Bürkholder Inequality

The last technical result is a Bürkholder Inequality that gives a norm relation between a martingale and its optional process. We refer to Liptser and Shiryayev (1989) p.75, for the proof of this result.

Theorem 6.9 (Bürkholder Inequality)**.** *If $M = (M_t, \mathcal{F}_t)_{t\geq 0}$ is a martingale, then there are universal constants γ_b and κ_b (independent of M) such that for every $t \geq 0$*

$$\gamma_b||\sqrt{[M]_t}||_2 \leq ||M_t||_2 \leq \kappa_b||\sqrt{[M]_t}||_2,$$

where $[M]_t$ is the quadratic variation of M_t.

This theorem is used to prove Lemma 6.6 and in the oracle inequalities of Theorem 4.1, the constants depend on κ_b.

6.2 Proofs of the main theorems

6.2.1 Proof of Theorem 4.1

In the following, we consider the sets Δ_1, Δ_2 and Ω defined by (25) and (26) and the set Ω_H^k defined by (27). For sake of simplicity in the notations, we denote $\aleph_k$ the intersection between the four sets: $\aleph_k = \Delta_1 \cap \Delta_2 \cap \Omega \cap \Omega_H^k$. We have the following decomposition:

$$\mathbb{E}[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_0||^2_{det}] \leq 2||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} + 2\mathbb{E}[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k}] + 2\mathbb{E}[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k^c}].$$

The first term is the usual bias term. From Proposition 6.3, we deduce that the last term is bounded by $\tilde{c}_1/n$. We now focus on the term $\mathbb{E}[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k}]$. From Lemma 6.2, for all $m \in \mathcal{M}_n$, the matrices $\boldsymbol{G}^{\hat{\beta}}_{\boldsymbol{m}}$ are invertible on $\Delta_1 \cap \Delta_2 \cap \Omega \cap \Omega_H^k$ as soon as $n \geq 16/(f_0 \mathrm{e}^{-3BR})^2$ and thus the estimator $\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}}$ of α_0 is well defined. From (22) and (24), with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, we have for all $m \in \mathcal{M}_n$,

$$\begin{aligned}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{rand(\hat{\beta})} \leq\, & 2\nu_n(\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}) + 2\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}, \alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}\rangle_{rand} \\ & + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\hat{\beta}}) + 2\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}, \alpha_m^{\boldsymbol{\beta_0}}\rangle_{rand} - 2\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}, \alpha_m^{\boldsymbol{\beta_0}}\rangle_{rand(\hat{\beta})},\end{aligned}$$

where the empirical process $\nu_n(.)$ is defined by Equation (23) and the random norm by (21). For $\mathcal{B}^{det}_{m,m'}(0,1)$ defined by (29), using the classical inequality $2xy \leq\; bx^2 + y^2/b$ with $b > 0$, we obtain

$$\begin{aligned}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{rand(\hat{\beta})} \leq\, & \frac{1}{16}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{rand} + 16||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||^2_{rand} + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\hat{\beta}}) \\ & + \frac{1}{16}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} + 16 \sup_{\alpha \in \mathcal{B}^{det}_{m,\hat{m}^{\hat{\beta}}}(0,1)} \nu_n^2(\alpha) \\ & + 2\Big(\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}, \alpha_m^{\boldsymbol{\beta_0}}\rangle_{rand} - \langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}, \alpha_m^{\boldsymbol{\beta_0}}\rangle_{rand(\hat{\beta})}\Big).\end{aligned}$$

Consequently, using the relations between the random norms $||.||_{rand(\hat{\beta})}$ and $||.||_{rand}$ and between the random norm $||.||_{rand}$ and the deterministic norm $||.||_{det}$ on $\aleph_k$, we obtain

$$\begin{aligned}\frac{1}{4}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} \leq \frac{3}{32}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} &+ 16||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\hat{\beta}}) \\ &+ \frac{1}{16}||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + 16 \sup_{\alpha\in\mathcal{B}^{det}_{m,\hat{m}^{\hat{\beta}}}(0,1)} \nu_n^2(\alpha) \\ &+ 2\Big(\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand} - \langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\hat{\beta})}\Big),\end{aligned}$$

also be rewritten for $p(m,m')$ defined by (30) for all $m' \in \mathcal{M}_n$, as

$$\begin{aligned}&\frac{3}{32}\mathbb{E}\Big[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{\aleph_k}\Big] \leq 16||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + 16p(m,\hat{m}^{\hat{\beta}}) \\ &+ \mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\hat{\beta}}) + 16\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left(\sup_{\alpha\in\mathcal{B}^{det}_{m,m'}(0,1)}\nu_n^2(\alpha) - p(m,m')\right)_+\mathbb{1}_{\aleph_k}\right) \\ &+ 2\mathbb{E}\Big[\Big(\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand} - \langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\hat{\beta})}\Big)\mathbb{1}_{\aleph_k}\Big].\end{aligned}$$

We fix $K_0 \geq 16\kappa$ such that $16p(m,m') \leq \mathrm{pen}(m) + \mathrm{pen}(m')$, for all m, m' in $\mathcal{M}_n$, so that

$$\begin{aligned}\frac{3}{32}\mathbb{E}\Big[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{\aleph_k}\Big] \leq 16||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} &+ 2\,\mathrm{pen}(m) \\ &+ 16\sum_{m'\in\mathcal{M}_n}\mathbb{E}\Big(\Big(\sup_{\alpha\in\mathcal{B}^{det}_{m,m'}(0,1)}\nu_n^2(\alpha) - p(m,m')\Big)_+\mathbb{1}_{\aleph_k}\Big) \\ &+ 2\mathbb{E}\Big[\Big(\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand} - \langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\hat{\beta})}\Big)\mathbb{1}_{\aleph_k}\Big],\end{aligned}$$

that is

$$\frac{3}{32}\mathbb{E}[||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{\aleph_k}] \leq 16||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + 2\,\mathrm{pen}(m) + A(m) + \mathbb{E}[B(m,\hat{m}^{\hat{\beta}})\mathbb{1}_{\aleph_k}] \tag{32}$$

where

$$A(m) = 16\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left(\sup_{\alpha\in\mathcal{B}^{det}_{m,m'}(0,1)}\nu_n^2(\alpha) - p(m,m')\right)_+\mathbb{1}_{\aleph_k}\right), \tag{33}$$

$$B(m,\hat{m}^{\hat{\beta}}) = 2\Big(\langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand} - \langle\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\hat{\beta})}\Big). \tag{34}$$

It remains to study the terms $A(m)$ and $B(m,\hat{m}^{\hat{\beta}})$.

Study of (33). According to Proposition 6.4, for n large enough

$$\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left(\sup_{\alpha\in\mathcal{B}^{det}_{m,m'}(0,1)}\nu_n^2(\alpha) - p(m,m')\right)_+\mathbb{1}_{\aleph_k}\right) \leq \frac{C_3}{n},$$

where $p(m,m')$ is defined by (30) and C_3 is a constant depending on f_0, $|\boldsymbol{\beta_0}|_1$, B, $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$, $||\alpha_0||_{\infty,\tau}$ and the choice of the basis. Hence, for $C_3' = 16C_3$, we conclude that

$$A(m) \leq \frac{C_3'}{n}. \tag{35}$$

Study of (34). Using again the classical inequality $2xy \le bx^2 + y^2/b$ with $b > 0$, we obtain

$$\langle \hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\beta_0}_m, \alpha^{\beta_0}_m \rangle_{rand} - \langle \hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\beta_0}_m, \alpha^{\beta_0}_m \rangle_{rand(\hat{\beta})} \le \frac{1}{32} || \hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}} - \alpha^{\beta_0}_m ||^2_{det}$$
$$+ 32 \sup_{\alpha \in \mathcal{B}^{det}_{m, \hat{m}^{\hat{\beta}}}(0,1)} \Big(\frac{1}{n} \sum_{i=1}^n \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) (\mathrm{e}^{\beta_0^T Z_i} - \mathrm{e}^{\hat{\beta}^T Z_i}) Y_i(t) \mathrm{d}t \Big)^2. \tag{36}$$

Now, from Assumption 3.5.(iii) and by definition (31) of $\mathcal{B}^{det}_n(0,1)$, we write that

$$\sup_{\alpha \in \mathcal{B}^{det}_{m, \hat{m}^{\hat{\beta}}}(0,1)} \left(\frac{1}{n} \sum_{i=1}^n \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) (\mathrm{e}^{\beta_0^T Z_i} - \mathrm{e}^{\hat{\beta}^T Z_i}) Y_i(t) \mathrm{d}t \right)^2$$

is less than

$$\sup_{\alpha \in \mathcal{B}^{det}_n(0,1)} \left(\frac{1}{n} \sum_{i=1}^n \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} (1 - \mathrm{e}^{\hat{\beta}^T Z_i - \beta_0^T Z_i}) Y_i(t) \mathrm{d}t \right)^2.$$

We have

$$\left| \frac{1}{n} \sum_{i=1}^n \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} (1 - \mathrm{e}^{\hat{\beta}^T Z_i - \beta_0^T Z_i}) Y_i(t) \mathrm{d}t \right|$$
$$\le \frac{1}{n} \sum_{i=1}^n \left| 1 - \mathrm{e}^{\hat{\beta}^T Z_i - \beta_0^T Z_i} \right| \left| \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} Y_i(t) \mathrm{d}t \right|.$$

Using the fact that $|\mathrm{e}^x - \mathrm{e}^y| \le |x - y| \mathrm{e}^{x \vee y}$ for all $(x, y) \in \mathbb{R}^2$ and applying Assumptions 2.2.(i) and Assumptions 3.1, we obtain that

$$\left| \frac{1}{n} \sum_{i=1}^n \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} (1 - \mathrm{e}^{\hat{\beta}^T Z_i - \beta_0^T Z_i}) Y_i(t) \mathrm{d}t \right|$$
$$\le \frac{1}{n} \sum_{i=1}^n |\hat{\beta}^T Z_i - \beta_0^T Z_i| \mathrm{e}^{|\hat{\beta}^T Z_i - \beta_0^T Z_i|} \left| \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} Y_i(t) \mathrm{d}t \right|$$
$$\le B \mathrm{e}^{2BR} |\hat{\beta} - \beta_0|_1 \left| \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} Y_i(t) \mathrm{d}t \right|.$$

Now, write

$$\sup_{\alpha \in \mathcal{B}^{det}_n(0,1)} \left(\frac{1}{n} \sum_{i=1}^n \int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} (1 - \mathrm{e}^{\hat{\beta}^T Z_i - \beta_0^T Z_i}) Y_i(t) \mathrm{d}t \right)^2$$
$$\le B^2 \mathrm{e}^{4BR} |\hat{\beta} - \beta_0|_1^2 \sup_{\alpha \in \mathcal{B}^{det}_n(0,1)} \frac{1}{n} \sum_{i=1}^n \left(\int_0^\tau \alpha(t) \alpha^{\beta_0}_m(t) \mathrm{e}^{\beta_0^T Z_i} Y_i(t) \mathrm{d}t \right)^2$$
$$\le B^2 \mathrm{e}^{4BR} |\hat{\beta} - \beta_0|_1^2 \sup_{\alpha \in \mathcal{B}^{det}_n(0,1)} \{ \eta_n(\alpha, \alpha^{\beta_0}_m) + D_n(\alpha, \alpha^{\beta_0}_m) \} \tag{37}$$

where $\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})$ is defined by

$$\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right)^2 - \mathbb{E}\left[\left(\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right)^2\right]\right],$$

and

$$D_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}}) = \mathbb{E}\left[\left(\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z}}Y(t)\mathrm{d}t\right)^2\right].$$

We first claim that the term $\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\{D_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})\}$ is bounded, by using that from the Cauchy-Schwarz Inequality,

$$\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)} \mathbb{E}\left[\left(\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z}}Y(t)\mathrm{d}t\right)^2\right] \leq ||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2.$$

Thus, gathering bounds (36) and (37, we obtain that

$$B(m, \hat{m}^{\hat{\boldsymbol{\beta}}}) \leq \frac{1}{16}||\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||_{det}^2 + 64\left[B^2\mathrm{e}^{4BR}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^2\Big(\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\{\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})\} + ||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\Big)\right].$$

So, taking the expectation and applying Proposition 6.5 to control

$$\mathbb{E}[\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}(\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}}))^2],$$

we get

$$\begin{aligned}
&\mathbb{E}[B(m, \hat{m}^{\hat{\boldsymbol{\beta}}})\mathbb{1}_{\aleph_k}] \leq \frac{1}{16}\mathbb{E}[||\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{1}_{\aleph_k}] \\
&+64B^2\mathrm{e}^{4BR}\left\{\mathbb{E}^{1/2}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^4\mathbb{1}_{\aleph_k}]\mathbb{E}^{1/2}\left[\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\{\eta_n^2(\alpha, \alpha_m^{\boldsymbol{\beta_0}})\}\right] + ||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{E}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^2\mathbb{1}_{\aleph_k}]\right\}. \qquad (38)
\end{aligned}$$

Finally, combining (32), (35) and (38) we conclude that

$$\begin{aligned}
\frac{1}{16}\mathbb{E}[||\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}} - \alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{1}_{\aleph_k}] &\leq 16||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||_{det}^2 + 2\,\mathrm{pen}(m) + \frac{C_3'}{n} \\
&+ 64B^2\mathrm{e}^{4BR}||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{E}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^2\mathbb{1}_{\aleph_k}] \\
&+ 64B^2\mathrm{e}^{4BR}\mathbb{E}^{1/2}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^4\mathbb{1}_{\aleph_k}]\frac{\mathbb{E}^{1/2}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]||\alpha_m^{\boldsymbol{\beta_0}}||_2^2}{\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}f_0}\frac{1}{\sqrt{n}}.
\end{aligned}$$

On $\Omega\cap\Omega_H^k$, using that, from definition (15) and Proposition 6.1, $||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2 \leq 2||\alpha_0||_{det} \leq \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0}^T Z}]\tau||\alpha_0||_{\infty,\tau}$, we have

$$64B^2\mathrm{e}^{4BR}||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{E}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^2\mathbb{1}_{\aleph_k}] \leq C(s, B, R, \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}], ||\alpha_0||_{\infty,\tau}, \tau)\frac{\log(pn^k)}{n},$$

and that

$$\begin{aligned}
&64B^2\mathrm{e}^{4BR}\mathbb{E}^{1/2}[|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1^4\mathbb{1}_{\aleph_k}]\frac{\mathbb{E}^{1/2}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]||\alpha_m^{\boldsymbol{\beta_0}}||_2^2}{\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}f_0}\frac{1}{\sqrt{n}} \\
&\qquad \leq \tilde{C}(s, B, |\boldsymbol{\beta_0}|_1, R, \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}], \mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}], ||\alpha_0||_{\infty,\tau}, \tau, f_0)\frac{\log(pn^k)}{n\sqrt{n}},
\end{aligned}$$

where s is the sparsiy index of $\boldsymbol{\beta_0}$ and

$$C(s, B, R, \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}], ||\alpha_0||_{\infty,\tau}, \tau) \quad \text{and} \quad \tilde{C}(s, B, |\boldsymbol{\beta_0}|_1, R, \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}], \mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}], ||\alpha_0||_{\infty,\tau}, \tau, f_0)$$

are constants depending on the elements in brackets. Combining the previous bounds with Proposition 6.3, we conclude that Theorem 4.1 is proved since

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}] \leq \kappa_0 \inf_{m\in\mathcal{M}_n}\{||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} + 2\,\mathrm{pen}(m)\} + \frac{C_1}{n} + C_2\frac{\log(pn)}{n},$$

where C_1 and C_2 are constants depending on the sparsity index s of $\boldsymbol{\beta_0}$, B, $|\boldsymbol{\beta_0}|_1$, $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$, $\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]$,$||\alpha_0||_{\infty,\tau}$, τ, f_0.

$\square$

6.2.2 Proof of Corollary 4.2

From Proposition 6.1 and the proof of Corollary 1 in Comte et al. (2011), we deduce that

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_2] \leq \frac{\mathrm{e}^{B|\boldsymbol{\beta_0}|_1}}{f_0}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}] \leq \tilde{C}_1 \inf_{m\in\mathcal{M}_n}\left\{D_m^{-2\gamma} + \frac{D_m}{n}\right\} + \tilde{C}_2(s)\frac{\log(np)}{n},$$

and since

$$\inf_{m\in\mathcal{M}_n}\left\{D_m^{-2\gamma} + \frac{D_m}{n}\right\} = n^{-\frac{2\gamma}{2\gamma+1}},$$

we finally get the corollary. $\square$

6.3 Proofs of the technical propositions and lemmas

6.3.1 Proof of Lemma 6.2

Let $m \in \mathcal{M}_n$ be fixed and let v be an eigenvalue of $\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}$. There exists $\boldsymbol{A_m} \neq 0$ with coefficients $(a_j)_j$ such that $\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}\boldsymbol{A_m} = v\boldsymbol{A_m}$ and thus $\boldsymbol{A_m^T}\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}\boldsymbol{A_m} = v\boldsymbol{A_m^T}\boldsymbol{A_m}$. Now, take $h := \sum_j a_j\varphi_j \in S_m$. We have $||h||^2_{rand(\hat{\boldsymbol{\beta}})} = \boldsymbol{A_m^T}\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}\boldsymbol{A_m^T}$ and $||h||^2_2 = \boldsymbol{A_m^T}\boldsymbol{A_m}$. Thus, on $\Delta_1 \cap \Delta_2$ defined in (25) and (26) and from Proposition 6.1:

$$\boldsymbol{A_m^T}\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}\boldsymbol{A_m^T} = ||h||^2_{rand(\hat{\boldsymbol{\beta}})} \geq \frac{1}{2}||h||^2_{rand} \geq \frac{1}{4}||h||^2_{det} \geq \frac{1}{4}f_0\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}||h||^2_2.$$

Therefore, on $\Delta_1 \cap \Delta_2$, for all $m \in \mathcal{M}_n$, we have $\min\mathrm{Sp}(\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}) \geq f_0\mathrm{e}^{-3BR}/4$. Moreover, on Ω, we have $f_0 \geq 2\hat{f}_0/3$ and $\max(\hat{f}_0\mathrm{e}^{-3BR}/6, n^{-1/2}) = \hat{f}_0\mathrm{e}^{-3BR}/6$ for $n \geq 36/(\hat{f}_0\mathrm{e}^{-3BR})^2$, which is equivalent on Ω to choose $n \geq 16/(f_0\mathrm{e}^{-3BR})^2$. $\square$

6.3.2 Proof of Proposition 6.3

We have the following decomposition :

$$\begin{aligned}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k^c}] \leq & \mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\Delta_1^c}] + \mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\Delta_2^c}] \\ & + \mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\Omega^c}] + \mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{(\Omega_H^k)^c}].\end{aligned}$$

We deduce that

$$\begin{aligned}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k^c}] \leq 2\Big(&\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}\mathbb{1}_{\Delta_1^c}] + \mathbb{E}[||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||^2_{det}\mathbb{1}_{\Delta_1^c}] \\ &+\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}\mathbb{1}_{\Delta_2^c}] + \mathbb{E}[||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||^2_{det}\mathbb{1}_{\Delta_2^c}] \\ &+\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}\mathbb{1}_{\Omega^c}] + \mathbb{E}[||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||^2_{det}\mathbb{1}_{\Omega^c}] \\ &+\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}\mathbb{1}_{(\Omega_H^k)^c}] + \mathbb{E}[||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||^2_{det}\mathbb{1}_{(\Omega_H^k)^c}]\Big).\end{aligned}$$

From definition (15) of $\alpha_m^{\boldsymbol{\beta_0}}$ and Proposition 6.1, we have $||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||^2_{det} \leq ||\alpha_0||^2_{det} \leq \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]||\alpha_0||_2^2$. From this relation and using Cauchy-Schwarz Inequality, we have

$$\begin{aligned}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k^c}] \leq 4\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]\Big[&\mathbb{E}^{1/2}(||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}||_2^4)\Big(\mathbb{P}^{1/2}(\Delta_1^c)+\mathbb{P}^{1/2}(\Delta_2^c) \\ &+\mathbb{P}^{1/2}(\Omega^c) + \mathbb{P}^{1/2}((\Omega_H^k)^c)\Big) + ||\alpha_0||_2^2(\mathbb{P}(\Delta_1^c) + \mathbb{P}(\Delta_2^c)+\mathbb{P}(\Omega^c) + \mathbb{P}((\Omega_H^k)^c))\Big].\end{aligned}$$

From Assumption 3.4, Proposition 3.2, Lemmas 6.6, 6.7 and 6.8 with $k = 6$, we conclude that

$$\begin{aligned}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det}\mathbb{1}_{\aleph_k^c}] &\leq 2\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]\Bigg[\sqrt{C_b n^4}\Bigg(\sqrt{\frac{C_6^{(\Delta_1)}}{n^6}} + \sqrt{\frac{C_6^{(\Delta_2)}}{n^6}} + \sqrt{\frac{C_0}{n^6}} + \sqrt{\frac{c}{n^6}}\Bigg) \\ &\qquad + ||\alpha_0||_2^2\Bigg(\frac{C_6^{(\Delta_1)}}{n^6} + \frac{C_6^{(\Delta_2)}}{n^6} + \frac{C_0}{n^6} + \frac{c}{n^6}\Bigg)\Bigg] \\ &\leq \frac{\tilde{c}_1}{n},\end{aligned}$$

which ends the proof of Proposition 6.3. □

6.3.3 Proof of Proposition 6.5

The proof is inspired from the paper of Brunel et al. (2010). If we denote $(\varphi_j)_{j\in\mathcal{K}_n}$ the orhonormal basis of the global nesting space $\mathcal{S}_n$ (see Assumption 3.5.(iii)), since α belongs to $\mathcal{B}_n^{det}(0,1) \subset \mathcal{S}_n$, we can write $\alpha(t) = \sum_{j\in\mathcal{K}_n} a_j\varphi_j(t)$, with $\dim \mathcal{S}_n = \mathcal{D}_n = |\mathcal{K}_n|$. With this definition, we obtain

$$\begin{aligned}\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}}) = \sum_{j,j'} a_j a_{j'}\frac{1}{n}\sum_{i=1}^n\Big(&\int_0^\tau \varphi_j(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\int_0^\tau \varphi_{j'}\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t \\ &- \mathbb{E}\Big[\int_0^\tau \varphi_j(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\int_0^\tau \varphi_{j'}\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\Big]\Big)\end{aligned}$$

For sake of simplicity, we introduce the notation

$$A^i_{j,j'} = \int_0^\tau \varphi_j(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\int_0^\tau \varphi_{j'}(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t.$$

Applying the Cauchy-Schwarz Inequality, we get

$$|\eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})| \leq \sqrt{\sum_{j,j'} a_j^2 a_{j'}^2}\sqrt{\sum_{j,j'}\Big(\frac{1}{n}\sum_{i=1}^n(A^i_{j,j'} - \mathbb{E}[A^i_{j,j'}])\Big)^2}.$$

From Proposition 6.1, we have

$$\sup_{\alpha \in \mathcal{B}_n^{det}(0,1)} \eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})^2 \leq \sup_{(a_j), \sum_j a_j^2 \leq 1} \frac{1}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1} f_0)^2} \sum_{j,j'} a_j^2 a_{j'}^2 \sum_{j,j'} \Big(\frac{1}{n} \sum_{i=1}^n (A_{j,j'}^i - \mathbb{E}[A_{j,j'}^i]) \Big)^2$$
$$\leq \frac{1}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1} f_0)^2} \sum_{j,j'} \Big(\frac{1}{n} \sum_{i=1}^n (A_{j,j'}^i - \mathbb{E}[A_{j,j'}^i]) \Big)^2.$$

Taking the expectation, it follows that

$$\mathbb{E}\left[\sup_{\alpha \in \mathcal{B}_n^{det}(0,1)} \eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})^2 \right] \leq \frac{1}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1} f_0)^2} \sum_{j,j'} \mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^n A_{j,j'}^i \right]$$
$$\leq \frac{1}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1} f_0)^2} \sum_{j,j'} \frac{1}{n} \mathbb{E}\Big[(A_{j,j'}^1)^2 \Big].$$

Thus, from the definition of $A_{j,j'}^1$, we obtain that $\mathbb{E}[\sup_{\alpha \in \mathcal{B}_n^{det}(0,1)} \eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})^2]$ is less than

$$\frac{1}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|} f_0)^2} \frac{1}{n} \sum_{j,j'} \mathbb{E}\left[\left(\int_0^\tau \varphi_j(t) \alpha_m^{\boldsymbol{\beta_0}}(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z}} Y(t) \mathrm{d}t \right)^2 \left(\int_0^\tau \varphi_{j'}(t) \alpha_m^{\boldsymbol{\beta_0}}(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z}} Y(t) \mathrm{d}t \right)^2 \right].$$

From Brunel et al. (2010) p.301, Equation (2.7), we have

$$\sum_{j \in \mathcal{K}_n} \left(\int_0^\tau \varphi_j(t) \alpha_m^{\boldsymbol{\beta_0}}(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z}} Y(t) \mathrm{d}t \right)^2 \leq \int_0^\tau (\alpha_m^{\boldsymbol{\beta_0}}(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z}} Y(t))^2 \mathrm{d}t \leq \mathrm{e}^{2\boldsymbol{\beta_0^T Z}} ||\alpha_m^{\boldsymbol{\beta_0}}||_2^2.$$

From this inequality, we obtain

$$\mathbb{E}\left[\sup_{\alpha \in \mathcal{B}_n^{det}(0,1)} \eta_n(\alpha, \alpha_m^{\boldsymbol{\beta_0}})^2 \right] \leq \frac{\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}] ||\alpha_m^{\boldsymbol{\beta_0}}||_2^4}{(\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1} f_0)^2} \frac{1}{n}. \quad \square$$

6.3.4 Proof of Lemma 6.6

From Assumption 3.1, we recall that $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq 2R$. On $\hat{\mathcal{H}}_{\hat{m}^{\hat{\beta}}}^{\hat{\boldsymbol{\beta}}}$, we have

$$||\hat{\alpha}_{\hat{m}^{\hat{\beta}}}^{\hat{\boldsymbol{\beta}}}||_2^2 = \sum_{j \in J_{\hat{m}^{\hat{\beta}}}} (\hat{a}_j^{\hat{m}^{\hat{\beta}}})^2 = ||\boldsymbol{A}_{\hat{\boldsymbol{m}}^{\hat{\beta}}}||_2^2 = ||(\boldsymbol{G}_{\hat{\boldsymbol{m}}^{\hat{\beta}}}^{\hat{\boldsymbol{\beta}}})^{-1} \boldsymbol{\Gamma}_{\hat{\boldsymbol{m}}^{\hat{\beta}}}||_2^2$$
$$\leq (\min \mathrm{Sp}(\boldsymbol{G}_{\hat{\boldsymbol{m}}^{\hat{\beta}}}^{\hat{\boldsymbol{\beta}}}))^{-2} ||\boldsymbol{\Gamma}_{\hat{\boldsymbol{m}}^{\hat{\beta}}}||_2^2$$
$$\leq \min\left(\frac{36}{\hat{f}_0^2 \mathrm{e}^{-2B|\beta_0|_1 - 2B|\boldsymbol{\beta_0} - \hat{\beta}|_1}}, n \right) \sum_{j \in J_{\hat{m}^{\hat{\beta}}}} \left(\frac{1}{n} \sum_{i=1}^n \int_0^\tau \varphi_j(t) \mathrm{d}N_i(t) \right)^2$$
$$\leq \min\left(\frac{36}{\hat{f}_0^2 \mathrm{e}^{-2B|\beta_0|_1 - 4BR}}, n \right) \frac{1}{n} \sum_{i=1}^n \sum_{j \in J_{\hat{m}^{\hat{\beta}}}} \left(\int_0^\tau \varphi_j(t) \mathrm{d}N_i(t) \right)^2.$$

So we have

$$||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}}||_2^4 \leq n^2 \frac{1}{n}\sum_{i=1}^n \left(\sum_{j\in J_{\hat{m}^{\hat{\beta}}}} \left(\int_0^\tau \varphi_j(t)\mathrm{d}N_i(t)\right)^2 \right)^2 \leq n^2 \frac{1}{n}\sum_{i=1}^n \left(\sum_{j\in \mathcal{K}_n} \left(\int_0^\tau \varphi_j(t)\mathrm{d}N_i(t)\right)^2 \right)^2,$$

where $\mathcal{K}_n$ is a set of indices of the global nesting space $\mathcal{S}_n$, defined in Assumption 3.5.(iii), and $\dim \mathcal{S}_n = \mathcal{D}_n = |\mathcal{K}_n|$. Thus, we deduce that

$$||\hat{\alpha}^{\hat{\beta}}_{\hat{m}^{\hat{\beta}}}||_2^4 \leq n^2 \mathcal{D}_n \frac{1}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \left(\int_0^\tau \varphi_j(t)\mathrm{d}N_i(t)\right)^4.$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \left(\int_0^\tau \varphi_j(t)\mathrm{d}N_i(t)\right)^4\right] &\leq \frac{2^3}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \mathbb{E}\left[\left(\int_0^\tau \varphi_j(t)\mathrm{d}M_i(t)\right)^4\right] \\
&\quad + \frac{2^3}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \mathbb{E}\left[\left(\int_0^\tau \varphi_j(t)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right)^4\right].
\end{aligned}$$

Using the Bürkholder Inequality (see Liptser and Shiryayev (1989)), we get

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \left(\int_0^\tau \varphi_j(t)\mathrm{d}M_i(t)\right)^4\right] &\leq \kappa_b \frac{1}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \mathbb{E}\left[\left(\int_0^\tau \varphi_j^2(t)\mathrm{d}N_i(t)\right)^2\right] \\
&\leq \kappa_b \frac{1}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \mathbb{E}\left[N_i(\tau)\sum_{s:\Delta N_i\neq 0}\varphi_j^4(s)\right] \\
&\leq \kappa_b \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[N_i(\tau)\sum_{s:\Delta N_i\neq 0}\sum_{j\in \mathcal{K}_n}\varphi_j^4(s)\right],
\end{aligned}$$

which is finally bounded from Assumption 3.5.(ii) by

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sum_{j\in \mathcal{K}_n} \left(\int_0^\tau \varphi_j(t)\mathrm{d}M_i(t)\right)^4\right] &\leq \kappa_b \phi^2 \mathcal{D}_n^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[N_i(\tau)\sum_{s:\Delta N_i\neq 0} 1\right] \\
&\leq \kappa_b \phi^2 \mathcal{D}_n^2 \mathbb{E}[N_1(\tau)^2].
\end{aligned}$$

Then, we can write that

$$\begin{aligned}
[N_1(\tau)]^2 &= \left[M_1(\tau) + \int_0^\tau \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z}}Y(t)\mathrm{d}t\right]^2 \\
&\leq 2(M_1(\tau))^2 + 2\left(\int_0^\tau \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z}}Y(t)\mathrm{d}t\right)^2,
\end{aligned}$$

and

$$\mathbb{E}[(M_1(\tau))^2] \leq \mathbb{E}\left[\int_0^\tau \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z}}Y(t)\mathrm{d}t\right] \leq \tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}],$$

so that

$$\mathbb{E}[(N_1(\tau))^2] \leq 2||\alpha_0||_{\infty,\tau}\tau\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}] + 2||\alpha_0||_{\infty,\tau}^2(\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}])^2\tau^2.$$

So, by using Cauchy-Schwarz Inequality, we obtain

$$\begin{aligned}
&\mathbb{E}\Bigg[\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{K}_n}\left(\int_0^\tau \phi_j(t)\mathrm{d}N_i(t)\right)^4\Bigg] \\
&\qquad \leq 8\kappa_b\phi^2\mathcal{D}_n^2\mathbb{E}[(N_1(\tau))^2] + 8\sum_{j\in\mathcal{K}_n}\mathbb{E}\Bigg[\left(\int_0^\tau \varphi_j(t)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z}}Y(t)\mathrm{d}t\right)^4\Bigg] \\
&\qquad\qquad \leq 8\kappa_b\phi^2\mathcal{D}_n^2\mathbb{E}[(N_1(\tau))^2] + 8||\alpha_0||_{\infty,\tau}^4\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]\tau^2\mathcal{D}_n.
\end{aligned}$$

Eventually, under Assumption 3.5.(i), we get

$$\begin{aligned}
\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}||_2^4] &\leq n^2\mathcal{D}_n\Big[8\kappa_b\phi^2\mathcal{D}_n^2\Big(2||\alpha_0||_{\infty\tau}\tau\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}] + 2||\alpha_0||_{\infty,\tau}^2(\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}])^2\tau^2\Big) \\
&\qquad + 8||\alpha_0||_{\infty,\tau}^4\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]\tau^2\mathcal{D}_n\Big] \\
&\leq C_b n^2\mathcal{D}_n^3 \\
&\leq C_b n^4,
\end{aligned}$$

where C_b is a constant that depends on κ_b, $||\alpha_0||_{\infty,\tau}$, τ, $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$ and $\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}]$ and on the choice of the basis. $\square$

6.3.5 Proof of Lemma 6.7

The event Δ_1 defined by (25) can be rewritten as

$$\Delta_1 = \left\{\omega\in\Omega, \forall\alpha\in\mathcal{S}_n\backslash\{0\} : \left|\frac{||\alpha||_{rand(\omega)}^2}{||\alpha||_{det}^2} - 1\right| \leq \frac{1}{2}\right\},$$

and consider

$$\vartheta_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big(\alpha(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t) - \mathbb{E}[\alpha(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)]\Big)\mathrm{d}t = ||\sqrt{\alpha}||_{rand}^2 - ||\sqrt{\alpha}||_{det}^2. \tag{39}$$

If $\omega\in(\Delta_1)^c$, then there exists α (which can depend on ω) such that

$$\left|\frac{||\alpha||_{rand(\omega)}^2}{||\alpha||_{det}^2} - 1\right| > \frac{1}{2}.$$

Taking $\gamma = \alpha/||\alpha||_{det}^2$, we have that

$$\gamma\in\mathcal{S}_n\backslash\{0\}, \quad ||\gamma||_{det}^2 = 1, \quad \text{and} \quad |||\gamma||_{rand(\omega)}^2 - 1| > \frac{1}{2}.$$

So, if $\omega\in(\Delta_1)^c$, then

$$\omega\in\left\{\omega\in\Omega : \sup_{\gamma\in\mathcal{S}_n\backslash\{0\}, ||\gamma||_{det}^2=1} |||\gamma||_{rand(\omega)}^2 - 1| > \frac{1}{2}\right\}$$

From this, we deduce that,

$$\mathbb{P}((\Delta_1)^c) \leq \mathbb{P}\left(\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)} |\vartheta_n(\alpha^2)| > 1 - \frac{1}{\rho_1}\right),$$

where $\mathcal{B}_n^{det}(0,1)$ is defined by (31). Since $\alpha \in \mathcal{B}_n^{det}(0,1) \subset \mathcal{S}_n$, then we can write $\alpha(t) = \sum_{j\in\mathcal{K}_n} a_j^m \varphi_j(t)$, where $\mathcal{K}_n$ is a set of indices of $\mathcal{S}_n$ and $\dim \mathcal{S}_n = \mathcal{D}_n = |\mathcal{K}_n|$. With this notation, we have

$$\vartheta_n(\alpha^2) = \sum_{j,k} a_j a_k \vartheta_n(\varphi_j\varphi_k).$$

From Proposition 6.1, we have

$$\sup_{\alpha\in\mathbb{B}_n^{det}(0,1)} |\vartheta_n(\alpha^2)| \leq \frac{1}{f_0 \mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}} \sup_{(a_j),\sum_{j\in\mathcal{K}_n} a_j^2\leq 1} \Big|\sum_{j,k} a_j a_k \vartheta_n(\varphi_j\varphi_k)\Big|.$$

Let consider the process $(U_i^{(j,k)})$ defined by

$$U_i^{(j,k)} = \int_0^\tau \varphi_j(t)\varphi_k(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} Y_i(t)\mathrm{d}t,$$

We have $|U_i^{(j,k)}| \leq \mathrm{e}^{B|\beta_0|_1}$ and from Cauchy-Schwarz Inequality, we have

$$(U_i^{(j,k)})^2 \leq \mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} \int_0^\tau \varphi_j^2(t)\mathrm{d}t \int_0^\tau \varphi_k^2(t)\mathrm{d}t \leq \mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}.$$

We can apply the standard Bernstein Inequality (see Massart (2007)) to the process $(U_i^{(j,k)})$, and we obtain

$$\mathbb{P}\Big(|\vartheta_n(\varphi_j\varphi_k)| \geq \mathrm{e}^{B|\beta_0|_1} x + \sqrt{2\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} x}\Big) \leq 2\mathrm{e}^{-nx}. \tag{40}$$

Let introduce

$$\Theta := \{\forall j,k, |\vartheta_n(\varphi_j\varphi_k)| \leq \mathrm{e}^{B|\boldsymbol{\beta_0}|_1} x + \mathrm{e}^{B|\boldsymbol{\beta_0}|_1}\sqrt{2x}\} \quad \text{and} \quad x := \frac{f_0^2 \mathrm{e}^{-2B|\beta_0|_1}}{16\mathcal{D}_n^2 \mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}}.$$

On Θ, we can write that $\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)} |\vartheta_n(\alpha^2)|$ is less than

$$\begin{aligned}
&\frac{1}{f_0 \mathrm{e}^{-B|\beta_0|_1}} \sup_{(a_j),\sum_{j\in\mathcal{K}_n} a_j^2\leq 1} \sum_{j,k} |a_j a_k| (\mathrm{e}^{B|\beta_0|_1} x + \mathrm{e}^{B|\beta_0|_1}\sqrt{2x}) \\
\leq &\frac{1}{f_0 \mathrm{e}^{-B|\beta_0|_1}} \sup_{(a_j),\sum_{j\in\mathcal{K}_n} a_j^2\leq 1} \Big(\sum_j |a_j|\Big)^2 (\mathrm{e}^{B|\beta_0|_1} x + \mathrm{e}^{B|\beta_0|_1}\sqrt{2x}),
\end{aligned}$$

which is less than

$$\begin{aligned}
\leq &\frac{1}{f_0 \mathrm{e}^{-B|\beta_0|_1}} D_m \left(\frac{\mathrm{e}^{B|\beta_0|_1} f_0^2 \mathrm{e}^{-2B|\beta_0|_1}}{16\mathcal{D}_n^2 \mathrm{e}^{2B|\beta_0|_1}} + \frac{\mathrm{e}^{B|\beta_0|_1}\sqrt{2} f_0 \mathrm{e}^{-B|\beta_0|_1}}{4\mathcal{D}_n \mathrm{e}^{B|\beta_0|_1}}\right) \\
\leq &\frac{1}{2}\left(\frac{1}{8}\frac{f_0}{\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}\mathcal{D}_n} + \frac{1}{\sqrt{2}}\right) \\
\leq &\frac{1}{2}\left(\frac{1}{4} + \frac{1}{\sqrt{2}}\right) \\
\leq &\frac{1}{2}.
\end{aligned} \tag{41}$$

From Inequality (41), we deduce that $\mathbb{P}((\Delta_1)^c) \leq \mathbb{P}(\Theta^c)$. So using Inequality (40), we can conclude that

$$\begin{aligned}
\mathbb{P}((\Delta_1)^c) &\leq \sum_{j,k} \mathbb{P}\Big(|\vartheta_n(\varphi_j\varphi_k)| > \mathrm{e}^{B|\boldsymbol{\beta}_0|_1} x + \mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\sqrt{2x}\Big) \\
&\leq 2\mathcal{D}_n^2 \exp\left(-\frac{n f_0^2 \mathrm{e}^{-2B|\boldsymbol{\beta}_0|_1}}{16\mathcal{D}_n^2 \mathrm{e}^{2B|\boldsymbol{\beta}_\mathbf{0}|_1}}\right) \\
&\leq 2n \exp\left(-\frac{f_0^2}{16\mathrm{e}^{4B|\boldsymbol{\beta}_0|_1}} \frac{n}{\mathcal{D}_n^2}\right) \\
&\leq 2n \exp\left(-\frac{f_0^2}{16\mathrm{e}^{4B|\boldsymbol{\beta}_0|_1}} \log n\right) \\
&\leq \frac{C_k^{\Delta_1}}{n^k}, \quad \forall k \geq 1,
\end{aligned}$$

as $\mathcal{D}_n \leq \sqrt{n}/\log n$ from Assumption 3.5.(iii), which ends the proof of Lemma 6.7 with $C_k^{\Delta_1}$ a constant depending on ρ_1, f_0, B and $|\boldsymbol{\beta}_0|_1$. □

6.3.6 Proof of Lemma 6.8

For $\rho_2 \geq 1$, let define

$$\Delta_2^{\rho_2} = \left\{\forall \alpha \in \mathcal{S}_n : \left|\frac{||\alpha||^2_{rand(\hat{\boldsymbol{\beta}})}}{||\alpha||^2_{rand}} - 1\right| \leq 1 - \frac{1}{\rho_2}\right\}.$$

Let consider

$$\tilde{\vartheta}_n(\alpha) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau (\alpha(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z}_i} Y_i(t) - \alpha(t)\mathrm{e}^{\boldsymbol{\beta}_\mathbf{0}^{\boldsymbol{T}}\boldsymbol{Z}_i} Y_i(t))\mathrm{d}t = ||\sqrt{\alpha}||^2_{rand(\hat{\boldsymbol{\beta}})} - ||\sqrt{\alpha}||^2_{rand}.$$

Following the same approach as in the proof of Lemma 6.7, we have

$$\mathbb{P}((\Delta_2^{\rho_2})^c) \leq \mathbb{P}\left(\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)} |\tilde{\vartheta}_n(\alpha^2)| > 1 - \frac{1}{\rho_2}\right), \tag{42}$$

where $\mathcal{B}_n^{det}(0,1) = \{\alpha \in \mathcal{S}_n : ||\alpha||_{det} \leq 1\}$. The process $\tilde{\vartheta}_n(\alpha^2)$ is bounded by

$$|\tilde{\vartheta}_n(\alpha^2)| \leq B\mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathbf{0}|_1||\alpha||_2^2 \leq |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathbf{0}|_1 \frac{B\mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}}{f_0\mathrm{e}^{-B|\boldsymbol{\beta}_\mathbf{0}|_1}}||\alpha||^2_{det}.$$

So we get

$$\sup_{\alpha\in\mathcal{B}_{\mathcal{S}_n}^{det}(0,1)} |\tilde{\vartheta}_n(\alpha^2)| \leq |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathbf{0}|_1 \frac{B\mathrm{e}^{2B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}}{f_0}.$$

From Proposition 3.2, we have with probability larger than $1 - cn^{-k}$

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\mathbf{0}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}}.$$

Then we have with probability larger than $1 - cn^{-k}$

$$\sup_{\alpha \in \mathcal{B}^{det}_{\mathcal{S}_n}(0,1)} |\tilde{\vartheta}_n(\alpha^2)| \leq C(s)\sqrt{\frac{\log(pn^k)}{n}} \frac{B\mathrm{e}^{2B|\beta_0|_1}\mathrm{e}^{2BR}}{f_0}.$$

Thus, by taking $1 - 1/\rho_2 = C(s)\sqrt{\frac{\log(pn^k)}{n}} \frac{B\mathrm{e}^{2B|\beta_0|_1}\mathrm{e}^{2BR}}{f_0}$ in (42), we obtain

$$\mathbb{P}((\Delta_2^{\rho_2})^c) \leq cn^{-k}.$$

From Assumption 3.3, we deduce that for n large enough,

$$1 - \frac{1}{\rho_2} < \frac{1}{2},$$

so that Δ_2 defined by (26) verifies $\mathbb{P}((\Delta_2)^c) \leq \mathbb{P}((\Delta_2^{\rho_2})^c) \leq C_k^{(\Delta_2)} n^{-k}$, with $C_k^{(\Delta_2)} = c > 0$. $\square$

A Prediction result on the Lasso estimator $\hat{\beta}$ of β_0 for unbounded counting processes

To obtain a non-asymptotic prediction bound on the Lasso estimator $\hat{\boldsymbol{\beta}}$ of the regression parameter in the Cox model, we rely on Theorem 3.1 of Huang et al. (2013), that we recall here.

Let consider the classical Lasso estimator $\hat{\boldsymbol{\beta}}$ defined by (3) when $p \gg n$.

We define $\dot{\boldsymbol{l}}^*_{\boldsymbol{n}}(\boldsymbol{\beta}) = (\dot{l}^*_{n,1}(\boldsymbol{\beta}), ..., \dot{l}^*_{n,p}(\boldsymbol{\beta}))^T = \partial l^*_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ the gradient of the Cox partial log-likelihood $l^*_n(\boldsymbol{\beta})$ defined by (4) and $\ddot{\boldsymbol{l}}^*_{\boldsymbol{n}}(\boldsymbol{\beta}) = \partial^2 l^*_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ the Hessian matrix.

Let us now describe the result of Huang et al. (2013), on which we rely for our study, starting with the notations. Let $\mathcal{O} = \{j : \beta_{0j} \neq 0\}$, $\mathcal{O}^c = \{j : \beta_{0j} = 0\}$ and $s = |\mathcal{O}|$ the cardinality of $\mathcal{O}$. 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, let us define the following condition:

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

This term corresponds to the compatibility factor introduced by van de Geer (2007). It is one of the classical condition used to obtain non-asymptotic oracle inequalities. See also Bühlmann and van de Geer (2009) for more details about this compatibility factor and the comparison of this criterion with other assumptions such as the Restricted Eigenvalue condition among other.

With these notations, we can state the following theorem established by Huang et al. (2013).

Theorem A.1 (Huang et al. (2013))**.** *Let $k > 0$ and $\nu = B(\xi+1)s\Gamma_{n,k}/\{2\kappa^2(\xi,\mathcal{O})\}$. Suppose Assumption 2.2.(i) holds and $\nu \leq 1/\mathrm{e}$. Then, on the event*

$$\widetilde{\Omega}^k_H = \left\{ |\dot{\boldsymbol{l}}^*_{\boldsymbol{n}}(\boldsymbol{\beta_0})|_\infty \leq \frac{\xi - 1}{\xi + 1}\Gamma_{n,k} \right\}, \quad \text{with} \quad \Gamma_{n,k} = C_0 B \frac{\xi+1}{\xi-1}\sqrt{2\frac{\log(pn^k)}{n}}, \tag{43}$$

we have

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq \frac{\mathrm{e}^{\eta}(\xi+1)s}{2\kappa^2(\xi,\mathcal{O})}\Gamma_{n,k},$$

where $\eta \leq 1$ *is the smaller solution of* $\eta\mathrm{e}^{-\eta} = \nu$ *and* $C_0 > \sqrt{\tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]}$.

We refer to Huang et al. (2013) for the proof of Theorem A.1. Huang et al. (2013) have calculated the probability of $\widetilde{\Omega}_H^k$ only in the case where $\max_{1\leq i\leq n}|N_i(\tau)| < +\infty$. We extend the result to the unbounded case in the following lemma.

Lemma A.2. *Let consider, for* $k > 0$, *the event* $\widetilde{\Omega}_H^k$ *defined by (43). Then, under Assumptions 2.2.(i) and (iv), there exists a constant* $c > 0$ *depending on* τ, $||\alpha_0||_{\infty,\tau}$ *and* $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$ *such that*

$$\mathbb{P}((\widetilde{\Omega}_H^k)^c) \leq cn^{-k}.$$

The proof of this lemma follows. From this lemma, we can rewrite Theorem A.1 as:

Corollary A.3. *Let* $\nu = B(\xi+1)s\Gamma_{n,k}/\{2\kappa^2(\xi,\mathcal{O})\}$, $k > 0$ *and* $c > 0$. *Suppose Assumptions 2.2.(i) and (iv) hold and* $\nu \leq 1/\mathrm{e}$. *Then, with probability larger than* $1 - cn^{-k}$

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq \frac{\mathrm{e}^{\eta}(\xi+1)s}{2\kappa^2(\xi,\mathcal{O})}\Gamma_{n,k} \quad \textit{with} \quad \Gamma_{n,k} = C_0 B\frac{\xi+1}{\xi-1}\sqrt{2\frac{\log(pn^k)}{n}},$$

where $\eta \leq 1$ *is the smaller solution of* $\eta\mathrm{e}^{-\eta} = \nu$ *and* $C_0 > \sqrt{\tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]}$.

From Corollary A.3 and Assumption 2.2.(i), we deduce a prediction inequality given by the following proposition.

Proposition A.4. *Let* $k > 0$ *and* $c > 0$. *Under Assumptions 2.2.(i) and 2.2.(iv), with probability larger than* $1 - cn^{-k}$, *we have*

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}}, \tag{44}$$

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

Remark A.5. *From Proposition A.4 and Definition (27) of* Ω_H^k, *we deduce that* $\widetilde{\Omega}_H^k \subset \Omega_H^k$.

Proof of Lemma A.2 To prove Lemma A.2, we start from Lemma 3.3. p.10 in the paper of Huang et al. (2013), that we enounce below.

Lemma A.6 (Lemma 3.3 from Huang et al. (2013))**.** *Suppose that Assumption 2.2.(i) is verified. Let* $\boldsymbol{\dot{l}_n^*(\beta)}$ *be the gradient of the* $l_n^*(\boldsymbol{\beta})$ *defined by (4). Then, for all* $C_0 > 0$,

$$\mathbb{P}\left(|\boldsymbol{\dot{l}_n^*(\beta_0)}|_\infty > C_0 Bx, \sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}N_i(t) \leq C_0^2 n\right) \leq 2p\mathrm{e}^{-nx^2/2}. \tag{45}$$

In particular, if $\max_{i\leq n} N_i(\tau) \leq 1$, *then* $\mathbb{P}(|\boldsymbol{\dot{l}_n^*(\beta_0)}|_\infty > Bx) \leq 2p\mathrm{e}^{-nx^2/2}$.

Before proving the lemma that is in interest, we recall the Bernstein Inequality for martingales (see van de Geer (1995)).

Lemma A.7 (Lemma 2.1 from van de Geer (1995))**.** *Let $\{M_t\}_{t\geq 0}$ be a locally square integrable martingale w.r.t. the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Denote the predictable variation of $\{M_t\}$ by $V_t = \langle M, M\rangle_t$, $t \geq 0$, and its jumps by $\Delta M_t = M_t - M_{t^-}$. Suppose that $|\Delta M(t)| \leq K$ for all $t > 0$ and some $0 \leq K < \infty$. Then for each $a > 0$, $b > 0$,*

$$\mathbb{P}(M_t \geq a \text{ and } V_t \leq b^2 \text{ for some } t\ ) \leq \exp\left[-\frac{a^2}{2(aK+b^2)}\right].$$

From Lemma A.6, to prove Lemma A.2, it remains to control

$$\mathbb{P}\left(\sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}N_i(t) > C_0^2 n\right),$$

Using the Doob-Meyer decomposition and since,

$$\sum_{i=1}^n \int_0^\tau Y_i(t)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t \leq n\tau||\alpha_0||_{\infty,\tau}\mathrm{e}^{B|\boldsymbol{\beta_0}|_1},$$

we obtain for $C_0 > \sqrt{\tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]}$,

$$\mathbb{P}\left(\sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}N_i(t) > C_0^2 n\right) \leq \mathbb{P}\left(\sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}M_i(t) > C_0^2 n - n\tau||\alpha_0||_{\infty,\tau}\mathrm{e}^{B|\boldsymbol{\beta_0}|_1}\right).$$

Then, we apply Lemma A.7 to the martingale $\sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}M_i(t)$, with $K = 1$ and

$$V_t = \mathbb{E}\Big[\sum_{i=1}^n \int_0^\tau Y_i^2(t)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\Big] \leq ||\alpha_0||_{\infty,\tau}\tau\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]n.$$

We obtain

$$\mathbb{P}\left(\sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}M_i(t) > C_0^2 n - n\tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]\right)$$
$$\leq \exp\left(-\frac{n(C_0^2 - \tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}])^2}{2C_0^2}\right).$$

Finally, we get

$$\mathbb{P}\Big(|\boldsymbol{\dot{l}_n^*(\beta_0)}|_\infty > C_0 B x\Big) \leq 2p\mathrm{e}^{-nx^2/2} + \exp\left(-\frac{n}{2C_0^2}(C_0^2 - \tau||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}])\right).$$

Taking $x = \sqrt{2\log(n^k p)/n}$, there exists a constant $c > 0$ depending on τ, $||\alpha_0||_{\infty,\tau}$ and $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$ such that

$$\mathbb{P}((\widetilde{\Omega}_H^k)^c) \leq cn^{-k},$$

which leads to the expected result of Lemma A.2. □

References

O. 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.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *In Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

P. K. Andersen, Ø. Borgan, R. D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0. URL http://dx.doi.org/10.1007/978-1-4612-4348-9.

Y. Baraud. A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli*, 16(4):1064–1085, 2010.

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

K. Bertin, E. Le Pennec, and V. Rivoirard. Adaptive Dantzig density estimation. *Annales de l'IHP, Probabilités et Statistiques*, 47(1):pp. 43–74, 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):pp. 1705–1732, 2009. ISSN 0090-5364. URL http://dx.doi.org/10.1214/08-AOS620.

L. Birgé and P. Massart. *From model selection to adaptive estimation*. Springer, 1997.

J. Bradic, Fan, J., and J. Jiang. Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):pp. 3092–3120, 2012.

J. Bradic and R. Song. Gaussian Oracle Inequalities for Structured Selection in Non-Parametric Cox Model. *arXiv preprint arXiv:1207.4510*, 2012.

E. Brunel and F. Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā: The Indian Journal of Statistics*, pages 441–475, 2005.

E Brunel, F Comte, and A. Guilloux. Nonparametric density estimation in presence of bias and censoring. *test*, 18(1):166–194, 2009.

E. Brunel, F. Comte, and C. Lacour. Minimax estimation of the conditional cumulative distribution function. *Sankhya A*, 72(2):293–330, 2010.

P. Bühlmann and S. van de Geer. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:pp. 1360–1392, 2009. ISSN 1935-7524. URL http://dx.doi.org/10.1214/09-EJS506.

F. Bunea, A. B. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:pp. 169–194, 2007a. ISSN 1935-7524. URL http://dx.doi.org/10.1214/07-EJS008.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via l1 penalized least squares. In *Proceedings of the 19th annual conference on Learning Theory*, COLT'06, pages 379–391, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35294-5, 978-3-540-35294-5. URL http://dx.doi.org/10.1007/11776420_29.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007b.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparse density estimation with l1 penalties. In *Learning theory*, pages 530–543. Springer, 2007c.

F. Comte, S. Gaïffas, and A. 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.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. (Methodological)*, 34:pp. 187–220, 1972. ISSN 0035-9246.

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

T.R. Fleming and D.P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

G. Grégoire. Least squares cross-validation for counting process intensities. *Scandinavian journal of statistics*, pages 343–360, 1993.

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

A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, pages 681–712, 2000.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, 28(5):1356–1378, 2000. URL http://dx.doi.org/10.1214/aos/1015957397.

S. Kong and B. Nan. Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Arxiv preprint arXiv:1204.1992*, 2012.

F. Letué. *Modèle de Cox : estimation par sélection de modèle et modèle de chocs bivarié*. PhD thesis, Université de Paris Sud, UFR scientifique d'Orsay, 2000.

R. Sh. Liptser and A. N. Shiryayev. *Theory of martingales*, volume 49 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989. ISBN 0-7923-0395-4. URL http://dx.doi.org/10.1007/978-94-009-2438-3. Translated from the Russian by K. Dzjaparidze [Kacha Dzhaparidze].

C.L. Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):pp. 1436–1462, 2006.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009.

A. Nemirovski. Topics in nonparametric statistics. *Ecole d'Ete de Probabilites de Saint-Flour XXVIII, 1998*, 28:85, 2000.

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

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466, 1983b.

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

M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563, 1996. ISSN 0020-9910. URL http://dx.doi.org/10.1007/s002220050108.

M. Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. ISBN 3-540-24518-9. Upper and lower bounds of stochastic processes.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN.

R. Tibshirani. The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):pp. 385–395, 1997. ISSN 1097-0258. URL http://dx.doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

S. van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *The Annals of Statistics*, 23(5):pp. 1779–1801, 1995. ISSN 00905364. URL http://www.jstor.org/stable/2242545.

S. van de Geer. The deterministic lasso. Rapport technique, ETH Zürich, Switzerland, Available at http://stat.ethz.ch/research/publ archive/2007/140., 2007.

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

N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

C.H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):pp. 1567–1594, 2008.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.