Estimation for counting processes with high-dimensional covariates

Sarah Lemler

Université d'Évry Val d'Essonne
Laboratoire de Mathématiques et Modélisation d'Évry
École doctorale 423 : des Génomes Aux Organismes

Thèse

présentée en première version en vue d'obtenir le grade de

Docteur de l'Université d'Évry Val d'Essonne

Spécialité : Mathématiques

par

Sarah Lemler

Estimation for counting processes with high-dimensional covariates

Thèse soutenue le 9 décembre 2014 devant le jury composé de :

Fabienne COMTE	Université Paris Descartes	(Examinatrice)
Jean-Yves DAUXOIS	Université de Toulouse	(Examinateur)
Cécile DUROT	Université Paris Ouest Nanterre la Défense	(Rapporteure)
Agathe GUILLOUX	Université Pierre et Marie Curie	(Directrice)
Sylvie HUET	INRA unité MIAGE	(Examinatrice)
Sophie LAMBERT-LACROIX	Université de Grenoble	(Rapporteure)
Marie-Luce TAUPIN	Université d'Évry Val d'Essonne	(Directrice)

Laboratoire de Mathématiques et Modélisation d'Évry (LaMME)
Université d'Évry Val d'Essonne (UEVE)
Institut de Biologie Génétique et BioInformatique (IBGBI)
4ème étage
23 boulevard de France, 91 037 Évry

Laboratoire de Statistique Théorique et Appliquée (LSTA)
Université Pierre & Marie Curie (UPMC)
Tour 25, couloirs 15-25 & 15-16
2ème étage
4, place Jussieu, 75252 Paris Cedex 05

À mes parents

Remerciements

Ces quelques lignes, sans doute trop courtes, pour remercier toutes les personnes qui m'ont permis de mener à bien cette aventure.

Mes premiers remerciements vont à mes deux directrices de thèse Agathe et Marie-Luce. J'ai pu mesurer au cours de ces trois ans et demi la chance que j'ai de vous avoir pour directrices de thèse et ces quelques lignes ne suffiront pas à vous exprimer toute ma gratitude. Marie-Luce, c'est d'abord à toi que je dois d'être là. Merci de m'avoir fait découvrir l'analyse de survie, un champ des mathématiques dont j'ignorais l'existence et dont les applications à la médecine m'ont convaincue de l'importance de la recherche théorique en mathématiques dans un domaine si essentiel. Merci surtout d'avoir cru en moi et de m'avoir dit cette seule phrase « si tu cherches un stage de M2 ou une thèse n'hésite pas à me contacter », qui a modifié mon parcours et m'a propulsée dans une sphère que je n'aurais jamais envisagé être en mesure de rejoindre. Je tiens aussi à te remercier pour ta réactivité à mes sollicitations, ta patience et ton soutien dans les moments de doute. Agathe, merci d'avoir suivi de si près tout mon travail pendant ces trois ans et demi et d'avoir partagé avec moi ta grande culture scientifique. Merci pour ta grande disponibilité, ta patience et tes encouragements. Je te remercie aussi de m'avoir poussée à programmer en R alors même que je faisais tout pour reculer le moment de m'y mettre et de m'avoir ainsi prouvé que j'en étais capable. J'espère un jour avoir ta rapidité à comprendre un problème et à trouver une solution pour le résoudre. Merci à toutes les deux pour la confiance que vous m'avez témoignée et surtout, pour tout le temps que vous m'avez consacré tout au long de ma thèse et particulièrement ces derniers mois de rédaction (parfois même jusqu'à 22h un vendredi ou un samedi soir pour me rassurer sur mes doutes ou répondre à mes questions). À toutes les deux, un immense merci et j'espère que notre histoire continuera (avec moins de sollicitations, c'est promis) !

Merci à mes deux rapporteures, Cécile Durot et Sophie Lambert-Lacroix, pour m'avoir fait l'honneur de rapporter ma thèse. Je vous remercie sincèrement toutes les deux pour le temps que vous avez consacré à mon travail et pour les commentaires enrichissants et précieux pour la suite. Cécile Durot, je tiens à vous exprimer toute ma gratitude pour la relecture si attentive et minutieuse que vous avez faite de ma thèse et pour m'avoir ainsi permis d'améliorer mon travail grâce à vos remarques pertinentes et à vos conseils. Sophie Lambert-Lacroix, merci pour votre relecture bienveillante et circonstanciée.

Merci à Fabienne Comte, Sylvie Huet et Jean-Yves Dauxois d'avoir si aimablement et rapidement accepté de faire partie de mon jury de thèse. Merci Fabienne pour m'avoir fait découvrir les statistiques non-paramétriques en Master 2 et surtout pour l'intérêt que vous avez continué à porter à mon travail pendant ma thèse. Merci pour toutes les fois où vous avez pris le temps de répondre à mes sollicitations et merci aussi de m'avoir conseillée lorsque j'avais des doutes. Sylvie, merci d'avoir suivi mon travail pendant ces trois ans et demi, à l'occasion d'un séminaire SSB, de mon comité de thèse ou d'un exposé à Fréjus. Merci pour

votre regard éclairant et bienveillant sur mon travail. Merci Jean-Yves d'avoir été le premier à accepter de faire partie de ce jury. Je suis ravie d'avoir fait votre connaissance à cette occasion.

Je tiens à remercier tous les membres de l'équipe Statistique et Génome. Quand je suis arrivée au laboratoire pour mon stage de Master 2, j'y ai été extrêmement bien accueillie, et si pendant ces trois années de thèse, je m'y suis sentie aussi bien, vous y êtes tous pour beaucoup. Merci Cyril, j'ai eu plaisir à partager avec toi les TDs de L2 bio pendant ces quatre années. Je te remercie aussi pour ton oreille attentive, tes conseils et pour m'avoir remonté le moral quand j'en ai eu besoin. Merci Maurice pour avoir tant de fois partagé mes plaintes lorsque le RER D faisait des siennes et égayé mes trajets à force de bons plans à Paris, de discussions sur les films, les livres, les expositions, les applications... Merci Catherine pour tes conseils, pour avoir toujours pris le temps de répondre à mes questions et pour m'avoir fait découvrir de grands chorégraphes contemporains ! Merci Carène pour ta bonne humeur et tes petites attentions qui font toujours plaisir, merci Yolande pour ton rire communicatif et vive la zumba ! Claudine, si tu n'existais pas il faudrait t'inventer, alors merci de mettre autant d'ambiance au labo ! Merci Michèle pour ton accueil si chaleureux à mon arrivée au laboratoire, pour cette journée au Château de Fontainebleau et à Barbizan et pour les nombreuses discussions. Merci aussi à Anne-Sophie pour ton cours sur le modèle linéaire et surtout pour tes encouragements dans les moments difficiles, à Bernard pour m'avoir aussi bien préparée à l'audition pour obtenir ma bourse de thèse, à Christophe pour toujours avoir fait en sorte que je puisse participer à toutes les conférences qui m'intéressaient, à Pierre pour ton calme à toute épreuve qui fait du bien, à Julien en souvenir des TDs de L1 pendant ma première année et à Guillem. Merci Valérie pour ton aide pour l'organisation de la soutenance. Merci à Camille et Marine pour avoir guidé mes pas dans le monde des doctorants. Marius, mon grand frère de thèse, merci de m'avoir supportée pendant ces trois années et demi, merci pour les nombreuses discussions de mathématiques et autres et pour ton soutien sans faille, merci aussi à Van Hanh pour son sourire si communicatif. Merci à Alia vive Maths En Jeans, Edith pour nos nombreuses discussions autour de la gym suédoise, la zumba ou toute autre chose, Margot pour tes encouragements et nos discussions pendant les trajets de RER, Morgane pour nos nombreuses discussions dans la cuisine, dans les couloirs ou dans le RER, merci pour toutes ces pauses bavardages qui font du bien, Virginie ma nouvelle co-bureau et compagnonne de Zumba, merci de m'avoir fait découvrir les codes RGB qui ouvrent un champ des possibles infini sur beamer, Jean-Michel pour toutes les fois où tu as voulu me faire sursauter de peur et où tu as échoué, mais aussi pour toutes les discussions à propos de maths ou pas, Benjamin que j'ai eu plaisir à retrouver au laboratoire deux ans après notre Master 2 et Quentin, merci à vous tous pour cette bonne entente !

Merci également à tous les thésards de Paris 6 et en particulier à mes compagnons de bureau Assia, Roxane, Baptiste et Matthieu ainsi qu'à Cécile, Benjamin, Erwan et Mokhtar pour m'avoir si gentiment intégrée à leur petite équipe !

Merci à Adeline Samson pour m'avoir si bien dirigée vers Marie-Luce lorsque je l'ai contactée pour mon stage de Magistère. Un grand merci également à Patricia Reynaud-Bouret et Franck Picard pour l'intérêt qu'ils m'ont porté au colloque Jeunes Probabilistes et Statisticiens à Forges-les-Eaux. Merci à tous les deux pour tous vos conseils et encouragements. Merci Patricia pour tes suggestions éclairantes sur mon travail.

Merci à tous les jeunes statisticiens que j'ai croisés pendant ces trois années. Gaëlle, je suis ravie d'avoir fait ta connaissance à Fréjus juste avant de commencer ma thèse, merci de

m'avoir aidée à comprendre comment faire mes voeux inter-académiques, merci de toujours répondre à mes questions de manière détaillée et claire, enfin, ta thèse, si bien rédigée m'a bien aidée à comprendre la méthode de Goldenshluger et Lepski. Merci aussi à Angelina et à tous mes compères de Saint-Flour avec qui j'ai vraiment passé de bons moments : Carole, Laure, Lucie, Ilaria, Mélisande, Andrés, Benjamin, Clément et Sébastien.

Je tiens à présent à remercier mes amis pour tous ces bons moments partagés au quotidien, en week end, en vacances ou lors de notre BNH annuel. Tout d'abord mes deux inconditionnels soutiens, Flo et Marion, merci d'être toujours là dans les bons moments, mais aussi dans les moments plus difficiles. Merci à vous deux de m'avoir soutenue pendant la période de rédaction, motivée avec vos messages et de ne jamais m'avoir reproché de ne pas avoir de temps pour vous pendant toute cette période. Merci surtout à chacune d'entre vous pour les innombrables bons moments partagés et nos interminables discussions ! Merci aussi à Marion pour toutes ces années d'amitié depuis le primaire et pour tes mails plein d'attention tout au long de la rédaction. Emilie, merci pour ta générosité, pour les soirées jeux et toutes nos discussions sur nos thèses, mais pas seulement, loin sans faut. Jo, merci de prendre toujours régulièrement de mes nouvelles et de toujours me remotiver dans mes moments de doute avec des smileys d'encouragement, toi qui sais si bien ce que c'est que de faire une thèse. Je remercie également Eleo (pour nos longues conversations téléphoniques moins fréquentes qu'au lycée, mais avec toujours autant de choses à se dire), Pauline, Alix, Laure (pour toutes tes invitations si conviviales pour tes anniversaires ou les fêtes), Alex, Hugo, Pierre, Vincent, Guillaume, Louis-Marie. Merci enfin à Laura, mon amie d'enfance qui, même si elle est loin, reste toujours présente.

Mes derniers remerciements et non les moindres vont à ma famille. Merci d'abord à mes grands-parents qui ont toujours été fiers de leurs petits-enfants et qui ne s'en sont jamais cachés : merci de croire autant en moi. Léna, ma sis', quelle chance précieuse j'ai que tu sois là ! Je ne pourrai jamais assez te dire à quel point ton soutien compte pour moi. Merci pour toutes tes attentions, pour les ondes positives que tu m'envoies régulièrement et pour bien plus encore ! Surtout, merci pour tous nos bons moments à deux. Enfin, mes plus profonds remerciements sont pour vous Maman et Papa. Ma reconnaissance pour tout ce que vous m'apportez va bien au-delà de ces quelques lignes et je ne saurais exprimer tout ce que je vous dois. En voici donc une infime partie. Je vous remercie tous les deux pour votre inébranlable soutien en toute épreuve et vos conseils avisés. Merci de m'avoir toujours soutenue dans mes choix, de m'avoir poussée à toujours viser plus loin, plus haut, merci de la confiance que vous avez en ma réussite. Merci aussi à tous les deux d'avoir pris le temps de relire les parties en français de ma thèse pour corriger les fautes d'orthographe. Maman, merci pour tout le temps que tu me consacres (heureusement que les forfaits téléphoniques sont illimités !), merci de toujours me simplifier les aspects matériels, merci pour tous les bons petits plats que tu me prépares pour ma semaine lorsque je rentre pour le week end, merci de prendre autant soin de moi. Papa, je me rappelle lorsque tu me faisais faire des « problèmes » au primaire et que tu m'as initiée aux logigrammes, je détestais ça et rien ne laissait alors présager que j'allais faire des maths plus tard ! Merci de pouvoir toujours aussi compter sur toi, merci pour tes conseils et merci surtout de me prouver très souvent qu'il faut que j'aie plus confiance en moi ! Merci pour cette chance inestimable que vous m'offrez, tous les deux, de pouvoir autant compter sur vous !

Organisation de la thèse

Cette thèse est consacrée à l'estimation de la fonction de risque d'un processus de comptage, en présence d'un vecteur de covariables de grande dimension. Nous proposons différentes procédures d'estimation. Pour chacun des estimateurs obtenus, nous établissons des inégalités oracles non-asymptotiques assurant leurs performances théoriques. Enfin, une implémentation pratique des différentes procédures est proposée. Cette thèse comporte quatre chapitres (hors introduction) répartis en deux parties.

- Le **Chapitre 1** est une introduction générale : nous y présentons le contexte, une description des méthodes d'estimation utilisées, les démarches de preuves et enfin nos contributions. Les méthodes d'estimation et les démarches de preuves sont dans un premier temps, décrites dans des modèles plus simples, le modèle de régression additive et le modèle d'estimation de densité, en mettant en évidence les similitudes et différence entre les méthodes. Puis nous expliquons notre cheminement dans ce travail et nos contributions dans chacune des deux parties de la thèse.
- La **Partie I** est consacrée à l'estimation non-paramétrique de l'intensité d'un processus de comptage, en présence d'un vecteur de covariables en grande dimension. Cette partie est constituée du **Chapitre 2**, suivie d'une annexe.
  - Dans le **Chapitre 2**, nous proposons d'estimer l'intensité par un estimateur de l'intensité dans un modèle de Cox, choisi automatiquement à partir des données. Cette estimation est basée sur une procédure Lasso appliquée simultanément aux deux fonctions intervenant dans le modèle de Cox : le risque de base et le risque relatif. En annexe, nous faisons une présentation mettant en perspective les similitudes et les différences entre l'étude du Lasso en régression additive et l'étude du Lasso dans notre modèle de processus de comptage.
- La **Partie II** est consacrée à l'estimation de la fonction du risque de base dans le modèle de Cox, en présence d'un vecteur de covariables en grande dimension. Cette partie est constituée d'une introduction commune et de trois chapitres. Dans cette partie, nous proposons deux procédures d'estimation du risque de base, établissons des inégalités oracles non asymptotiques, et enfin procédons à une comparaison pratique des deux procédures d'estimation.
  - L'**Introduction**, qui est commune aux **Chapitres 3**, **4** et **5**, présente le cadre de travail commun à chacun de ces chapitres.

— Dans le **Chapitre 3**, nous proposons d'estimer le risque de base avec une procédure de sélection de modèles.

— Dans le **Chapitre 4**, nous proposons d'estimer le risque de base en utilisant un estimateur à noyau dont le paramètre de lissage est choisi de façon adaptative par la méthode de Goldenshluger et Lepski.

— Enfin, au **Chapitre 5** nous menons une étude par simulations. Cette étude présente une comparaison des performances pratiques des procédures d'estimation du risque de base proposées aux deux chapitres précédents. Nous terminons cette étude, en appliquant ces procédures d'estimation à un jeu de données réelles sur le cancer du sein.

Nous avons aussi comparé les méthodes de sélection de modèles et de sélection de fenêtres dans notre modèle, ainsi que les résultats obtenus pour ces estimateurs aux **Chapitres 3** et **4**.

Chaque chapitre est précédé d'un court résumé permettant de le situer dans son contexte.

Table des matières

Remerciements v
Organisation de la thèse ix

Table des matières **xi**

Liste des figures **xiii**

Liste des tableaux **xiv**

1 Introduction **1**
1.1 Cadre de travail 2
1.2 Généralités sur les méthodes d'estimation et les résultats attendus . . . 11
1.3 Cheminement et principaux résultats 32

I Estimation of the non-parametric intensity of a counting process by a Cox model with high-dimensional covariates **49**

2 High-dimensional estimation of counting process intensities **51**
2.1 Introduction 53
2.2 Estimation procedure 56
2.3 Oracle inequalities for the Cox model for a known baseline function . . 60
2.4 Oracle inequalities for general intensity 65
2.5 An empirical Bernstein's inequality 68
2.6 Technical results 71
2.7 Proofs 73
Appendices 84
A Connection between the weighted norm and the Kullback divergence . . 84
B An other empirical Bernstein's inequality 87
C Weighted Lasso procedure in the specific case of the Cox model 92
D Detailed comparison with the additive regression model 97

II Estimation of the baseline function in the Cox model with high-dimensional covariates **105**

Introduction **107**

3 Adaptive estimation of the baseline hazard function in the Cox model by model selection **113**
3.1 Introduction . . . 115
3.2 Estimation procedure . . . 116
3.3 Non-asymptotic oracle inequalities . . . 123
3.4 Proofs . . . 125
Appendices . . . 144
A Prediction result on the Lasso estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta_0}$ for unbounded counting processes . . . 144

4 Adaptive kernel estimation of the baseline function in the Cox model 147
4.1 Introduction . . . 149
4.2 Estimation procedure . . . 150
4.3 Non-asymptotic bounds for kernel estimators . . . 155
4.4 Proofs . . . 158
Appendices . . . 172
A Technical lemma . . . 172
B Unbounded case . . . 173

5 Simulations **179**
5.1 Simulated data . . . 181
5.2 Step 1: Estimation of the regression parameter $\boldsymbol{\beta_0}$ . . . 184
5.3 Step 2: Estimation of the baseline function α_0 . . . 187
5.4 Measuring the quality of the estimators . . . 193
5.5 Results in the case of simulated data . . . 194
5.6 Application to a real dataset on breast cancer . . . 205
Appendices . . . 210
A The rectangle method . . . 210
B Calibration of the constants . . . 210
C Description of the real data . . . 212

Conclusion **215**

A Appendices **219**
A.1 Un peu de théorie sur les processus de comptage . . . 220
A.2 Construction de la vraisemblance de Cox . . . 224
A.3 Quelques inégalités de concentration . . . 227

Bibliographie **229**

Liste des figures

1.1 Courbes des fonctions de risque d'un individu pris dans les conditions standard et d'un individu ayant des prédispositions génétiques. 4
1.2 Illustration de la pénalisation ℓ_1 13
1.3 Courbes du vrai risque de base (en rouge), de l'estimateur à noyau obtenu par validation croisée (en vert), de l'estimateur par minimum de contraste pénalisé (en bleu) et de l'estimateur à noyau adaptatif (en noir). Les courbes ont été obtenues pour : $n = 500$, $p = \lfloor\sqrt{n}\rfloor$, $\alpha_0 \sim \ln\mathcal{N}(1/4, 0)$, 20% de censure 47

1.0.1 Plots of the baseline hazard function for different parameters of a log-normal distribution (left) and of a Weibull distribution (right). 183
5.1.1 Median, first and third quartiles of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ (on the left) and for $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_2$ (on the right), $\hat{\boldsymbol{\beta}}$ being either $\hat{\boldsymbol{\beta}}^{ML}$, either $\hat{\boldsymbol{\beta}}^{Lasso}$ or $\hat{\boldsymbol{\beta}}^{AdapL}$. 196
5.2.1 Plots of the true baseline function (in red), the cross-validated kernel estimator (in green), the penalized contrast estimator (in blue) and the adaptive kernel estimator (in black). The plots have been obtain for: $n = 500$, $p = \lfloor\sqrt{n}\rfloor$, $\alpha_0 \sim \ln\mathcal{N}(1/4, 0)$, $d = 4.5$ 199
5.2.2 Median, first and third quartiles of the random MISEs of the kernel estimator with a bandwidth selected by cross-validation (5.2.2a), of the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method (5.2.2b), of the penalized contrast estimator in the case of histograms (5.2.2c) and of the penalized contrast estimator in the trigonometric case (5.2.2d), with the adaptive Lasso estimator of the regression parameter in the case of a Weibull distribution $\mathcal{W}(1.5, 1)$. 200
6.2.1 Kernel estimators with a bandwidth selected either by cross-validation (in green) or by the Goldenshluger and Lepski method (in black) and by model selection estimator in the histogram basis (in blue). The first column is associated to the group of untreated patients and the second column corresponds to the group of Tamoxifen patients for $p = 10$ (first line), $p = 100$ (second line) and $p = 1000$ (third line). 207
6.2.2 Kernel estimator with a bandwidth selected either by cross-validation (in green) or by the Goldenshluger and Lepski method (in black) and model selection estimator in the trigonometric basis (in blue). The first column is associated to the group of untreated patients and the second column corresponds to the group of Tamoxifen patients for $p = 10$ (first line), $p = 100$ (second line) and $p = 1000$ (third line). 208

Liste des tableaux

1.1 Tableau comparatif pour la procédure Lasso dans le modèle de régression additif et dans le modèle à intensité multiplicative d'Aalen non-paramétrique. 37
1.2 Tableau comparatif des méthodes de sélection de modèles et de Goldenshluger et Lepski pour l'estimation du risque de base α_0 en grande dimension. 43

5.1.1 ℓ_1-norm and ℓ_2 norm of estimation errors of the estimators $\hat{\boldsymbol{\beta}}^{ML}$, $\hat{\boldsymbol{\beta}}^{Lasso}$ and $\hat{\boldsymbol{\beta}}^{AdapL}$ for $n = 200$ and $p = 5, 15, 200$, and $n = 500$ and $p = 5, 22, 500$. 195
5.1.2 ℓ_1-norm and ℓ_2 norm of prediction errors of the estimators $\hat{\boldsymbol{\beta}}^{ML}$, $\hat{\boldsymbol{\beta}}^{Lasso}$ and $\hat{\boldsymbol{\beta}}^{AdapL}$ for $n = 200$ and $p = 5, 15, 200$, and $n = 500$ and $p = 5, 22, 500$. 197
5.1.3 Specificities (SPEC) and sensitivities (SENS) for the Lasso and the adaptive Lasso (AdapL) for $n = 200$ and $p = 200$. 198
5.1.4 ℓ_1-norm of the estimation errors on the support of $\boldsymbol{\beta_0}$ and its complementary for the maximum Cox partial likelihood and the Lasso estimators for $n = 200$ and $p = 15$. 198
5.2.1 Random empirical MISE of the kernel estimator with a bandwidth selected by the Goldenshluger et Lepski method (5.2.1a) and of the penalized contrast estimator in a histogram basis (5.2.1b) obtained from an adaptive Lasso estimator of the regression parameter given two rate of censoring: 20% and 50% of censoring. 202
5.2.2 Random empirical MISE for the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method (5.2.2a) and for the penalized contrast estimator in a histogram basis (5.2.2b), with an adaptive Lasso estimator of the regression parameter, for three different Weibull distributions of the survival times. 203
5.2.3 Random empirical MISE of the kernel estimators with a bandwidth selected by cross-validation (KernelCV) and by the Goldenshluger and Lepski method (KernelGL) and of the penalized contrast estimators in a histogram basis (MShist) and in a trigonometric basis (MStrigo) for an adaptive Lasso estimator of the regression parameter and a Weibull distributions $\mathcal{W}(1.5, 1)$ for the survival times. 204
6.1.1 Selected variables in the two groups of patients. We precise the name of the clinical selected variables and only give the number of selected genes (e.g.=gene expression). 206
C.0.1 Table of the clinical information 213

Chapitre 1

Introduction

Sommaire

1.1 Cadre de travail 2
1.1.1 Problématique générale 2
1.1.2 Formalisation 5
1.1.3 Estimation des paramètres du modèle de Cox 6
1.1.4 Et quand la dimension augmente... 8
1.2 Généralités sur les méthodes d'estimation et les résultats attendus 11
1.2.1 La procédure Lasso 11
1.2.2 La sélection de modèles 18
1.2.3 Estimateurs à noyaux et sélection de fenêtres 23
1.2.4 Bibliographie en analyse de survie 29
1.3 Cheminement et principaux résultats 32
1.3.1 Partie I : Estimation de l'intensité complète d'un processus de comptage par un modèle de Cox 33
1.3.2 Partie II : Estimation adaptative de la fonction de base dans le modèle de Cox 40

1.1 Cadre de travail

1.1.1 Problématique générale

En analyse de données de survie, deux questions sont souvent classiquement posées. La première est de déterminer les facteurs qui influencent la durée de survie. La seconde est de prédire pour un individu donné, la durée de survie, au vu des conditions ou facteurs auxquels il est soumis. Dans cette thèse, nous portons un intérêt tout particulier à la deuxième question, portant sur la prédiction. Cette question nous intéresse tout particulièrement dans un contexte dit de grande dimension.

Pour plus de clarté, commençons par définir les notions qui interviennent en analyse de survie. La durée de survie, notée T, est la durée entre un instant initial et la survenue d'un événement donné, appelé événement terminal. Un exemple classique de durée de survie en épidémiologie est la durée entre le début de la prise d'un traitement et le décès du patient, qui est dans ce cas l'événement terminal. Malgré la terminologie, un événement terminal n'est pas forcément le décès de l'individu, il peut aussi définir la rechute ou la rémission par exemple. Pour plus de clarté et sans perte de généralité, nous utiliserons dans la suite le terme de décès plutôt que d'événement terminal. L'analyse de données de survie est alors l'étude statistique de données avec pour variable d'intérêt la durée de survie T.

Une des particularités des durées de survie est qu'elles ne sont pas toujours entièrement observées, on parle de données censurées. En effet, il se peut que le décès d'un individu ne soit pas observé à la fin de l'étude, que l'individu sorte de l'étude avant qu'on ait pu observer sa durée de survie ou encore que l'individu décède d'une autre cause que de la maladie considérée. On observe alors un temps inférieur à la durée de survie T : c'est la censure aléatoire à droite. D'autres types de difficultés d'observation existent tels que les troncatures par exemple. Ces données censurées demandent un traitement particulier et doivent être prises en compte dans l'analyse statistique des données de survie.

Dans le contexte de la première question évoquée, à savoir les facteurs qui influencent la durée de survie, on peut s'intéresser par exemple à l'effet d'un traitement. Ces facteurs, appelés aussi variables explicatives ou covariables, sont observées naturellement, cliniquement (âge, sexe,...) ou encore à l'aide de nouvelles technologies telles que les puces à ADN, pour mesurer des niveaux d'expression de gènes. Les covariables peuvent être en grand nombre, on parle de grande dimension. Pour illustrer ce cadre de grande dimension, considérons l'exemple sur lequel nous avons travaillé dans cette thèse, issu de l'étude de Loi et al. (2007). Ces données concernent 414 individus atteints d'un cancer du sein. Ces patients sont divisés en deux groupes : les patients traités et les patients non-traités. Pour chaque patient, nous disposons de quelques variables cliniques (âge, taille de la tumeur,...) et de 44 928 niveaux d'expression de

gènes. Nous sommes dans cet exemple dans le cas dit de l'ultra grande dimension que nous détaillerons par la suite.

Pour répondre à la deuxième question, qui sera au coeur de cette thèse, on modélise la relation entre la durée de survie T et les p covariables $Z_1, ..., Z_p$. Le modèle de Cox, introduit par Cox (1972), est un modèle classique pour l'étude des données de survie. Dans ce modèle, la fonction de risque, qui modélise la relation entre la durée de survie et les covariables, a la forme suivante :

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \exp(\boldsymbol{\beta_0^T Z}), \tag{1.1}$$

où $\boldsymbol{Z} = (Z_1, ..., Z_p)^T \in \mathbb{R}^p$ est un vecteur de covariables p-dimensionnel. C'est un *modèle à risque proportionnel*, i.e. le rapport de la fonction de risque (1.1) de deux individus i et j ne dépend pas du temps :

$$\forall t \in [0, \tau], \quad \frac{\lambda_0(t, \boldsymbol{Z_i})}{\lambda_0(t, \boldsymbol{Z_j})} = \frac{\exp(\boldsymbol{\beta_0^T Z_i})}{\exp(\boldsymbol{\beta_0^T Z_j})} = \exp(\boldsymbol{\beta_0^T (Z_i - Z_j)}).$$

De manière générale, la fonction de risque est définie par

$$\lambda_0(s, \boldsymbol{Z}) = \frac{f(t|\boldsymbol{Z})}{S(t|\boldsymbol{Z})} = \lim_{h \to 0} \frac{1}{h} \mathbb{P}(t \leq T < t + h | T \geq t, \boldsymbol{Z}), \tag{1.2}$$

où f est la densité de T conditionnellement au vecteur de covariables $\boldsymbol{Z} = (Z_1, ..., Z_p) \in \mathbb{R}^p$ et pour tout $t \geq 0$, $S(t|\boldsymbol{Z}) = \mathbb{P}(T > t|\boldsymbol{Z})$ sa fonction de survie conditionnelle. De la même manière que la densité ou la fonction de survie, la fonction de risque caractérise la distribution de la durée de survie T. La fonction de risque $\lambda_0(t, z)$ peut être interprétée comme le risque instantané de décès au temps t pour un individu pour lequel $\boldsymbol{Z} = z$, sachant qu'il est vivant. Ainsi, cette fonction a un intérêt, en particulier lorsqu'on cherche à prédire la durée de survie et elle se généralise de plus à des contextes plus généraux que la survie, auxquels nous nous sommes aussi intéressés dans ce travail.

Dans le modèle de Cox (1.1) deux paramètres sont inconnus : le paramètre de régression $\boldsymbol{\beta_0} \in \mathbb{R}^p$ associé aux covariables et la fonction du temps α_0, appelée risque de base. La plupart des études s'intéressent à l'estimation du paramètre de régression $\boldsymbol{\beta_0}$ (voir Section 1.1.3), ce qui permet de répondre à la première question, concernant les facteurs pronostiques, mais pas à la deuxième, relative à la prédiction de la durée de survie. En effet, la probabilité pour que la durée de survie T d'un individu soit supérieure à cinq ans conditionnellement aux covariables est donnée par

$$\mathbb{P}(T > 60 \text{ mois} | \boldsymbol{Z}) = \exp\left(- \int_0^{60} \alpha_0(s) \mathrm{e}^{\boldsymbol{\beta_0^T Z}} \mathrm{d}s \right).$$

Dans cette formule, les deux paramètres inconnus du modèles de Cox (1.1) apparaissent. Lorsque nous cherchons à prédire la durée de survie d'un individu pour une

maladie donnée, nous avons donc besoin de connaître la fonction de risque complète. De manière plus spécifique, connaître la fonction d'un individu pour une maladie telle que le cancer du sein par exemple, permet de déterminer à quel moment commencer les examens de dépistage. En effet, si l'on compare les courbes des fonctions de risque d'un individu pris dans les conditions standard et d'un individu qui a une prédisposition génétique au cancer du sein (voir Figure 1.1), la fonction de risque augmente beaucoup plus vite pour l'individu à risque familial et un suivi plus précoce est alors préconisé pour cet individu. Pour répondre à la deuxième question, nous nous sommes donc intéressés à l'estimation de la fonction de risque λ_0.

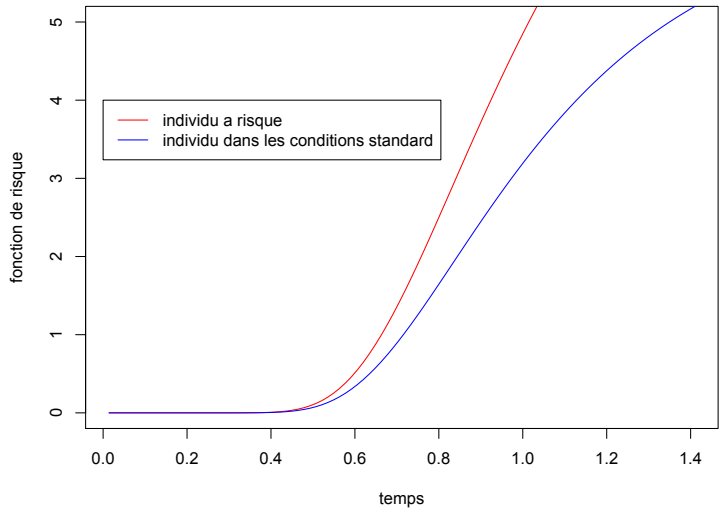


FIGURE 1.1 – *Courbes des fonctions de risque d'un individu pris dans les conditions standard et d'un individu ayant des prédispositions génétiques.*

Pour l'estimation de la fonction de risque complète en présence d'un grand nombre de covariables, nous avons considéré deux approches :

- ➢ la première approche consiste à estimer la fonction de risque complète par un modèle de Cox à l'aide d'une procédure d'estimation spécifique à la grande dimension appliquée simultanément aux deux paramètres du modèle de Cox,
- ➢ dans la seconde approche, nous avons considéré des procédures en deux étapes pour estimer les paramètres du modèle de Cox dans un cadre de covariables en grande dimension : nous avons d'abord estimé le paramètre de régression $\boldsymbol{\beta_0}$ par une procédure spécifique à la grande dimension, puis dans un deuxième temps, à partir de l'estimateur de $\boldsymbol{\beta_0}$, nous avons proposé des procédures d'estimation du risque de base α_0 non spécifiques à la grande dimension.

1.1.2 Formalisation

Nous utilisons dans la suite le formalisme général des processus de comptage. Pour $i = 1, ..., n$, nous définissons N_i un processus de comptage marqué [1] et Y_i un processus aléatoire à valeurs dans $[0, 1]$. Nous considérons l'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et la filtration $(\mathcal{F}_t)_{t\geq 0}$ définie par

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

où $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ est le vecteur aléatoire de covariables $\mathcal{F}_0$-mesurable de l'individu i. Notons Λ_i le compensateur du processus N_i par rapport à $(\mathcal{F}_t)_{t\geq 0}$, tel que $M_i = N_i - \Lambda_i$ soit une $(\mathcal{F}_t)_{t\geq 0}$-martingale. Nous supposons que le processus N_i satisfait le modèle à intensité multiplicative d'Aalen : pour tout $t \geq 0$,

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i}) Y_i(s) \mathrm{d}s, \tag{1.3}$$

où λ_0 est la fonction de risque (1.2) inconnue à estimer.

Nous observons les variables indépendantes et identiquement distribuées (i.i.d.) $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$, où $[0, \tau]$ est l'intervalle de temps entre le début et la fin de l'étude.

Ce cadre général, introduit par Aalen (1980), inclut plusieurs contextes tels que les données censurées, les processus de Poisson marqués et les processus de Markov. Nous renvoyons à Andersen et al. (1993) pour plus de détails.

Cas particulier des données censurées : Dans le cas spécifique de la censure à droite, nous définissons $(T_i)_{i=1,...,n}$ les durées de survie i.i.d. et $(C_i)_{i=1,...,n}$ les durées de censure i.i.d. de n individus. Nous observons $\{(X_i, \boldsymbol{Z_i}, \delta_i)\}_{i=1,...,n}$ où $X_i = \min(T_i, C_i)$ est la durée jusqu'à la survenue d'un évènement (décès ou censure), $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T$ est le vecteur de covariables et $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ est l'indicateur de censure. La durée de survie T_i est supposée indépendante de la durée de censure conditionnellement au vecteur de covariables $\boldsymbol{Z_i}$ pour $i = 1, ..., n$. Avec ces notations, les processus $(\mathcal{F}_t)$-adaptés Y_i et N_i sont respectivement définis par l'indicateur de présence à risque $Y_i(t) = \mathbb{1}_{\{X_i \geq t\}}$ et le processus de comptage $N_i(t) = \mathbb{1}_{\{X_i \leq t, \delta_i = 1\}}$ qui saute lorsque le ième individu décède.

Dans la Section 1.1 de l'introduction, nous considérons le cas particulier des données censurées pour simplifier, mais pour toute la suite nous nous plaçons dans le cadre plus général des processus de comptage décrit ci-dessus.

Dans ce cadre de travail, nous avons considéré plusieurs modèles de risque instantané λ_0.

1. Nous renvoyons à l'Annexe A.1 pour une définition d'un processus de comptage marqué et pour des rappels généraux sur les processus de comptage

Modèle 1 : Risque non-paramétrique

$$\text{aucune forme n'est imposée a priori sur } \lambda_0 : \mathbb{R}^+ \times \mathbb{R}^p \to \mathbb{R}. \tag{1.4}$$

Modèle 2 : Modèle de Cox

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \mathrm{e}^{f_0(\boldsymbol{Z_i})}, \tag{1.5}$$

où $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ est le vecteur de covariables p-dimensionnel de l'individu i. Dans ce modèle nous avons considéré :

- ▷ le modèle de Cox non-paramétrique : aucune forme a priori n'est imposée ni au risque de base α_0, ni à la fonction de régression $f_0 : \mathbb{R}^p \to \mathbb{R}$,
- ▷ le modèle de Cox semi-paramétrique : aucune forme a priori n'est imposée au risque de base, mais la fonction de régression est linéaire en les covariables $f_0(\boldsymbol{Z}) = \boldsymbol{\beta_0^T Z}$, où $\boldsymbol{\beta_0} \in \mathbb{R}^p$. Ce cas particulier correspond au modèle de Cox classique (1.1).

Le modèle de Cox non-paramétrique (1.5) a déjà été proposé par Hastie & Tibshirani (1986) pour une variable unidimensionnelle et Letué (2000) dans un travail en commun avec Castellan, l'a considéré dans le cas où $p < n$. Le modèle de Cox semi-paramétrique (1.1) a été considéré pour la première fois avec le formalisme des processus de comptage par Andersen & Gill (1982). Nous appellerons modèle de Cox non-paramétrique, le modèle (1.5) et simplement modèle de Cox, le modèle classique (1.1).

1.1.3 Estimation des paramètres du modèle de Cox

Chacun des deux paramètres du modèle de Cox a une interprétation propre. Le paramètre $\boldsymbol{\beta_0}$ du modèle de Cox traduit le poids des variables explicatives $Z_{i,1}, ..., Z_{i,p}$ de l'individu i. Lorsque le vecteur de covariables est nul pour l'individu i, la fonction de risque $\lambda_0(., \boldsymbol{Z_i})$ de cet individu est alors égale à $\alpha_0(.)$. Le risque de base $\alpha_0(.)$, comme son nom l'indique, représente donc le risque instantané de décès conditionnel d'un individu pris dans des conditions standard, avec les valeurs 0 (de référence) pour les covariables. Nous allons présenter les méthodes classiques d'estimation de ces deux paramètres lorsque $p < n$.

1. **Estimation du paramètre de régression $\boldsymbol{\beta_0}$:**

   Le paramètre de régression $\boldsymbol{\beta_0}$ est estimé en minimisant l'opposé de la pseudo-log-vraisemblance de Cox ou log-vraisemblance partielle de Cox introduite par Cox (1972) et définie par

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \log \left(\frac{e^{\boldsymbol{\beta^T Z_i}}}{S_n(t, \boldsymbol{\beta})} \right) \mathrm{d}N_i(t), \quad \text{avec } S_n(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{e}^{\boldsymbol{\beta^T Z_i}} Y_i(t) \tag{1.6}$$

et $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ le vecteur de covariables de l'individu i pour $i = 1, ..., n$. Nous renvoyons à l'Annexe A.2 pour une description de la construction de la log-vraisemblance partielle de Cox à partir de la définition de la log-vraisemblance pour les processus de comptage. La vraisemblance partielle de Cox ne fait pas intervenir le risque de base α_0. Elle permet donc d'estimer $\boldsymbol{\beta_0}$ sans avoir à connaître α_0. L'estimateur de $\boldsymbol{\beta_0}$ est alors défini par

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\{-l_n^*(\boldsymbol{\beta})\}. \tag{1.7}$$

Le problème (1.7) peut se réécrire sous forme matricielle. Si nous notons $\boldsymbol{U(\beta)}$ la fonction score, c'est-à-dire le vecteur $p \times 1$ des dérivées premières de $l_n^*(\boldsymbol{\beta})$, $\boldsymbol{\beta}$ est solution de l'équation $\boldsymbol{U(\beta)} = \vec{\boldsymbol{0}}$. Il y a en tout, p équations, une pour chacune des p variables. En général, il n'y a pas de solution explicite. En pratique, on utilise des algorithmes d'optimisation itératifs.

Andersen & Gill (1982) ont prouvé, en utilisant la théorie des processus de comptage, la consistance et la normalité asymptotique de cet estimateur.

2. **Estimation du risque de base α_0 :**

L'estimateur du risque de base est obtenu à partir d'un estimateur du risque de base cumulé. Le risque de base cumulé $A_0(t) = \int_0^\tau \alpha_0(t)\mathrm{d}t$ est estimé, pour un vecteur $\boldsymbol{\beta} \in \mathbb{R}^p$ fixé, par l'estimateur de Nelson-Aalen défini par

$$\hat{A}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u, \boldsymbol{\beta})}\mathrm{d}\bar{N}(u),$$

où $\bar{Y} = (1/n)\sum_{i=1}^n Y_i$ et $\bar{N} = (1/n)\sum_{i=1}^n N_i$. Cet estimateur, introduit par Aalen (1975; 1978), généralise l'estimateur de l'intensité cumulée empirique proposé indépendamment par Nelson (1969; 1972) et Altshuler (1970) dans le cas de durées de survie censurées.

L'estimateur de Breslow du risque de base cumulé $A_0(.)$ est une extension de l'estimateur de Nelson-Aalen : on remplace dans l'estimateur de Nelson-Aalen, le paramètre $\boldsymbol{\beta} \in \mathbb{R}^p$ fixé, par l'estimateur $\hat{\boldsymbol{\beta}}$ défini par (1.7). L'estimateur de Breslow est alors défini par

$$\hat{A}_0(t, \hat{\boldsymbol{\beta}}) = \int_0^t \frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u, \hat{\boldsymbol{\beta}})}\mathrm{d}\bar{N}(u). \tag{1.8}$$

En lissant les incréments de l'estimateur de Breslow (1.8), Ramlau-Hansen (1983b) a proposé un estimateur à noyau du risque de base α_0 défini par

$$\hat{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{t-u}{h}\Big)\frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u, \hat{\boldsymbol{\beta}})}\mathrm{d}N_i(u), \tag{1.9}$$

où $K : \mathbb{R} \to \mathbb{R}$ est une fonction d'intégrale 1 appelée noyau et $h > 0$ est la fenêtre. Nous renvoyons à la Section 1.2.3 pour plus de détails sur les estimateurs à noyau. L'estimateur du risque de base dépend donc de l'estimateur du paramètre de régression $\boldsymbol{\beta_0}$.

1.1.4 Et quand la dimension augmente...

Les procédures d'estimation et les résultats que nous avons mentionnés en Section 1.1.3 ne sont valides que lorsque $p < n$ et même p largement inférieur à n. Lorsque p, le nombre de covariables, grandit, et devient du même ordre de grandeur que n, la taille de l'échantillon, voire plus grand, de nombreux problèmes, maintenant bien connus, apparaissent. En effet, les résultats de la Section 1.1.3, permettent d'écrire que $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_2^2 = O_{\mathbb{P}}(\frac{p}{n})$. Ainsi, dès que $p>$n, l'estimateur $\hat{\boldsymbol{\beta}}$ n'est plus consistant.

Comme nous l'avons déjà dit, le paramètre $\boldsymbol{\beta_0} = (\beta_{0_1}, ..., \beta_{0,p})$ traduit le poids des variables explicatives $(Z_1, ..., Z_p)$ sur la durée de survie T. Lorsque le nombre de variables explicatives est important, un objectif serait d'évaluer la contribution de chaque variable, d'éliminer les variables non pertinentes et donc de sélectionner un sous-ensemble des variables explicatives pertinent, qui permet d'expliquer la durée de survie T : c'est la *sélection de variables*. La sélection de variables revient donc à la sélection des coordonnées non nulles dans le vecteur $\boldsymbol{\beta_0}$. L'idée va donc être de choisir le meilleur modèle, i.e. le meilleur ensemble de coordonnées non nulles de $\boldsymbol{\beta_0}$. Le modèle choisi sera le meilleur au sens d'un critère fixé a priori.

Un critère de qualité classique de choix de modèle, est la valeur de l'opposé de la log-vraisemblance partielle de Cox (1.6). Plus cette valeur est petite, meilleur est sensé être le modèle. Autrement dit, on choisira le vecteur qui minimisera cet opposé de la log-vraisemblance partielle. La vraisemblance (ou la vraisemblance partielle de Cox) est une fonction croissante du nombre de covariables (soit aussi du nombre de coordonnées non nulles de $\boldsymbol{\beta_0}$). Ainsi, minimiser sur un ensemble de modèles, l'opposée de la log-vraisemblance partielle de Cox, en supposant que la minimisation soit possible, conduira immanquablement à choisir le plus grand modèle, autrement dit à choisir un estimateur dont aucune des coordonnées n'est nulle. Outre que cet estimateur peut ne pas être défini, même quand il l'est, il restera difficilement interprétable. Pour remédier à ces problèmes, une solution classique consiste à pénaliser le critère d'estimation, c'est-à-dire à minimiser l'opposé de la log-vraisemblance partielle de Cox à laquelle on aura ajouté un terme de pénalité, ce qui aura intuitivement tendance à inciter à choisir un modèle « plus petit »:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \{-l_n^*(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\beta})\},$$

où pen : $\mathbb{R}^p \to \mathbb{R}^+$. Dans ce contexte, les pénalités usuelles sont des pénalités qui sont des fonctions croissantes de la dimension du vecteur covariables. Nous renvoyons

à Bickel et al. (2006) pour une réflexion sur le concept de pénalités et sur les différentes méthodes autour de ce concept. Le choix de la pénalité dépend de la finalité de l'étude considérée comme nous allons le voir dans les exemples que nous allons donner par la suite. L'objectif de ce type d'approche est de fournir des estimateurs interprétables, c'est-à-dire avec peu de coordonnées non nulles correspondant aux variables qui influencent la durée de survie.

Procédures ℓ_0

Un type de pénalité classique, qui permet d'annuler des coordonnées en faisant intervenir la semi-norme ℓ_0 du vecteur $\boldsymbol{\beta}$, est ce qu'on appelle une *pénalité* ℓ_0 :

$$\text{pen}(\boldsymbol{\beta}) = \Gamma_n |\boldsymbol{\beta}|_0,$$

où $|\boldsymbol{\beta}|_0$ est définie par $|\boldsymbol{\beta}|_0 = \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}}$. La définition et la valeur de Γ_n sont propres au critère considéré. Deux exemples de critères classiques faisant intervenir la pénalisation ℓ_0 sont les critères AIC et BIC. Le critère AIC, introduit par Akaike (1973) est de la forme

$$\text{pen}^{\text{AIC}}(\boldsymbol{\beta}) := \frac{2}{n} |\boldsymbol{\beta}|_0 \tag{1.10}$$

et le critère BIC, introduit par Schwarz (1978) est défini par

$$\text{pen}^{BIC}(\boldsymbol{\beta}) := \frac{\log p}{n} |\boldsymbol{\beta}|_0, \tag{1.11}$$

où p est le nombre de covariables.

Le passage à la grande dimension...

Lorsque le nombre de variables explicatives augmente, les procédures de sélection de variables basées sur les critères AIC et BIC ne sont pas utilisables en pratique. En effet, la complexité algorithmique de ces méthodes est telle, qu'elles sont difficiles à implémenter, même pour des p modestes. Lorsque p augmente et que l'on choisit de faire une comparaison exhaustive de tous les modèles, le nombre de modèles à comparer augmente en 2^p et une étude exhaustive de tous les modèles devient alors impossible. Breiman (1995) a ainsi montré que si les problèmes régularisés par une norme ℓ_0 conduisent à des modèles parcimonieux, les estimateurs sont, quant à eux, instables lorsque p grandit.

Les méthodes de sélection de variables considérées précédemment ont été introduites dans un cadre classique d'estimation, où la dimension p est raisonnable devant la taille de l'échantillon. Nous avons tendance à parler de grande dimension lorsque $p > n$. En réalité, déjà lorsque p est de l'ordre de $\sqrt{n}$, les algorithmes habituels ne convergent plus. Nous pouvons donc définir trois régimes : le régime classique, dit de petite dimension, lorsque les algorithmes habituels fonctionnent (généralement pour

$p \leq \sqrt{n}$), le régime modéré lorsque $p \geq \sqrt{n}$ mais reste de l'ordre de n et le régime de l'ultra grande dimension défini par $p \gg n$ (d'après Verzelen (2012), le régime de l'ultra grande dimension est atteint dès que $s \log(p/s)/n \geq 1/2$, où s est le nombre de coordonnées non nulles du vecteur $\boldsymbol{\beta_0}$). Nous parlerons de grande dimension dès que $p \geq \sqrt{n}$, c'est-à-dire dès que les méthodes habituelles ne fonctionnent plus. Dans ce manuscrit, nous considérons des problèmes en grande dimension.

Remarque 1.1 (Le cas de l'ultra grande dimension en pratique)**.** Un domaine où l'ultra grande dimension, i.e. $p \gg n$, est par nature très présente, est le contexte de la génomique. En effet, les biotechnologies récentes permettent d'acquérir des données de très grande dimension pour chaque individu. Par exemple, les puces à ADN mesurent les niveaux d'expression de dizaines de milliers de gènes simultanément. Plus récemment, les NGS pour *Séquençage Nouvelle Génération* permettent de séquencer l'ensemble du génome d'un individu en l'espace de quelques semaines. Toutefois, ces techniques coûtent encore relativement cher et dans les études classiques en génomique, le nombre d'individus ne dépasse souvent pas quelques centaines. Si on reprend l'exemple du jeu de données sur le cancer du sein, issu de l'étude de Loi et al. (2007), nous disposons de 44 928 niveaux d'expression de gènes en plus des quatre variables cliniques (âge, taille de la tumeur...) pour 414 patients divisés en deux groupes, les patients traités et les patients non traités. Dans cet exemple, le nombre de covariables est très important du fait du nombre de gènes pour lesquels on a mesuré le niveau d'expression, mais aussi du fait du nombre de variables cliniques. Cet exemple correspond au cas de ce que nous avons appelé le cas de l'ultra grande dimension ($p \gg n$). Cependant, en pratique, il n'est pas raisonnable de considérer l'ensemble des covariables lorsque leur nombre est trop important devant la taille de l'échantillon. Un travail de screening est alors souvent effectué pour faire une première sélection parmi l'ensemble des covariables et sortir de l'ultra grande dimension. Généralement, on se ramène à $p = O(n)$.

Avant de présenter notre contribution, nous présentons les différentes méthodes d'estimation ainsi que les résultats attendus dans des modèles classiques.

1.2 Généralités sur les méthodes d'estimation et les résultats attendus

Dans cette section, nous présentons trois procédures d'estimation : la procédure Lasso, la sélection de modèles et l'estimation à noyaux. Ces procédures ont toutes été largement étudiées en régression additive ou en densité. Dans un souci pédagogique, nous les présentons chacune dans l'un de ces deux cas simples. Les procédures, les outils et les résultats obtenus dans ces modèles, sont décrits de façon à illustrer notre démarche, les difficultés rencontrées et la manière dont nous les avons contournées.

1.2.1 La procédure Lasso

La procédure Lasso est l'une des procédures classiques développée en grande dimension. Elle a été introduite par Tibshirani (1996) dans le cas d'un modèle linéaire. L'estimateur Lasso a ensuite été très largement étudié en régression linéaire (voir Knight & Fu (2000), Efron et al. (2004), Donoho et al. (2006), Meinshausen & Bühlmann (2006), Zhao & Yu (2006), Zhang & Huang (2008a) et Meinshausen & Yu (2009)) et plus généralement dans le cas d'un modèle de régression additive non-paramétrique (voir Juditsky & Nemirovski (2000), Nemirovski (2000), Bunea et al. (2004; 2006; 2007a;b), Greenshtein & Ritov (2004) ou encore Bickel et al. (2009)). Enfin, le Lasso ou le Dantzig, estimateur proche du Lasso (nous renvoyons à Bickel et al. (2009) pour comparaison de ces deux estimateurs en régression additive) ont été considérés pour estimer une densité (voir respectivement Bunea et al. (2007c) et Bertin et al. (2011)).

Dans cette sous-section, nous présentons le Lasso dans un cas classique de régression additive, en insistant sur la construction de l'estimateur Lasso, les inégalités oracles non-asymptotiques recherchées, ainsi que les hypothèses requises.

Modèle illustratif : Dans un modèle de régression additif, soient $(\boldsymbol{Z_1},Y_1),...,(\boldsymbol{Z_n},Y_n)$ un échantillon de n couples indépendants $(\boldsymbol{Z_i}, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ tels que

$$Y_i = f_0(\boldsymbol{Z_i}) + W_i, \quad i = 1, ..., n, \tag{1.12}$$

où $f_0 : \mathbb{R}^p \to \mathbb{R}$ est une fonction de régression à estimer, les $\boldsymbol{Z_i}$ sont déterministes et les erreurs de régression W_i sont centrées et de variance σ^2. Les erreurs de régression $(W_i)_{i\in\{1,...,n\}}$, peuvent être de différents types : des erreurs de régression gaussiennes si pour tout i, $W_i \sim \mathcal{N}(0, \sigma^2)$, des erreurs de régression bornées ou plus généralement des erreurs de régression sous-exponentielles.

Remarque. Si $f_0(\boldsymbol{Z_i}) = \boldsymbol{\beta_0^T Z_i}$, on obtient le modèle de régression linéaire.

Critère d'estimation : La procédure Lasso est une procédure d'estimation par minimum de contraste pénalisé, fondée sur un *critère empirique* noté C_n qui dépend des

observations $Z_1, ..., Z_n$. En régression additive, le critère d'estimation habituellement considéré est le critère des moindres carrés. Il est défini par

$$C_n(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{Z_i}))^2. \tag{1.13}$$

Dictionnaire de fonction : L'estimation de la fonction de régression f_0 non linéaire repose sur l'idée qu'elle peut être correctement approchée par une combinaison linéaire d'un petit nombre de fonctions. On introduit donc un dictionnaire de fonctions $\mathbb{F}_M = \{f_1, ..., f_M\}$, c'est-à-dire une collection de fonctions $f_j : \mathbb{R}^p \to \mathbb{R}$, pour $j = 1, ..., M$, à partir desquelles on va construire un estimateur de f_0. Le dictionnaire $\mathbb{F}_M$ est constitué de fonctions de base telles que les ondelettes, les splines, les fonctions en escaliers, les fonctions coordonnées, etc. Les fonctions f_j peuvent aussi être des estimateurs obtenus avec différents paramètres de régularisation.

Fonctions candidates : Les fonctions candidates pour estimer f_0 sont des combinaisons linéaires des fonctions du dictionnaire. Pour $\boldsymbol{\beta} \in \mathbb{R}^M$, on définit la fonction candidate $f_{\boldsymbol{\beta}}$ telle que

$$f_{\boldsymbol{\beta}}(\boldsymbol{Z_i}) = \sum_{j=1}^{M}\beta_j f_j(\boldsymbol{Z_i}). \tag{1.14}$$

Dans le cas particulier où $f_j(\boldsymbol{Z_i}) = Z_{i,j}$, le modèle de régression est linéaire avec $p = M$ et il s'agit d'un problème de sélection de variables. Ainsi, introduire un dictionnaire permet simplement de se placer dans un cadre plus général que celui de la sélection de variables. Typiquement, la taille du dictionnaire $\mathbb{F}_M$ utilisé pour estimer une fonction des covariables en grande dimension est grande, i.e. $M \gg n$. En considérant des fonctions candidates sous la forme (1.14), on suppose que f_0 admet une approximation sparse dans $\mathbb{F}_M$, c'est-à-dire qu'on peut l'approcher par une combinaison linéaire d'un petit nombre de fonctions de $\mathbb{F}_M$. En pratique, le choix du dictionnaire $\mathbb{F}_M$ est important car f_0 peut admettre une bonne approximation sparse dans certaines bases de fonctions et pas dans d'autres.

Procédure Lasso : En minimisant le critère (1.13) sur le dictionnaire $\mathbb{F}_M$, on se ramène alors à un problème d'estimation paramétrique : le paramètre à estimer est le paramètre $\boldsymbol{\beta}$ de la combinaison linéaire (1.14). L'estimateur Lasso du paramètre $\boldsymbol{\beta}$ est en fait l'estimateur du minimum de contraste sous une contrainte de type ℓ_1 :

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \begin{cases} \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{C_n(f_{\boldsymbol{\beta}})\}, \\ \text{s.c. } \sum_{j=1}^{p}|\beta_j| \leq s \end{cases} \tag{1.15}$$

où s est un paramètre positif. Nous pouvons réécrire (1.15) sous sa forme pénalisée :

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{C_n(f_{\boldsymbol{\beta}}) + \Gamma_n|\boldsymbol{\beta}|_1\}, \tag{1.16}$$

où $\Gamma_n > 0$ est un paramètre de régularisation à calibrer. Les deux notations (1.15) et (1.16) sont équivalentes [2], mais la forme la plus usuelle est celle définie en (1.16) et c'est celle-ci que nous considérerons dans la suite.

Le problème ainsi posé est convexe en $\boldsymbol{\beta}$ (voir Bunea et al. (2007b) par exemple) et les procédures standard d'optimisation convexe peuvent donc être utilisées pour calculer $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$. Nous renvoyons à Efron et al. (2004), Friedman et al. (2007) et Meier et al. (2008) pour une discussion détaillée sur ces problèmes d'optimisation. L'un des avantages de la norme ℓ_1 est qu'elle fournit un estimateur sparse (ou parcimonieux). En effet, la boule de norme ℓ_1 contrainte permet d'annuler un certain nombre de coordonnées du paramètre $\boldsymbol{\beta}$. On définit $J(\boldsymbol{\beta}) = \{j \in \{1, ..., M\} : \beta_j \neq 0\}$ l'*ensemble de sparsité* du vecteur $\boldsymbol{\beta} \in \mathbb{R}^M$ et $|J(\boldsymbol{\beta})|$ son *indice de sparsité*.

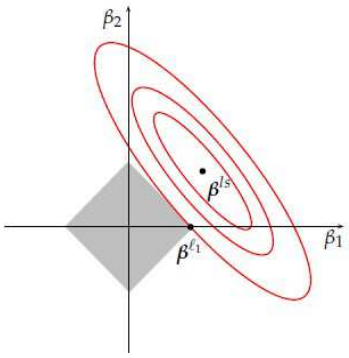


FIGURE 1.2 – *Illustration de la pénalisation ℓ_1*

La figure 1.2 ci-dessus illustre l'intérêt d'utiliser la norme ℓ_1 pour obtenir un estimateur parcimonieux. Ici, $\hat{\boldsymbol{\beta}}^{ls}$ représente l'estimateur non contraint des moindres carrés. Les ellipses rouges représentent les contours de la fonction de perte quadratique autour de l'estimateur $\hat{\boldsymbol{\beta}}^{ls}$. Cette figure représente l'estimateur obtenu pour le problème (1.15) soumis à une contrainte ℓ_1, lorsque $\hat{\boldsymbol{\beta}}^{ls}$ n'appartient pas au domaine admissible. L'estimateur Lasso $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$, solution de la minimisation (1.16), est noté $\hat{\boldsymbol{\beta}}^{\ell_1}$ sur la figure 1.2. La deuxième composante de $\hat{\boldsymbol{\beta}}^{\ell_1}$ est annulée, car l'ellipse atteint la région admissible sur l'angle situé sur l'axe $\beta_2 = 0$. L'estimateur obtenu par cette méthode est donc facilement interprétable, y compris quand $p > n$.

Selon le choix du paramètre de régularisation Γ_n, la procédure Lasso annule plus ou moins de coordonnées du paramètre $\boldsymbol{\beta}$. Lorsque $\Gamma_n = 0$, nous retrouvons l'estimateur

2. voir la thèse de Hebiri (2009)

des moindres carrés non-contraint et aucune des coordonnées de $\boldsymbol{\beta}$ n'est annulée. En revanche, pour de très grandes valeurs de Γ_n, l'estimateur obtenu a toutes ses coordonnées nulles, on obtient donc $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \vec{0}$. Il existe plusieurs méthodes permettant de choisir Γ_n, mais la plus connue est la validation croisée (voir Tibshirani (1996)).

Remarque. Théoriquement, le paramètre de régularisation Γ_n est de l'ordre de $\sqrt{\log M/n}$ (voir Bühlmann & van de Geer (2009) par exemple). On peut aussi considérer une procédure Lasso pondéré au sens où la pénalité est en norme ℓ_1 pondérée du paramètre $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\left\{ C_n(f_{\boldsymbol{\beta}}) + \text{pen}(\boldsymbol{\beta}) \right\}, \quad \text{avec } \text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j|, \tag{1.17}$$

et les $(\omega_j)_{j\in\{1,\dots,M\}}$ sont des poids à déterminer à partir des données (*data-driven weights* en anglais). Bickel et al. (2009) ont, par exemple, considéré des poids $\omega_j = \Gamma_n ||f_j||_n$, où Γ_n est d'ordre $\sqrt{\log M/n}$.

Estimateur final : En résolvant (1.17), on obtient l'estimateur Lasso de f_0, défini par

$$f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}(\boldsymbol{Z_i}) = \sum_{j=1}^{M} \hat{\beta}_{L,j} f_j(\boldsymbol{Z_i}), \quad \text{pour } i = 1, ..., n.$$

Qualité d'estimation : Pour mesurer la qualité d'un estimateur, on cherche à obtenir des bornes de risque (*risk bound*) soit avec grande probabilité pour la norme empirique associée aux moindres carrés définie pour toute fonction $f : \mathbb{R}^p \to \mathbb{R}$ par

$$||f_0 - f||_n^2 = \frac{1}{n}\sum_{i=1}^{n} (f_0(\boldsymbol{Z_i}) - f(\boldsymbol{Z_i}))^2,$$

soit en espérance, c'est-à-dire pour la fonction de perte définie par

$$\ell(f_0, f) = \mathbb{E}[C_n(f)] - \mathbb{E}[C_n(f_0)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} (f_0(\boldsymbol{Z_i}) - f(\boldsymbol{Z_i}))^2 \right].$$

Pour l'estimateur Lasso, on trouve par exemple des résultats qui bornent la norme empirique avec grande probabilité (voir Bickel et al. (2009)) ou son espérance comme ce qu'ont fait par exemple Massart & Meynet (2011).

Nature des résultats :

Trois types de résultats sont classiquement considérés quand $p \geq n$. Nous ne mentionnons ici que les résultats non-asymptotiques.

— Lorsqu'on cherche à fournir la meilleure approximation du vecteur $\boldsymbol{\beta_0^T Z}$ dans le cas linéaire ou de $f_0(\boldsymbol{Z})$ dans le cas non-paramétrique, on parle de prédiction. Les résultats en prédiction visent à obtenir une majoration pour $||\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{X}\boldsymbol{\beta_0}||_n$, en régression linéaire, où $\boldsymbol{X}$ est la matrice de design $\boldsymbol{X} = (Z_{i,j})_{i,j}$ pour $i = 1, ..., n$ et $j = 1, ..., p$, et pour $||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n$ dans le cas non-paramétrique. De très nombreux articles se sont intéressés à obtenir des résultats en prédiction dans le modèle de régression linéaires ou additives. Nous en donnons quelques exemples. Zhang & Huang (2008b) et Bickel et al. (2009) ont obtenu des bornes non-asymptotiques en régression linéaire. Dans le cas non-paramétrique, les performances en prédiction du Lasso sont mesurées en établissant des inégalités oracles non-asymptotiques (voir ci-dessous). Bunea et al. (2007a), Bunea et al. (2007b), Bickel et al. (2009) ou encore Massart & Meynet (2011) ont établi des inégalités oracles pour l'estimateur Lasso de la fonction de régression.

— Lorsqu'on s'intéresse plutôt à l'estimation du paramètre de régression $\boldsymbol{\beta_0}$ dans le cas linéaire, on parle d'estimation. Dans un cadre non-asymptotique, les résultats d'estimation sont souvent exprimés en majorant $|\boldsymbol{\beta_0} - \hat{\boldsymbol{\beta}}_{\boldsymbol{L}}|_1$ ou $|\boldsymbol{\beta_0} - \hat{\boldsymbol{\beta}}_{\boldsymbol{L}}|_2$ ou plus rarement $|\boldsymbol{\beta_0} - \hat{\boldsymbol{\beta}}_{\boldsymbol{L}}|_q$. Bunea et al. (2007a), Bunea (2008) et Bickel et al. (2009) ont ainsi obtenu, sous certaines hypothèses, des inégalités en estimation en norme ℓ_1, et Bickel et al. (2009), sous des hypothèses plus restrictives, ont aussi établi une inégalité en sélection en norme ℓ_q pour $1 < q \leq 2$. Des inégalités en estimation sont obtenues par Meinshausen & Yu (2009) en norme ℓ_2 et par Zhang & Huang (2008a) en norme ℓ_q avec $q \geq 1$.

— Si enfin, on veut identifier le support $J(\boldsymbol{\beta_0}) = \{j, \beta_{0_j} \neq 0\}$ de $\boldsymbol{\beta_0}$, ou encore le vecteur signe de $\boldsymbol{\beta_0}$, on parle de sélection. Les premiers résultats non asymptotiques de sélection de variables en grande dimension ont été établis dans le cadre de la régression linéaire par Zhao & Yu (2006).

Nous nous intéressons plus particulièrement aux résultats non-asymptotiques en prédiction et présentons les inégalités oracles ainsi que la démarche pour les établir.

(a) Inégalités oracles pour le Lasso dans le modèle de régression additif

Pour une constante $\zeta > 0$ fixée et un paramètre de régularisation Γ_n de l'ordre de $\sqrt{\log(M)/n}$, on définit l'inégalité oracle non-asymptotique pour l'estimateur Lasso par

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq (1+\zeta) \inf_{\boldsymbol{\beta} \in \mathbb{R}^M} \{||f_{\boldsymbol{\beta}} - f_0||_n^2 + R_{n,M}(\boldsymbol{\beta})\}, \tag{1.18}$$

où $||f_{\boldsymbol{\beta}} - f_0||_n^2$ est un terme de biais et $R_{n,M}(\boldsymbol{\beta})$ est un terme de variance. Ce dernier donne la vitesse de convergence de l'estimateur $f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}$ vers la vraie fonction

f_0. Il est d'ordre $\sqrt{\log M/n}$ ou $\log M/n$ selon que la convergence de l'estimateur $f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}$ vers la vraie fonction f_0 soit lente ou rapide respectivement. Nous parlerons alors d'inégalité oracle lente ou rapide. Les termes de biais et de variance sont inversement proportionnels à la taille du dictionnaire $\mathbb{F}_M$: plus le dictionnaire de fonction $\mathbb{F}_M$ sera grand, plus le terme de biais sera petit et plus le terme de variance sera grand. L'un des objectifs de notre travail est d'établir des inégalités oracles du type (1.18) avec grande probabilité.

(b) Démarche pour obtenir une inégalité oracle lente

Le début de la démarche pour obtenir des inégalités oracles lentes ou rapides pour l'estimateur Lasso est classique. Par définition du Lasso, on a pour tout $\boldsymbol{\beta} \in \mathbb{R}^M$,

$$C_n(f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) + \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) \leq C_n(f_{\boldsymbol{\beta}}) + \text{pen}(\boldsymbol{\beta}).$$

Par définition (1.13) du critère des moindres carrés combiné à un simple jeu d'écriture, on obtient finalement

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq ||f_0 - f_{\boldsymbol{\beta}}||_n^2 + \frac{2}{n}\sum_{i=1}^{n}(f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_{\boldsymbol{\beta}})(\boldsymbol{Z_i})W_i + \text{pen}(\boldsymbol{\beta}) - \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}).$$

Pour tout $\boldsymbol{\beta} \in \mathbb{R}^M$, on a $f_{\boldsymbol{\beta}} = \sum_{j=1}^{M} \beta_j f_j$, en remplaçant $f_{\boldsymbol{\beta}}$ et $f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}$ dans l'inégalité précédente on obtient donc finalement

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq ||f_\beta - f_0||_n^2 + 2\sum_{j=1}^{M}(\hat{\beta}_L - \beta)_j V_n(f_j) + \text{pen}(\boldsymbol{\beta}) - \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}),$$

avec

$$V_n(f_j) = \frac{1}{n}\sum_{i=1}^{n} f_j(\boldsymbol{Z_i})W_i. \tag{1.19}$$

A ce stade, il reste à contrôler le processus $V_n(f_j)$ pour obtenir une inégalité oracle lente. Nous renvoyons au paragraphe (d) pour une discussion sur le contrôle de ce terme. Pour établir une inégalité oracle rapide, on a besoin d'une hypothèse supplémentaire.

(c) Hypothèse

L'inégalité oracle rapide nécessite une hypothèse supplémentaire sur la matrice de Gram $\boldsymbol{\Psi_n} = \boldsymbol{X^T X}/n$ où $\boldsymbol{X} = (f_j(\boldsymbol{Z_i}))_{i,j}$ avec $i \in \{1, ..., n\}$ et $j \in \{1, ..., M\}$. Lorsque $M \gg n$, la matrice de Gram $\boldsymbol{\Psi_n}$ est dégénérée, ce qui s'écrit

$$\min_{\boldsymbol{b} \in \mathbb{R}^M \setminus \{0\}} \frac{(\boldsymbol{b^T \Psi_n b})^{1/2}}{|\boldsymbol{b}|_2} = \min_{\boldsymbol{b} \in \mathbb{R}^M \setminus \{0\}} \frac{|\boldsymbol{Xb}|_2}{\sqrt{n}|\boldsymbol{b}|_2} = 0.$$

Le problème ordinaire de minimisation des moindres carrés n'a pas de solution unique, puisqu'il nécessite la définie positivité de la matrice de Gram, c'est-à-dire

$$\min_{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\}} \frac{|\boldsymbol{X}\boldsymbol{b}|_2}{\sqrt{n}|\boldsymbol{b}|_2} > 0.$$

La procédure Lasso nécessite une hypothèse plus faible que la définie positivité de la matrice de Gram.

Nous renvoyons à Bühlmann & van de Geer (2009) et à Bickel et al. (2009) pour des discussions détaillées sur les différentes hypothèses que l'on peut considérer pour établir une inégalité oracle rapide. Une des hypothèses les plus faibles est la condition aux valeurs propres restreintes (Restricted Eigenvalue condition), introduite par Bickel et al. (2009) pour le modèle de régression additif.

Hypothèse 1.1 (Condition $\mathbf{RE}(s, c_0)$ pour le modèle de régression additif)**.** *Pour un entier $s \in \{1, ..., M\}$ et une constante $c_0 > 0$, on dit que la condition aux valeurs propres restreintes (Restricted Eigenvalue Condition)* $\mathbf{RE}(s, c_0)$ *est vérifiée si*

$$0 < \boldsymbol{\kappa}(s, c_0) = \min_{\substack{J\subset\{1,...,M\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\leq c_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{\Psi}_{\boldsymbol{n}}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2}, \tag{1.20}$$

où $\boldsymbol{b}_J$ désigne le vecteur $\boldsymbol{b}$ restreint à l'ensemble J : $(b_J)_j = b_j$ si $j \in J$ et $(b_J)_j = 0$ si $j \in J^c$, où $J^c = \{1, ..., M\}\setminus J$.

La condition aux valeurs propres restreintes assure que la plus petite valeur propre de la matrice de Gram définie comme

$$\min_{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{\Psi}_{\boldsymbol{n}}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}|_2},$$

restreinte aux ensembles de sparsités J, tels que $|J| \leq s$, soit strictement positive. Autrement dit, la condition **RE** assure la définie positivité de la matrice $\boldsymbol{\Psi}_{\boldsymbol{n}}$ uniquement pour les vecteurs $\boldsymbol{b} \in \mathbb{R}^M\setminus\{0\}$ qui vérifient l'inégalité $|\boldsymbol{b}_{J^c}|_1 \leq c_0|\boldsymbol{b}_J|_1$. L'entier s joue ici le rôle d'une borne supérieure pour l'indice de sparsité $|J(\boldsymbol{\beta})|$ associé à un vecteur $\boldsymbol{\beta}$, de sorte que les sous-matrices carrées de taille inférieure à $2s$ de la matrice de Gram $\boldsymbol{\Psi}_{\boldsymbol{n}}$ soient définies positives. Dans la démonstration permettant d'obtenir l'inégalité oracle rapide, l'hypothèse **RE** permet de relier la norme ℓ_2 de $|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})_{J(\boldsymbol{\beta})}|_2$ à la norme empirique $||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_{\boldsymbol{\beta}}||_n$ pour tout $\boldsymbol{\beta} \in \mathbb{R}^M$.

(d) Le rôle des inégalités de concentration

Pour que les inégalités oracles soient lentes ou rapides, elles nécessitent le contrôle uniforme du processus $V_n(f_j)$ défini par (1.19) (voir paragraphe (b)), souvent à

l'aide d'inégalités de concentration. Bickel et al. (2009) ont utilisé une inégalité de déviation propre aux variables aléatoires gaussiennes pour contrôler $V_n(f_j)$. Lorsque les erreurs de régression W_i pour $i = 1, ..., n$ sont centrées et bornées, on peut appliquer une inégalité de Hoeffding. Dans le cas sous-exponentiel, on utilise une inégalité de Bernstein standard (voir Annexe A.3.1) pour contrôler le processus centré $V_n(f_j)$. Nous renvoyons à l'Annexe D du Chapitre 2, où nous présentons une version de la preuve de Bickel et al. (2009) permettant d'obtenir une inégalité oracle lente pour une erreur de régression sous-exponentielle plus générale. Le contrôle du processus $V_n(f_j)$ par une inégalité de Bernstein y est détaillé.

1.2.2 La sélection de modèles

La sélection de modèles a d'abord été introduite par Akaike (1973) et Mallows (1973), puis formalisée par Birgé & Massart (1997) pour l'estimation de la densité et par Barron et al. (1999) dans le cadre plus général d'une fonction inconnue à estimer (densité ou fonction de régression). Elle a été étudiée en régression non-paramétrique (voir entre autres Baraud (2000), Yang (1999), Birgé & Massart (2001), Baraud (2002) ou encore Wegkamp (2003)) et en densité (voir Birgé & Massart (1998) et Birgé (2008)). Nous renvoyons aussi à Massart (2007) comme un ouvrage de référence sur la sélection de modèles.

Dans cette sous-section, nous expliquons le principe de la sélection de modèles pour l'estimation d'une densité, en détaillant la construction de la procédure de sélection de modèles, les inégalités oracles non-asymptotiques recherchées, ainsi que les hypothèses requises.

Modèle : Soient $Z_1, ..., Z_n$ un n-échantillon de densité f_0 inconnue. On suppose que f_0 appartient à l'ensemble $\mathbb{L}^2(\mathbb{R})$, muni du produit scalaire usuel $\langle ., . \rangle$ et de la norme $||.||$ usuelle.

Critère d'estimation et fonction de perte : La procédure de sélection de modèles est basée, comme la procédure Lasso, sur la minimisation d'un critère d'estimation. Considérons le critère des moindres carrés défini pour la densité par

$$C_n(f) = ||f||_2^2 - \frac{2}{n}\sum_{i=1}^{n} f(Z_i). \tag{1.21}$$

La pertinence de ce choix est assurée puisque $\mathbb{E}[C_n(f)] = ||f||_2^2 - 2\langle f, f_0 \rangle_2 = ||f - f_0||_2^2 - ||f_0||_2^2$, qui est minimum en $f = f_0$. La fonction de perte associée est définie pour tout $f \in \mathbb{L}^2(\mathbb{R})$ par

$$\ell(f_0, f) = ||f_0 - f||_2^2 = \int_{\mathbb{R}} (f_0(x) - f(x))^2 \mathrm{d}x.$$

Bases de fonctions et collection de modèles : Soit $(\varphi_j^m)_{j\in J}$, où $J \subset \mathbb{N}^*$, une base de $\mathbb{L}^2(\mathbb{R})$. La densité f_0 se décompose de manière unique sous la forme

$$f_0 = \sum_{j\in J} a_j\varphi_j, \quad \text{avec } a_j = \langle f_0, \varphi_j\rangle\varphi_j.$$

Pour une collection finie d'indices $\mathcal{M}_n \subset \mathbb{N}$, on considère la collection de modèles $(S_m)_{m\in\mathcal{M}_n}$, telle que chaque modèle S_m est engendré par la base $(\varphi_j)_{j\in J_m}$, avec $J_m \subset J$: $S_m = \text{Vect}\{\varphi_j, j \in J_m\}$. La dimension de S_m est alors donnée par $D_m := \text{Card}(J_m)$.

Collection d'estimateurs : Associée à cette collection de modèles, on définit la collection d'estimateurs $(\hat{f}_m)_{m\in\mathcal{M}_n}$ telle que

$$\hat{f}_m = \arg\min_{f\in S_m}\{C_n(f)\}. \tag{1.22}$$

L'estimateur $\hat{f}_m$ est défini de manière unique par $\hat{f}_m = \sum_{j\in J_m}\hat{a}_j\varphi_j$, avec $\hat{a}_j = (1/n)\sum_{i=1}^n \varphi_j(Z_i)$. Pour $m \in \mathcal{M}_n$, notons

$$f_m = \arg\min_{f\in S_m}\{\ell(f_0, f)\},$$

la projection orthogonale de f_0 sur le modèle S_m. Avec la définition (1.21) de $C_n(.)$, on en déduit que $\mathbb{E}[\hat{f}_m] = f_m$. Ainsi, $\hat{f}_m$ est un estimateur sans biais de la projection de f_0 sur le modèle S_m.

Procédure de sélection de modèles : L'objectif est de sélectionner le modèle S_m parmi l'ensemble des modèles, pour lequel le risque de l'estimateur associé $\mathbb{E}[\ell(f_0, \hat{f}_m)]$ est aussi proche que possible du meilleur des risques que l'on peut obtenir dans la collection $(\hat{f}_m)_{m\in\mathcal{M}_n}$. Par le théorème de Pythagore, l'erreur quadratique se décompose en deux termes :

$$\mathbb{E}[||\hat{f}_m - f_0||_2^2] = ||f_0 - f_m||_2^2 + \mathbb{E}[||\hat{f}_m - f_m||_2^2]. \tag{1.23}$$

Le premier terme est le terme de biais, il représente l'erreur d'approximation de f_0 par f_m. Il est d'autant plus petit que S_m est grand. Le second est le terme de variance provenant de l'erreur commise en remplaçant a_j par une version empirique $\hat{a}_j$. Contrairement au terme de biais, il est en général d'autant plus grand que le modèle S_m est grand. Un choix de modèle optimal en terme d'excès de risque nécessite donc de trouver un compromis entre ces deux termes, i.e., choisir S_m assez grand pour diminuer le biais, mais pas trop pour éviter que la variance n'augmente trop : il s'agit du *compromis biais-variance*. L'oracle m^* est alors l'indice $m \in \mathcal{M}_n$ qui réalise le meilleur compromis biais-variance. Il minimise l'excès de risque :

$$m^* = \arg\min_{m\in\mathcal{M}_n}\{\mathbb{E}[||\hat{f}_m - f_0||_2^2]\}.$$

L'estimateur oracle associé est alors $\hat{f}_{m^*}$. ll est incalculable en pratique (car m^* dépend de f_0, qui est inconnue) et l'objectif est donc de sélectionner un estimateur $\hat{f}_{\hat{m}}$ ayant des performances similaires à celles de l'oracle. Il s'agit donc de définir un critère de sélection du modèle imitant le comportement biais-variance (1.25) du risque de l'estimateur. Le terme de variance est majoré par

$$\mathbb{E}[||\hat{f}_m - f_m||_2^2] \leq \frac{1}{n}\sum_{j\in J_m}\mathbb{E}[\varphi_j^2(Z_i)] \leq \frac{||f_0||_\infty}{n}\int_{\mathbb{R}}\sum_{j\in J_m}\varphi_j^2(x)\mathrm{d}x = ||f_0||_\infty\frac{D_m}{n}. \quad (1.24)$$

On déduit de (1.23) et (1.24), l'inégalité suivante

$$\mathbb{E}[||\hat{f}_m - f_0||_2^2] \leq ||f_0 - f_m||_2^2 + ||f_0||_\infty\frac{D_m}{n}. \quad (1.25)$$

On cherche le modèle S_m, sur lequel $||f_0 - f_m||_2^2 + ||f_0||_\infty D_m/n$ est minimum. Cela revient à chercher le minimum de $-||f_m||_2^2 + ||f_0||_\infty D_m/n$. Une heuristique que l'on doit à Mallows (1973) et connue sous le nom d'*heuristique de Mallows* consiste à remplacer $-||f_m||_2^2$, qui est inconnu, par son estimateur $-||\hat{f}_m||_2^2 + ||f_0||_\infty D_m/n$, et à prendre $\hat{m} \in \mathcal{M}_n$ qui minimise en m le critère

$$-||\hat{f}_m||_2^2 + 2||f_0||_\infty\frac{D_m}{n} = C_n(\hat{f}_m) + 2||f_0||_\infty\frac{D_m}{n}. \quad (1.26)$$

De manière plus générale, le modèle $\hat{m}$ est sélectionné par la procédure suivante :

$$\hat{m} = \underset{m\in\mathcal{M}_n}{\arg\min}\{C_n(\hat{f}_m) + \mathrm{pen}(m)\}, \quad (1.27)$$

où pen : $\mathcal{M}_n \to \mathbb{R}^+$ est de l'ordre du terme de variance $\mathbb{E}[||\hat{f}_m - f_m||_2^2]$.

Remarque. Dans le cas de la sélection de variables, la taille du modèle correspond au terme $|\boldsymbol{\beta}|_0$ défini dans la Section 1.1.4. La sélection de modèles avec une pénalité $\mathrm{pen}(m) = ||f_0||_\infty D_m/n$, proportionnelle à la taille du modèle appartient donc bien à la famille des procédures pénalisées en norme ℓ_0.

Estimateur final : L'estimateur final, noté $\hat{f}_{\hat{m}}$, est alors obtenu à partir de (1.22) et (1.27). On parle d'*estimateur par minimum de contraste pénalisé*.

Nature des résultats :

Les inégalités oracles assurent que l'estimateur obtenu par sélection de modèles est performant. Nous renvoyons à l'ouvrage de Massart (2007) pour plus de détails. Nous présentons les inégalités oracles en sélection de modèles ainsi que la démarche pour les établir.

(a) Inégalités oracles en sélection de modèles

La qualité de l'estimateur $\hat{f}_{\hat{m}}$ est établie via des inégalités oracles de la forme

$$\mathbb{E}[||f_0 - \hat{f}_{\hat{m}}||_2^2] \leq C \inf_{m \in \mathcal{M}_n} \{||f_0 - f_m||_2^2 + \text{pen}(m)\} + \frac{C_1}{n} \tag{1.28}$$

où C est une constante universelle et C_1/n un terme négligeable devant l'infimum (voir Birgé & Massart (1997)). La constante C_1 dépend généralement du cardinal de $\mathcal{M}_n$ et de la complexité de la famille de modèles. Le terme de biais $||f_0 - f_m||_2^2$ décroît lorsque D_m croît, il dépend de la régularité de f_0, inconnue, et est d'autant plus petit que f_0 est régulière. Le terme de variance $\text{pen}(m)$ croît avec D_m. Il est d'ordre D_m/n et correspond à l'ordre de la variance sur un modèle. Ainsi à la constante C près et lorsque le terme résiduel est négligeable, $\hat{f}_{\hat{m}}$ réalise le meilleur compromis biais-variance. Le terme résiduel C_1/n provient du contrôle uniforme du processus empirique (1.31), défini ci-dessous, par l'inégalité de déviation (1.33) (voir les paragraphes (b) et (d) ci-dessous).

(b) Démarche pour obtenir une inégalité oracle

Par définition, l'estimateur $\hat{f}_{\hat{m}}$ satisfait pour tout $m \in \mathcal{M}_n$ l'inégalité suivante :

$$C_n(\hat{f}_{\hat{m}}) + \text{pen}(\hat{m}) \leq C_n(\hat{f}_m) + \text{pen}(m) \leq C_n(f_m) + \text{pen}(m). \tag{1.29}$$

Pour tout $f \in \mathbb{L}^2(\mathbb{R})$, le critère des moindres carrés (1.21) se décompose en

$$C_n(f) = ||f||_2^2 - 2\langle f_0, f\rangle_2 - \nu_n(t) = ||f - f_0||_2^2 - ||f_0||_2^2 - \nu_n(t), \tag{1.30}$$

où $\nu_n(.)$ est un processus centré défini par

$$\nu_n(f) = C_n(f) - \mathbb{E}[C_n(f)] = -\frac{2}{n}\sum_{i=1}^{n} f(Z_i) + 2\int_{\mathbb{R}} f(x)f_0(x)\mathrm{d}x. \tag{1.31}$$

En combinant (1.29) et (1.30), on obtient l'inégalité suivante

$$||f_0 - \hat{f}_{\hat{m}}||_2^2 \leq ||f_0 - f_m||_2^2 + \text{pen}(m) + \nu_n(\hat{f}_{\hat{m}}) - \nu_n(f_m) - \text{pen}(\hat{m}). \tag{1.32}$$

pour tout $m \in \mathcal{M}_n$. Toute la difficulté réside ensuite dans le contrôle uniforme du processus $\nu_n(\hat{f}_{\hat{m}}) - \nu_n(f_m) = \nu_n(\hat{f}_{\hat{m}} - f_m)$. Nous renvoyons au paragraphe (d) pour une discussion sur le contrôle de ce processus.

(c) Hypothèses sur les modèles

Les hypothèses suivantes sur les modèles sont standard.

Hypothèse 1.2.

(i) *Pour tout* $m \in \mathcal{M}_n$, $D_m \leq n$.

(ii) *Il existe une constante* $\phi > 0$ *telle que pour tout* f *dans* S_m,

$$||f||_\infty^2 \leq \phi D_m ||f||_2^2.$$

(iii) *Les modèles sont emboîtés :* $D_{m_1} \leq D_{m_2} \Rightarrow S_{m_1} \subset S_{m_2}$.

Nous renvoyons à Birgé & Massart (1998), Barron et al. (1999), Baraud (2002) et Massart (2007) pour des détails sur les hypothèses. L'hypothèse 1.2.(i) permet d'assurer que la taille des modèles n'est pas trop grande par rapport à la taille de l'échantillon. Si l'on reprend l'exemple de la sélection de variables, la taille d'un modèle correspond au nombre de coefficients inconnus à estimer et l'hypothèse 1.2.(i) impose donc que ce nombre soit inférieur au nombre d'observations. L'hypothèse 1.2.(ii) compare la norme $\mathbb{L}^2$ standard à la norme infinie. Cette hypothèse permet de remplacer, dans la majoration de la variance (1.24), le terme inconnu $||f_0||_\infty$ par ϕ^2. Elle est équivalente à l'inégalité suivante (voir le Lemme 1 de Birgé & Massart (1998)) :

$$\left|\left| \sum_{j \in J_m} \varphi_j \right|\right|_\infty^2 \leq \phi^2 D_m.$$

L'hypothèse 1.2.(iii) est relativement forte. Elle peut dans certains cas être relâchée et remplacée par l'hypothèse de l'existence d'un espace englobant tous les modèles de la collection (voir par exemple l'hypothèse $\mathcal{H}_2$ Brunel & Comte (2005)). Cette hypothèse permet de ne pas avoir à parcourir tous les modèles, lors de la procédure de sélection de modèles. En sélection de variables, le nombre de modèles à comparer augmente en 2^p lorsque p augmente. Considérer des modèles emboîtés permet alors de réduire la complexité algorithmique de la procédure. Les modèles étant emboîtés, on peut considérer le réarrangement suivant : $J = \mathbb{N}\backslash\{0\}$, $J_m = 1, ..., D_m$ et $S_m = \text{Vect}\{\varphi_1, ..., \varphi_{D_m}\}$.

Plusieurs bases de fonctions de $\mathbb{L}^2(\mathbb{R})$ vérifient les hypothèses précédentes et peuvent être considérées : la base trigonométrique, la base des polynômes par morceaux, la base d'ondelettes à support compact ou encore la base d'histogrammes, entre autres. Nous renvoyons à Birgé & Massart (1998), Barron et al. (1999) et Baraud (2002) pour une description détaillée des différentes bases.

(d) Le rôle des inégalités de concentration

D'après l'inégalité (1.32), le processus à contrôler est le processus centré $\nu_n(\hat{f}_{\hat{m}} - f_m)$, pour $m \in \mathcal{M}_n$ et $\nu_n(.)$ défini par (1.31). L'idée pour contrôler

ce processus est de choisir une pénalité qui permette de majorer les variations de $\nu_n(\hat{f}_{\hat{m}} - f_m)$. Cependant, ce processus n'est pas facile à contrôler et nécessite d'utiliser des techniques spécifiques aux processus empirique. Plus précisément,

$$\nu_n(\hat{f}_{\hat{m}} - f_m) \leq ||\hat{f}_{\hat{m}} - f_m||_2 \sup_{f \in \mathcal{B}_{m,\hat{m}}} \nu_n(f),$$

où $\mathcal{B}_{m,m'} = \{f \in S_m + S_{m'} : ||f||_2 \leq 1\}$ pour tout $m, m' \in \mathcal{M}_n$. En remplaçant dans l'inégalité (1.32) et en utilisant l'inégalité classique $2xy \leq b^{-1}x^2/ + by^2$, pour $b > 0$, on obtient

$$||f_0 - \hat{f}_{\hat{m}}||_2^2 \leq ||f_0 - f_m||_2^2 + \text{pen}(m) + b^{-1}||\hat{f}_{\hat{m}} - f_m||_2^2 + b \sup_{f \in \mathcal{B}_{m,\hat{m}}} \nu_n^2(f) - \text{pen}(\hat{m}).$$

L'inégalité de Talagrand définie à l'Annexe A.3.3 est alors l'une des inégalités les plus utilisées pour contrôler en espérance le supremum d'un processus empirique (voir Massart (2007)). On en déduit une inégalité de déviation de la forme :

Inégalité de déviation 1.1.

$$\sum_{m' \in \mathcal{M}_n} \mathbb{E}\Big(\Big(\sup_{f \in \mathcal{B}_{m,m'}} \nu_n^2(f) - (\text{pen}(m) + \text{pen}(m'))\Big)_+\Big) \leq \frac{C}{n}. \tag{1.33}$$

Cette inégalité de déviation précise comment, dans l'heuristique de Mallows (1.26), l'estimateur du biais $||\hat{f}_m||_2^2$ se concentre autour de D_m/n uniformément en $m \in \mathcal{M}_n$.

1.2.3 Estimateurs à noyaux et sélection de fenêtres

Les estimateurs à noyaux sont construits à partir d'une fonction $K : \mathbb{R} \to \mathbb{R}$, d'intégrale 1, appelée *noyau* et d'un paramètre réel $h > 0$, appelé fenêtre. La méthode des noyaux a été introduite par Rosenblatt (1956) pour estimer une densité f_0. Le choix du noyau se fait parmi un ensemble de noyaux fréquemment utilisés (voir Parzen (1962) et Tsybakov & Zaiats (2009), qui définit six noyaux classiques). Le choix de la fenêtre est crucial et a fait l'objet de nombreuses études. Nous présentons deux méthodes de sélection de fenêtres : la validation croisée et la méthode plus récente de Goldenshluger et Lepski.

Dans cette sous-section, nous expliquons l'idée de la méthode des noyaux dans le cas classique d'estimation d'une densité, nous présentons les deux méthodes de sélection de fenêtres, puis nous décrivons les inégalités oracles obtenues avec la méthode de Goldenshluger et Lepski et la démarche pour les obtenir.

Modèle : Soient $Z_1, ..., Z_n$ des variables aléatoires de densité $f_0 \in \mathbb{L}^2(\mathbb{R})$ inconnue à estimer.

Noyau : Le noyau est une fonction $K : \mathbb{R} \to \mathbb{R}$, intégrable sur $\mathbb{R}$ et d'intégrale 1. Rosenblatt (1956) a considéré un noyau rectangulaire, défini par $K(u) = \mathbb{1}_{]-1;1[}(u)/2$. Les noyaux classiques sont le noyau gaussien $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ et le noyau d'Epanechnikov $K(u) = \frac{3}{4}(1-u^2)\ 1_{(|u|\leq 1)}$, parmi les plus utilisés.

Pour tout réel $h > 0$, on note K_h la fonction définie par $K_h : x \mapsto K(x/h)/h$. Le réel $h > 0$ est appelé la *fenêtre* et on introduit la grille de fenêtres $\mathcal{H}_n$.

Collection d'estimateurs : La méthode des noyaux repose sur le principe selon lequel la famille $(K_h)_{h>0}$ forme une approximation de l'unité pour le produit de convolution (voir Komornik (2002) par exemple), i.e. pour tout $z \in \mathbb{R}$, la convolée

$$K_h * f_0(z) := \int_{\mathbb{R}} K_h(z - u) f_0(u) du$$

converge vers $f_0(z)$ en norme $\mathbb{L}^2$, quand h tend vers 0. On peut donc approcher f_0 par $K_h * f_0$ et comme par définition, $K_h * f_0(z) = \mathbb{E}[K_h(z - Z_1)]$, nous en déduisons l'estimateur à noyau de la fonction de densité f_0 : pour $h > 0$ fixé et $z \in \mathbb{R}$,

$$\hat{f}_h(z) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{h}\right). \tag{1.34}$$

On construit ainsi une famille d'estimateurs $(\hat{f}_h)_{h\in\mathcal{H}_n}$.

Procédure de sélection de fenêtres : Le choix de la fenêtre h (*bandwidth* en anglais) est très important pour la qualité de l'estimation. La fenêtre optimale h^* est celle qui minimise l'excès de risque

$$h^* = \arg\min_{h\in\mathcal{H}_n} \mathbb{E}[||\hat{f}_h - f_0||_2^2],$$

c'est donc la fenêtre $h \in \mathcal{H}_n$ qui réalise le meilleur compromis biais-variance dans la décomposition

$$\mathbb{E}[||\hat{f}_h - f_0||_2^2] = ||f_0 - f_h||_2^2 + \mathbb{E}[||\hat{f}_h - f_h||_2^2]. \tag{1.35}$$

Cependant, cette fenêtre n'est pas calculable en pratique puisqu'elle dépend de f_0, qui est inconnue. Plusieurs méthodes existent pour sélectionner cette fenêtre. Nous en présentons ici deux : la méthode de validation croisée et la méthode de Goldenshluger et Lepski, récemment proposée par Goldenshluger & Lepski (2011).

- **Sélection de fenêtres par validation croisée :**

  Une des méthodes classiquement utilisée en pratique pour choisir la fenêtre h d'un estimateur à noyau est la méthode de validation croisée (*cross-validation* en anglais). Cette méthode de sélection de fenêtres a été introduite par Rudemo

(1982) and Bowman (1984) pour l'estimation de la densité par la méthode des noyaux.

L'idée de la validation croisée est de proposer un estimateur de l'excès de risque à partir des données. On a

$$h^* = \arg\min_{h\in\mathcal{H}_n} \mathbb{E}[||\hat{f}_h - f_0||_2^2] = \arg\min_{h\in\mathcal{H}_n} \mathbb{E}\int \hat{f}_h^2(z)\mathrm{d}z - 2\mathbb{E}\int \hat{f}_h(z)f_0(z)\mathrm{d}z.$$

Considérons l'estimateur

$$\mathrm{CV}(h) = \int \hat{f}_h^2(z)\mathrm{d}z - \frac{2}{n}\sum_i \hat{f}_{h,-i}(Z_i), \tag{1.36}$$

où

$$\hat{f}_{h,-i}(Z_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\Big(\frac{z-Z_j}{h}\Big).$$

L'espérance de la somme dans (1.36) vaut

$$\mathbb{E}\left[\frac{1}{n}\sum_i \hat{f}_{h,-i}(Z_i)\right] = \mathbb{E}\left[\int \hat{f}_{h,-n}(z)f_0(z)\mathrm{d}z\right] = \mathbb{E}\left[\int \hat{f}_h(z)f_0(z)\mathrm{d}z\right].$$

On en déduit que $\mathbb{E}[\mathrm{CV}(h)] = \mathbb{E}[||\hat{f}_h - f_0||_2^2] - \int f_0^2(x)\mathrm{d}x$ et donc que $\mathrm{CV}(h) + \int f_0^2(x)\mathrm{d}x$ est un estimateur sans biais de l'erreur quadratique moyenne intégrée $\mathbb{E}[||\hat{f}_h - f_0||_2^2]$. La procédure de sélection de fenêtres par validation croisée est alors la suivante

$$\hat{h}^{CV} = \arg\min_h \mathrm{CV}(h).$$

L'estimateur à noyau de la densité avec une fenêtre choisie par validation croisée est alors défini par $\hat{f}_{\hat{h}^{CV}}$.

Les propriétés établies pour l'estimateur à noyau $\hat{f}_{\hat{h}^{CV}}$ sont asymptotiques. Pour l'estimation d'une densité, Marron & Padgett (1987) ont montré que la fenêtre sélectionnée par validation croisée est optimale au sens où le ratio entre l'erreur quadratique intégrée (ISE pour Integrated Squared Estimator) obtenue par validation croisée et l'infimum de l'ISE obtenu sur toutes les fenêtres converge vers 1 presque sûrement. Hall & Marron (1987) ont donné la vitesse de convergence de l'ISE associée à l'estimateur à noyau obtenu par validation croisée. Ils ont aussi prouvé l'optimalité uniforme de ces vitesses pour une classe assez large de fonctions de densité.

Mais, à notre connaissance, aucun résultat non-asymptotique n'a été démontré pour l'estimateur à noyau avec une fenêtre sélectionnée par validation croisée.

- **La méthode de Goldenshluger et Lepski**

Récemment, Goldenshluger & Lepski (2011) ont proposé une procédure de sélection de fenêtres pour l'estimation à noyau de la densité d'une variable aléatoire multivariée et ont établi des inégalités oracles non-asymptotiques pour l'estimateur ainsi obtenu. Cette procédure a, depuis, été adaptée à différents cadres de travail. Ainsi, elle a par exemple été considérée par Doumic et al. (2012) dans le contexte d'estimation d'un taux de division d'une population structurée en taille dans un cadre non-paramétrique, par Bertin et al. (2013) pour estimer une densité conditionnelle et par Chagny (2014) pour estimer une fonction réelle à l'aide de noyaux déformés.

Décrivons cette méthode dans le cas plus simple de l'estimation d'une densité f_0. Définissons

$$f_h(z) := \mathbb{E}[\hat{f}_h(z)] = \frac{1}{h}\int_{\mathbb{R}} K\Big(\frac{z-u}{h}\Big) f_0(u)\mathrm{d}u = K_h * f_0(z), \tag{1.37}$$

où $K_h(z) = (1/h)K(z/h)$ pour tout $z \in \mathbb{R}$ et $*$ est le produit de convolution usuel. Comme pour la procédure de sélection de modèles, le critère de sélection de fenêtres proposé par Goldenshluger & Lepski (2011) est basé sur une imitation de la décomposition biais-variance (1.35) : il s'agit d'estimer chacun des deux termes de la décomposition biais-variance.

D'après la définition (1.37) de f_h, le terme de variance est facilement majoré par

$$\mathbb{E}[||\hat{f}_h - f_h||_2^2] \leq \frac{1}{nh}\int_{\mathbb{R}} K^2(u)\mathrm{d}u \leq \frac{||K||_2^2}{nh}. \tag{1.38}$$

On cherche un estimateur $\hat{f}_{\hat{h}}$ dont le risque est aussi proche que possible de $\min_{h\in\mathcal{H}_n}\{||f_0 - f_h||_2^2 + ||K||_2^2/nh\}$. Pour approcher le terme de variance, on pose pour tout $h \in \mathcal{H}_n$,

$$V(h) = \kappa\frac{||K||_2^2}{nh}, \tag{1.39}$$

où $\kappa > 0$ est une constante universelle. La spécificité de la méthode de Goldenshluger & Lepski (2011) réside dans la façon d'estimer le biais à partir d'un estimateur intermédiaire faisant intervenir deux fenêtres. Plus précisément, le terme de biais est estimé, pour tout $h \in \mathcal{H}_n$, par

$$A(h) = \max_{h'\in\mathcal{H}_n}\Big\{||\hat{f}_{h'} - \hat{f}_{h,h'}||_2^2 - V(h')\Big\}_+, \tag{1.40}$$

où $\hat{f}_{h,h'}(t) = K_{h'} * \hat{f}_h(t)$ est l'estimateur intermédiaire. On peut expliquer cette estimateur du biais de manière heuristique. L'idée est la suivante : pour être le plus proche possible du terme de biais $||f_0 - K_h * f_0||_2^2$, on remplace la densité inconnue f_0 par l'estimateur $\hat{f}_{h'}$ pour une fenêtre $h' \in \mathcal{H}_n$ fixée et nous obtenons $||\hat{f}_{h'} - K_h * \hat{f}_{h'}||_2^2$. Mais contrairement au biais, cette quantité contient de la

variabilité, on lui retire donc un terme $V(h')$ de l'ordre de la variance. Enfin, comme il n'y a aucune raison de choisir une fenêtre h' plutôt qu'une autre, on balaye l'ensemble de la grille $\mathcal{H}_n$.

La procédure de sélection de la fenêtre est définie par

$$\hat{h} = \underset{h \in \mathcal{H}_n}{\arg\min}\{A(h) + V(h)\}. \tag{1.41}$$

Dans la suite, nous ferons référence à cette méthode sous l'appellation « méthode de Goldenshluger et Lepski ».

Estimateur final : L'estimateur à noyau final obtenu avec la méthode de Goldenshluger et Lepski est défini, à partir de (1.34) et (1.41), par $\hat{f}_{\hat{h}}$.

Nature des résultats :

Des inégalités oracles non-asymptotiques ont été établies par Goldenshluger & Lepski (2011) (pour la norme ℓ_q, $q \geq 1$) et par Bertin et al. (2013) pour l'estimateur $\hat{f}_{\hat{h}}$ de la densité, dont la fenêtre $\hat{h}$ a été sélectionnée par la méthode de Goldenshluger et Lepski. Nous présentons les inégalités oracles obtenues pour cet estimateur ainsi que la démarche pour les établir.

(a) Inégalités oracles pour l'estimateur à noyau avec une fenêtre sélectionnée par la méthode de Goldenshluger et Lepski

L'objectif dans ce contexte est d'établir des inégalités oracles de la forme

$$\mathbb{E}[||f_0 - \hat{f}_{\hat{h}}||_2^2] \leq C \inf_{h \in \mathcal{H}_n}\{||f_0 - \hat{f}_h||_2^2 + V(h)\} + R_n \tag{1.42}$$

où C est une constante universelle et R_n un terme de reste négligeable devant l'infimum. Dans cette inégalité, le terme de biais croît avec h, tandis que le terme de variance $V(h)$ décroît lorsque h augmente.

(b) Démarche pour obtenir une inégalité oracle

On donne ici le début de la démarche pour obtenir une inégalité oracle dans le cas d'un estimateur à noyau $\hat{f}_{\hat{h}}$, obtenu avec la méthode de Goldenshluger et Lepski.

D'après la définition (1.40) de $A(h)$, on peut écrire

$$\begin{aligned} ||\hat{f}_{\hat{h}} - f_0||_2^2 &\leq 3||\hat{f}_{\hat{h}} - \hat{f}_{h,\hat{h}}||_2^2 + 3||\hat{f}_{h,\hat{h}} - \hat{f}_h||_2^2 + 3||\hat{f}_h - f_0||_2^2 \\ &\leq 3A(h) + 3V(\hat{h}) + 3A(\hat{h}) + 3V(h) + 3||\hat{f}_h - f_0||_2^2. \end{aligned}$$

De la définition (1.41) de $\hat{h}$, on en déduit que

$$\mathbb{E}[||\hat{f}_{\hat{h}} - f_0||_2^2] \leq 6\mathbb{E}[A(h)] + 6V(h) + 3\mathbb{E}[||\hat{f}_h - f_0||_2^2]. \tag{1.43}$$

Le dernier terme de cette inégalité est majoré d'après (1.35) et (1.38) par un terme de biais et un terme de variance du même ordre que $V(h)$. Il reste donc à contrôler $\mathbb{E}[A(h)]$ pour obtenir une inégalité oracle de la forme (1.42). Nous renvoyons au paragraphe (d) pour des détails sur le contrôle de $\mathbb{E}[A(h)]$.

(c) Hypothèses sur les noyaux et sur les fenêtres

Pour obtenir une inégalité oracle de la forme (1.42), des hypothèses sur le noyau K et la fenêtre h sont requises.

Hypothèse 1.3.

(i) $||K||_\infty = \sup_{u\in[-1,1]} |K(u)| < \infty$.

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

Les Hypothèses 1.3.(i) et 1.3.(ii) sont standard pour l'estimation à noyau d'une densité (voir Tsybakov (2004) et Goldenshluger & Lepski (2011)). Elles sont satisfaites par la plupart des noyaux standard.

(d) Le rôle des inégalités de concentration

Pour obtenir une inégalité oracle de la forme (1.42) à partir de (1.43), il reste à contrôler $\mathbb{E}[A(h)]$. On décompose $A(h)$ de la façon suivante

$$3\max_{h'\in\mathcal{H}_n}\left\{||\hat{f}_{h'} - f_{h'}||_2^2 - \frac{V(h')}{6}\right\} + 3\max_{h'\in\mathcal{H}_n}\left\{||f_{h,h'} - \hat{f}_{h,h'}||_2^2 - \frac{V(h')}{6}\right\} + 3\max_{h'\in\mathcal{H}_n}||f_{h'} - f_{h,h'}||_2^2.$$

Le dernier terme est facilement majoré. Le contrôle en espérance de chacun des deux premiers termes se fait en utilisant la propriété selon laquelle, pour tout $v \in \mathbb{L}^2(\mathbb{R})$, $||v||_2 = \sup_{f\in\mathcal{B}}\langle v, f\rangle$. On définit alors $\nu_{n,h'}(f) = \langle \hat{f}_{h'} - f_{h'}, f\rangle$, et on en déduit que $||\hat{f}_{h'} - f_{h'}||_2^2 = \sup_{f\in\mathcal{B}} \nu_{n,h'}^2(f)$. En considérant une inégalité de Talagrand définie à l'Annexe A.3.3, nous obtenons, pour l'espérance de chacun des deux premiers termes de la décomposition de $A(h)$, une inégalité de déviation de la forme :

Inégalité de déviation 1.2.

$$\sum_{h'\in\mathcal{H}_n} \mathbb{E}\left[\left\{\sup_{f\in\mathcal{B}} \nu_{n,h'}^2(f) - V(h')\right\}_+\right] \leq \frac{c}{n}. \tag{1.44}$$

Cette inégalité est comparable à l'inégalité de déviation (1.33) en sélection de modèles.

Comparaison des méthodes de sélection de modèles et de Goldenshluger et Lepski :

Chagny (2013), dans sa thèse, a mis en évidence les similitudes entre la sélection de modèles et la méthode de Goldenshluger et Lepski. Nous renvoyons à la Table 1.1 de la thèse de Chagny (2013) pour un tableau récapitulatif des éléments clés de chacune des méthodes dans le cas très simple de l'estimation d'une densité, afin de mettre en évidence ces ressemblances.

1.2.4 Bibliographie en analyse de survie

Nous présentons dans cette sous-section, un état de l'art pour les trois méthodes présentées ci-dessus (Lasso, sélection de modèles et estimation à noyaux) dans le cadre de l'analyse de survie.

1. **La procédure Lasso dans le modèle de Cox usuel**

   Tibshirani (1997) a proposé une procédure Lasso appliquée à la vraisemblance partielle de Cox pour estimer le paramètre de régression $\boldsymbol{\beta_0}$ du modèle de Cox. Cependant, peu de résultats théoriques existent pour cet estimateur Lasso. Nous commençons par citer un premier résultat asymptotique : Bradic et al. (2012) ont établi une inégalité asymptotique en estimation en norme ℓ_2 pour l'estimateur Lasso du paramètre de régression du modèle de Cox semi-paramétrique (1.1). Les résultats non-asymptotiques pour l'estimateur Lasso dans le modèle de Cox n'ont été établis que très récemment. Kong & Nan (2012) and Bradic & Song (2012) ont obtenus des résultats de prédiction sous forme d'inégalités oracles pour la fonction de régression du modèle de Cox non-paramétrique (1.5). Huang et al. (2013) ont, quant à eux, obtenu un résultat en prédiction et des résultats en estimations en normes ℓ_1, ℓ_2 et en norme ℓ_q avec $q \geq 1$ pour le modèle de Cox semi-paramétrique (1.1).

   Des extensions du Lasso ont aussi été proposées pour estimer la partie paramétrique du modèle de Cox (1.1). L'adaptive Lasso a été considéré par Zhang & Lu (2007), qui ont obtenu un résultat asymptotique d'estimation et par Zou (2008), qui a assuré un résultat de consistance en sélection. Antoniadis et al. (2010) ont établi une inégalité asymptotique en estimation pour l'estimateur de Dantzig dans le modèle de Cox semi-paramétrique (1.1) (nous renvoyons à Bickel et al. (2009) pour une comparaison entre l'estimateur Lasso et l'estimateur de Dantzig dans le modèle de régression additif (1.12)).

   Plus généralement, en analyse de survie, le Lasso a aussi été utilisé dans le cas d'un modèle additif d'Aalen (cf. Andersen et al. (1993) pour une description de ce modèle) par Martinussen & Scheike (2009) et Gaïffas & Guilloux (2012). Ces derniers ont obtenu une inégalité oracle non-asymptotique pour la fonction des

covariables du modèle.

2. **La sélection de modèles**

   La méthode d'estimation non-paramétrique par sélection de modèles a été adaptée à l'analyse de survie dans différents modèles. Letué (2000) a adapté cette méthode pour estimer en petite dimension ($p < n$), la fonction de régression f_0 du modèle de Cox non-paramétrique (1.5). Pour cela, elle a considéré le contraste de la log-vraisemblance partielle de Cox associé à une fonction de perte du type information de Kullback-Leibler et elle a obtenu pour son estimateur une inégalité oracle de la forme (1.42). Plus récemment, Brunel & Comte (2005), Brunel et al. (2009) et Brunel et al. (2010) ont considéré la méthode de sélection de modèles pour estimer une densité dans le cas de données censurées.

   Dans un contexte plus proche du nôtre, la méthode de sélection de modèles a été utilisée pour estimer l'intensité d'un processus de comptage dans un modèle à intensité multiplicative d'Aalen (1.3). Reynaud-Bouret (2006) a établi des inégalités oracles pour l'estimateur par minimum de contraste pénalisé de la fonction de risque sans covariable. Comte et al. (2011) ont aussi considéré la sélection de modèles pour estimer une fonction de risque non-paramétrique (en petite dimension). Ils ont appliqué la sélection de modèles en deux directions, selon le temps et selon les covariables, et ont obtenu une inégalité oracle pour leur estimateur.

3. **L'estimation à noyaux**

   L'estimateur à noyau pour la fonction de risque a d'abord été proposé par Watson & Leadbetter (1964a;b) dans le cas de données non-censurées, ses propriétés ont ensuite été étudiées par Rice & Rosenblatt (1976). Ramlau-Hansen (1983b) a introduit l'estimateur à noyau pour estimer l'intensité d'un processus de comptage et a étendu son utilisation à des données de survie censurées. Il a ainsi proposé l'estimateur à noyau (1.9) du risque de base dans le modèle de Cox.

   Comme pour l'estimation d'une densité, le choix de la fenêtre est crucial. Mais très peu de résultats existent en survie. Essentiellement deux méthodes de calibration adaptatif de fenêtre ont été considérées en analyse de survie. La plus ancienne est basée sur la validation croisée. La seconde est basée sur la méthode de Goldenshluger et Lepski. La méthode de validation croisée dans un modèle à intensité multiplicative d'Aalen pour les processus de comptage a été suggérée par Ramlau-Hansen (1981). Grégoire (1993) a prouvé la consistance en $1/n$ de l'estimateur de l'erreur quadratique intégrée associé à l'estimateur à noyau obtenu par validation croisée de l'intensité d'un processus de comptage dans le modèle à intensité multiplicative d'Aalen. Cependant, à notre connaissance, aucun résultat non-asymptotique n'a été établi pour l'estimateur à noyau obtenu

par validation croisée de l'intensité. La méthode de Goldenshluger et Lepski a été considérée par Bouaziz et al. (2013) pour estimer l'intensité de processus de comptage des évènements récurrents (cadre différent de celui qui nous intéresse ici). Ils ont obtenu une inégalité oracle non-asymptotique pour l'estimateur obtenu.

Nous décrivons dans la suite de l'introduction, nos propres contributions.

1.3 Cheminement et principaux résultats

Rappelons le cadre de travail. Nous observons n copies indépendantes $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$, où $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ est le vecteur de covariables, N_i un processus de comptage marqué, Y_i un processus aléatoire à valeurs dans $[0, 1]$ et $[0, \tau]$ un intervalle de temps. À partir de ces observations, nous cherchons à estimer la fonction de risque $\lambda_0(., \boldsymbol{Z_i})$.

I. La Partie I est consacrée à l'estimation en grande dimension de la fonction de risque non-paramétrique λ_0, satisfaisant un modèle à intensité multiplicative d'Aalen défini, pour tout $t \geq 0$, par

$$\Lambda(t) = \int_0^t \lambda_0(s, \boldsymbol{Z}) Y(s) \mathrm{d}s,$$

où $\boldsymbol{Z} \in \mathbb{R}^p$ est le vecteur générique de covariables de même loi que les vecteurs de covariables $\boldsymbol{Z_i}$ pour chaque individu i. Nous estimons cette fonction de risque par un modèle de Cox non-paramétrique constitué de deux paramètres fonctionnels inconnus. Nous proposons de l'estimer à l'aide d'une procédure Lasso simultanée.

II. La Partie II est consacrée à l'estimation du risque de base dans un modèle de Cox semi-paramétrique usuel, où la fonction de risque s'écrit

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z}}.$$

Dans ce modèle, le paramètre $\boldsymbol{\beta_0}$ a une dimension p supérieure à la taille de l'échantillon n. Comme nous l'avons vu dans la Section 1.1.3, l'estimation du risque de base requiert l'estimation préalable du paramètre de régression. Nous avons ainsi proposé des procédures d'estimation du risque de base constituée de deux étapes :

- la première étape, commune aux deux procédures considérées, porte sur l'estimation du paramètre de régression $\boldsymbol{\beta_0}$ à l'aide d'une procédure Lasso, telle que celle décrite par Huang et al. (2013),
- la deuxième étape porte sur la construction d'un estimateur du risque de base α_0. Nous avons proposé deux méthodes d'estimation de α_0. Ces estimateurs appartiennent chacun à l'une des deux grandes familles d'estimateurs non-paramétriques : estimateur par sélection de modèles et estimateur à noyaux. Le choix de la fenêtre de l'estimateur à noyau découle d'une adaptation de la méthode de Goldenshluger et Lepski.

Nous avons implémenté les procédures d'estimation de la Partie II, afin de comparer les deux méthodes d'estimation du risque de base.

Dans les deux parties, nous avons établi des inégalités oracles non-asymptotiques.

D'après les Sections 1.2.1 et 1.2.2, les procédures Lasso et de sélection de modèles requièrent le choix d'un critère d'estimation. Nous définissons les critères de la vraisemblance et des moindres carrés usuels pour les processus de comptage.

La vraisemblance empirique totale (utilisée au Chapitre 2)
D'après la formule de Girsanov ou de Jacod (1973; 1975) (voir également Andersen et al. (1993)), le critère de la log-vraisemblance pour une fonction $\lambda : [0, \tau] \times \mathbb{R}^p \to \mathbb{R}$, est donné par

$$C_n(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \log\lambda(t, \boldsymbol{Z_i})\mathrm{d}N_i(t) - \int_0^\tau \lambda(t, \boldsymbol{Z_i})Y_i(t)\mathrm{d}t\right\}. \tag{1.45}$$

Ce critère d'estimation est utilisé dans le Chapitre 2. Il présente l'avantage de permettre l'estimation de la fonction de risque complète. La fonction de perte classiquement associée à ce critère est la divergence de Kullback (nous renvoyons à la Section 1.3.1 pour des détails concernant cette fonction de perte). Notons que la vraisemblance partielle de Cox, définie par (1.6), est obtenue à partir de la vraisemblance (1.45). Nous renvoyons à l'Annexe A.2 pour les détails concernant les liens entre les deux.

Le contraste des moindres carrés (utilisé au Chapitre 3)
Le critère des moindres carrés pour les processus de comptage a été considéré par Reynaud-Bouret (2006) et Comte et al. (2011). Il est défini pour une fonction $\lambda : [0, \tau] \times \mathbb{R}^p \to \mathbb{R}$ par

$$C_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \lambda^2(t, \boldsymbol{Z_i})Y_i(t)\mathrm{d}t - \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \lambda(t, \boldsymbol{Z_i})\mathrm{d}N_i(t). \tag{1.46}$$

Nous aurions pu utiliser ce critère pour estimer la fonction de risque complète, cependant le processus empirique qui en découle n'est pas linéaire en λ et est donc difficile à contrôler. Nous l'avons plutôt considéré, dans une version modifiée, pour estimer le risque de base du modèle de Cox au Chapitre 3 (voir Section 1.3.2 pour les détails). Le critère des moindres carrés est relié à la norme empirique plus facilement interprétable.

1.3.1 Partie I : Estimation de l'intensité complète d'un processus de comptage par un modèle de Cox

1. Le modèle

Considérons le cadre général des processus de comptage décrit à la Section 1.1 avec pour $i = 1, ..., n$, le processus de comptage marqué N_i satisfaisant le modèle

à intensité multiplicative d'Aalen (1.3). Nous ne faisons aucune hypothèse sur la forme de la fonction de risque λ_0, en particulier, nous ne supposons pas que λ_0 vérifie un modèle de Cox.

D'après la décomposition de Doob-Meyer, on peut heuristiquement écrire que pour tout $i \in \{1, ..., n\}$,

$$\mathrm{d}N_i(t) = \lambda_0(t, \boldsymbol{Z_i})Y_i(t)\mathrm{d}t + \mathrm{d}M_i(t).$$

Si l'on compare au Modèle de régression additif (1.12), $\mathrm{d}M_i(.)$ peut être interprété comme un terme de bruit de type sous-exponentiel.

Notre objectif est d'estimer de manière non-paramétrique la fonction de risque λ_0 par un modèle de Cox non-paramétrique, c'est-à-dire par une fonction de la forme $\lambda(t, \boldsymbol{Z}) = \alpha(t)\mathrm{e}^{f(\boldsymbol{Z})}$, où la fonction du temps α et la fonction des covariables f sont à estimer.

2. La procédure d'estimation

En suivant les étapes du paragraphe 1.2.1 pour le modèle de régression additif, nous décrivons la mise en oeuvre de la procédure Lasso dans le modèle à intensité multiplicative d'Aalen.

Choix du critère d'estimation : Nous avons choisi le critère de la vraisemblance empirique totale (1.45). C'est le critère qui s'est avéré le plus pertinent pour estimer la fonction de risque et obtenir des inégalités oracles pour l'estimateur obtenu. En effet, contrairement au critère des moindres carrés, à partir duquel nous obtenons un processus empirique non linéaire en les fonctions que l'on cherche à estimer, les processus empiriques qui découlent du critère de la vraisemblance total sont linéaires en les fonctions à estimer (nous renvoyons au paragraphe **5.** pour la définition de ces processus).

Dictionnaires de fonctions : Notre modèle fait intervenir, non pas un paramètre à estimer, comme c'est le cas dans le modèle de régression additif ou dans le modèle de Cox non-paramétrique lorsque l'on ne s'intéresse qu'à la fonction de régression, mais deux paramètres à estimer. Nous estimons ces deux paramètres fonctionnels par des combinaisons linéaires de fonctions. Nous introduisons donc deux dictionnaires de fonctions : un dictionnaire de fonctions des covariables $\mathbb{F}_M = \{f_1, ..., f_M\}$, où $f_j : \mathbb{R}^p \to \mathbb{R}$ pour $j = 1, ..., M$ et un dictionnaire des fonctions du temps $\mathbb{G}_N = \{\theta_1, ..., \theta_N\}$, où $\theta_k : \mathbb{R}^+ \to \mathbb{R}$ pour $k = 1, ..., N$. La taille du dictionnaire $\mathbb{F}_M$ utilisé pour estimer la fonction des covariables en grande dimension est grande , i.e. $M \gg n$, alors que pour estimer une fonction sur $\mathbb{R}^+$, nous considérons un dictionnaire $\mathbb{G}_N$ de taille de l'ordre de la taille de l'échantillon n.

Fonctions candidates : Ensuite, comme dans le modèle de régression additif, nous déterminons des fonctions candidates pour estimer l'intensité complète λ_0, mais dans notre cas, non plus à partir d'un seul dictionnaire, mais à partir des deux dictionnaires $\mathbb{F}_M$ et $\mathbb{G}_N$. Comme nous cherchons à estimer la fonction de risque par un modèle de Cox, nous considérons des fonctions candidates de la forme

$$\lambda_{\beta,\gamma}(t, \boldsymbol{Z_i}) = \alpha_\gamma(t)\mathrm{e}^{f_\beta(\boldsymbol{Z_i})},$$

avec $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N$ et

$$\log \alpha_\gamma = \sum_{k=1}^{N} \gamma_k \theta_k \quad \text{et} \quad f_\beta = \sum_{j=1}^{M} \beta_j f_j.$$

Procédure Lasso simultanée : Nous estimons les deux paramètres vectoriels $\boldsymbol{\beta}$ et $\boldsymbol{\gamma}$ simultanément à l'aide d'une procédure Lasso pondéré. Nous avons pour cela considéré la procédure suivante :

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathbb{R}^M\times\mathbb{R}^N}{\arg\min} \{C_n(\lambda_{\beta,\gamma}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma})\}, \tag{1.47}$$

avec

$$\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j| \text{ and } \mathrm{pen}(\boldsymbol{\gamma}) = \sum_{k=1}^{N} \delta_k |\gamma_k|,$$

$(\omega_j)_{j\in\{1,\dots,M\}}$ et $(\delta_k)_{k\in\{1,\dots,N\}}$ sont des poids à déterminer.

Estimateur final : En appliquant la procédure (1.47) au critère de la vraisemblance (1.45), nous obtenons finalement l'estimateur Lasso suivant :

$$\lambda_{\hat{\beta}_L,\hat{\gamma}_L}(t, \boldsymbol{Z_i}) = \alpha_{\hat{\gamma}_L}(t)\mathrm{e}^{f_{\hat{\beta}_L}(\boldsymbol{Z_i})}.$$

3. Les fonctions de perte

La fonction de perte associée au critère d'estimation (1.45) est la divergence de Kullback empirique définie par

$$\widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left(\log \lambda_0(t, \boldsymbol{Z_i}) - \log \lambda_{\beta,\gamma}(t, \boldsymbol{Z_i})\right) \lambda_0(t, \boldsymbol{Z_i}) Y_i(t)\mathrm{d}t \tag{1.48}$$

$$- \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left(\lambda_0(t, \boldsymbol{Z_i}) - \lambda_{\beta,\gamma}(t, \boldsymbol{Z_i})\right) Y_i(t)\mathrm{d}t. \tag{1.49}$$

Dans cette définition, le premier terme (1.48) correspondrait à la divergence de Kullback usuelle si $\lambda_0(t, \boldsymbol{Z_i})$ et $\lambda(t, \boldsymbol{Z_i})$ étaient des densités. Cependant, comme la fonction de risque n'est pas une densité, le terme résiduel (1.49) intervient.

Nous renvoyons à van de Geer (1995) et Senoussi (1988) pour des définitions similaires de divergences de Kullback empiriques.
Les qualités des estimateurs sont également évaluées par la fonction de perte définie pour toute fonction h sur $[0,\tau]\times\mathbb{R}^p$ par

$$||h||_{n,\Lambda}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (h(t,\boldsymbol{Z_i}))^2 \mathrm{d}\Lambda_i(t)}. \tag{1.50}$$

Dans cette définition, plus l'intensité du processus N_i est grande, plus la contribution de l'individu i dans cette norme empirique est importante. Nous avons montré que cette norme est reliée à la divergence de Kullback empirique.

Pour plus de clarté, nous avons résumé dans un tableau comparatif (voir Table 1.1) les éléments clés de la démarche associée à la procédure Lasso pour le Modèle de régression additif (1.12) et pour notre Modèle à intensité multiplicative d'Aalen (1.3) non-paramétrique.

4. Les hypothèses

Comme dans le cas de la régression additive, une hypothèse sur la matrice de Gram est nécessaire afin d'établir une inégalité oracle rapide. Dans notre contexte, la matrice de design est définie par

$$\tilde{\boldsymbol{X}}(t) \;=\; \left[\begin{array}{c|ccc} & \theta_1(t) & \dots & \theta_N(t) \\ \boldsymbol{X} & \vdots & & \vdots \\ & \theta_1(t) & \dots & \theta_N(t) \end{array}\right] \in \mathbb{R}^{n\times(M+N)},$$

où $\boldsymbol{X}=(f_j(\boldsymbol{Z_i}))_{i,j}$ pour $i=1,...,n$ et $j=1,...,M$. Elle fait intervenir les fonctions des deux dictionnaires. La matrice de Gram associée est donnée par

$$\tilde{\boldsymbol{G}}_{\boldsymbol{n}}=\frac{1}{n}\int_0^\tau \tilde{\boldsymbol{X}}(t)^T\tilde{\boldsymbol{C}}(t)\tilde{\boldsymbol{X}}(t)\mathrm{d}t,$$

avec $\tilde{\boldsymbol{C}}(t)=(\mathrm{diag}(\lambda_0(t,\boldsymbol{Z_i})Y_i(t)))_{1\le i\le n}, \forall t\ge 0$. Habituellement, on applique la condition aux valeurs propres restreintes (1.20) à la matrice de Gram. Cependant, la matrice $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ est aléatoire, même conditionnellement aux covariables. Nous imposons donc plutôt une condition aux valeurs propres restreintes à $\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})$, où l'espérance est prise conditionnellement aux covariables.

Hypothèse 1.4. *Pour un entier $s\in\{1,...,M+N\}$ et une constante $r_0>0$, $\mathbb{E}[\tilde{\boldsymbol{G}}_{\boldsymbol{n}}]$ satisfait la condition aux valeurs propres restreintes $\widetilde{\mathbf{RE}}(s,r_0)$ si :*

$$0<\tilde{\boldsymbol{\kappa}}_{\mathbf{0}}(s,r_0)=\min_{\substack{J\subset\{1,...,M+N\},\\ |J|\le s}}\min_{\substack{\boldsymbol{b}\in\mathbb{R}^{M+N}\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\le r_0|\boldsymbol{b}_J|_1}}\frac{(\boldsymbol{b^T}\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2}.$$

	Modèle de régression additif	Modèle de Cox non-paramétrique
Modèle	$\boldsymbol{Y} = f_0(\boldsymbol{Z}) + \boldsymbol{W}$	$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t)\mathrm{e}^{f_0(\boldsymbol{Z})}$
Paramètres inconnus	f_0 fonction des covariables	f_0 fonction des covariables α_0 fonction du temps
Dictionnaires	$\mathbb{F}_M = \{f_1, ..., f_j\}, \quad f_j : \mathbb{R}^p \to \mathbb{R}$	$\mathbb{F}_M = \{f_1, ..., f_j\}, \quad f_j : \mathbb{R}^p \to \mathbb{R}$ $\mathbb{G}_N = \{\theta_1, ..., \theta_N\}, \quad \theta_k : \mathbb{R}^+ \to \mathbb{R}$
Candidats	$f_{\boldsymbol{\beta}}(\boldsymbol{Z}) = \sum_{j=1}^{M} \beta_j f_j(\boldsymbol{Z})$	$\lambda_{\beta,\gamma}(t, \boldsymbol{Z}) = \alpha_\gamma(t)\mathrm{e}^{f_\beta(\boldsymbol{Z})}, \ \log\alpha_\gamma = \sum_{k=1}^{N} \gamma_k\theta_k$ et $f_\beta = \sum_{j=1}^{M} \beta_j f_j$
Critères d'estimation	Moindres carrés $C_n(f_{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\boldsymbol{\beta}}(\boldsymbol{Z_i}))^2$	Log-vraisemblance empirique totale $C_n(\lambda_{\beta,\gamma}) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \log\lambda_{\beta,\gamma}(t, \boldsymbol{Z_i})\mathrm{d}N_i(t) - \int_0^\tau \lambda_{\beta,\gamma}(t, \boldsymbol{Z_i})Y_i(t)\mathrm{d}t\Big\}$
Estimateurs	$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\Big\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f_{\boldsymbol{\beta}}(\boldsymbol{Z_i}))^2 + \mathrm{pen}(\boldsymbol{\beta})\Big\}$	$(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathbb{R}^M\times\mathbb{R}^N}{\arg\min}\{C_n(\lambda_{\beta,\gamma}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma})\}$
Perte empirique	$\lvert\lvert f_{\hat{\boldsymbol{\beta}}} - f_0\rvert\rvert_n^2 = \frac{1}{n}\sum_{i=1}^{n}(f_{\hat{\boldsymbol{\beta}}} - f_0)^2(\boldsymbol{Z_i})$	$\lvert\lvert\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}} - \lambda_0\rvert\rvert_{n,\Lambda}^2 = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}} - \lambda_0)^2(t, \boldsymbol{Z_i})^2\mathrm{d}\Lambda_i(t)$

TABLE 1.1 – *Tableau comparatif pour la procédure Lasso dans le modèle de régression additif et dans le modèle à intensité multiplicative d'Aalen non-paramétrique.*

Nous avons montré, que si la condition $\widetilde{\mathbf{RE}}$ est vérifiée pour $\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})$, alors la version empirique de la condition **RE** appliquée à $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ est satisfaite avec grande probabilité. Cette hypothèse nous permet de relier la norme ℓ_2 à la norme empirique pondérée $||.||_{n,\Lambda}$.
Cette version modifiée de la condition aux valeurs propres restreintes est nouvelle, et c'est la première fois à notre connaissance qu'une inégalité oracle non-asymptotique rapide est établie sous une telle hypothèse.

5. Les processus empiriques et les inégalités de concentration

En suivant la même démarche de preuve que celle détaillée pour le modèle de régression additif, on a pour tout $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in $\mathbb{R}^M \times \mathbb{R}^N$,

$$C_n(\lambda_{\hat{\beta}_L, \hat{\gamma}_L}) + \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) + \text{pen}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) \leq C_n(\lambda_{\beta,\gamma}) + \text{pen}(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\gamma}),$$

et les processus à contrôler sont définis par

$$\eta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t f_j(\boldsymbol{Z_i}) \mathrm{d}M_i(s) \text{ et } \nu_{n,t}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \theta_k(s) \mathrm{d}M_i(s).$$

Pour contrôler ces processus, nous avons établi de nouvelles inégalités de Bernstein empiriques pour les martingales à sauts.

Ces inégalités de Bernstein sont cruciales pour déterminer les poids ω_j et δ_k, de sorte que $|\eta_{n,\tau}(f_j)| \leq \omega_j$ (respectivement $|\nu_{n,\tau}(\theta_k)| \leq \delta_k$) avec grande probabilité. Nous obtenons des poids de l'ordre de

$$\omega_j \approx \sqrt{\frac{\log M}{n} \hat{V}_{n,\tau}(f_j)} \quad \text{et} \quad \delta_k \approx \sqrt{\frac{\log N}{n} \hat{R}_{n,\tau}(\theta_k)}.$$

L'introduction de poids dans la procédure Lasso en analyse de survie avait déjà été considérée par Gaïffas & Guilloux (2012) pour le modèle additif d'Aalen, mais jamais pour un modèle multiplicatif.

6. Le résultat principal

Nous avons établi les premières inégalités oracles non-asymptotiques rapides en divergence de Kullback empirique et en norme empirique pondérée pour l'estimateur Lasso simultanée de la fonction de risque complète en grande dimension sous une nouvelle condition aux valeurs propres restreintes. Nous en donnons ici une version simplifiée.

Résultat 1.1 (Théorème 2.4 dans Lemler (2012)). *Supposons l'hypothèse* $\widetilde{\mathbf{RE}}$ *satisfaite pour certaines constantes par la matrice* $\mathbb{E}[\tilde{\boldsymbol{G}}_{\boldsymbol{n}}]$. *Alors, avec grande probabilité,*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\beta}_L, \hat{\gamma}_L}) \leq (1+\zeta) \inf_{(\boldsymbol{\beta},\boldsymbol{\gamma})} \Big\{ \widetilde{K}_n(\lambda_0, \lambda_{\beta,\gamma}) + \widetilde{C}(\boldsymbol{\beta},\boldsymbol{\gamma}) \max_{\substack{1\leq j\leq M \\ 1\leq k\leq N}} \{\omega_j^2, \delta_k^2\} \Big\},$$

et

$$\begin{aligned} &|| \log\lambda_0 - \log \lambda_{\hat{\beta}_L, \hat{\gamma}_L} ||^2_{n,\Lambda} \\ &\quad \leq (1+\zeta) \inf_{(\boldsymbol{\beta},\boldsymbol{\gamma})} \Big\{ || \log\lambda_0 - \log \lambda_{\beta,\gamma} ||^2_{n,\Lambda} + \widetilde{C}'(\boldsymbol{\beta},\boldsymbol{\gamma}) \max_{\substack{1\leq j\leq M \\ 1\leq k\leq N}} \{\omega_j^2, \delta_k^2\} \Big\}, \end{aligned}$$

où $\widetilde{C} > 0$ *et* $\widetilde{C}' > 0$ *sont des constantes.*

Nous obtenons une inégalité oracle non-asymptotique rapide en prédiction. En effet, la vitesse de convergence de cette inégalité est d'ordre

$$\Big(\max_{\substack{1\leq j\leq M \\ 1\leq k\leq N}} \{\omega_j, \delta_k\} \Big)^2 \approx \max \Big\{ \frac{\log M}{n}, \frac{\log N}{n} \Big\}. \tag{1.51}$$

Cette vitesse de convergence fait apparaître les deux vitesses de convergence que nous aurions attendu si nous avions considéré l'estimation des deux parties du modèle de Cox séparément. Ici, l'estimation de la fonction de risque par un modèle de Cox non-paramétrique implique deux parties différentes : la première partie est le risque de base $\alpha_{\boldsymbol{\gamma}} : \mathbb{R} \to \mathbb{R}$ et la seconde partie est la fonction des covariables $f_{\boldsymbol{\beta}} : \mathbb{R}^p \to \mathbb{R}$. La double pénalisation ℓ_1 considérée vise à estimer simultanément la fonction f_0 des covariables en grande dimension et la fonction non-paramétrique α_0. Comme les deux paramètres f_0 et α_0 sont estimés simultanément, nous obtenons une vitesse de convergence qui fait apparaître les vitesses de convergence de chacune des deux parties de l'estimation. Cependant, si nous considérons le point de vue de Bertin et al. (2011), nous pouvons nous attendre à ce qu'un choix de N d'ordre n permette d'estimer correctement la fonction du temps. Par conséquent, dans un contexte de très grande dimension, le terme dominant dans (1.51) sera d'ordre $\log M/n$, ce qui est la vitesse caractéristique d'une inégalité oracle rapide.

Les travaux de la Partie I font l'objet d'un article soumis pour publication, et accepté aux *Annales de l'Institut Henri Poincaré* :
S. Lemler. (2012), *Oracle inequalities for the lasso in the high-dimensional multiplicative Aalen intensity model.* arXiv preprint arXiv:1206.5628. `http://arxiv.org/pdf/1206.5628v4.pdf`.

1.3.2 Partie II : Estimation adaptative de la fonction de base dans le modèle de Cox

1. **Le modèle**

Nous considérons le modèle de Cox : pour tout $t \in [0, \tau]$

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}},$$

où $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ et $i = 1, ..., n$, et où le vecteur de covariables est de grande dimension (cf. Section 1.1.4). Le risque de base est une fonction du temps de $\mathbb{R}^+$ dans $\mathbb{R}$. Nous avons établi un premier résultat en considérant une procédure Lasso pour estimer les deux paramètres du modèle de Cox, il est aussi intéressant de considérer une procédure en deux étapes en estimant le paramètre de régression $\boldsymbol{\beta_0}$ par une procédure Lasso et en considérant des procédures d'estimation du risque de base qui ne soient pas spécifiques à la grande dimension.

2. **Les procédures d'estimation**

Première étape : estimation du paramètre de régression $\boldsymbol{\beta_0}$

Pour estimer le paramètre de régression $\boldsymbol{\beta_0}$ du modèle de Cox en grande dimension, nous considérons une procédure Lasso appliquée à la vraisemblance partielle de Cox (1.6) :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \{-l_n^*(\boldsymbol{\beta}) + \Gamma_n |\boldsymbol{\beta}|_1\}, \tag{1.52}$$

où Γ_n est d'ordre $\sqrt{\log(pn)/n}$. À partir d'un résultat de Huang et al. (2013), nous avons montré la proposition suivante.

Proposition 1.1. *Soit $k > 0$ et $c > 0$. Supposons que les covariables $Z_{i,j}$ sont bornées pour tout $i \in \{1, ..., n\}$ et $j \in \{1, ..., p\}$. Alors, avec probabilité supérieure à $1 - cn^{-k}$, nous avons*

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

où $C(s) > 0$ est une constante qui dépend de l'indice de sparsité $s = |J_0|$ de $\boldsymbol{\beta_0}$.

Cette proposition donne une borne non-asymptotique en prédiction cruciale pour établir les inégalités oracles pour les deux estimateurs du risque de base.

Deuxième étape : les procédures d'estimation du risque de base α_0

(a) Sélection de modèles :
Nous avons considéré la procédure classique de sélection de modèles.

Critère d'estimation et fonction de perte : Nous avons choisi de travailler avec un critère type moindres carrés :

$$C_n(\alpha, \boldsymbol{\beta}) = -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \alpha(t)\mathrm{d}N_i(t) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha^2(t)\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}Y_i(t)\mathrm{d}t. \quad (1.53)$$

Ce critère ne correspond pas tout à fait au critère des moindres carrés (1.46) appliqué à la fonction de risque $\lambda(t, \boldsymbol{Z_i}) = \alpha(t)\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}$ puisque nous avons enlevé dans chaque terme du contraste C_n, une exponentielle. Nous avons montré qu'il est pertinent de considérer ce critère pour estimer α_0. Le critère (1.53) fait apparaître les deux paramètres du modèle de Cox α et $\boldsymbol{\beta}$. Nous considérons ce critère en $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ défini par (1.52), ceci explique le fait que l'estimateur du risque de base dépend de l'estimateur du paramètre de régression. Associée à ce critère, nous avons considéré la norme déterministe suivante :

$$||\alpha||^2_{det} = \int_0^\tau \alpha^2(t)\mathbb{E}[e^{\boldsymbol{\beta_0^T Z_1}}Y_1(t)]\mathrm{d}t. \quad (1.54)$$

Comme nous l'avons vu, les inégalités oracles reposent sur des comparaisons de norme. Ici, une partie du travail consiste donc à comparer la norme déterministe (1.54) à la norme $\mathbb{L}^2$ standard.

Bases de fonctions et collection de modèles : Soit $\mathcal{M}_n$ un ensemble d'indices et $\{S_m, m \in \mathcal{M}_n\}$ une collection de modèles définis par

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

où $(\varphi_j^m)_{j\in J_m}$ est une base orthonormale de $(\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$. On note D_m le cardinal de S_m, i.e. $|J_m| = D_m$.

Collection d'estimateurs : La collection d'estimateurs $(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}})_{m\in\mathcal{M}_n}$ sur une collection de modèles $\{S_m : m \in \mathcal{M}_n\}$, est telle que :

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

Procédure de sélection de modèles : La procédure de sélection de modèles est donnée par :

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

Estimateur final : L'estimateur par minimum de contraste pénalisé est alors défini par $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$.

(b) Estimateur à noyau et méthode de Goldenshluger et Lepski :

Collection d'estimateurs : Considérons l'estimateur à noyau usuel, introduit par Ramlau-Hansen (1983b), pour estimer le risque de base :

$$\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}N_i(u), \tag{1.55}$$

avec

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad \text{et} \quad S_n(u,\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}} Y_i(u),$$

dans lequel nous injectons l'estimateur Lasso $\hat{\boldsymbol{\beta}}$ défini par (1.52). Nous obtenons ainsi une collection d'estimateurs $(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h)_{h\in\mathcal{H}_n}$, où $\mathcal{H}_n$ est une grille de fenêtres, choisie judicieusement.

Procédure de sélection de fenêtres : Nous proposons une procédure de sélection de fenêtres par la méthode de Goldenshluger et Lepski :

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

avec

$$A^{\hat{\boldsymbol{\beta}}}(h) = \sup_{h'\in\mathcal{H}_n}\Big\{||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}||^2_{2,\varepsilon} - V(h')\Big\}_+$$

et $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(t) = K_{h'} * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h(t)$. La norme $||.||_{2,\varepsilon}$ correspond à la norme sur $\mathbb{L}^2([\varepsilon, \tau - \varepsilon])$, où $\varepsilon = \max\{h : h \in \mathcal{H}_n\}$.

Estimateur final : L'estimateur à noyau avec une fenêtre sélectionnée par la méthode de Goldenshluger et Lepski est alors défini par $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$.

Nous avons contruit un tableau récapitulatif et comparatif des procédures de sélection de modèles et de Goldenshluger et Lepski dans le cas de l'estimation du risque de base du modèle de Cox en grande dimension (voir Table 1.2). Nous renvoyons à la thèse de Chagny (2013) pour un tableau équivalent dans le cas de la densité.

	Méthode de sélection de modèles	Méthode de Goldenshluger et Lepski
Indices	$\mathcal{M}_n$ une collection d'indices m	$\mathcal{H}_n$ une collection de fenêtres h
Collection d'estimateurs	Estimateur par minimum de contraste : $m \in \mathcal{M}_n$, $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}} = \underset{\alpha \in S_m}{\arg\min}\{C_n(\alpha, \hat{\boldsymbol{\beta}})\}$	Estimateur à noyaux : $h \in \mathcal{H}_n$, $\hat{\alpha}_h = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}N_i(u)$
Biais estimé	$C_n(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}, \hat{\boldsymbol{\beta}})$	$A^{\hat{\boldsymbol{\beta}}}(h) = \max_{h' \in \mathcal{H}_n}\Big\{\|\|\hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}} - \hat{\alpha}_{h,h'}^{\hat{\boldsymbol{\beta}}}\|\|_{2,\varepsilon}^2 - V(h')\Big\}_+$
Variance estimée	$\mathrm{pen}(m) = \kappa(1 + \|\|\alpha_0\|\|_{\infty,\tau})\frac{D_m}{n}$	$V(h) = \kappa\|\|\alpha_0\|\|_{\infty,\tau}\|\|K\|\|_2^2\frac{1}{nh}$
Indice sélectionné	$\hat{m} = \underset{m \in \mathcal{M}_n}{\arg\min}\{C_n(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}, \hat{\boldsymbol{\beta}}) + \mathrm{pen}(m)\}$	$\hat{h} = \underset{h \in \mathcal{H}_n}{\arg\min}\{A^{\hat{\boldsymbol{\beta}}}(h) + V(h)\}$
Inégalité de déviation	$\sum_{m' \in \mathcal{M}_n}\mathbb{E}\Big[\Big\{\sup_{\alpha \in \mathcal{B}_{m,m'}^{det}(0,1)}\nu_n^2(\alpha) - p(m,m'))\Big\}_+\mathbb{1}_{\Delta_1}\Big] \leq \frac{c}{n}$	$\sum_{h' \in \mathcal{H}_n}\mathbb{E}\Big[\Big\{\sup_{\alpha \in \mathcal{B}_\tau(h')}\nu_{n,h'}^2(\alpha) - V(h')\Big\}_+\Big] \leq \frac{c}{n}$

TABLE 1.2 – *Tableau comparatif des méthodes de sélection de modèles et de Goldenshluger et Lepski pour l'estimation du risque de base α_0 en grande dimension.*

3. Les résultats théoriques des Chapitres 3 et 4

Nous avons établi des inégalités oracles non-asymptotiques pour chacun des estimateurs.

Résultat 1.2 (Théorème 3.2 du Chapitre 3)**.** *Définissons la pénalité par*

$$\text{pen}(m) := \kappa(1 + ||\alpha_0||_{\infty,\tau})\frac{D_m}{n}, \tag{1.56}$$

où κ est une constante numérique. Sous certaines hypothèses classiques on a

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det}] \leq \kappa_0 \inf_{m\in\mathcal{M}_n}\{||\alpha_0 - \alpha_m||^2_{det} + \text{pen}(m)\} + C_1(s)\frac{\log(np)}{n}, \tag{1.57}$$

où κ_0, $C_1(s)$ est une constante qui dépend de l'indice de sparsité s de $\boldsymbol{\beta_0}$.

Résultat 1.3 (Théorème 4.2 du Chapitre 4)**.** *Pour $h \in \mathcal{H}_n$, où $\mathcal{H}_n$ est une grille de fenêtres, on définit*

$$V(h) = \kappa\frac{||K||_2^2}{nh}, \tag{1.58}$$

où κ est une constante numérique. Sous certaines hypothèses classiques, pour une constante $a > 0$, on a

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{2,\varepsilon}] \leq \tilde{\kappa}_0 \inf_{h\in\mathcal{H}_n}\Big\{||\alpha_h - \alpha_0||^2_{2,\varepsilon} + V(h)\Big\} + C_2(s)\frac{\log^a(n)\log(np)}{n}, \tag{1.59}$$

où $a \geq 0$, $\tilde{\kappa}_0$, $C_2(s)$ est une constante qui dépend de l'indice de sparsité s de $\boldsymbol{\beta_0}$.

Ces deux résultats sont des versions simplifiées des Théorèmes 3.2 du Chapitre 3 et 4.2 du Chapitre 4. Les hypothèses permettant d'établir le Résultat 1.2 sont similaires à celles supposées à la Section 1.2.2, elles sont précisées et commentées au Chapitre 3. L'expression exacte de $V(h)$ dans le Résultat 1.3 est donnée au Chapitre 4 et les hypothèses classiques sous lesquelles l'inégalité (1.59) est établie, y sont détaillées. Commentons les différents termes apparaissant dans ces deux inégalités oracles. Les termes de biais $||\alpha_0 - \alpha_m||^2_{det}$ et $||\alpha_h - \alpha_0||^2_{2,\varepsilon}$ décroissent respectivement lorsque D_m et $1/h$ croissent. Ils dépendent de la régularité de α_0, qui est inconnue, et sont d'autant plus petits que α_0 est régulière. Les termes de variance $\text{pen}(m)$ et $V(h)$ croissent respectivement avec D_m et $1/h$. Ils sont respectivement d'ordres D_m/n et $1/nh$, qui correspondent à l'ordre de la variance sur un modèle ou pour une fenêtre fixée. Ainsi, aux constantes κ_0 et $\tilde{\kappa}_0$ près et lorsque les termes résiduels sont négligeables, on retrouve la même majoration du risque sur un modèle ou pour une fenêtre fixée et le meilleur compromis biais-variance est alors réalisé pour un estimateur de la famille de modèles ou un estimateur avec une fenêtre dans la grille $\mathcal{H}_n$. Les vitesses de convergence

sont alors les mêmes que celles obtenues sur un modèle ou pour une fenêtre, il n'y a pas de perte lorsqu'on applique la procédure de sélection de modèles ou de sélection de fenêtres. Le terme résiduel en $\log(np)/n$, qui apparaît dans ces deux inégalités oracles, provient de l'estimation en grande dimension en les covariables du paramètre de régression $\boldsymbol{\beta_0}$ du modèle de Cox, et plus précisément du contrôle de $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ donné par la Proposition 1.1. Nous renvoyons à la Remarque 1.2, pour une discussion sur le rôle de ce terme sur la vitesse de convergence de l'inégalité, selon l'odre de p par rapport à n. La perte en $\log^a(n)$ qui apparaît dans l'inégalité (1.59) provient du contrôle entre l'estimateur à noyau (1.55), qui dépend de $\hat{\boldsymbol{\beta}}$ et un « pseudo-estimateur », qu'on introduit et qui ne dépend pas de $\hat{\boldsymbol{\beta}}$ (nous renvoyons au Chapitre 4 pour des détails sur ce pseudo-estimateur). Ces deux inégalités sont les premières inégalités oracles établies pour un estimateur du risque de base dans le modèle de Cox lorsque le nombre de covariables est possiblement grand. De plus, l'inégalité (1.59), établit le premier résultat non-asymptotique pour un estimateur à noyau du risque de base. Notons que les Résultats 1.2 et 1.3 restent vrais pour tout p, aussi bien pour un grand nombre de covariables que lorsque $p < n$.

Remarque 1.2. Le terme résiduel en $\log(pn)/n$, commun aux inégalités (1.57) et (1.59), a plus ou moins de poids dans les inégalités, selon l'ordre de p par rapport à n. Il donne l'ordre de grandeur de la perte selon la dimension de p. On retrouve pour p petit les résultats classiques. La vitesse de convergence de ces inégalités oracles est ralentie lorsque le régime de l'ultra grande dimension est atteint, c'est-à-dire lorsque $p \gg n$.

4. **Chapitre 5 : illustration pratique de ces deux méthodes**

Au Chapitre 5, nous avons mené une étude comparative des procédures d'estimation du risque de base, sur des données simulées.

(a) **Les données.** Nous avons généré des données simulées selon un cadre de censure aléatoire à droite. Nous avons considéré deux distributions pour la durée de survie : une distribution de Weibull et une distribution log-normale. Nous avons, de plus, fait varier les paramètres de ces distributions de manière à avoir différentes formes pour le risque de base α_0.

(b) **Première étape : estimation du paramètre de régression.** Pour estimer le paramètre de régression $\boldsymbol{\beta_0}$ du modèle de Cox, nous avons considéré quatre procédures selon la dimension de p et de n :

— lorsque $p \leq \sqrt{n}$, nous avons considéré la procédure classique (1.7) de minimisation de la log-vraisemblance partielle de Cox,

— lorsque $p > n$, nous avons considéré la procédure Lasso (1.52) appliquée à la log-vraisemblance partielle de Cox,

— pour améliorer l'estimation de $\boldsymbol{\beta_0}$, nous avons aussi considéré une procédure Lasso pour sélectionner les variables non nulles, auxquelles nous avons ensuite appliqué la procédure classique (1.7), lorsque le nombre de variables non nulles ne dépasse pas $\sqrt{n}$,

— enfin, nous avons considéré une procédure adaptive Lasso.

Nous avons comparé les erreurs de prédiction et d'estimation de ces différentes méthodes.

(c) **Deuxième étape : estimation du risque de base.** Nous avons ensuite implémenté trois procédures d'estimation du risque de base :

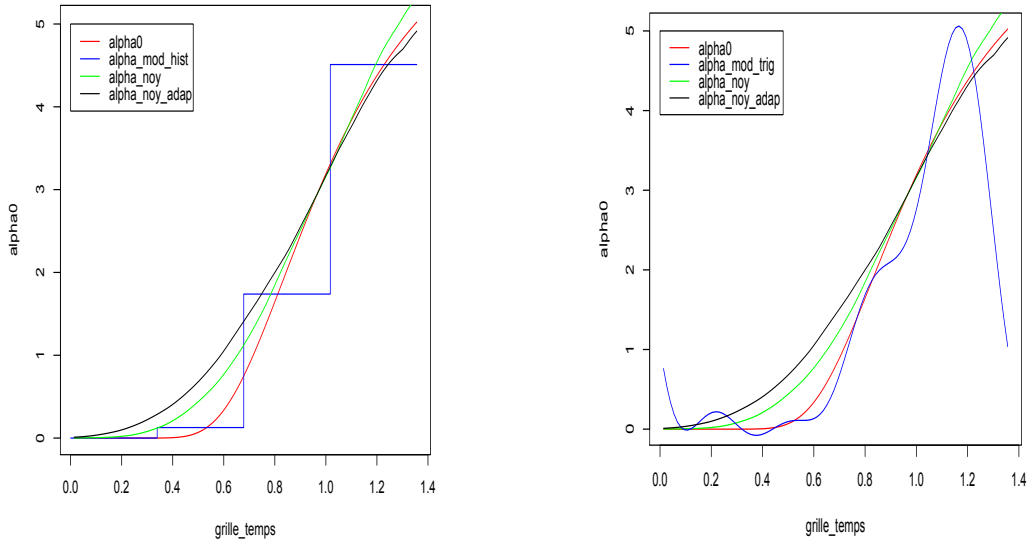
— la sélection de modèles,

— l'estimateur à noyau (1.55) et la méthode de Goldenshluger et Lepski,

— la validation croisée pour sélectionner la fenêtre de l'estimateur à noyau (1.55).

Nous avons calculé les erreurs quadratiques moyennes intégrées (MISE) pour chacun de ces estimateurs et effectué une procédure de Monte Carlo avec 500 réplications, de manière à comparer ces différentes procédures d'estimation. Les courbes de la Figure 1.3 donnent un premier aperçu des performances pratiques de nos estimateurs. Nous avons considéré la base d'histogrammes (à gauche) et la base trigonométrique (à droite) pour la sélection de modèles.

Pour finir, nous avons appliqué nos procédures à un jeu de données réelles sur le cancer du sein, issu de l'étude de Loi et al. (2007).



(a) *Base d'histogrammes.* (b) *Base trigonométrique.*

FIGURE 1.3 – *Courbes du vrai risque de base (en rouge), de l'estimateur à noyau obtenu par validation croisée (en vert), de l'estimateur par minimum de contraste pénalisé (en bleu) et de l'estimateur à noyau adaptatif (en noir). Les courbes ont été obtenues pour :* $n = 500$, $p = \lfloor \sqrt{n} \rfloor$, $\alpha_0 \sim \ln \mathcal{N}(1/4, 0)$, *20% de censure*

Première partie

Estimation of the non-parametric intensity of a counting process by a Cox model with high-dimensional covariates

Chapitre 2

High-dimensional estimation of counting process intensities

Sommaire

2.1 Introduction 53
2.1.1 Framework 53
2.1.2 Previous results 54
2.1.3 Our contribution 55
2.2 Estimation procedure 56
2.2.1 The estimation criterion and the loss function 56
2.2.2 Weighted Lasso estimation procedure 57
2.3 Oracle inequalities for the Cox model for a known baseline function 60
2.3.1 A slow oracle inequality 60
2.3.2 A fast oracle inequality 61
2.3.3 Particular case : variable selection in the Cox model 64
2.4 Oracle inequalities for general intensity 65
2.4.1 A slow oracle inequality 65
2.4.2 A fast oracle inequality 66
2.5 An empirical Bernstein's inequality 68
2.6 Technical results 71
2.6.1 Bernstein concentration inequality 72
2.6.2 Connection between the weighted norm and the Kullback divergence 72
2.7 Proofs 73
Appendices 84
A Connection between the weighted norm and the Kullback divergence 84
B An other empirical Bernstein's inequality 87
C Weighted Lasso procedure in the specific case of the Cox model . 92
D Detailed comparison with the additive regression model 97

Abstract. In a general counting process setting, we consider the problem of obtaining a prognostic on the survival time adjusted on covariates in high-dimension. Towards this end, we construct an estimator of the whole conditional intensity. We estimate it by the best Cox proportional hazards model given two dictionaries of functions. The first dictionary is used to construct an approximation of the logarithm of the baseline hazard function and the second to approximate the relative risk. We introduce a new data-driven weighted Lasso procedure to estimate the unknown parameters of the best Cox model approximating the intensity. We provide non-asymptotic oracle inequalities for our procedure in terms of an appropriate empirical Kullback divergence. Our results rely on an empirical Bernstein's inequality for martingales with jumps and properties of modified self-concordant functions.

Résumé Dans le cadre général d'un processus de comptage, nous nous intéressons à la façon d'obtenir un pronostic sur la durée de survie en fonction des covariables en grande dimension. Pour ce faire, nous construisons un estimateur de l'intensité conditionnelle. Nous l'estimons par le meilleur modèle de Cox étant donné deux dictionnaires de fonctions. Le premier dictionnaire est utilisé pour construire le logarithme du risque de base et le second, pour approximer le risque relatif. Nous introduisons une nouvelle procédure Lasso pondéré avec une pondération basée sur les données pour estimer les paramètres inconnus du meilleur modèle de Cox approximant l'intensité. Nous établissons une inégalité oracle non-asymptotique en divergence de Kullback empirique, qui est la fonction de perte la plus appropriée à notre procédure. Nos résultats reposent sur une inégalité de Bernstein pour les martingales à sauts et sur des propriétés des fonctions self-concordantes.

2.1 Introduction

We consider one of the statistical challenges brought by the recent advances in biomedical technology to clinical applications. For example, in Dave et al. (2004), the considered data relate 191 patients with follicular lymphoma. The observed variables are the survival time, that can be right-censored, clinical variables, as the age or the disease stage, and 44 929 levels of gene expression. In this high-dimensional right-censored setting, there are two clinical questions. One is to determine prognostic biomarkers, the second is to predict the survival from follicular lymphoma adjusted on covariates. We focus our interest on the second (see Gourlay et al. (2012) and Steyerberg et al. (2005)). As a consequence, we consider the statistical question of estimating the whole conditional intensity. To adjust on covariates, the most popular semi-parametric regression model is the Cox proportional hazards model (see Cox (1972)) : the conditional hazard rate function of the survival time T given the vector of covariates $\boldsymbol{Z} = (Z_1, ..., Z_p)^T$ is defined by

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

where $\boldsymbol{\beta_0} = (\beta_{0_1}, ..., \beta_{0_p})^T$ is the vector of regression coefficients and α_0 is the baseline hazard function. The unknown parameters of the model are $\boldsymbol{\beta_0} \in \mathbb{R}^p$ and the function α_0. To construct an estimator of λ_0, one usually considers the partial likelihood introduced by Cox (1972) to derive an estimator of $\boldsymbol{\beta_0}$ and then plug this estimator to obtain the well-known Breslow estimator of α_0. We propose in this chapter an alternative one-step strategy.

2.1.1 Framework

Before describing our strategy, let us clarify our framework. We consider the general setting of counting processes. For $i = 1, ..., n$, let N_i be a marked counting process and Y_i a predictable random process with values in $[0, 1]$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ be the filtration defined by

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

where $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ is the $\mathcal{F}_0$-measurable random vector of covariates of individual i. Let $\Lambda_i(t)$ be the compensator of the process $N_i(t)$ with respect to $(\mathcal{F}_t)_{t\geq 0}$, so that $M_i(t) = N_i(t) - \Lambda_i(t)$ is a $(\mathcal{F}_t)_{t\geq 0}$-martingale.

Assumption 2.1. *The process N_i satisfies the Aalen multiplicative intensity model : for all $t \geq 0$,*

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

where λ_0 is an unknown nonnegative function called intensity.

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).

Remark 2.1 (Censoring case)**.** In the specific case of right censoring, let $(T_i)_{i=1,...,n}$ be independent and identically distributed (i.i.d.) survival times of n individuals and $(C_i)_{i=1,...,n}$ their i.i.d. censoring times. We observe $\{(X_i, \boldsymbol{Z_i}, \delta_i)\}_{i=1,...,n}$ where $X_i = \min(T_i, C_i)$ is the event time, $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T$ is the vector of covariates and $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ is the censoring indicator. The survival times T_i are supposed to be conditionally independent of the censoring times C_i given some vector of covariates $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ for $i = 1, ..., n$. With these notations, the $(\mathcal{F}_t)$-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.

We observe the 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.

Assumption 2.2. *We assume that*

$$A_0 = \sup_{1 \leq i \leq n} \left\{ \int_0^\tau \lambda_0(s, \boldsymbol{Z_i}) \mathrm{d}s \right\} < \infty.$$

This is the standard assumption in statistical estimation of intensities of counting processes, see Andersen et al. (1993) for instance. We also precise that, in the following, we work conditionally to the covariates and from now on, all probabilities $\mathbb{P}$ and expectations $\mathbb{E}$ are conditional to the covariates. Our goal is to estimate λ_0 nonparametrically in a high-dimensional setting, i.e. when the number of covariates p is larger than the sample size n ($p \gg n$).

2.1.2 Previous results

In high-dimensional regression, the benchmarks for results are the ones obtained in the additive regression model. In this setting, Tibshirani (1996) has introduced the Lasso procedure, which consists in minimizing an ℓ_1-penalized criterion. The Lasso estimator has been widely studied for this model, with consistency results (see Meinshausen & Bühlmann (2006)) and variable selection results (see Zhao & Yu (2007), Zhang & Huang (2008a)). Recently, attention has been directed on establishing nonasymptotic oracle inequalities for the Lasso (see Bunea et al. (2006; 2007a), Bickel et al. (2009), Massart & Meynet (2011), Bartlett et al. (2012) and Koltchinskii (2011) among others).

In the setting of survival analysis, the Lasso procedure has been first considered by Tibshirani (1997) and applied to the partial log-likelihood. More generally, other procedures have been introduced for the parametric part of the Cox model : the adaptive Lasso, the smooth clipped absolute deviation penalizations and the Dantzig selector

are respectively considered in Zou (2008), Zhang & Lu (2007), Fan & Li (2002) and Antoniadis et al. (2010). Non parametric approaches are considered in Letué (2000), Hansen et al. (2012) and Comte et al. (2011). Lasso procedures for the alternative Aalen additive model have been introduced in Martinussen & Scheike (2009) and Gaïffas & Guilloux (2012).

All of the existing results in the Cox model are based on the partial log-likelihood, which does not answer the clinical question associated with a prognosis. Antoniadis et al. (2010) have established asymptotic estimation inequalities in the Cox proportional hazard model for the Dantzig estimator (see Bickel et al. (2009) for a comparison between these two estimators in an additive regression model). In Bradic et al. (2012), asymptotic estimation inequalities for the Lasso estimator have also been obtained in the Cox model. More recently, Kong & Nan (2012) and Bradic & Song (2012) have established non-asymptotic oracle inequalities for the Lasso in the generalized Cox model

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \exp(f_0(\boldsymbol{Z})), \tag{2.1}$$

where α_0 is the baseline hazard function and f_0 a function of the covariates. However, the focus in both papers is on the Cox partial log-likelihood, the obtained results are either on $f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0$ or on $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}$ for $f_0(\boldsymbol{Z}) = \boldsymbol{\beta_0^T Z}$ and the problem of estimating the whole intensity λ_0 is not considered, as needed for the prediction of the survival time.

2.1.3 Our contribution

The first motivation of the present chapter is to address the problem of estimating λ_0 defined in (2.1) regardless of an underlying model. We use an agnostic learning approach, see Kearns et al. (1994), to construct an estimator that mimics the performance of the best Cox model, whether this model is true or not. More precisely, we will consider candidates for the estimation of λ_0 of the form

$$\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z}) = \alpha_\gamma(t) \mathrm{e}^{f_{\boldsymbol{\beta}}(\boldsymbol{Z})}, \quad \text{for } (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N,$$

where f_β and α_γ are respectively linear combinations of functions of two dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$. The estimator of λ_0 is defined as the candidate which minimizes a weighted ℓ_1-penalized total log-likelihood as opposed to the Cox partial log-likelihood. The second motivation of the chapter is to obtain non-asymptotic oracle inequalities for Lasso estimators of the complete intensity λ_0. Indeed, in practice, one can not consider that the asymptotic regime has been reached, cf. in Dave et al. (2004) for example. In addition, Comte et al. (2011) established non-asymptotic oracle inequalities for the whole intensity but not in a high-dimensional setting and to the best of our knowledge, no non-asymptotic results for the estimation of the whole intensity in high dimension exist in the literature.

Towards this end, we will proceed in two steps. In a first step, we assume that λ_0 verifies Model (2.1), where α_0 is assumed to be known. In this particular case, the only

nonparametric function to estimate is f_0 and we estimate it by a linear combination of functions of the dictionary $\mathbb{F}_M$. In this setting, we obtain non-asymptotic oracle inequalities for the Cox model when α_0 is supposed to be known. In a second step, we consider the general problem of estimating the whole intensity λ_0. We state non-asymptotic oracle inequalities both in terms of empirical Kullback divergence and weighted empirical quadratic norm for our Lasso estimators, thanks to properties of modified self-concordant functions (see Bach (2010)).

These results are obtained via three ingredients : a new Bernstein's inequality, a modified Restricted Eigenvalue condition on the expectation of the weighted Gram matrix and modified self-concordant functions. Let us be more precise. We establish empirical versions of Bernstein's inequality involving the optional variation for martingales with jumps (see Gaïffas & Guilloux (2012) and Hansen et al. (2012) for related results). This allows us to define a fully data-driven weighted ℓ_1-penalization. For the resulting estimator, we work under a modified Restricted Eigenvalue condition according to which the expectation of a weighted Gram matrix fullfilled the Restricted Eigenvalue condition (see Bickel et al. (2009)). This new version of the Restricted Eigenvalue condition is both new and weaker than the comparable condition in the Cox model. Finally, we extend the notion of self-concordance (see Bach (2010)) to the problem at hands in order to connect our weighted empirical quadratic norm and our empirical Kullback divergence. In this context, we state the first fast non-asymptotic oracle inequality for the whole intensity.

The chapter is organized as follows. In Section 2.2, we describe the framework and the Lasso procedure for estimating the intensity. The estimation risk that we consider and its associated loss function are presented. In Section 2.3, prediction and estimation oracle inequalities in the particular Cox model with known baseline hazard function are stated. In Section 2.4, non-asymptotic oracle inequalities with different convergence rates are given for a general intensity. Section 2.5 is devoted to statement of empirical Bernstein's inequalities associated with our processes. In Section 2.6, we enounce some technical results used in the proofs, which are gathered in section 2.7.

2.2 Estimation procedure

2.2.1 The estimation criterion and the loss function

To estimate the intensity λ_0, we consider the total empirical log-likelihood. By Jacod's Formula (see Andersen et al. (1993)), the log-likelihood based on the data $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$ is given by

$$C_n(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \log\lambda(t, \boldsymbol{Z_i})\mathrm{d}N_i(t) - \int_0^\tau \lambda(t, \boldsymbol{Z_i})Y_i(t)\mathrm{d}t\right\}. \tag{2.2}$$

Our estimation procedure is based on the minimization of this empirical risk. To this empirical risk, we associate the empirical Kullback divergence defined by

$$\widetilde{K}_n(\lambda_0,\lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\log\lambda_0(t,\boldsymbol{Z_i}) - \log\lambda(t,\boldsymbol{Z_i})\right)\lambda_0(t,\boldsymbol{Z_i})Y_i(t)\mathrm{d}t$$
$$-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\lambda_0(t,\boldsymbol{Z_i}) - \lambda(t,\boldsymbol{Z_i})\right)Y_i(t)\mathrm{d}t. \tag{2.3}$$

We refer to van de Geer (1995) and Senoussi (1988) for close definitions. We notice in addition, that this loss function is closed to the Kullback-Leibler information considered in the density framework (see Stone (1994) and Le Pennec & Cohen (2013)). The following proposition justifies the choice of this criterion.

Proposition 2.1. *The empirical Kullback divergence* $\widetilde{K}_n(\lambda_0,\lambda)$ *is nonnegative and equals zero if and only if* $\lambda = \lambda_0$ *almost surely on the interval* $[0, \tau \wedge \sup\{t : \exists i \in \{1,...,n\}, Y_i(t) \neq 0\}]$.

Remark 2.2 (Censoring case)**.** In the specific case of right censoring, the proposition holds true on $[0, \tau \wedge \max_{1\leq i\leq n}(X_i)]$. In this case, we can specify that $\mathbb{P}([0,\tau] \subset [0, \max_{1\leq i\leq n}(X_i)]) = 1 - (1 - S_T(\tau))^n(1 - S_C(\tau))^n$, where S_T and S_C are the survival functions of the survival time T and the censoring time C respectively. From Assumption 2.2, $S_T(\tau) > 0$ and if τ is such that $S_C(\tau) > 0$, then $\mathbb{P}([0,\tau] \subset [0, \max_{1\leq i\leq n}(X_i)])$ is large. See Gill (1983) for a discussion on the role of τ.

In the following, we consider that we estimate $\lambda_0(t)$ for t in $[0, \tau \wedge \sup\{t : \exists i \in \{1,...,n\}, Y_i(t) \neq 0\}]$. Let introduce the weighted empirical quadratic norm defined for all function h on $[0,\tau] \times \mathbb{R}^p$ by

$$||h||_{n,\Lambda} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (h(t,\boldsymbol{Z_i}))^2\mathrm{d}\Lambda_i(t)}, \tag{2.4}$$

where Λ_i is defined in Assumption 2.1. Notice that, in this definition, the higher the intensity of the process N_i is, the higher the contribution of individual i to the empirical norm is. This norm is connected to the empirical Kullback divergence, as it will be shown in Proposition 2.5. Finally, for a vector $\boldsymbol{b}$ in $\mathbb{R}^M$, we define, $|\boldsymbol{b}|_1 = \sum_{j=1}^M |b_j|$ and $|\boldsymbol{b}|_2^2 = \sum_{j=1}^M b_j^2$.

2.2.2 Weighted Lasso estimation procedure

The estimation procedure is based on the choice of two finite sets of functions, called dictionaries. Let $\mathbb{F}_M = \{f_1,...,f_M\}$ where $f_j : \mathbb{R}^p \to \mathbb{R}$ for $j = 1,...,M$, and $\mathbb{G}_N = \{\theta_1,...,\theta_N\}$ where $\theta_k : \mathbb{R}_+ \to \mathbb{R}$ for $k = 1,...,N$, be two dictionaries. Typically

the size of the dictionary $\mathbb{F}_M$ used to estimate the function of the covariates in a high-dimensional setting is large, i.e. $M \gg n$, whereas to estimate a function on $\mathbb{R}_+$, we consider a dictionary $\mathbb{G}_N$ with size N of the order of n. The sets $\mathbb{F}_M$ and $\mathbb{G}_N$ can be collections of functions such as wavelets, splines, step functions, coordinate functions etc. They can also be collections of several estimators computed using different tuning parameters. To make sure that no identification problems appear by using two dictionaries, it is assumed that only the dictionary $\mathbb{G}_N = \{\theta_1, ..., \theta_N\}$ can contain the constant function, not $\mathbb{F}_M = \{f_1, ..., f_M\}$. The candidates for the estimator of λ_0 are of the form

$$\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z_i}) = \alpha_\gamma(t) \mathrm{e}^{f_{\boldsymbol{\beta}}(\boldsymbol{Z_i})}, \quad \text{with } \log \alpha_\gamma = \sum_{k=1}^{N} \gamma_k \theta_k \text{ and } f_{\boldsymbol{\beta}} = \sum_{j=1}^{M} \beta_j f_j,$$

where $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N$.

The dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$ are chosen such that the two following assumptions are fulfilled.

Assumption 2.3.

(i) *For all j in $\{1, ..., M\}$, $||f_j||_{n,\infty} = \max_{1 \le i \le n} |f_j(Z_i)| < \infty$.*

(ii) *For all k in $\{1, ..., N\}$, $||\theta_k||_\infty = \max_{t \in [0,\tau]} |\theta_k(t)| < \infty$.*

We consider a weighted Lasso procedure for estimating λ_0.

Definition 2.1. *The Lasso estimator of λ_0 is defined by $\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}$, where*

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N}{\arg\min} \{C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma})\},$$

with

$$\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j| \text{ and } \mathrm{pen}(\boldsymbol{\gamma}) = \sum_{k=1}^{N} \delta_k |\gamma_k|.$$

The positive weights $\omega_j = \omega(f_j, n, M, \nu, x)$, $j = 1, ..., M$ and $\delta_k = \delta(\theta_k, n, N, \tilde{\nu}, y)$, $k = 1, ..., N$ are defined as follows. Let $x > 0$, $y > 0$, $\varepsilon > 0$, $\tilde{\varepsilon} > 0$, $c = 2\sqrt{2(1+\varepsilon)}$, $\tilde{c} = 2\sqrt{2(1+\tilde{\varepsilon})}$ and $(\nu, \tilde{\nu}) \in (0, 3)^2$ such that $\nu > \Phi(\nu)$ and $\tilde{\nu} > \Phi(\tilde{\nu})$, where $\Phi(u) = \exp(u) - u - 1$. With these notations, the weigths are defined by

$$\omega_j = c\sqrt{\frac{\hat{W}_n^\nu(f_j)(x + \log M)}{n}} + 2\frac{x + \log M}{3n} ||f_j||_{n,\infty}, \tag{2.5}$$

$$\delta_k = \tilde{c}\sqrt{\frac{\hat{T}_n^{\tilde{\nu}}(\theta_k)(y + \log N)}{n}} + 2\frac{y + \log N}{3n} ||\theta_k||_\infty, \tag{2.6}$$

for

$$\hat{W}_n^{\nu}(f_j) = \frac{\nu/n}{\nu/n - \Phi(\nu/n)}\hat{V}_n(f_j) + \frac{x/n}{\nu/n - \Phi(\nu/n)}||f_j||_{n,\infty}^2, \tag{2.7}$$

$$\hat{T}_n^{\tilde{\nu}}(\theta_k) = \frac{\tilde{\nu}/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}\hat{R}_n(\theta_k) + \frac{y/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}||\theta_k||_{\infty}^2, \tag{2.8}$$

where $\hat{V}_n(f_j)$ and $\hat{R}_n(\theta_k)$ are the "observable" empirical variance of f_j and θ_k respectively, given by

$$\hat{V}_n(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}(f_j(\boldsymbol{Z_i}))^2 \mathrm{d}N_i(s) \text{ and } \hat{R}_n(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}(\theta_k(s))^2 \mathrm{d}N_i(s).$$

Remark 2.3. The general Lasso estimator for $\boldsymbol{\beta}$ is classically defined by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\{C_n(\lambda_{\boldsymbol{\beta}}) + \Gamma\sum_{j=1}^{M}|\beta_j|\},$$

with $\Gamma > 0$ a smoothing parameter. Usually, Γ is of order $\sqrt{\log M/n}$ (see Massart & Meynet (2011) for the usual additive regression model and Antoniadis et al. (2010) for the Cox model among other). The Lasso penalization for $\boldsymbol{\beta}$ corresponds to the simple choice $\omega_j = \Gamma$ where $\Gamma > 0$ is a smoothing parameter. Our weights could be compared with those of Bickel et al. (2009) in the case of an additive regression model with a gaussian noise. They have considered a weighted Lasso with a penalty term of the form $\Gamma\sum_{j=1}^{M}||f_j||_n|\beta_j|$, with Γ of order $\sqrt{\log M/n}$ and $||.||_n$ the usual empirical norm. We can deduce from the weights ω_j defined by (2.5) higher suitable weights that can be written $\Gamma_{n,M}^1\tilde{\omega}_j$ with $\tilde{\omega}_j = \sqrt{\hat{W}_n^{\nu}(f_j)}$ defined by (2.7), which is of order $\sqrt{\hat{V}_n(f_j)}$ and

$$\Gamma_{n,M}^1 = c\sqrt{\frac{x+\log M}{n}} + 2\frac{x+\log M}{3n}\max_{1\leq j\leq M}\frac{||f_j||_{n,\infty}}{\sqrt{\hat{W}_n^{\nu}(f_j)}}.$$

The regularization parameter $\Gamma_{n,M}^1$ is still of order $\sqrt{\log M/n}$. The weights $\tilde{\omega}_j$ correspond to the estimation of the weighted empirical norm $||.||_{n,\Lambda}$ that is not observable and play the same role than the empirical norm $||f_j||_n$ in Bickel et al. (2009). These weights are also of the same form as those of van de Geer (2008) for the logistic model.

The idea of adding some weights in the penalization comes from the adaptive Lasso, although it is not the same procedure. Indeed, in the adaptive Lasso (see Zou (2006)) one chooses $\omega_j = |\tilde{\beta}_j|^{-a}$ where $\tilde{\beta}_j$ is a preliminary estimator and $a > 0$ a constant. The idea behind this is to correct the bias of the Lasso in terms of variables selection accuracy (see Zou (2006) and Zhang (2010) for regression analysis and Zhang & Lu (2007) for the Cox model). The weights ω_j can also be used to scale each variable at the same level, which is suitable when some variables have a large variance compared to the others.

Remark 2.4 (Towards practical issues)**.** The actual computation of the estimator $\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}$, although of the greatest interest, is beyond the scope of the present chapter. However, we give here the principal steps to get it. Two types of algorithms could be considered : the cyclical coordinate descent or the proximal gradient descent. As far as we know, maximal algorithms have not yet been implemented for the Cox model (neither for the partial likelihood nor for the total likelihood). On the other hand, cyclical coordinate descent is implemented for the Cox model, e.g. in the R function glmnet, but only for the partial likelihood. In addition, our sum of weighted ℓ_1-penalizations is not usual and require attention when applying the proximal operator in both cyclical coordinate descend and proximal algorithm. Finally, the cross validation procedure will have to consider regularization parameters on a squared grid. This done, we would be able to compare unweighted to weighted procedures in terms of selection, estimation or prediction accuracies. Following the example in an other context of Hansen et al. (2012), we shall expect our weighted procedure to outscore the unweighted one.

2.3 Oracle inequalities for the Cox model for a known baseline function

As a first step, we suppose that the intensity satisfies the generalization of the Cox model (2.1) with a known baseline function α_0. In this context, only f_0 has to be estimated and λ_0 is estimated by

$$\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}(t,\boldsymbol{Z_i}) = \alpha_0(t)\mathrm{e}^{f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}(\boldsymbol{Z_i})} \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^M}{\arg\min}\{C_n(\lambda_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta})\}. \tag{2.9}$$

In this section, we state non-asymptotic oracle inequalities for the prediction loss of the Lasso in terms of the Kullback divergence. These inequalities allow us to compare the prediction error of the estimator and the best approximation of the regression function by a linear combination of the functions of the dictionary in a non-asymptotic way.

2.3.1 A slow oracle inequality

In the following theorem, we state an oracle inequality in the Cox model with slow rate of convergence, i.e. with a rate of convergence of order $\sqrt{\log M/n}$. This inequality is obtained under a very light assumption on the dictionary $\mathbb{F}_M$.

Proposition 2.2. *Consider Model (2.1) with known α_0. Let $x > 0$ be fixed, ω_j be defined by (2.5) and for $\boldsymbol{\beta} \in \mathbb{R}^M$,*

$$\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j|\beta_j|.$$

Let Assumption 2.2 and Assumption 2.3.(i) be satisfied. Then, for a numerical positive constant $A_{\varepsilon,\nu,n}(x)$ defined by

$$A_{\varepsilon,\nu,n}(x) = \frac{2}{\log(1+\varepsilon)} \log\left(2 + \frac{A_0(\nu/n + \Phi(\nu/n))}{x/n}\right) + 1, \tag{2.10}$$

where $\Phi(u) = \mathrm{e}^u - u - 1$, we have with a probability larger than $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x}$,

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) \leq \inf_{\beta \in \mathbb{R}^M} \left(\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\,\mathrm{pen}(\boldsymbol{\beta})\right). \tag{2.11}$$

This proposition states a non-asymptotic oracle inequality in prediction on the conditional hazard rate function in the Cox model. The ω_j are the order of $\sqrt{\log M/n}$ and the penalty term is of order $|\boldsymbol{\beta}|_1\sqrt{\log M/n}$. This variance order is usually referred as a slow rate of convergence in high dimension (see Bickel et al. (2009) for the additive regression model, Bertin et al. (2011) and Bunea et al. (2010) for density estimation).

Remark 2.5. From the definition (2.10) of $A_{\varepsilon,\nu,n}(x)$, when x depends on n and tends towards infinity with n, then $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x}$ goes to 1.

2.3.2 A fast oracle inequality

Now, we are interested in obtaining a non-asymptotic oracle inequality with a fast rate of convergence of order $\log M/n$ and we need further assumptions in order to prove such result. In this subsection, we shall work locally, for $\mu > 0$, on the set $\Gamma_M(\mu) = \{\boldsymbol{\beta} \in \mathbb{R}^M : ||\log \lambda_{\boldsymbol{\beta}} - \log \lambda_0||_{n,\infty} \leq \mu\}$, simply denoted $\Gamma(\mu)$ to simplify the notations and we consider the following assumption :

Assumption 2.4. *There exists $\mu > 0$, such that $\Gamma(\mu)$ contains a non-empty open set of $\mathbb{R}^M$.*

This assumption has already been considered by van de Geer (2008) or Kong & Nan (2012). Roughly speaking, it means that one can find a set where we can restrict our attention for finding good estimator of f_0. This assumption is needed in order to connect, via the notion of self-concordance (see Bach (2010)), the weighted empirical quadratic norm and the empirical Kullback divergence (see Proposition 2.3).

The weighted Lasso estimator becomes

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta} \in \Gamma(\mu)}{\arg\min}\{C_n(\lambda_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta})\}. \tag{2.12}$$

By definition, this weighted Lasso estimator is obtained on a ball centered around the true function λ_0. However in Assumption 2.4, we can always consider a large radius μ, which weakens it. This could not change the rate of convergence in the oracle inequalities ($\sim \log M/n$) but only the range of a constant. In the particular case in

which $\log \lambda_{\boldsymbol{\beta}}$ for all $\boldsymbol{\beta} \in \mathbb{R}^M$ and $\log \lambda_0$ are bounded, there exists $\mu > 0$ such that $|| \log \lambda_{\boldsymbol{\beta}} - \log \lambda_0||_{n,\infty} \leq || \log \lambda_\beta||_{n,\infty} + || \log \lambda_0||_{n,\infty} \leq \mu$.

To achieve a fast rate of convergence, one needs an additional assumption on the Gram matrix. See Bühlmann & van de Geer (2009) and Bickel et al. (2009) for detailed discussions on the different assumptions required for fast oracle inequalities. One of the weakest assumption is the Restricted Eigenvalue condition introduced by Bickel et al. (2009). We choose to work under this Restricted Eigenvalue condition. Let us first introduce further notations :

$$\boldsymbol{\Delta} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}} - \boldsymbol{\beta}) \text{ with } \boldsymbol{\beta} \in \Gamma(\mu) \text{ and } \boldsymbol{D} = (\text{diag}(\omega_j))_{1 \leq j \leq M},$$

$$\boldsymbol{X} = (f_j(\boldsymbol{Z_i}))_{i,j}, \text{ with } i \in \{1, ..., n\} \text{ and } j \in \{1, ..., M\},$$

$$\boldsymbol{G_n} = \frac{1}{n} \boldsymbol{X^T C X} \text{ with } \boldsymbol{C} = (\text{diag}(\Lambda_i(\tau)))_{1 \leq i \leq n}. \tag{2.13}$$

In the matrix $\boldsymbol{G_n}$, the covariates of individual i are re-weighted by its cumulative risk $\Lambda_i(\tau)$, which is consistent with the definition of the empirical norm in (2.4). Let also $J(\boldsymbol{\beta})$ be the sparsity set of vector $\boldsymbol{\beta} \in \Gamma(\mu)$ defined by $J(\boldsymbol{\beta}) = \{j \in \{1, ..., M\} : \beta_j \neq 0\}$, and the sparsity index is then given by $|J(\boldsymbol{\beta})| = \text{Card}\{J(\boldsymbol{\beta})\}$. For $J \subset \{1, ..., M\}$, we denote by $\boldsymbol{\beta}_J$ the vector $\boldsymbol{\beta}$ restricted to the set $J : (\beta_J)_j = \beta_j$ if $j \in J$ and $(\beta_J)_j = 0$ if $j \in J^c$ where $J^c = \{1, ..., M\} \setminus J$.

Usually, in order to obtain a fast oracle inequality, we need to assume a Restricted Eigenvalue condition on the Gram matrix $\boldsymbol{G_n}$. However, since $\boldsymbol{G_n}$ is random in our case, we impose the Restricted Eigenvalue condition to $\mathbb{E}(\boldsymbol{G_n})$, where the expectation is taken conditionally to the covariates.

Assumption 2.5. *For some integer $s \in \{1, ..., M\}$ and a constant $a_0 > 0$, the following condition holds :*

$$0 < \boldsymbol{\kappa}_0(s, a_0) = \min_{\substack{J \subset \{1,...,M\}, \\ |J| \leq s}} \min_{\substack{\boldsymbol{b} \in \mathbb{R}^M \setminus \{0\}, \\ |\boldsymbol{b}_{J^c}|_1 \leq a_0 |\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^T \mathbb{E}(\boldsymbol{G_n}) \boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2}. \tag{$\mathbf{RE(s, a_0)}$}$$

The integer s here plays the role of an upper bound on the sparsity $|J(\boldsymbol{\beta})|$ of a vector of coefficients $\boldsymbol{\beta}$, so that the square submatrices of size less than $2s$ of the expectation of the weighted Gram matrix are positive definite.

This assumption is weaker than the classical one and the following lemma implies that if the Restricted Eigenvalue condition is verified for $\mathbb{E}(\boldsymbol{G_n})$, then the empirical version of the Restricted Eigenvalue condition applied to $\boldsymbol{G_n}$ holds true with large probability. This modified Restricted Eigenvalue condition is new and this is the first time to our best knowledge that a fast-non asymptotic oracle inequality has been established under such a condition.

Lemma 2.1. *Let $L > 0$ such that $\max_{1\le j\le M}\max_{1\le i\le n}|f_j(\boldsymbol{Z_i})| \le L$. Under Assumptions 2.2, 2.3.(i) and $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$, we have*

$$0 < \boldsymbol{\kappa} \le \min_{\substack{J\subset\{1,\dots,M\},\\ |J|\le s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\le a_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b^T G_n b})^{1/2}}{|\boldsymbol{b}_J|_2}, \text{ and } \boldsymbol{\kappa} = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s, a_0), \tag{2.14}$$

with probability larger than $1 - \pi_n$, where

$$\pi_n = 2M^2 \exp\Big[- \frac{n\boldsymbol{\kappa}^4}{2L^2(1+a_0)^2 s(L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2/3)}\Big].$$

Lemma 2.1 assures that the empirical Restricted Eigenvalue condition holds true on an event of large probability, on which we establish a fast non-asymptotic oracle inequality.

Remark 2.6 (Censoring case)**.** In the particular case of the right censoring (see Remark 2.1), we obtain a better version of Lemma 2.1. Indeed, in this case, $\Lambda_i([0,\tau])$ is exponentially distributed with rate parameter 1 and since $\Lambda_i([0,\tau]) \le \Lambda_i([0,\infty])$ almost surely, its expectation is then just less than 1, so that we obtain (2.14) with probability larger than

$$1 - 2M^2 \exp\Big(- \frac{n\boldsymbol{\kappa}^4}{2L^2(1+a_0)^2 s(L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2)}\Big) \text{ with } \boldsymbol{\kappa} = (1/\sqrt{2})\boldsymbol{\kappa_0}(s, a_0).$$

Theorem 2.1. *Consider Model (2.1) with known α_0 and for $x > 0$, let ω_j be defined by (2.5) and $\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}$ be defined by (2.12). Let $\zeta > 0$ and $s \in \{1, ..., M\}$ be fixed. Let Assumptions 2.2, 2.3.(i), 2.4 and $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$ be satisfied with $a_0 = (3 + 4/\zeta)$ and let $\boldsymbol{\kappa} = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s, a_0)$. Then, for a numerical positive constant $A_{\varepsilon,\nu,n}(x)$ defined by (2.10), with a probability larger than $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} - \pi_n$, the following inequality holds*

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \le (1+\zeta) \inf_{\substack{\boldsymbol{\beta}\in\Gamma(\mu)\\ |J(\boldsymbol{\beta})|\le s}} \left\{ \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + C(\zeta,\mu)\frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2}(\max_{1\le j\le M}\omega_j)^2 \right\}, \tag{2.15}$$

where $C(\zeta, \mu) > 0$ is a constant depending on ζ and μ.

This result allows to compare the prediction error of the estimator and the best sparse approximation of the regression function by an oracle that knows the truth, but is constrained by sparsity. The Lasso estimator approaches the best approximation in the dictionary with a fast error term of order $\log M/n$.

Thanks to Proposition 2.3, which states a connection between the empirical Kullback divergence (2.3) and the weighted empirical quadratic norm (2.4), we deduce from Theorem 2.1 a non-asymptotic oracle inequality in weighted empirical quadratic norm.

Corollary 2.1. *Under the assumptions of Theorem 2.1, with a probability larger than* $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} - \pi_n$,

$$\begin{aligned}&||\log\lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}} - \log\lambda_0||^2_{n,\Lambda}\\ &\quad\leq (1+\zeta)\inf_{\substack{\beta\in\Gamma(\mu)\\ |J(\boldsymbol{\beta})|\leq s}}\left\{||\log\lambda_{\boldsymbol{\beta}} - \log\lambda_0||^2_{n,\Lambda} + \tilde{c}(\zeta,\mu)\frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2}(\max_{1\leq j\leq M}\omega_j)^2\right\},\end{aligned}$$

where $\tilde{c}(\zeta,\mu)$ *is a positive constant depending on* ζ *and* μ.

Note that for α_0 supposed to be known, this oracle inequality is also equivalent to

$$||f_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}} - f_0||^2_{n,\Lambda} \leq (1+\zeta)\inf_{\substack{\beta\in\Gamma(\mu)\\ |J(\boldsymbol{\beta})|\leq s}}\left\{||f_{\boldsymbol{\beta}} - f_0||^2_{n,\Lambda} + \tilde{c}(\zeta,\mu)\frac{|J(\boldsymbol{\beta})|}{\boldsymbol{\kappa}^2}(\max_{1\leq j\leq M}\omega_j)^2\right\}.$$

2.3.3 Particular case : variable selection in the Cox model

We now consider the case of variable selection in the Cox model (2.1) with $f_0(Z_i) = \boldsymbol{\beta_0^T Z_i}$. In this case, $M = p$ and the functions of the dictionary are such that for $i = 1,...,n$ and $j = 1,...,p$

$$f_j(\boldsymbol{Z_i}) = Z_{i,j} \text{ and } f_{\boldsymbol{\beta}}(\boldsymbol{Z_i}) = \sum_{j=1}^{p}\beta_j Z_{i,j} = \boldsymbol{\beta^T Z_i}.$$

Let $\boldsymbol{X} = (Z_{i,j})_{\substack{1\leq i\leq n\\ 1\leq j\leq p}}$ be the design matrix and for $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$ defined by (2.9), let

$$\boldsymbol{\Delta_0} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}), \boldsymbol{D} = (\mathrm{diag}(\omega_j))_{1\leq j\leq M}, J_0 = J(\boldsymbol{\beta_0}) \text{ and } |J_0| = \mathrm{Card}\{J_0\}.$$

We now state non-asymptotic inequalities for prediction on $\boldsymbol{X\beta_0}$ and for estimation on $\boldsymbol{\beta_0}$. In this subsection, we don't need to work locally on the set $\Gamma(\mu)$ to obtain Proposition 2.4 and instead of considering Assumption 2.4, we only have to introduce the following assumption to connect the empirical Kullback divergence and the weighted empirical quadratic norm.

Assumption 2.6. *Let* R *be a positive constant, such that* $\max_{i\in\{1,...,n\}}|Z_i|_2 \leq R$.

We consider the Lasso estimator defined with the regularization parameter $\Gamma_1 > 0$:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{C_n(\lambda_{\boldsymbol{\beta}}) + \Gamma_1\sum_{j=1}^{p}\omega_j|\beta_j|\},$$

Theorem 2.2. *Consider Model (2.1) with known α_0. For $x > 0$, let ω_j be defined by (2.5) and denote $\boldsymbol{\kappa'} = (1/\sqrt{2A_0})\boldsymbol{\kappa_0}(s,3)$. Let Assumptions 2.2, 2.3.(i), 2.6 and $\mathbf{RE}(\mathbf{s},\mathbf{a_0})$ with $a_0 = 3$ be satisfied. Let Γ_1 be such that*

$$\Gamma_1 \leq \frac{1}{48Rs}\frac{\min\limits_{1\leq j\leq M}\omega_j^2}{\max\limits_{1\leq j\leq M}\omega_j^2}\frac{\boldsymbol{\kappa'^2}}{\max\limits_{1\leq j\leq M}\omega_j}.$$

Then, for a numerical positive constant $A_{\varepsilon,\nu,n}(x)$ defined by (2.10), we have with a probability larger than $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-\Gamma_1 x} - \pi_n$,

$$||\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})||^2_{n,\Lambda} \leq \frac{4}{\xi^2}\frac{|J_0|}{\boldsymbol{\kappa'^2}}\Gamma_1^2(\max_{1\leq j\leq p}\omega_j)^2 \tag{2.16}$$

and

$$|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_1 \leq 8\frac{\max\limits_{1\leq j\leq p}\omega_j}{\min\limits_{1\leq j\leq p}\omega_j}\frac{|J_0|}{\xi\boldsymbol{\kappa'^2}}\Gamma_1\max_{1\leq j\leq p}\omega_j. \tag{2.17}$$

This theorem gives non-asymptotic upper bounds for two types of loss functions. Inequality (2.16) gives a non-asymptotic bound on prediction loss with a rate of convergence in $\log M/n$, while Inequality (2.17) states a bound on $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}$.

2.4 Oracle inequalities for general intensity

In the previous section, we have assumed α_0 known and have obtained results on the relative risk. Now, we consider a general intensity λ_0 that does not rely on an underlying model. Oracle inequalities are established under different assumptions with slow and fast rates of convergence.

2.4.1 A slow oracle inequality

The slow oracle inequality for a general intensity is obtained under light assumptions that concern only the construction of the two dictionaries $\mathbb{F}_M$ and $\mathbb{G}_N$.

Theorem 2.3. *For $x > 0$ and $y > 0$, let ω_j and δ_k be defined by (2.5) and (2.6) respectively and $(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}})$ be defined in Estimation procedure 2.1. Let Assumption 2.2 and Assumption 2.3.(i)-(ii) be satisfied. Then, for two numerical positive constants $A_{\varepsilon,\nu,n}(x)$ and $B_{\tilde{\varepsilon},\tilde{\nu},n}(y)$ defined respectively by (2.10) and*

$$B_{\tilde{\varepsilon},\tilde{\nu},n}(y) = \frac{2}{\log(1+\tilde{\varepsilon})}\log\Big(2 + \frac{A_0(\tilde{\nu}/n + \Phi(\tilde{\nu}/n))}{y/n}\Big) + 1, \tag{2.18}$$

where $\Phi(u) = \mathrm{e}^u - u - 1$, we have with probability larger than $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} - B_{\tilde{\varepsilon},\tilde{\nu},n}(y)\mathrm{e}^{-y}$,

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) \leq \inf_{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathbb{R}^M\times\mathbb{R}^N}\{\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + 2\,\mathrm{pen}(\boldsymbol{\beta}) + 2\,\mathrm{pen}(\boldsymbol{\gamma})\}. \tag{2.19}$$

We have chosen to estimate the complete intensity, which involves two different parts : the first part is the baseline function $\alpha_{\boldsymbol{\gamma}} : \mathbb{R} \to \mathbb{R}$ and the second part is the function of the covariates $f_{\boldsymbol{\beta}} : \mathbb{R}^p \to \mathbb{R}$. The double ℓ_1-penalization considered here is tuned to concurrently estimate the function f_0 depending on high-dimensional covariates and the non-parametric function α_0. As f_0 and α_0 are estimated at once, the resulting rate of convergence is the sum of the two expected rates in both situations considered separately ($\sim \sqrt{\log M/n} + \sqrt{\log N/n}$). Nevertheless, from Bertin et al. Bertin et al. (2011), we expect that a choice of N of order n would suitably estimate α_0. As a consequence, in a very high-dimensional setting the leading error term in (2.19) would be of order $\sqrt{\log M/n}$, which again is the classical slow rate of convergence in a regression setting.

2.4.2 A fast oracle inequality

We are now interested in obtaining the fast non-asymptotic oracle inequality and as usual, we need to introduce further notations and assumptions. In this subsection, we shall again work locally for $\rho > 0$ on the set $\widetilde{\Gamma}_{M,N}(\rho) = \{(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^M \times \mathbb{R}^N : ||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty} \leq \rho\}$, simply denoted $\widetilde{\Gamma}(\rho)$ and we consider the following assumption:

Assumption 2.7. *There exists $\rho > 0$, such that $\widetilde{\Gamma}(\rho)$ contains a non-empty open set of $\mathbb{R}^M \times \mathbb{R}^N$.*

On $\widetilde{\Gamma}(\rho)$, we define the weighted Lasso estimator as

$$(\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)}{\arg\min} \{C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \text{pen}(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\gamma})\}.$$

Let us give the additional notations. Let $\tilde{\boldsymbol{D}} = \text{diag}(\omega_1, ..., \omega_M, \delta_1, ..., \delta_N)$, $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$, and set $\tilde{\boldsymbol{\Delta}}$ be

$$\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{D}} \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma} \end{pmatrix} \in \mathbb{R}^{M+N}.$$

Let $\mathbf{1}_{n \times N}$ be the matrix $n \times N$ with all coefficients equal to one,

$$\begin{aligned} \tilde{\boldsymbol{X}}(t) &= \left[(f_j(\boldsymbol{Z_i}))_{1 \leq j \leq M, 1 \leq i \leq n} \quad \mathbf{1}_{n \times N}(\text{diag}(\theta_k(t)))_{1 \leq k \leq N}\right] \\ &= \left[\begin{array}{c|ccc} & \theta_1(t) & \dots & \theta_N(t) \\ \boldsymbol{X} & \vdots & & \vdots \\ & \theta_1(t) & \dots & \theta_N(t) \end{array}\right] \in \mathbb{R}^{n \times (M+N)} \end{aligned}$$

and

$$\tilde{\boldsymbol{G}}_{\boldsymbol{n}} = \frac{1}{n} \int_0^\tau \tilde{\boldsymbol{X}}(t)^T \tilde{\boldsymbol{C}}(t) \tilde{\boldsymbol{X}}(t) \mathrm{d}t,$$

with $\tilde{\boldsymbol{C}}(t) = (\text{diag}(\lambda_0(t, \boldsymbol{Z_i})Y_i(t)))_{1\leq i\leq n}, \forall t \geq 0$. Let also $J(\boldsymbol{\beta})$ and $J(\boldsymbol{\gamma})$ be the sparsity sets of vectors $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$ respectively defined by

$$J(\boldsymbol{\beta}) = \{j \in \{1, ..., M\} : \beta_j \neq 0\} \text{ and } J(\boldsymbol{\gamma}) = \{k \in \{1, ..., N\} : \gamma_k \neq 0\},$$

and the sparsity indexes are then given by

$$|J(\boldsymbol{\beta})| = \sum_{j=1}^{M} \mathbb{1}_{\{\beta_j \neq 0\}} = \text{Card}\{J(\boldsymbol{\beta})\} \text{ and } |J(\boldsymbol{\gamma})| = \sum_{k=1}^{N} \mathbb{1}_{\{\gamma_k \neq 0\}} = \text{Card}\{J(\boldsymbol{\gamma})\}.$$

To obtain the fast non-asymptotic oracle inequality, we consider the Restricted Eigenvalue condition applied to the matrix $\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})$.

Assumption 2.8. *For some integer $s \in \{1, ..., M + N\}$ and a constant $r_0 > 0$, we assume that $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ satisfies:*

$$0 < \tilde{\boldsymbol{\kappa}}_{\mathbf{0}}(s, r_0) = \min_{\substack{J\subset\{1,...,M+N\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^{M+N}\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\leq r_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2}. \qquad (\widetilde{\mathbf{RE}}(\mathbf{s}, \mathbf{r_0}))$$

The condition on the matrix $\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})$ is rather strong because the block matrix involves both functions of the covariates of $\mathbb{F}_M$ and functions of time which belong to $\mathbb{G}_N$. This is the price to pay for an oracle inequality on the full intensity. If we had instead considered two restricted eigenvalue assumptions on each block, we would have established an oracle inequality on the sum of the two unknown parameters α_0 and f_0 and not on λ_0. As in Lemma 2.1, we can show that under Assumption $\widetilde{\mathbf{RE}}(\mathbf{s}, \mathbf{r_0})$, we have an empirical Restricted Eigenvalue condition on the matrix $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$.

Lemma 2.2. *Let L defined as in Lemma 2.1. Under Assumptions 2.2, 2.3.(i)-(ii) and $\widetilde{\mathbf{RE}}(\mathbf{s}, \mathbf{r_0})$, we have*

$$0 < \tilde{\boldsymbol{\kappa}} \leq \min_{\substack{J\subset\{1,...,M\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^{M}\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\leq r_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2} \text{ and } \tilde{\boldsymbol{\kappa}} = (1/\sqrt{2A_0})\tilde{\boldsymbol{\kappa}}_{\mathbf{0}}(s, r_0), \qquad (2.20)$$

with probability larger than $1 - \tilde{\pi}_n$, where

$$\tilde{\pi}_n = 2M^2 \exp\Big[-\frac{n\tilde{\boldsymbol{\kappa}}^4}{2L^2(1 + r_0)^2 s(L^2(1 + r_0)^2 s + \tilde{\boldsymbol{\kappa}}^2/3)}\Big].$$

Theorem 2.4. *For $x > 0$ and $y > 0$, let ω_j and δ_k be defined by (2.5) and (2.6) respectively. Let $\zeta > 0$ and $s \in \{1, ..., M + N\}$ be fixed. Let Assumptions 2.2, 2.3.(i)-(ii), 2.7 and $\widetilde{\mathbf{RE}}(\mathbf{s}, \mathbf{r_0})$ be satisfied with*

$$r_0 = \left(3 + 8\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)/\zeta\right),$$

and let $\tilde{\boldsymbol{\kappa}} = (1/\sqrt{2A_0})\tilde{\boldsymbol{\kappa}}_{\mathbf{0}}(s, r_0)$. *Then, for two numerical positive constants* $A_{\varepsilon,\nu,n}(x)$ *and* $B_{\tilde{\varepsilon},\tilde{\nu},n}(y)$ *defined respectively by (2.10) and (2.18), we have with probability larger than* $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} - B_{\tilde{\varepsilon},\tilde{\nu},n}(y)\mathrm{e}^{-y} - \tilde{\pi}_n$,

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}}) \leq (1+\zeta) \inf_{\substack{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\widetilde{\Gamma}(\rho) \\ \max(|J(\boldsymbol{\beta})|,|J(\boldsymbol{\gamma})|)\leq s}} \Big\{ \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \widetilde{C}(\zeta,\rho) \frac{(|J(\boldsymbol{\beta})| \vee |J(\boldsymbol{\gamma})|)}{\tilde{\boldsymbol{\kappa}}^2} \max_{\substack{1\leq j\leq M \\ 1\leq k\leq N}} \{\omega_j^2, \delta_k^2\} \Big\}, \quad (2.21)$$

and

$$||\log\lambda_0 - \log\lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}}||^2_{n,\Lambda} \leq (1+\zeta) \inf_{\substack{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\widetilde{\Gamma}(\rho) \\ \max(|J(\boldsymbol{\beta})|,|J(\boldsymbol{\gamma})|)\leq s}} \Big\{ ||\log\lambda_0 - \log\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}||^2_{n,\Lambda} + \widetilde{C}'(\zeta,\rho) \frac{(|J(\boldsymbol{\beta})| \vee |J(\boldsymbol{\gamma})|)}{\tilde{\boldsymbol{\kappa}}^2} \max_{\substack{1\leq j\leq M \\ 1\leq k\leq N}} \{\omega_j^2, \delta_k^2\} \Big\}, \quad (2.22)$$

where $\widetilde{C}(\zeta,\rho) > 0$ *and* $\widetilde{C}'(\zeta,\rho) > 0$ *are constants depending only on* ζ *and* ρ.

We obtain a non-asymptotic fast oracle inequality in prediction. Indeed, the rate of convergence of this oracle inequality is of order

$$\Big(\max_{\substack{1\leq j\leq M \\ 1\leq k\leq N}} \{\omega_j, \delta_k\}\Big)^2 \approx \max\Big\{\frac{\log M}{n}, \frac{\log N}{n}\Big\},$$

namely, if we choose $\mathbb{G}_N$ of size n, the rate of convergence of this oracle inequality is then of order $\log M/n$ (see Subsection 2.4.1 for more details). While Estimation procedure 2.1 allows to derive a prediction for the survival time through the conditional intensity, Theorem 2.4 measures the accuracy of this prediction. In that sense, the clinical problem of establishing a prognosis has been addressed at this point. To our best knowledge, this oracle inequality is the first non-asymptotic oracle inequality in prediction for the whole intensity with a fast rate of convergence of order $\log M/n$.

For the part depending on the covariates, recent results establish non-asymptotic oracle inequalities for the Lasso estimator of f_0 in the usual Cox model (see Bradic & Song (2012) and Kong & Nan (2012)). We cannot compare our results to theirs, since we estimate the whole intensity with the total empirical log-likelihood whereas both of them consider the partial log-likelihood.

The remaining part of the chapter is devoted to the technical results and proofs

2.5 An empirical Bernstein's inequality

The main ingredient of Proposition 2.2 and Theorems 2.1, 2.3 and 2.4 are Bernstein's concentration inequalities that we present in this section. To clarify the relation

between the stated oracle inequalities and the Bernstein's inequality, we sketch here the proof of Theorem 2.3. Using the Doob-Meyer decomposition $N_i = M_i + \Lambda_i$, we can easily show that for all $\boldsymbol{\beta} \in \mathbb{R}^M$ and for all $\boldsymbol{\gamma} \in \mathbb{R}^N$

$$\begin{aligned} C_n(\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) - C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) &= \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) - \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) \\ &\quad - (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}} - (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}, \end{aligned} \tag{2.23}$$

where

$$\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \vec{\boldsymbol{f}}(\boldsymbol{Z_i}) \mathrm{d}M_i(t) \quad \text{and} \quad \boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \vec{\boldsymbol{\theta}}(t) \mathrm{d}M_i(t), \tag{2.24}$$

with $\vec{\boldsymbol{f}} = (f_1, ..., f_M)^T$ and $\vec{\boldsymbol{\theta}} = (\theta_1, ..., \theta_N)^T$. By definition of the Lasso estimator, we have for all $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in $\mathbb{R}^M \times \mathbb{R}^N$

$$C_n(\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) + \mathrm{pen}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) \leq C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma}),$$

and we finally obtain

$$\begin{aligned} \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) &+ (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}} + \mathrm{pen}(\boldsymbol{\gamma}) - \mathrm{pen}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) \\ &+ (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}). \end{aligned}$$

Consequently, $\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}})$ is bounded by

$$\begin{aligned} \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) &+ \sum_{j=1}^{M} (\hat{\beta}_{L,j} - \beta_j)\eta_{n,\tau}(f_j) + \sum_{j=1}^{M} \omega_j(|\beta_j| - |\hat{\beta}_{L,j}|) \\ &+ \sum_{k=1}^{N} (\hat{\gamma}_{L,k} - \gamma_k)\nu_{n,\tau}(\theta_k) + \sum_{k=1}^{N} \delta_k(|\gamma_k| - |\hat{\gamma}_{L,k}|), \end{aligned}$$

with

$$\eta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t f_j(\boldsymbol{Z_i}) \mathrm{d}M_i(s) \text{ and } \nu_{n,t}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n} \int_0^t \theta_k(s) \mathrm{d}M_i(s). \tag{2.25}$$

We will control $\eta_{n,\tau}(f_j)$ and $\nu_{n,\tau}(\theta_k)$ respectively by ω_j and δ_k. More precisely, the weights ω_j (respectively δ_k) will be chosen such that $|\eta_{n,\tau}(f_j)| \leq \omega_j$ (respectively $|\nu_{n,\tau}(\theta_k)| \leq \delta_k$) and $\mathbb{P}(|\eta_{n,\tau}(f_j)| > \omega_j)$ (respectively $\mathbb{P}(|\nu_{n,\tau}(\theta_k)| > \delta_k)$ large. As $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ involve martingales, we could directly apply classical Bernstein's inequalities for martingales (see van de Geer (1995), Lemma 2.1) and following the beginning of the proof of Theorem 3 in Gaïffas & Guilloux (2012), with $x > 0$ and $y > 0$ and any $0 < w < \infty$, $0 < \tilde{w} < \infty$, we would obtain

$$\mathbb{P}\left[\eta_{n,t}(f_j) \geq \sqrt{\frac{2wx}{n}} + \frac{x}{3n}, V_{n,t}(f_j) \leq w\right] \leq \mathrm{e}^{-x}$$

$$\mathbb{P}\left[\nu_{n,t}(\theta_k) \geq \sqrt{\frac{2\tilde{w}y}{n}} + \frac{y}{3n}, R_{n,t}(\theta_k) \leq \tilde{w}\right] \leq \mathrm{e}^{-y},$$

where the predictable variations $V_{n,t}(f_j)$ and $R_{n,t}(\theta_k)$ of $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ are respectively defined by

$$V_{n,t}(f_j) = n < \eta_n(f_j) >_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (f_j(\boldsymbol{Z_i}))^2 \lambda_0(t, \boldsymbol{Z_i}) Y_i(s)\mathrm{d}s,$$
$$R_{n,t}(\theta_k) = n < \nu_n(\theta_k) >_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (\theta_k(t))^2 \lambda_0(t, \boldsymbol{Z_i}) Y_i(s)\mathrm{d}s.$$

Applying these inequalities, the weights of Definition 2.1 would have the forms $\omega_j = \sqrt{2V_{n,t}(f_j)x/n} + x/3n$ and $\delta_k = \sqrt{2R_{n,t}(\theta_k)y/n} + y/3n$. As $V_{n,t}(f_j)$ and $R_{n,t}(\theta_k)$ both depend on λ_0, this would not result a statistical procedure. We propose to replace in the Bernstein's inequality the predictable variations by the optional variations of the processes $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ defined by

$$\hat{V}_{n,t}(f_j) = n[\eta_n(f_j)]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (f_j(\boldsymbol{Z_i}))^2 \mathrm{d}N_i(s) \tag{2.26}$$

and

$$\hat{R}_{n,t}(\theta_k) = n[\nu_n(\theta_k)]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (\theta_k(t))^2 \mathrm{d}N_i(s). \tag{2.27}$$

This ensures that the weights ω_j and δ_k will depends on $\hat{V}_{n,t}(f_j)$ and $\hat{R}_{n,t}(\theta_k)$ respectively. Equivalent strategies in different models have been considered in Gaïffas & Guilloux (2012) or Hansen et al. (2012). The following theorem states the resulting Bernstein's inequalities.

Theorem 2.5. *Let Assumption 2.2 be satisfied. For any numerical constant $\varepsilon > 0$, $\tilde{\varepsilon} > 0$, $c = \sqrt{2(1+\varepsilon)}$ and $\tilde{c} = \sqrt{2(1+\tilde{\varepsilon})}$, the following holds for any $x > 0$, $y > 0$:*

$$\mathbb{P}\left[|\eta_{n,t}(f_j)| \geq c\sqrt{\frac{\hat{W}_n^{\nu}(f_j)x}{n}} + \frac{x}{3n}||f_j||_{n,\infty}\right] \leq \left(\frac{2}{\log(1+\varepsilon)}\log\left(2 + \frac{A_0(\nu/n + \Phi(\nu/n))}{x/n}\right) + 1\right)\mathrm{e}^{-x}, \tag{2.28}$$

$$\mathbb{P}\left[|\nu_{n,t}(\theta_k)| \geq \tilde{c}\sqrt{\frac{\hat{T}_n^{\tilde{\nu}}(\theta_k)y}{n}} + \frac{y}{3n}||\theta_k||_{\infty}\right] \leq \left(\frac{2}{\log(1+\tilde{\varepsilon})}\log\left(2 + \frac{A_0(\tilde{\nu}/n + \Phi(\tilde{\nu}/n))}{y/n}\right) + 1\right)\mathrm{e}^{-y}, \tag{2.29}$$

where

$$W_n^{\nu}(f_j) = \frac{\nu/n}{\nu/n - \Phi(\nu/n)}\hat{V}_n(f_j) + \frac{x/n}{\nu/n - \Phi(\nu/n)}||f_j||^2_{n,\infty}, \tag{2.30}$$

$$T_n^{\tilde{\nu}}(\theta_k) = \frac{\tilde{\nu}/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}\hat{R}_n(\theta_k) + \frac{y/n}{\tilde{\nu}/n - \Phi(\tilde{\nu}/n)}||\theta_k||^2_{\infty}, \tag{2.31}$$

for real numbers $(\nu, \tilde{\nu}) \in (0,3)^2$ *such that* $\nu > \Phi(\nu)$ *and* $\tilde{\nu} > \Phi(\tilde{\nu})$*, where* $\Phi(u) = \exp(u) - u - 1$.

We deduce the weights ω_j and δ_k defined in (2.5) and (2.6) respectively, from Theorem 2.5. These empirical Bernstein's inequalities hold true for martingales with jumps, when the predictable variation is not observable.

Remark 2.7. Theorem 2.5 is closed to Theorem 3 in Hansen et al. (2012), although in our version the event bounding $\hat{W}_n^{\nu}(f_j)$ and $\hat{T}_n^{\tilde{\nu}}(\theta_k)$ has been removed from the probability (see the proof of Theorem 2.5).

Other weights can also be obtained from empirical Bernstein's inequalities that are closer to those obtained by Gaïffas & Guilloux (2012) in Theorem 3. We refer to Appendix B, in which we provide the other Bernstein's inequalities adapted from Gaïffas & Guilloux (2012) to our processes, and also to an other version of the paper (see Lemler (2012)), in which these weights appear. Their forms are less simple than those defined in (2.5) and (2.6), but they do not depend on tuning parameters ν and $\tilde{\nu}$ to determine for the applications. An interesting perspective would be to determine which one of those two forms of weights gives the best results in the applications.

Remark 2.8 (Censoring case)**.** In the specific case of right censoring, since $\max_{1\leq i\leq n}|N_i(\tau)| \leq 1$, we can directly apply the Bernstein type inequality for martingales of Hansen et al. (2012) to get quite simpler right term in Inequality (2.28). Indeed in this case, for real numbers $(u, v) \in (0,3)^2$ such that $u > \phi(u)$ and $v > \phi(v)$, where $\phi(z) = \exp(z) - z - 1$, and $c_{1,\varepsilon} = \sqrt{2(1+\varepsilon)}$, we would get

$$\mathbb{P}\Big[\eta_{n,t}(f_j) \geq c\sqrt{\frac{\hat{W}_n^u(f_j)x}{n}} + \frac{x}{3n}||f_j||_{n,\infty}\Big] \leq 4\Big(\frac{\log(1+u/x)}{\log(1+\varepsilon)} + 1\Big)\mathrm{e}^{-x}. \tag{2.32}$$

2.6 Technical results

In this section, we present the technical results, that are not useful for a first reading of the chapter but useful for a better understanding of the theory and used in the proofs of Section 2.7. Associated proofs are in Appendix A.

2.6.1 Bernstein concentration inequality

We recall here the classical Bernstein concentration inequality (see Proposition 2.9 in Massart (2007)).

Theorem 2.6. *Let $\zeta_1, ..., \zeta_n$ be n independent real valued random variables. Assume that there exist some positive numbers v and c such that for all integers $m \geq 2$*

$$\sum_{i=1}^{n} \mathbb{E}[|\zeta_i|^m] \leq m! v c^{m-2}. \tag{2.33}$$

For any positive x we have

$$\mathbb{P}(\sum_{i}^{n} (\zeta_i - \mathbb{E}(\zeta_i)) \geq x) \leq \exp\Big(- \frac{x^2}{2(v + cx)}. \Big) \tag{2.34}$$

Note that if the variables ζ_i are bounded, $|\zeta_i| \leq b$ for all i in $\{1, ..., n\}$, then assumption (2.33) is satisfied with

$$v = \sum_{i=1}^{n} \mathbb{E}[\zeta_i^2] \quad \text{and} \quad c = b/3.$$

2.6.2 Connection between the weighted norm and the Kullback divergence

The following propositions connect the empirical Kullback divergence (2.3) to the weighted empirical norm (2.4) in the different cases considered in the chapter.

Proposition 2.3 holds true when the intensity verifies Model 2.1 with a known baseline hazard function α_0.

Proposition 2.3. *Under Assumption 2.4, for all $\boldsymbol{\beta} \in \Gamma(\mu)$,*

$$\mu' || \log \lambda_{\boldsymbol{\beta}} - \log \lambda_0 ||_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) \leq \mu'' || \log \lambda_{\boldsymbol{\beta}} - \log \lambda_0 ||_{n,\Lambda}^2,$$

where $\mu' = \phi(\mu)/\mu^2$, $\mu'' = \phi(-\mu)/\mu^2$ and $\phi(t) = \mathrm{e}^{-t} + t - 1$.

Proposition 2.3 can be rewriten in the particular case of variable selection in the Cox model as follows :

Proposition 2.4. *Under Assumptions 2.6 and $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$ with $a_0 = 3$, there exist two positive numerical constants ξ and ξ' such that*

$$\xi || (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^{\boldsymbol{T}} \boldsymbol{X} ||_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) \leq \xi' || (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^{\boldsymbol{T}} \boldsymbol{X} ||_{n,\Lambda}^2.$$

In the general case, when the intensity does not rely on an underlying model, the connection between the weighted empirical norm (2.4) and the empirical Kullback divergence (2.3) is given by the following proposition.

Proposition 2.5. *Under Assumption 2.7, for all $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$,*

$$\rho' || \log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0 ||_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) \leq \rho'' || \log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0 ||_{n,\Lambda}^2,$$

where $\rho' = \phi(\rho)/\rho^2$, $\rho'' = \phi(-\rho)/\rho^2$ and $\phi(t) = \mathrm{e}^{-t} + t - 1$.

2.7 Proofs

2.7.1 Proof of Proposition 2.1

Following the proof of Theorem 1 in Senoussi (1988), we rewrite the empirical Kullback divergence (2.3) as

$$\begin{aligned}\widetilde{K}_n(\lambda_0,\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\log\lambda_0(t,\boldsymbol{Z_i}) - \log\lambda(t,\boldsymbol{Z_i}) - \left(1-\frac{\lambda(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})}\right)\right]\lambda_0(t,\boldsymbol{Z_i})Y_i(t)\mathrm{d}t \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\exp\left(\log\frac{\lambda(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})}\right) - \log\frac{\lambda(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})} - 1\right]\lambda_0(t,\boldsymbol{Z_i})Y_i(t)\mathrm{d}t.\end{aligned}$$

Since $t \to \mathrm{e}^t - t - 1 > 0$, except for $t = 0$, we deduce that, except for $\lambda = \lambda_0$,

$$\exp\left(\log\frac{\lambda(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})}\right) - \log\frac{\lambda(t,\boldsymbol{Z_i})}{\lambda_0(t,\boldsymbol{Z_i})} - 1 > 0.$$

Thus $\widetilde{K}_n(\lambda_0,\lambda)$ is positive and vanishes only if $(\log\lambda_0 - \log\lambda)(t,\boldsymbol{Z_i}) = 0$ almost surely, namely if $\lambda_0 = \lambda$ almost surely. □

2.7.2 Proof of Proposition 2.2

According to the definition (2.9) of $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$, for all $\boldsymbol{\beta}$ in $\mathbb{R}^M$, we have

$$C_n(\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) \le C_n(\lambda_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta}).$$

Here α_0 is assumed to be known. Hence applying (2.23), we obtain

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) \le \widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}). \tag{2.35}$$

It remains to control the term $(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}$. For ω_j defined in (2.5), set

$$\mathcal{A} = \bigcap_{j=1}^{M}\left\{|\eta_{n,\tau}(f_j)| \le \frac{\omega_j}{2}\right\}. \tag{2.36}$$

On $\mathcal{A}$, we have

$$|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T\boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}}| \le \sum_{j=1}^{M}\omega_j|(\hat{\beta}_L - \beta)_j|. \tag{2.37}$$

The result (2.11) follows since $\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M}\omega_j|\beta_j|$. It remains to bound up $\mathbb{P}(\mathcal{A}^c)$. By applying Theorem 2.5

$$\mathbb{P}(\mathcal{A}^c) \le \sum_{j=1}^{M}\mathbb{P}\Big(|\eta_{n,\tau}(f_j)| > \frac{\omega_j}{2}\Big) \le A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x},$$

with

$$A_{\varepsilon,\nu,n}(x) = \frac{2}{\log(1+\varepsilon)} \log\left(2 + \frac{A_0(\nu/n + \Phi(\nu/n))}{x/n}\right) + 1. \tag{2.38}$$

We conclude that $\mathbb{P}(\mathcal{A}) \geq 1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x}$, which ends up the proof of Theorem 2.2. □

2.7.3 Proof of Lemma 2.1

We show with high probability, that under $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$, for all $J \subset \{1, ..., M\}$ such that $|J| \leq s$ and for all $b \in \mathbb{R}^M \backslash \{0\}$ such that $|b_{J^c}|_1 \leq a_0 |b_J|_1$,

$$\frac{\boldsymbol{b^T G_n b}}{|\boldsymbol{b}_J|_2^2} > \boldsymbol{\kappa}^2, \quad \text{with} \quad \boldsymbol{\kappa} = (1/\sqrt{2A_0})\boldsymbol{\kappa_0}(s, a_0),$$

and A_0 defined in Assumption 2.2. Let consider the set

$$\Omega_{G_n,t} = \{|(\boldsymbol{G_n} - \mathbb{E}(\boldsymbol{G_n}))_{j,k}| \leq t, \forall (j,k) \in \{1, ..., M\}^2\},$$

with a fixed $t \geq 0$. Under $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$, on $\Omega_{G_n,t}$, for all $J \subset \{1, ..., M\}$ such that $|J| \leq s$ and for all $\boldsymbol{b} \in \mathbb{R}^M \backslash \{0\}$ such that $|\boldsymbol{b}_{J^c}|_1 \leq a_0 |\boldsymbol{b}_J|_1$, we have

$$\boldsymbol{b^T G_n b} = \boldsymbol{b}^T(\boldsymbol{G_n} - \mathbb{E}(\boldsymbol{G_n}))\boldsymbol{b} + \boldsymbol{b}^T\mathbb{E}(\boldsymbol{G_n})\boldsymbol{b} \geq \boldsymbol{b}^T(\boldsymbol{G_n} - \mathbb{E}(\boldsymbol{G_n}))\boldsymbol{b} + \boldsymbol{\kappa_0}(s, a_0)^2 |\boldsymbol{b}_J|_2^2.$$

On $\Omega_{G_n,t}$, under $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$ we deduce that

$$\boldsymbol{b^T G_n b} \geq -\sum_{i,j} t|b_i||b_j| + \boldsymbol{\kappa_0}(s, a_0)^2 |\boldsymbol{b}_J|_2^2 \geq (-2t(1+a_0)^2 s + \boldsymbol{\kappa_0}(s, a_0)^2)|\boldsymbol{b}_J|_2^2.$$

We choose $t = A_0\boldsymbol{\kappa}^2/2(1+a_0)^2 s$ with $\boldsymbol{\kappa} = \boldsymbol{\kappa_0}(s, a_0)/\sqrt{2A_0}$ to get $\boldsymbol{b^T G_n b} \geq \boldsymbol{\kappa}^2 |\boldsymbol{b}_J|_2^2$.

It remains to calculate $\mathbb{P}(\Omega_{G_n,t})$. The coefficient (j,k) of the matrix $\boldsymbol{G_n} - \mathbb{E}(\boldsymbol{G_n})$ is given by

$$\frac{1}{n}\sum_{i=1}^{n} (\Lambda_i(\tau) - \mathbb{E}[\Lambda_i(\tau)]) f_j(\boldsymbol{Z_i}) f_k(\boldsymbol{Z_i}).$$

Under Assumptions 2.2 and 2.3.(i), we can apply Bernstein's inequality (2.34) to get

$$\mathbb{P}\left(|(\boldsymbol{G_n} - \mathbb{E}(\boldsymbol{G_n}))_{i,j}| > \frac{A_0\boldsymbol{\kappa}^2}{(1+a_0)^2 s}\right) \leq 2\exp\left(-\frac{n\boldsymbol{\kappa}^4}{2(1+a_0)^2 s L^2 (L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2/3)}\right).$$

So the probability of $\Omega^c_{G_n,t}$ with $t = A_0\boldsymbol{\kappa}^2/2(1+a_0)^2 s$ is given by

$$\mathbb{P}(\Omega^c_{G_n,t}) \leq 2M^2 \exp\left(-\frac{n\boldsymbol{\kappa}^4}{2(1+a_0)^2 s L^2 (L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2/3)}\right),$$

via an union bound and by denoting

$$\pi_n = 2M^2 \exp\left(-\frac{n\boldsymbol{\kappa}^4}{2(1+a_0)^2 s L^2 (L^2(1+a_0)^2 s + \boldsymbol{\kappa}^2/3)}\right),$$

we finally get (2.14) with probability larger than $1 - \pi_n$. □

2.7.4 Proof of Theorem 2.1

Let us introduce the event

$$\boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa}) = \left\{ 0 < \boldsymbol{\kappa} = \min_{\substack{J\subset\{1,\dots,M\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\leq a_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{G}_{\boldsymbol{n}}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2} \right\}. \tag{2.39}$$

From Inequality (2.35), on $\mathcal{A}$ defined by (2.36), for $\boldsymbol{\beta} \in \Gamma(\mu)$, it follows that

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}^{\mu}_L - \beta)_j| \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + \sum_{j=1}^{M} \omega_j (|(\hat{\beta}^{\mu}_L - \beta)_j| + |\beta_j| - |(\hat{\beta}^{\mu}_L)_j|).$$

On $J(\boldsymbol{\beta})^c$, $|(\hat{\beta}^{\mu}_L - \beta)_j| + |\beta_j| - |(\hat{\beta}^{\mu}_L)_j| = 0$, so on $\mathcal{A}$ we obtain

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}^{\mu}_L - \beta)_j| \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2 \sum_{j\in J(\boldsymbol{\beta})} \omega_j |(\hat{\beta}^{\mu}_L - \beta)_j|. \tag{2.40}$$

We apply Cauchy-Schwarz Inequality to the second right hand side of (2.40) to get

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) + \sum_{j=1}^{M} \frac{\omega_j}{2} |(\hat{\beta}^{\mu}_L - \beta)_j| \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\sqrt{|J(\boldsymbol{\beta})|} \sqrt{\sum_{j\in J(\boldsymbol{\beta})} \omega_j^2 |\hat{\beta}^{\mu}_L - \beta|_j^2}. \tag{2.41}$$

With the notations $\boldsymbol{\Delta} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}} - \boldsymbol{\beta})$ and $\boldsymbol{D} = (\mathrm{diag}(\omega_j))_{1\leq j\leq M}$ introduced in Subsection 2.3.2 , Inequalities (2.40) and (2.41) become respectively

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) + \frac{1}{2}|\boldsymbol{\Delta}|_1 \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1, \tag{2.42}$$

and

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\sqrt{|J(\boldsymbol{\beta})|}|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_2. \tag{2.43}$$

We fix some $\zeta > 0$ and we consider the following set

$$\mathcal{A}_1 = \{\zeta \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) \leq 2|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1\}. \tag{2.44}$$

Here, we could take $\zeta = 1$, but this parameter ζ allows to have more freedom. The smaller ζ is, the higher $\mathbb{P}(\mathcal{A}_1)$ is, but the smaller $\mathbb{P}(\boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa}))$ is. So ζ realizes a compromise between these two probabilities. On $\mathcal{A}\bigcap\mathcal{A}_1^c$, the result of the theorem follows immediately from (2.42). As soon as, $|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})^c}|_1 \leq (3 + 4/\zeta)\,|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1$, on $\boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa})$, with $a_0 = (3 + 4/\zeta)$ and $\boldsymbol{\kappa} = (1/\sqrt{2A_0})\boldsymbol{\kappa}_0(s, a_0)$ we get

$$\boldsymbol{\kappa}^2 |\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_2^2 \leq \boldsymbol{\Delta}^T \boldsymbol{G}_{\boldsymbol{n}} \boldsymbol{\Delta}.$$

So, initially we will assume that $|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})^c}|_1 \leq (3 + 4/\zeta)\,|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1$, and we will verify later that this inequality holds. Since

$$\boldsymbol{\Delta}^T \boldsymbol{G}_{\boldsymbol{n}} \boldsymbol{\Delta} \leq (\max_{1\leq j\leq M} \omega_j)^2 || \log \lambda_{\hat{\beta}^{\mu}_L} - \log \lambda_{\boldsymbol{\beta}}||^2_{n,\Lambda},$$

Inequality (2.43) becomes on $\mathcal{A} \cap \mathbf{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa})$

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\sqrt{|J(\boldsymbol{\beta})|}(\max_{1\leq j\leq M} \omega_j)\boldsymbol{\kappa}^{-1}(||\log \lambda_{\hat{\beta}^{\mu}_{L}} - \log \lambda_0||_{n,\Lambda} + ||\log \lambda_0 - \log \lambda_{\boldsymbol{\beta}}||_{n,\Lambda}).$$

Now, applying Proposition 2.3 to connect the weighted empirical norm to the empirical Kullback divergence, it follows that

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\sqrt{|J(\boldsymbol{\beta})|}(\max_{1\leq j\leq M} \omega_j)\frac{\boldsymbol{\kappa}^{-1}}{\sqrt{\mu'}}\left(\sqrt{\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}})} + \sqrt{\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}})}\right).$$

We now use the elementary inequality $2uv \leq bu^2 + \frac{v^2}{b}$ with $b > 0$, $u = \sqrt{|J(\boldsymbol{\beta})|}(\max_{1\leq j\leq M} \omega_j)\boldsymbol{\kappa}^{-1}$ and v being either $\sqrt{\frac{1}{\mu'}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}})}$ or $\sqrt{\frac{1}{\mu'}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}})}$. Consequently

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2b|J(\boldsymbol{\beta})|(\max_{1\leq j\leq M} \omega_j)^2\boldsymbol{\kappa}^{-2} + \frac{1}{b\mu'}\widetilde{K}_n(\lambda_0, \boldsymbol{\lambda}_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) + \frac{1}{b\mu'}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}).$$

Hence,

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \leq \frac{b\mu'+1}{b\mu'-1}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + 2\frac{b^2\mu'}{b\mu'-1}|J(\boldsymbol{\beta})|(\max_{1\leq j\leq M} \omega_j)^2\boldsymbol{\kappa}^{-2}.$$

We take $\frac{b\mu'+1}{b\mu'-1} = 1+\zeta$ and $C(\zeta,\mu) = 2\frac{b^2\mu'}{b\mu'+1}$ a constant depending on ζ and μ. It follows that for any $\boldsymbol{\beta} \in \Gamma(\mu)$:

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}}}) \leq (1+\zeta)\Big\{\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) + C(\zeta,\mu)|J(\boldsymbol{\beta})|(\max_{1\leq j\leq M} \omega_j)^2\boldsymbol{\kappa}^{-2}\Big\}.$$

Finally, taking the infimum over all $\boldsymbol{\beta} \in \Gamma(\mu)$ such that $|J(\boldsymbol{\beta})| \leq s$, we obtain (2.15).

We have now to verify that $|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})^c}|_1 \leq (3+4/\zeta)\,|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1$. On $\mathcal{A}\bigcap\mathcal{A}_1$, applying (2.42) we get that

$$|\boldsymbol{\Delta}|_1 \leq 4\left(1+\frac{1}{\zeta}\right)|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1,$$

so by splitting $\boldsymbol{\Delta} = \boldsymbol{\Delta}_{J(\boldsymbol{\beta})} + \boldsymbol{\Delta}_{J(\boldsymbol{\beta})^c}$, we finally obtain

$$|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})^c}|_1 \leq \left(3+\frac{4}{\zeta}\right)|\boldsymbol{\Delta}_{J(\boldsymbol{\beta})}|_1.$$

Finally, Lemma 2.1 ensures that $\mathbb{P}(\mathcal{A}^c \cup \mathbf{\Omega}^c_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa})) \leq A_{\varepsilon,\nu,n}(x)e^{-x} + \pi_n$, which achieves the proof of Theorem 2.1. □

2.7.5 Proof of Theorem 2.2

Before we begin the proof, we recall some notations in the particular case of variable selection. In this case, $M = p$ and the functions of the dictionary are such that for $i = 1, ..., n$ and $j = 1, ..., p$

$$f_j(\boldsymbol{Z_i}) = Z_{i,j} \text{ and } f_{\boldsymbol{\beta}}(\boldsymbol{Z_i}) = \sum_{j=1}^{p} \beta_j Z_{i,j} = \boldsymbol{\beta^T Z_i}.$$

Let $\boldsymbol{X} = (Z_{i,j})_{\substack{1\leq i\leq n \\ 1\leq j\leq p}}$ be the design matrix and for $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$ defined by (2.9), let

$$\boldsymbol{\Delta_0} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}), \boldsymbol{D} = (\text{diag}(\omega_j))_{1\leq j\leq M}, J_0 = J(\boldsymbol{\beta_0}) \text{ and } |J_0| = \text{Card}\{J_0\}.$$

We also define $\boldsymbol{\Delta_0}_{J_0}$ as the vector $\boldsymbol{\Delta_0}$ restricted to the set $J_0 : (\Delta_{0_{J_0}})_j = \Delta_{0,j}$ if $j \in J_0$ and $(\Delta_{0_{J_0}})_j = 0$ if $j \in J_0^c$ where $J_0^c = \{1, ..., p\} \setminus J_0$.

To prove Inequality (2.16) of Theorem 2.2, we start from (2.40) with $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$ defined by (2.9). Consequently $\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta}}) = 0$. Applying Proposition 2.4 with $\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}$ and $\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}(t, \boldsymbol{Z_i}) = \alpha_0(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}^{\boldsymbol{T}} \boldsymbol{Z_i}}$, we obtain that, on

$$\mathcal{A}_{\Gamma_1} = \bigcap_{j=1}^{p} \left\{ |\eta_{n,\tau}(f_j)| \leq \Gamma_1 \frac{\omega_j}{2} \right\},$$

$$\xi ||(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^{\boldsymbol{T}} \boldsymbol{X}||_{n,\Lambda}^2 + \Gamma_1 \sum_{j=1}^{p} \frac{\omega_j}{2} |\hat{\beta}_L - \beta_0|_j \leq 2\Gamma_1 \sum_{j\in J_0} \omega_j |\hat{\beta}_L - \beta_0|_j. \tag{2.45}$$

From this inequality, we deduce

$$\xi ||\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})||_{n,\Lambda}^2 \leq 2\Gamma_1 \sum_{j\in J_0} \omega_j |\hat{\beta}_L - \beta_0|_j \leq 2\sqrt{|J_0|}\Gamma_1 |\boldsymbol{\Delta}_{0,J_0}|_2. \tag{2.46}$$

From (2.45), we also have

$$\sum_{j=1}^{p} \omega_j |\hat{\beta}_L - \beta_0|_j \leq 4 \sum_{j\in J_0} \omega_j |\hat{\beta}_L - \beta_0|_j$$

and we obtain $|\boldsymbol{\Delta_0}|_1 \leq 4|\boldsymbol{\Delta_0}_{J_0}|_1$. We then split $|\boldsymbol{\Delta_0}|_1 = |\boldsymbol{\Delta_0}_{J_0}|_1 + |\boldsymbol{\Delta_0}_{J_0^c}|_1$ to get

$$|\boldsymbol{\Delta_0}_{J_0^c}|_1 \leq 3|\boldsymbol{\Delta_0}_{J_0}|_1. \tag{2.47}$$

On $\boldsymbol{\Omega_{RE_n}}_{(s,a_0)}(\boldsymbol{\kappa'})$, defined by (2.39), with $a_0 = 3$ and $\boldsymbol{\kappa'} = (1/\sqrt{2A_0})\boldsymbol{\kappa_0}(s, 3)$ we get

$$||\boldsymbol{X}\boldsymbol{\Delta_0}||_{n,\Lambda}^2 \geq \boldsymbol{\kappa'}^2 |\boldsymbol{\Delta}_{0,J_0}|_2^2. \tag{2.48}$$

According to (2.46), we conclude that on $\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa'})$

$$\xi ||\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})||_{n,\Lambda}^2 \leq 2\sqrt{|J_0|}\Gamma_1 \max_{1\leq j\leq p} \omega_j \frac{||\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})||_{n,\Lambda}}{\boldsymbol{\kappa'}},$$

which entails that

$$||\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})||_{n,\Lambda}^2 \leq \frac{4|J_0|}{\xi^2 \boldsymbol{\kappa'}^2} \Gamma_1^2 (\max_{1\leq j\leq p} \omega_j)^2,$$

with $\mathbb{P}(\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa'})) \geq 1 - A_{\varepsilon,\nu,n}(x)e^{-\Gamma_1 x} - \pi_n$.

Let us come to the proof of Inequality (2.17) in Theorem 2.2. On $\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa'})$, with $a_0 = 3$, Inequality (2.46) becomes

$$\xi \frac{\boldsymbol{\kappa'}^2}{\max_{1\leq j\leq M} \omega_j^2} |\boldsymbol{\Delta}_{0,J_0}|_2^2 \leq 2\sqrt{|J_0|}\Gamma_1 |\boldsymbol{\Delta}_{0,J_0}|_2. \tag{2.49}$$

According to (2.47) and thanks to Cauchy-Schwarz Inequality, we have

$$|\boldsymbol{\Delta_0}|_1 = |\boldsymbol{\Delta_0}_{J_0}|_1 + |\boldsymbol{\Delta_0}_{J_0^c}|_1 \leq 4|\boldsymbol{\Delta_0}_{J_0}|_1 \leq 4\sqrt{|J_0|}|\boldsymbol{\Delta_0}_{J_0}|_2.$$

From (2.49), we get

$$\frac{|\boldsymbol{\Delta_0}|_1}{4\sqrt{|J_0|}} \leq \frac{2\sqrt{|J_0|}}{\xi \boldsymbol{\kappa'}^2} \Gamma_1 \max_{1\leq j\leq p} \omega_j^2,$$

and finally

$$|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_1 \leq 8 \frac{|J_0|}{\xi \boldsymbol{\kappa'}^2} \Gamma_1 \frac{\max_{1\leq j\leq p} \omega_j^2}{\min_{1\leq j\leq p} \omega_j},$$

with $\mathbb{P}(\mathcal{A}_{\Gamma_1} \cap \boldsymbol{\Omega}_{\mathbf{RE}_{\boldsymbol{n}}(s,a_0)}(\boldsymbol{\kappa'})) \geq 1 - A_{\varepsilon,\nu,n}(x)e^{-\Gamma_1 x} - \pi_n$. □

2.7.6 Proof of Theorem 2.3

The proof is very similar to the one of Proposition 2.2. We start from (2.23) and (2.24), and write

$$\begin{aligned}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) &+ (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}} - \boldsymbol{\gamma})^T \boldsymbol{\nu}_{\boldsymbol{n},\boldsymbol{\tau}} + \mathrm{pen}(\boldsymbol{\gamma}) - \mathrm{pen}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) \\ &+ (\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T \boldsymbol{\eta}_{\boldsymbol{n},\boldsymbol{\tau}} + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}).\end{aligned} \tag{2.50}$$

Consider the set $\mathcal{A}$ defined by (2.36) and let define similarly the set $\mathcal{B}$ such that

$$\mathcal{B} = \bigcap_{k=1}^{N} \Big\{ |\nu_{n,\tau}(\theta_k)| \leq \frac{\delta_k}{2} \Big\}. \tag{2.51}$$

Applying Theorem 2.5, we obtain that

$$\mathbb{P}(\mathcal{A}^c) \leq A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} \text{ and } \mathbb{P}(\mathcal{B}^c) \leq B_{\tilde{\varepsilon},\tilde{\nu},n}(y)\mathrm{e}^{-y},$$

with $A_{\varepsilon,\nu,n}(x)$ defined by (2.10) and

$$B_{\tilde{\varepsilon},\tilde{\nu},n}(y) = \frac{2}{\log(1+\tilde{\varepsilon})}\log\Big(2 + \frac{A_0(\tilde{\nu}/n + \Phi(\tilde{\nu}/n))}{y/n}\Big) + 1,$$

and thus

$$\mathbb{P}[(\mathcal{A}\cap\mathcal{B})^c] = \mathbb{P}(\mathcal{A}^c\cup\mathcal{B}^c) \leq \mathbb{P}(\mathcal{A}^c) + \mathbb{P}(\mathcal{B}^c) \leq A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} + B_{\tilde{\varepsilon},\tilde{\nu},n}(y)\mathrm{e}^{-y}. \quad (2.52)$$

On $\mathcal{A}\cap\mathcal{B}$ arguing as in the proof of Theorem 2.2, with probability larger than $1 - A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} - B_{\tilde{\varepsilon},\tilde{\nu},n}(y)\mathrm{e}^{-y}$, we finish the proof by writing (2.19). □

2.7.7 Proof of Theorem 2.4

Let introduce the event

$$\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_{\boldsymbol{n}}(s,r_0)}(\tilde{\boldsymbol{\kappa}}) = \Big\{0 < \tilde{\boldsymbol{\kappa}} = \min_{\substack{J\subset\{1,\dots,M\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^M\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\leq r_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2}\Big\}.$$

From Inequality (2.50), on $\mathcal{A}\cap\mathcal{B}$ defined in (2.36) and (2.51), for $(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$, we obtain

$$\begin{aligned}\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}^{\boldsymbol{\rho}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}^{\boldsymbol{\rho}}}) + \sum_{j=1}^M \frac{\omega_j}{2}|(\hat{\beta}_L^\rho - \beta)_j| + \sum_{k=1}^N \frac{\delta_k}{2}|(\hat{\gamma}_L^\rho - \gamma)_k| \\ \leq \widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + 2\sum_{j\in J(\boldsymbol{\beta})}\omega_j|(\hat{\beta}_L^\rho - \beta)_j| + 2\sum_{k\in J(\boldsymbol{\gamma})}\delta_k|(\hat{\gamma}_L^\rho - \gamma)_k|. \quad (2.53)\end{aligned}$$

We then apply Cauchy-Schwarz inequality to the second right-term of (2.53) and obtain

$$\begin{aligned}&\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}^{\boldsymbol{\rho}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}^{\boldsymbol{\rho}}}) + \sum_{j=1}^M \frac{\omega_j}{2}|(\hat{\beta}_L^\rho - \beta)_j| + \sum_{k=1}^N \frac{\delta_k}{2}|(\hat{\gamma}_L^\rho - \gamma)_k| \\ &\leq \widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + 2\sqrt{|J(\boldsymbol{\beta})|}\sqrt{\sum_{j\in J(\boldsymbol{\beta})}\omega_j^2|\hat{\beta}_L^\rho - \beta|_j^2} + 2\sqrt{|J(\boldsymbol{\gamma})|}\sqrt{\sum_{k\in J(\boldsymbol{\gamma})}\delta_k^2|\hat{\gamma}_L^\rho - \gamma|_k^2}.\end{aligned} \quad (2.54)$$

With the notations of Subsection 2.4.2, Inequality (2.53) is rewritten as:

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}^{\boldsymbol{\rho}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}^{\boldsymbol{\rho}}}) + \frac{1}{2}|\tilde{\boldsymbol{\Delta}}|_1 \leq \widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + 2|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_1, \quad (2.55)$$

where $\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})} = \tilde{\boldsymbol{D}}\begin{pmatrix}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}^{\boldsymbol{\rho}} - \boldsymbol{\beta})_{J(\boldsymbol{\beta})}\\ (\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}^{\boldsymbol{\rho}} - \boldsymbol{\gamma})_{J(\boldsymbol{\gamma})}\end{pmatrix}$. In the same way, Inequality (2.54) becomes :

$$\widetilde{K}_n(\lambda_0,\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}^{\boldsymbol{\rho}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}^{\boldsymbol{\rho}}}) \leq \widetilde{K}_n(\lambda_0,\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + 4\max\Big(\sqrt{|J(\boldsymbol{\beta})|},\sqrt{|J(\boldsymbol{\gamma})|}\Big)|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_2. \quad (2.56)$$

Consider

$$\tilde{\mathcal{A}}_1 = \{\zeta \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) \leq 2|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_1\}. \tag{2.57}$$

On $\mathcal{A} \cap \mathcal{B} \cap \tilde{\mathcal{A}}_1$, Inequality (2.21) in Theorem 2.4 follows immediately from (2.55). As soon as, $|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}|_1 \leq \left(3 + 8\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)/\zeta\right)|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_1$, on $\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_{\boldsymbol{n}}(s,r_0)}(\tilde{\boldsymbol{\kappa}})$, with

$$\tilde{\boldsymbol{\kappa}} = (1/\sqrt{2})\tilde{\boldsymbol{\kappa}}_{\mathbf{0}}(s, r_0) \quad \text{and} \quad r_0 = \left(3 + 8\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)/\zeta\right),$$

we get that

$$\tilde{\boldsymbol{\kappa}}^2|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_2^2 \leq \tilde{\boldsymbol{\Delta}}^{\boldsymbol{T}}\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\tilde{\boldsymbol{\Delta}}$$

with

$$\tilde{\boldsymbol{\Delta}}^{\boldsymbol{T}}\tilde{\boldsymbol{G}}_{\boldsymbol{n}}\tilde{\boldsymbol{\Delta}} \leq \max_{\substack{1\leq j\leq M\\1\leq k\leq N}} \{\omega_j, \delta_k\}|| \log \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}} - \log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}||^2_{n,\Lambda}.$$

On $\mathcal{A}\cap\mathcal{B}\cap\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_{\boldsymbol{n}}(s,r_0)}(\tilde{\boldsymbol{\kappa}})$, from Equation (2.56) and applying Proposition 2.5 to connect the weighted empirical norm to the empirical Kullback divergence, we obtain that

$$\begin{aligned}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}}) &\leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) \\ &\quad + 4\Big(\sqrt{|J(\boldsymbol{\beta})|} \vee \sqrt{|J(\boldsymbol{\gamma})|}\Big)\max_{\substack{1\leq j\leq M\\1\leq k\leq N}} \{\omega_j, \delta_k\}\frac{\tilde{\boldsymbol{\kappa}}^{-1}}{\sqrt{\rho'}}\left(\sqrt{\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}})} + \sqrt{\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}})}\right).\end{aligned}$$

Using again $2uv \leq bu^2 + \dfrac{v^2}{b}$ with $b > 0$, $u = 2\left(\sqrt{|J(\boldsymbol{\beta})|} \vee \sqrt{|J(\boldsymbol{\gamma})|}\right)\max\limits_{\substack{1\leq j\leq M\\1\leq k\leq N}} \{\omega_j, \delta_k\}\tilde{\boldsymbol{\kappa}}^{-1}$ and v being either $\sqrt{\dfrac{1}{\rho'}\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}})}$ or $\sqrt{\dfrac{1}{\rho'}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}})}$, we obtain

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}^{\boldsymbol{\rho}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}^{\boldsymbol{\rho}}_{\boldsymbol{L}}}) \leq \frac{b\rho'+1}{b\rho'-1}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + 8\frac{b^2\rho'}{b\rho'-1}(|J(\boldsymbol{\beta})| \vee |J(\boldsymbol{\gamma})|)\Big(\max_{\substack{1\leq j\leq M\\1\leq k\leq N}} \{\omega_j, \delta_k\}\Big)^2\frac{\tilde{\boldsymbol{\kappa}}^{-2}}{\rho'}. \tag{2.58}$$

Finally, taking $\dfrac{b\rho'+1}{b\rho'-1} = 1+\zeta$ and $\widetilde{C}(\zeta,\rho) = 8\dfrac{b^2\rho'}{b\rho'+1}$ a constant depending on ζ and ρ, and taking the infimum over all $(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$ such that $\max(|J(\boldsymbol{\beta})|, |J(\boldsymbol{\gamma})|) \leq s$, we obtain Inequality (2.21).

Inequality (2.22) follows by applying Proposition 2.3 with $b = \dfrac{(1+\zeta)\rho' + \rho''}{(1+\zeta)\rho' - \rho''}$ in (2.58).

We have now to verify that

$$|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}|_1 \leq \left(3 + 8\max\left(\sqrt{|J(\boldsymbol{\beta})|}, \sqrt{|J(\boldsymbol{\gamma})|}\right)/\zeta\right)|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_1.$$

We deduce from (2.55) that, on $\mathcal{A}\cap\mathcal{B}\cap\tilde{\mathcal{A}}_1$, by splitting $\tilde{\boldsymbol{\Delta}} = \tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})} + \tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}$

$$|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta})^c,J(\boldsymbol{\gamma})^c}|_1 \leq \left(3+\frac{8}{\zeta}\max\left(\sqrt{|J(\boldsymbol{\beta})|},\sqrt{|J(\boldsymbol{\gamma})|}\right)\right)|\tilde{\boldsymbol{\Delta}}_{J(\boldsymbol{\beta}),J(\boldsymbol{\gamma})}|_1.$$

To achieve the proof of Theorem 2.4, we combine Equation (2.52) with Lemma 2.2 to conclude

$$\mathbb{P}\Big[\Big(\mathcal{A}\cap\mathcal{B}\cap\boldsymbol{\Omega}_{\widetilde{\mathbf{RE}}_{\boldsymbol{n}}(s,r_0)}(\tilde{\boldsymbol{\kappa}})\Big)^c\Big] \leq A_{\varepsilon,\nu,n}(x)\mathrm{e}^{-x} + B_{\tilde{\varepsilon},\tilde{\nu},n}(y)\mathrm{e}^{-y} + \tilde{\pi}_n.$$

□

2.7.8 Proof of Theorem 2.5

The proofs of (2.28) and (2.29) are quite similar, so we only present the one of (2.28). To prove (2.29), it suffices to replace $\eta_{n,t}(f_j)$ by the process $\nu_{n,t}(\theta_k)$ throughout the following. Denote by $U_{n,t}$ and $H_i(f_j)$ the quantities

$$U_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^n\int_0^t H_i(f_j)\mathrm{d}M_i(s) \quad \text{and} \quad H_i(f_j) := \frac{f_j(\boldsymbol{Z_i})}{\max\limits_{1\leq i\leq n}|f_j(\boldsymbol{Z_i})|}.$$

Since $H_i(f_j)$ is a bounded predictable process with respect to $\mathcal{F}_t$, $U_{n,t}(f_j)$ is a square integrable martingale. Its predictable variation is given by

$$\vartheta_{n,t}(f_j) = n < U_n(f_j) >_t = \frac{1}{n}\sum_{i=1}^n\int_0^t (H_i(f_j))^2\mathrm{d}\Lambda_i(s)$$

and the optional variation of $U_{n,t}(f_j)$ is

$$\hat{\vartheta}_{n,t}(f_j) = n[U_n(f_j)]_t = \frac{1}{n}\sum_{i=1}^n\int_0^t (H_i(f_j))^2\mathrm{d}N_i(s).$$

We also define

$$\hat{\mathcal{W}}_n^\nu(f_j) = \frac{\nu/n}{\nu/n-\Phi(\nu/n)}\hat{\vartheta}_{n,t}(f_j) + \frac{x/n}{\nu/n-\Phi(\nu/n)}, \tag{2.59}$$

for $\nu\in(0,3)$ such that $\nu>\Phi(\nu)$ with $\Phi(u)=\mathrm{e}^u-u-1$.

From Inequality (7.12) in Hansen et al. (2012), for any $0<v<\omega<+\infty$, we have

$$\mathbb{P}\left(U_{n,t}(f_j)\geq\sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}}+\frac{x}{3n}, v\leq\hat{\mathcal{W}}_n^\nu(f_j)\leq\omega\right) \leq 2\left(\frac{\log(\omega/v)}{\log(1+\varepsilon)}+1\right)\mathrm{e}^{-x}. \tag{2.60}$$

We focus now on removing the event $\{v \leq \hat{\mathcal{W}}_n^\nu(f_j) \leq \omega\}$ in (2.60). Let us consider the martingale given $\mathcal{F}_t$

$$\begin{aligned}\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j) &= \frac{1}{n}\sum_{i=1}^n \int_0^t (H_i(f_j))^2 \Big(\mathrm{d}N_i(s) - \mathrm{d}\Lambda_i(s)\Big) \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^t (H_i(f_j))^2 \mathrm{d}M_i(s),\end{aligned}$$

and let

$$S_{\nu,t}(f_j) = \sum_{i=1}^n \int_0^t \Phi\big(\frac{\nu}{n} H_i^2(f_j)\Big)\mathrm{d}\Lambda_i(s).$$

From van de Geer (1995), we know that

$$\exp(\nu(\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)) - S_{\nu,t}(f_j))$$

is a supermartingale. Now from Markov Inequality, for any $\nu, x > 0$, we obtain that

$$\mathbb{P}\Big[|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \geq \frac{S_{\nu,t}(f_j)}{\nu} + \frac{x}{\nu}\Big] \leq 2\mathrm{e}^{-x}. \tag{2.61}$$

For any $0 < h < 1$ and $x > 0$, $\Phi(xh) \leq h^2\Phi(x)$. This combined with the fact that $0 < H_i^2(f_j) < 1$, we get

$$S_{\nu,t}(f_j) \leq \Phi(\nu/n)\sum_{i=1}^n \int_0^t H_i^4(f_j)\mathrm{d}\Lambda_i(s) \leq \Phi(\nu/n) n \vartheta_{n,t}(f_j). \tag{2.62}$$

Combining (2.61) and (2.62), we deduce that

$$\mathbb{P}\Big[|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \geq \frac{\Phi(\nu/n)}{\nu/n}\vartheta_{n,t}(f_j) + \frac{x}{\nu}\Big] \leq 2\mathrm{e}^{-x}. \tag{2.63}$$

Now, under Assumption 2.2, we have $\vartheta_{n,t}(f_j) \leq A_0$, so the events

$$\Omega_n^\nu = \Big\{\frac{x/n}{\nu/n - \Phi(\nu/n)} \leq \hat{\mathcal{W}}_n^\nu(f_j)\Big\} \cap \{\vartheta_{n,t}(f_j) \leq A_0\}$$

is of probability one and thus

$$\begin{aligned}&\mathbb{P}\left(U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}\right) \\ &\qquad \leq \mathbb{P}\left(\left\{U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}\right\} \cap \Omega_n^\nu\right).\end{aligned}$$

From (2.63), we have

$$\mathbb{P}\Big[\hat{\vartheta}_{n,t}(f_j) \geq \vartheta_{n,t}(f_j)\Big(1 + \frac{\Phi(\nu/n)}{\nu/n}\Big) + \frac{x}{\nu}\Big] \leq \mathrm{e}^{-x},$$

and if we denote E_n^ν the event

$$E_n^\nu = \Big\{\hat{\vartheta}_{n,t}(f_j) \leq \vartheta_{n,t}(f_j)\Big(1 + \frac{\Phi(\nu/n)}{\nu/n}\Big) + \frac{x}{\nu}\Big\},$$

we get

$$\begin{aligned}\mathbb{P}\Bigg[U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}\Bigg] \\ \leq \mathrm{e}^{-x} + \mathbb{P}\Bigg[\Bigg\{U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}\Bigg\} \cap \Omega_n^\nu \cap E_n^\nu\Bigg].\end{aligned}$$

On the event $E_n^\nu \cap \Omega_n^\nu$, from the definition of $\hat{\mathcal{W}}_n^\nu(f_j)$ given by (2.59), we have

$$\begin{aligned}\hat{\mathcal{W}}_n^\nu(f_j) &\leq \frac{\nu/n}{\nu/n - \Phi(\nu/n)}\Big(\vartheta_{n,t}(f_j)\Big(1 + \frac{\Phi(\nu/n)}{\nu/n}\Big) + \frac{x}{\nu}\Big) + \frac{x/n}{\nu/n - \Phi(\nu/n)} \\ &\leq A_0\frac{\nu/n + \Phi(\nu/n)}{\nu/n - \Phi(\nu/n)} + 2\frac{x/n}{\nu/n - \Phi(\nu/n)}. \end{aligned} \tag{2.64}$$

From (2.64), we obtain

$$\begin{aligned}&\mathbb{P}\Bigg[\Bigg\{U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}\Bigg\} \cap \Omega_n^\nu \cap E_n^\nu\Bigg] \\ \leq &\mathbb{P}\Bigg[U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}, \\ &\qquad \frac{x/n}{\nu/n - \Phi(\nu/n)} \leq \hat{\mathcal{W}}_n^\nu(f_j) \leq A_0\frac{\nu/n + \Phi(\nu/n)}{\nu/n - \Phi(\nu/n)} + 2\frac{x/n}{\nu/n - \Phi(\nu/n)}\Bigg].\end{aligned}$$

We now apply Inequality (2.60) with

$$v = \frac{x/n}{\nu/n - \Phi(\nu/n)} \quad \text{and} \quad \omega = A_0\frac{\nu/n + \Phi(\nu/n)}{\nu/n - \Phi(\nu/n)} + 2\frac{x/n}{\nu/n - \Phi(\nu/n)},$$

$$\begin{aligned}\mathbb{P}\Bigg[U_{n,t}(f_j) \geq \sqrt{\frac{2(1+\varepsilon)\hat{\mathcal{W}}_n^\nu(f_j)x}{n}} + \frac{x}{3n}\Bigg] \\ \leq \left(\frac{2}{\log(1+\varepsilon)}\log\left(2 + \frac{A_0(\nu/n + \Phi(\nu/n))}{x/n}\right) + 1\right)\mathrm{e}^{-x}.\end{aligned}$$

Now it suffices to multiply both sides of the inequality inside the probability by $||f_j||_{n,\infty} = \max\limits_{1\leq i\leq n}|f_j(\boldsymbol{Z_i})|$ to end up the proof of Theorem 2.5. □

Appendix

A Connection between the weighted norm and the Kullback divergence

A.1 Proof of Proposition 2.5

The proofs of Proposition 2.3 and Proposition 2.5 are similar. So we only prove Proposition 2.5 which corresponds to the general case. To compare the empirical Kullback divergence (2.3) and the weighted empirical norm (2.4), we use Lemma 1 in Bach (2010), that we recall here:

Lemma 2.3. *Let g be a convex three times differentiable function $g : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$, $|g'''(t)| \leq Sg''(t)$, for some $S \geq 0$. Then, for all $t \geq 0$:*

$$\frac{g''(0)}{S^2}\phi(St) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{S^2}\phi(-St) \text{ with } \phi(u) = \mathrm{e}^{-u} + u - 1$$

This Lemma gives upper and lower Taylor expansions for some convex and three times differentiable function. It has been introduced to extend tools from self-concordant functions (i.e. which verify $|g'''(t)| \leq 2g''(t)^{3/2}$) and provide simple extensions of theoretical results for the square loss for logistic regression.

Let h be a function on $[0, \tau] \times \mathbb{R}^p$ and define

$$G(h) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau h(s, \boldsymbol{Z_i})\mathrm{d}\Lambda_i(s) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \mathrm{e}^{h(s,\boldsymbol{Z_i})}Y_i(s)\mathrm{d}s.$$

Consider the function $g : \mathbb{R} \to \mathbb{R}$ defined by $g(t) = G(h + tk)$, where h and k are two functions defined on $\mathbb{R}^p$. By differentiating G with respect to t we get :

$$g'(t) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau k(s, \boldsymbol{Z_i})\mathrm{d}\Lambda_i(s) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau k(s, \boldsymbol{Z_i})\mathrm{e}^{h(s,\boldsymbol{Z_i})+tk(s,\boldsymbol{Z_i})}Y_i(s)\mathrm{d}s,$$

$$g''(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (k(s, \boldsymbol{Z_i}))^2\mathrm{e}^{h(s,\boldsymbol{Z_i})+tk(s,\boldsymbol{Z_i})}Y_i(s)\mathrm{d}s,$$

$$g'''(t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (k(s, \boldsymbol{Z_i}))^3\mathrm{e}^{h(s,\boldsymbol{Z_i})+tk(s,\boldsymbol{Z_i})}Y_i(s)\mathrm{d}s.$$

It follows that

$$|g'''(t)| \leq ||k||_{n,\infty}g''(t).$$

Applying Lemma 2.3 with $S = ||k||_{n,\infty}$, we obtain for all $t \geq 0$,

$$\frac{g''(0)}{||k||^2_{n,\infty}}\phi(t||k||_{n,\infty}) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{||k||^2_{n,\infty}}\phi(-t||k||_{n,\infty}).$$

Take $t = 1$, $h(s, \boldsymbol{Z_i}) = \log \lambda_0(s, \boldsymbol{Z_i})$ and for $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$, $k(s, \boldsymbol{Z_i}) = \log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(s, \boldsymbol{Z_i}) - \log \lambda_0(s, \boldsymbol{Z_i})$. We obtain

$$g''(0)\frac{\phi(||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty})}{||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\infty}} \leq G(\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) - G(\log \lambda_0) - g'(0) \tag{2.65}$$

$$g''(0)\frac{\phi(-||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty})}{||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\infty}} \geq G(\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) - G(\log \lambda_0) - g'(0). \tag{2.66}$$

Now straightforward calculations show that $g'(0) = 0$ and $g''(0) = ||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\Lambda}$. Replacing $g'(0)$ and $g''(0)$ by their expressions in (2.65) and (2.66) and noting that $G(\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) - G(\log \lambda_0) = \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}})$, we get

$$\frac{\phi(||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty})}{||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\infty}}||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\Lambda} \leq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}})$$

$$\frac{\phi(-||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty})}{||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\infty}}||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\Lambda} \geq \widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}).$$

According to Assumption 2.4 for $(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \widetilde{\Gamma}(\rho)$, $||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty} \leq \rho$. Since $\phi(t)/t^2$ is decreasing and bounded below by 0, we can deduce that

$$\frac{\phi(||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty})}{||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\infty}} \geq \frac{\phi(\rho)}{\rho^2}$$

and

$$\frac{\phi(-||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||_{n,\infty})}{||\log \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}} - \log \lambda_0||^2_{n,\infty}} \leq \frac{\phi(-\rho)}{\rho^2}.$$

Take $\rho' := \phi(\rho)/\rho^2 > 0$ and $\rho'' := \phi(-\rho)/\rho^2 > 0$ to finish the proof. □

A.2 Proof of Proposition 2.4

Similarly to the proof of Proposition 2.5, under Assumption 2.6, when considering

$$G(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log(\alpha_0(s)\mathrm{e}^{\boldsymbol{\beta^T Z_i}})\mathrm{d}\Lambda_i(s) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha_0(s)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(s)\mathrm{d}s,$$

and $g(t) = G(\boldsymbol{\beta} + t\boldsymbol{\eta})$, we obtain

$$||(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^T\boldsymbol{X}||^2_{n,\Lambda}\frac{\phi(R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2)}{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2} \leq \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) \tag{2.67}$$

$$||(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^T\boldsymbol{X}||^2_{n,\Lambda}\frac{\phi(-R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2)}{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2} \geq \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}). \tag{2.68}$$

Now, we will show that $R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2$ is bounded. From Equation (2.42) with $\hat{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{L}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$ and $\boldsymbol{\beta} = \boldsymbol{\beta_0}$, we can deduce that

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_L}) \leq \frac{3}{2}\Gamma_1|\boldsymbol{\Delta}_0|_1,$$

where $\boldsymbol{\Delta_0} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})$ and $\boldsymbol{D} = (\mathrm{diag}(\omega_j))_{1\leq j\leq M}$. From (2.67) and (2.68), we have

$$\widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_L}) \geq \frac{||(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^{\boldsymbol{T}} \boldsymbol{X}||^2_{n,\Lambda}}{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2} \phi(R|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})|_2)$$

We apply Assumption $\mathbf{RE}(\mathbf{s}, \mathbf{a_0})$ with $a_0 = 3$ and $\boldsymbol{\kappa'} = \boldsymbol{\kappa'}(s, 3)$ and we infer that

$$\boldsymbol{\kappa'^2}|\boldsymbol{\Delta}_{0,J_0}|_2^2 \leq |\boldsymbol{\Delta_0^T X}|^2_{n,\Lambda}.$$

So we have,

$$\frac{\boldsymbol{\kappa'^2}|\boldsymbol{\Delta}_{0,J_0}|_2^2}{\max\limits_{1\leq j\leq M} \omega_j^2} \frac{\phi(R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2)}{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2} \leq \frac{3}{2}\Gamma_1 ||\boldsymbol{\Delta_0}|_1.$$

We can now use, with $s = |J_0|$, $|\boldsymbol{\Delta_0}|_2 \leq |\boldsymbol{\Delta_0}|_1 \leq 4|\boldsymbol{\Delta}_{0,J_0}|_1 \leq 4\sqrt{s}|\boldsymbol{\Delta}_{0,J_0}|_2$ to get

$$\begin{aligned}\boldsymbol{\kappa'^2}\phi(R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2) &\leq \frac{3}{2}\Gamma_1 \frac{\max\limits_{1\leq j\leq M} \omega_j^2}{\min\limits_{1\leq j\leq M} \omega_j^2} \max_{1\leq j\leq M} \omega_j \frac{(4\sqrt{s}|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})_{J_0}|_2)^2 R^2 |\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2}{|(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})_{J_0}|_2^2} \\ &\leq 24\Gamma_1 \frac{\max\limits_{1\leq j\leq M} \omega_j^2}{\min\limits_{1\leq j\leq M} \omega_j^2} \max_{1\leq j\leq M} \omega_j s R^2 |\boldsymbol{\Delta_0}|_2.\end{aligned}$$

A short calculation shows that for all $k \in (0, 1]$:

$$\mathrm{e}^{-2k(1-k)^{-1}} + (1-k)2k(1-k)^{-1} - 1 \geq 0,$$

(see Bach (2010) for more details). So by taking $2k(1-k)^{-1} = R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2$, we have

$$\mathrm{e}^{-R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2} + R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2 - 1 \geq \frac{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2}{2 + R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2}$$

and we deduce that

$$\frac{\boldsymbol{\kappa'^2} R^2 |\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2}{2 + R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2} \leq 24\Gamma_1 \frac{\max\limits_{1\leq j\leq M} \omega_j^2}{\min\limits_{1\leq j\leq M} \omega_j^2} \max_{1\leq j\leq M} \omega_j s R^2 |\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2.$$

This implies that

$$R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2 \leq \frac{\dfrac{48\Gamma_1 Rs}{\boldsymbol{\kappa'^2}} \dfrac{\max\limits_{1\leq j\leq M} \omega_j^2}{\min\limits_{1\leq j\leq M} \omega_j^2} \max\limits_{1\leq j\leq M} \omega_j^2}{1 - \dfrac{24\Gamma_1 Rs}{\boldsymbol{\kappa'^2}} \dfrac{\max\limits_{1\leq j\leq M} \omega_j^2}{\min\limits_{1\leq j\leq M} \omega_j^2} \max\limits_{1\leq j\leq M} \omega_j^2} \leq 2$$

as soon a

$$\Gamma_1 \leq \frac{1}{48Rs} \frac{\min\limits_{1\leq j\leq M} \omega_j^2}{\max\limits_{1\leq j\leq M} \omega_j^2} \frac{\boldsymbol{\kappa'^2}}{\max\limits_{1\leq j\leq M} \omega_j}.$$

Since $\phi(t)/t^2$ is decreasing and bounded below by 0, we can deduce that

$$\frac{\phi(R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2)}{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2} \geq \frac{\phi(2)}{4}$$

and

$$\frac{\phi(-R|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2)}{R^2|\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0}|_2^2} \leq \frac{\phi(-2)}{4}.$$

Take $\xi := \phi(2)/4 > 0$ and $\xi' := \phi(-2)/4 > 0$ and conclude that

$$\xi||(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^{\boldsymbol{T}}\boldsymbol{X}||_{n,\Lambda}^2 \leq \widetilde{K}_n(\lambda_0, \lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) \leq \xi'||(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta_0})^{\boldsymbol{T}}\boldsymbol{X}||_{n,\Lambda}^2.$$

□

B An other empirical Bernstein's inequality

In this appendix, we establish an other version of the Bernstein's inequalities for martingales with jumps stated in Theorem 2.5. These inequalities are close to those obtained by Gaïffas & Guilloux (2012) in Theorem 3. We refer to Section 2.5 for the definitions of the processes $\eta_{n,t}(f_j)$ and $\nu_{n,t}(\theta_k)$ (see Equations (2.99)) and of their optional variations $\hat{V}_{n,t}(f_j)$ and $\hat{R}_{n,t}(\theta_k)$ (see Equations (2.26) and (2.27)).

Theorem 2.7. *For any numerical constant $c_\ell > 1$, $c'_\ell > 1$, $\varepsilon > 0$, $\varepsilon' > 0$ and $c_0 > 0$, $c'_0 > 0$ such that $ec_0 > 2(4/3+\varepsilon)c_\ell$ and $ec'_0 > 2(4/3+\varepsilon')c'_\ell$, the following holds for any $x > 0$, $y > 0$:*

$$\mathbb{P}\left[|\eta_{n,t}(f_j)| \geq c_{1,\varepsilon}\sqrt{\frac{x + \hat{\ell}_{n,x}(f_j)}{n}\hat{V}_{n,t}(f_j)} + c_{2,\varepsilon}\frac{x + 1 + \hat{\ell}_{n,x}(f_j)}{n}||f_j||_{n,\infty}\right] \leq c_{3,\varepsilon,c_\ell}\mathrm{e}^{-x}, \tag{2.69}$$

$$\mathbb{P}\left[|\nu_{n,t}(\theta_k)| \geq c'_{1,\varepsilon'}\sqrt{\frac{y + \hat{\ell}'_{n,y}(\theta_k)}{n}\hat{R}_{n,t}(\theta_k)} + c'_{2,\varepsilon'}\frac{y + 1 + \hat{\ell}'_{n,y}(\theta_k)}{n}||\theta_k||_{n,\infty}\right] \leq c'_{3,\varepsilon',c'_\ell}\mathrm{e}^{-y}, \tag{2.70}$$

where $||f_j||_{n,\infty} = \max_{i=1,\dots,n}|f_j(\boldsymbol{Z_i})|$ and $||\theta_k||_{n,\infty} = \max_{t\in[0,\tau]}|\theta_k(t)|$,

$$\hat{\ell}_{n,x}(f_j) = c_\ell \log\log\left(\frac{2en\hat{V}_{n,t}(f_j) + 8e(4/3+\varepsilon)x||f_j||_{n,\infty}^2}{4(ec_0 - 2(4/3+\varepsilon)c_\ell)||f_j||_{n,\infty}^2} \vee e\right),$$

$$\hat{\ell}'_{n,y}(\theta_k) = c'_\ell \log\log\left(\frac{2en\hat{R}_{n,t}(\theta_k) + 8e(4/3+\varepsilon')y||\theta_k||_{n,\infty}^2}{4(ec'_0 - 2(4/3+\varepsilon')c'_\ell)||\theta_k||_{n,\infty}^2} \vee e\right),$$

and where

$$c_{1,\varepsilon} = 2\sqrt{1+\varepsilon}, \quad c_{2,\varepsilon} = 2\sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon))} + 2/3,$$

and $c_{3,\varepsilon,c_\ell} = 8 + 6(\log(1+\varepsilon))^{-c_\ell} \sum_{k\geq 1} k^{-c_\ell}$,

$$c'_{1,\varepsilon'} = 2\sqrt{1+\varepsilon'}, \quad c'_{2,\varepsilon'} = 2\sqrt{2\max(c'_0, 2(1+\varepsilon')(4/3+\varepsilon'))} + 2/3,$$

and $c'_{3,\varepsilon',c'_\ell} = 8 + 6(\log(1+\varepsilon'))^{-c'_\ell} \sum_{k\geq 1} k^{-c'_\ell}$.

In this theorem, $\hat{\ell}_{n,x}(f_j)$ and $\hat{\ell}'_{n,y}(\theta_k)$ are small technical terms coming out of the proof. There are data-driven terms that does not depend on unknown parameters as it is the case in $W_n^\nu(f_j)$ and $T_n^{\tilde{\nu}}(\theta_k)$ (see Equations (2.30) and (2.31)). From Theorem 2.7, we deduce the following weights:

$$\omega_j = c_{1,\varepsilon}\sqrt{\frac{x+\log M+\hat{\ell}_{n,x}(f_j)}{n}\hat{V}_n(f_j)} + c_{2,\varepsilon}\frac{x+1+\log M+\hat{\ell}_{n,x}(f_j)}{n}||f_j||_{n,\infty}$$

and

$$\delta_k = c'_{1,\varepsilon'}\sqrt{\frac{y+\log N+\hat{\ell}'_{n,y}(\theta_k)}{n}\hat{R}_n(\theta_k)} + c'_{2,\varepsilon'}\frac{y+1+\log N+\hat{\ell}'_{n,y}(\theta_k)}{n}||\theta_k||_\infty,$$

which are of order $\sqrt{(\log M/n)\hat{V}_n(f_j)}$ and $\sqrt{(\log N/n)\hat{R}_n(\theta_k)}$ respectively, so as the weights ω_j and δ_k defined by (2.5) and (2.6).

The proof of Theorem 2.7 is inspired from the proof of Theorem 3 in Gaïffas & Guilloux (2012).

Proof of Theorem 2.7:

The proofs of (2.69) and (2.70) are quite similar, so we only present the one of (2.69). To prove (2.70), it suffices to replace $\eta_{n,t}(f_j)$ by the process $\nu_{n,t}(\theta_k)$ throughout the following. We use the same notation than to prove Theorem 2.5, denoting by $U_{n,t}$ and $H_i(f_j)$ the quantities

$$U_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^n \int_0^t H_i(f_j)\mathrm{d}M_i(s) \text{ and } H_i(f_j) := \frac{f_j(\boldsymbol{Z_i})}{\max_{1\leq i\leq n}|f_j(\boldsymbol{Z_i})|}.$$

Since $H_i(f_j)$ is a bounded predictable process with respect to $\mathcal{F}_t$, $U_{n,t}(f_j)$ is a square integrable martingale. Its predictable variation is given by

$$\vartheta_{n,t}(f_j) = n < U_n(f_j) >_t = \frac{1}{n}\sum_{i=1}^n \int_0^t (H_i(f_j))^2 \mathrm{d}\Lambda_i(s)$$

and the optional variation of $U_{n,t}(f_j)$ is

$$\hat{\vartheta}_{n,t}(f_j) = n[U_n(f_j)]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (H_i(f_j))^2 \mathrm{d}N_i(s).$$

The standard Bernstein's inequality (see Uspensky (1937) or Massart (2007) for the classical Bernstein's inequality and van de Geer (1995) for the Bernstein's inequality for martingales) states that

$$\mathbb{P}\left[U_{n,t}(f_j) \geq \sqrt{\frac{2\omega x}{n}} + \frac{x}{3n}, \vartheta_{n,t}(f_j) \leq \omega\right] \leq \mathrm{e}^{-x}. \tag{2.71}$$

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

$$\begin{aligned}\left\{U_{n,t}(f_j) \geq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}\right\} \cap \{v < \vartheta_{n,t}(f_j) \leq \omega\} \\ \subset \left\{U_{n,t}(f_j) \geq \sqrt{\frac{2\omega x}{n}} + \frac{x}{3n}\right\} \cap \{v < \vartheta_{n,t}(f_j) \leq \omega\}.\end{aligned}$$

Consequently,

$$\mathbb{P}\left[U_{n,t}(f_j) \geq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, v < \vartheta_{n,t}(f_j) \leq \omega\right] \leq \mathrm{e}^{-x}. \tag{2.72}$$

Now, we aim at replacing $\vartheta_{n,t}(f_j)$ which is non observable by the observable $\hat{\vartheta}_{n,t}(f_j)$ in Bound (2.72). Let us consider the martingale given $\mathcal{F}_t$

$$\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (H_i(f_j))^2\Big(\mathrm{d}N_i(s) - \mathrm{d}\Lambda_i(s)\Big) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (H_i(f_j))^2 \mathrm{d}M_i(s).$$

From (2.72), we deduce that

$$\mathbb{P}\left[|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \geq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, v < \vartheta_{n,t}(f_j) \leq \omega\right] \leq 2\mathrm{e}^{-x}. \tag{2.73}$$

If $\vartheta_{n,t}(f_j)$ satisfies

$$|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \leq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}, \tag{2.74}$$

thanks to the fact that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ for any $a, A, b > 0$, taking $A = \vartheta_{n,t}(f_j)$, $a = \dfrac{2\omega x}{vn}$ and $b = \hat{\vartheta}_{n,t}(f_j) + \dfrac{x}{3n}$, we obtain

$$\vartheta_{n,t}(f_j) \leq 2\hat{\vartheta}_{n,t}(f_j) + 2\left(\frac{\omega}{v} + \frac{1}{3}\right)\frac{x}{n}. \tag{2.75}$$

Combining Inequality (2.75) and Inequality (2.74), we get

$$\hat{\vartheta}_{n,t}(f_j) \leq \vartheta_{n,t}(f_j) + \sqrt{\frac{2\omega x}{vn}\Big(2\hat{\vartheta}_{n,t}(f_j) + 2\Big(\frac{\omega}{v}+\frac{1}{3}\Big)\frac{x}{n}\Big)} + \frac{x}{3n}.$$

Once again using that $A \leq b + \sqrt{aA}$ entails $A \leq a + 2b$ with

$$A = \hat{\vartheta}_{n,t}(f_j), a = \frac{4\omega x}{vn} \text{ and } b = \vartheta_{n,t}(f_j) + \sqrt{\frac{4\omega x}{vn}\Big(\frac{\omega}{v}+\frac{1}{3}\Big)\frac{x}{n}} + \frac{x}{3n},$$

we obtain $\hat{\vartheta}_{n,t}(f_j) \leq 2\vartheta_{n,t}(f_j) + 2\Big(\frac{1}{3} + 2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v}+\frac{1}{3}\Big)} + \frac{2\omega}{v}\Big)\frac{x}{n}$. Now, we deduce from Inequality (2.75) that

$$\begin{aligned}\Big\{U_{n,t}(f_j) \leq &\sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}\Big\} \cap \Big\{|\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \leq \sqrt{\frac{2\omega\vartheta_{n,t}(f_j)x}{vn}} + \frac{x}{3n}\Big\} \\ &\subset \Big\{U_{n,t}(f_j) \leq 2\sqrt{\frac{\omega x}{vn}\hat{\vartheta}_{n,t}(f_j)} + \Big(2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v}+\frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}\Big\}. \qquad (2.76)\end{aligned}$$

Using (2.72) and (2.73), we finally obtain

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq 2\sqrt{\frac{\omega x}{vn}\hat{\vartheta}_{n,t}(f_j)} + \Big(2\sqrt{\frac{\omega}{v}\Big(\frac{\omega}{v}+\frac{1}{3}\Big)} + \frac{1}{3}\Big)\frac{x}{n}, v < \vartheta_{n,t}(f_j) \leq \omega\Big] \leq 3\mathrm{e}^{-x}. \qquad (2.77)$$

It remains to remove the event $\{v \leq \vartheta_{n,t}(f_j) < \omega\}$ in (2.77). For $k \geq 0$, set $v_k = c_0\frac{x+1}{n}(1+\varepsilon)^k$, and use the following decomposition into disjoint sets :

$$\{\vartheta_{n,t}(f_j) > c_0 x/n\} = \bigcup_{k\geq 0}\{v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\}. \qquad (2.78)$$

Instead of considering the event $\{v < \vartheta_{n,t}(f_j) \leq \omega\}$, we calculate the probabilities on $\{\vartheta_{n,t}(f_j) > v_0\}$ and on its complementary to finally get the expected probability. According to (2.77)

$$\mathbb{P}\Big[U_{n,t}(f_j) \geq c_{1,\epsilon}\sqrt{\frac{x}{n}\hat{\vartheta}_{n,t}(f_j)} + c_{2,\varepsilon}\frac{x}{n}, v_k < \vartheta_{n,t}(f_j) \leq v_{k+1}\Big] \leq 3\mathrm{e}^{-x},$$

with

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

Set for some constant $c_\ell > 1$,

$$\ell = c_\ell \log\log\left(\frac{\vartheta_{n,t}(f_j)}{v_0} \vee e\right).$$

On the event

$$D_{n,\ell,\varepsilon} = \left\{ |\hat{\vartheta}_{n,t}(f_j) - \vartheta_{n,t}(f_j)| \le \sqrt{\frac{2(1+\varepsilon)\vartheta_{n,t}(f_j)(x+\ell)}{n}} + \frac{x+\ell}{3n} \right\}$$

applying (2.75) with $\frac{\omega}{v} = 1+\varepsilon$ and replacing x by $x+\ell$, we have

$$\vartheta_{n,t}(f_j) \le 2\hat{\vartheta}_{n,t}(f_j) + 2(4/3+\varepsilon)\frac{x}{n} + \frac{2(4/3+\varepsilon)c_\ell}{n}\log\log\left(\frac{\vartheta_{n,t}(f_j)}{v_0} \vee e\right).$$

We now use the fact that $\log\log(x) \le x/e - 1$ for any $x \ge e$, and since $ec_0 > 2(4/3+\varepsilon)c_\ell$ we get

$$\vartheta_{n,t}(f_j) \le \frac{ec_0}{ec_0 - 2(4/3+\varepsilon)c_\ell}\left(2\hat{\vartheta}_{n,t}(f_j) + 2(4/3+\varepsilon)\frac{x}{n}\right).$$

Combining the last inequality with (2.76), we obtain the following embeddings :

$$\begin{aligned}\left\{ U_{n,t}(f_j) \le \sqrt{\frac{2(1+\varepsilon)\vartheta_{n,t}(f_j)(x+\ell)}{n}} + \frac{x+\ell}{3n} \right\} \cap D_{n,\ell,\varepsilon} \\ \subset \left\{ U_{n,t}(f_j) \le c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n} \right\}, \qquad (2.79)\end{aligned}$$

where $\hat{\ell} = c_\ell \log\log\Big(\frac{2en\hat{\vartheta}_{n,t}(f_j) + 2e(4/3+\varepsilon)x}{ec_0 - 2(4/3+\varepsilon)c_\ell} \vee e\Big)$. From (2.78), we have

$$\begin{aligned}&\mathbb{P}\left[U_{n,t}(f_j) \ge c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, \vartheta_{n,t}(f_j) > v_0 \right] \\ \le &\sum_{k\ge 0}\mathbb{P}\left[U_{n,t}(f_j) \ge c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, v_k < \vartheta_{n,t}(f_j) \le v_{k+1} \right].\end{aligned}$$

Then, we write

$$\begin{aligned}&\sum_{k\ge 0}\mathbb{P}\left[U_{n,t}(f_j) \ge c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, v_k < \vartheta_{n,t}(f_j) \le v_{k+1} \right] \\ =&\sum_{k\ge 0}\mathbb{P}\left[U_{n,t}(f_j) \ge c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, D_{n,\ell,\varepsilon}, v_k < \vartheta_{n,t}(f_j) \le v_{k+1} \right] \\ +&\sum_{k\ge 0}\mathbb{P}\left[U_{n,t}(f_j) \ge c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}} + c_{2,\varepsilon}\frac{x+\hat{\ell}}{n}, D^c_{n,\ell,\varepsilon}, v_k < \vartheta_{n,t}(f_j) \le v_{k+1} \right],\end{aligned}$$

where $D^c_{n,\ell,\varepsilon}$ is the complementary of $D_{n,\ell,\varepsilon}$. Applying (2.79), we get

$$\begin{aligned}
&\sum_{k\geq 0}\mathbb{P}\left[U_{n,t}(f_j)\geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}}+c_{2,\varepsilon}\frac{x+\hat{\ell}}{n},v_k<\vartheta_{n,t}(f_j)\leq v_{k+1}\right]\\
\leq&\sum_{k\geq 0}\mathbb{P}\left[U_{n,t}(f_j)\geq \sqrt{\frac{2(1+\varepsilon)\vartheta_{n,t}(f_j)(x+\ell)}{n}}+\frac{x+\ell}{3n},v_k<\vartheta_{n,t}(f_j)\leq v_{k+1}\right]\\
+&\sum_{k\geq 0}\mathbb{P}\left[D^c_{n,\ell,\varepsilon},v_k<\vartheta_{n,t}(f_j)\leq v_{k+1}\right].
\end{aligned}$$

Gathering (2.72) and (2.73), we finally obtain

$$\begin{aligned}
&\mathbb{P}\left[U_{n,t}(f_j)\geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}}+c_{2,\varepsilon}\frac{x+\hat{\ell}}{n},\vartheta_{n,t}(f_j)>v_0\right]\\
&\leq 3\left(\mathrm{e}^{-x}+\sum_{j\geq 1}\mathrm{e}^{-(x+c_\ell\log\log(v_j/v_0))}\right)\\
&=3(1+(\log(1+\varepsilon))^{-c_\ell}\sum_{j\geq 1}j^{-c_\ell})\mathrm{e}^{-x}.
\end{aligned}$$

We apply Inequality (2.71) with $\omega=v_0=c_0(x+1)/n$ to get

$$\mathbb{P}\left[U_{n,t}(f_j)\geq\left(\sqrt{2c_0}+\frac{1}{3}\right)\frac{x+1}{n},\vartheta_{n,t}(f_j)\leq v_0\right]\leq\mathrm{e}^{-x}. \tag{2.80}$$

According to (2.80), for $c_{3,\varepsilon}=\sqrt{2\max(c_0,2(1+\varepsilon)(4/3+\varepsilon))}+1/3$, we get

$$\mathbb{P}\left[U_{n,t}(f_j)\geq c_{1,\varepsilon}\sqrt{\frac{\hat{\vartheta}_{n,t}(f_j)(x+\hat{\ell})}{n}}+c_{3,\varepsilon}\frac{x+1+\hat{\ell}}{n}\right]\leq\left(4+3(\log(1+\varepsilon))^{-c_\ell}\sum_{j\geq 1}j^{-c_\ell}\right)\mathrm{e}^{-x}.$$

Now it suffices to multiply both sides of the inequality inside the probability by $||f_j||_{n,\infty}=\max_{i=1,\ldots,n}|f_j(\boldsymbol{Z_i})|$ to end up the proof of Theorem 2.7. □

C Weighted Lasso procedure in the specific case of the Cox model

In this appendix, we propose a weighted Lasso procedure to estimate the function of the covariates in the Cox model considering the partial log-likelihood criterion. We establish a slow non-asymptotic oracle inequality for the obtained estimator based on the same method than the one proposed in Section 2.3. This follows of Bradic & Song (2012).

Framework

Let consider a non-parametric Cox model defined for $i = 1, ..., n$ as

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t)\mathrm{e}^{f_0(\boldsymbol{Z_i})}, \tag{2.81}$$

where $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ is the vector of covariates of individual i, α_0 the baseline function and f_0 the regression function. We consider a high dimensional setting, i.e. $p \gg n$.

Let us then define a dictionary of functions of the covariates $\mathbb{F}_M = \{f_1, ..., f_M\}$, where $f : \mathbb{R}^p \to \mathbb{R}$. We refer to Subsection 2.2.2 for precisions on dictionaries. In high-dimension, we recall that $M \gg n$. We assume in the following that f_0 belongs to $\mathbb{F}_M$:

Assumption 2.9. *There exists* $\boldsymbol{\beta_0} \in \mathbb{R}^M$, *such that* $f_0 = f_{\boldsymbol{\beta_0}} = \sum_{j=1}^{M} \beta_{0_j} f_j$.

This assumption is rather strong but is needed if we want to establish an oracle inequality for the Lasso estimator based on the partial log-likelihood criterion. Indeed, this assumption allows to simplify the process that we have to control into a martingale (see Equation (2.87)).

Unlike in Section 2.2, where we estimate both parameters of a Cox model to estimate any intensity of a counting process, we focus here on the estimation of the regression function f_0 in the Cox model. Thus, instead of the total empirical log-likelihood 2.2, we consider the partial log-likelihood specific to the Cox model. This criterion introduced by Cox (1972) allows the estimation of the regression function f_0 without the knowledge of the baseline function α_0. It is defined for the non-parametric Cox model (2.81) as

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{f_{\boldsymbol{\beta}}(\boldsymbol{Z_i})}}{S_n(t, f_{\boldsymbol{\beta}})}\right)\mathrm{d}t, \quad \text{with } S_n(t, f_{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{k=1}^{n} = Y_k(t)\mathrm{e}^{f_{\boldsymbol{\beta}}(\boldsymbol{Z_k})},$$

and $f_{\boldsymbol{\beta}} \in \mathbb{F}_M$, so that $f_{\boldsymbol{\beta}} = \sum_{j=1}^{M} \beta_j f_j$.

From Assumption 2.9, the parameter to estimate is $\boldsymbol{\beta_0} \in \mathbb{R}^M$. We apply the following Lasso procedure:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{M}}}{\arg\min}\{-l_n^*(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \quad \text{with } \mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M}\omega_j|\beta_j|,$$

where $\omega = (\omega_1, ..., \omega_M)$ are data-driven weights to be determined.

Notations

We introduce the following two classical notations (see Andersen et al. (1993) for example):

$$E_n(t, \boldsymbol{\beta}) = \frac{S^{(1)}(t, f_{\boldsymbol{\beta}})}{S_n(t, f_{\boldsymbol{\beta}})} \quad \text{and} \quad V_n(t, \boldsymbol{\beta}) = \frac{S^{(2)}(t, f_{\boldsymbol{\beta}})}{S_n(t, f_{\boldsymbol{\beta}})} - \left(\frac{S^{(1)}(t, f_{\boldsymbol{\beta}})}{S_n(t, f_{\boldsymbol{\beta}})}\right)^{\otimes 2}, \tag{2.82}$$

where $\otimes$ denotes the outer product and for $l = 1, 2$,

$$S_n^{(l)}(t, f_{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{k=1}^{n} \vec{\boldsymbol{f}}^{\otimes l}(\boldsymbol{Z_k}) Y_k(t) \mathrm{e}^{f_{\boldsymbol{\beta}}(\boldsymbol{Z_k})}, \quad \text{where } \vec{\boldsymbol{f}} = (f_1, ..., f_M)^T.$$

With the notations (2.82) at hand, the score vector and the hessian of the partial log-likelihood have the following representations respectively

$$-\nabla l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau (E_n(\boldsymbol{\beta}, t) - \vec{\boldsymbol{f}}(\boldsymbol{Z_i})) dN_i(t), \text{ where } \vec{\boldsymbol{f}} = (f_1, ..., f_M)^T \tag{2.83}$$

$$-\nabla^2 l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau V_n(\boldsymbol{\beta}, t) dN_i(t) \tag{2.84}$$

Lastly, let us define the empirical functional norm $||.||_{n,b*}$ of all functions $f_b : \mathbb{R}^p \to \mathbb{R}$ for any $b \in \mathbb{R}^p$ and a fixed $b^* \in \mathbb{R}^p$ as

$$||f_b||_{n,b^*}^2 = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau Y_i(t) \varpi_i(t, \boldsymbol{b^*}) f_b^2(\boldsymbol{Z_i}) d\bar{N}(t) - \left[\frac{1}{n}\sum_{i=1}^{n} \int_0^\tau Y_i(t) \varpi_i(t, \boldsymbol{b^*}) f_b(\boldsymbol{Z_i}) d\bar{N}(t)\right]^2,$$

with $\varpi_i(t, \boldsymbol{b^*}) = \dfrac{e^{f_{b^*}(\boldsymbol{Z_i})}}{S_n^{(0)}(t, b^*)}$ and $\bar{N} = n^{-1}\sum_{i=1}^n N_i$, and we can rewrite this norm

$$||f_b||_{n,b^*}^2 = \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau Y_i(t) \varpi_i(t, \boldsymbol{b^*}) (f_b(\boldsymbol{Z_i}) - \bar{f}_{b^*}(t))^2 d\bar{N}(t),$$

where $\bar{f}_{b^*}(t) = \dfrac{1}{n}\sum_{i=1}^{n} Y_i(t) \varpi_i(t, \boldsymbol{b^*}) f_b(\boldsymbol{Z_i})$.

Relation between the partial log-likelihood and the empirical functional norm

Let denote $C_n(\boldsymbol{\beta}) = -l_n^*(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathbb{R}^M$. We can write $C_n(f_{\boldsymbol{\beta}}) = -l_n^*(\boldsymbol{\beta}) + l_n^*(\boldsymbol{\beta_0}) - l_n^*(\boldsymbol{\beta_0})$. From a Taylor expansion around $\boldsymbol{\beta_0}$, there exists $c \in (0, 1)$ and $\boldsymbol{\beta^*} = c\boldsymbol{\beta} + (1 - c)\boldsymbol{\beta_0}$ such that

$$C_n(\boldsymbol{\beta}) = -(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T \nabla l_n^*(\boldsymbol{\beta_0}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T \nabla^2 l_n^*(\boldsymbol{\beta^*})(\boldsymbol{\beta} - \boldsymbol{\beta_0}) - l_n^*(\boldsymbol{\beta_0}). \tag{2.85}$$

Now, from the definition the hessian (2.84) of the partial log-likelihood, we can rewrite (2.85) as

$$C_n(\boldsymbol{\beta}) - C_n(\boldsymbol{\beta_0}) = -(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T \nabla l_n^*(\boldsymbol{\beta_0}) - \frac{1}{2n}\sum_{i=1}^{n} \int_0^\tau (\boldsymbol{\beta} - \boldsymbol{\beta_0})^T V_n(\boldsymbol{\beta^*}, t)(\boldsymbol{\beta} - \boldsymbol{\beta_0}) dN_i(t)$$

With the notations introduced just above in the previous paragraph and by simple algebraic manipulations, the following quadratic representation of the hessian matrix holds:

$$-\boldsymbol{\beta}^T \nabla^2 l_n^*(\boldsymbol{\beta^*})\boldsymbol{\beta} = ||f_{\boldsymbol{\beta}}||^2_{n,\beta^*}.$$

So for all $\boldsymbol{\beta} \in \mathbb{R}^M$, there exists $c \in (0,1)$ and $\boldsymbol{\beta^*} = c\boldsymbol{\beta} + (1-c)\boldsymbol{\beta_0}$ such that

$$C_n(\boldsymbol{\beta}) = -(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T h_n(\boldsymbol{\beta_0}) + \frac{1}{2}||f_{\boldsymbol{\beta}} - f_0||^2_{n,\beta^*} - l_n^*(\boldsymbol{\beta_0}), \tag{2.86}$$

with

$$h_n(\boldsymbol{\beta_0}) = \nabla l_n^*(\boldsymbol{\beta_0}) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau (E_n(\boldsymbol{\beta_0}, t) - \vec{\boldsymbol{f}}(\boldsymbol{Z_i}))dN_i(t).$$

From Assumption 2.9 and using the Doob-Meyer decomposition (**??**) we deduce that

$$h_n(\boldsymbol{\beta_0}) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau (E_n(\boldsymbol{\beta_0}, t) - \vec{\boldsymbol{f}}(\boldsymbol{Z_i}))\mathrm{d}M_i(t) \tag{2.87}$$

Bernstein Inequality and oracle inequality

Now, as in Section 2.5, let us sketch the proof to obtain the slow oracle inequality. By definition, for all $\boldsymbol{\beta} \in \mathbb{R}^M$,

$$C_n(f_{\hat{\beta}_L}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) \leq C_n(f_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta})$$

Thus replacing $C_n(f_{\hat{\beta}_L})$ with Equation (2.86), we obtain

$$||f_{\hat{\beta}_L} - f_0||^2_{n,\boldsymbol{\beta^*}} \leq ||f_{\boldsymbol{\beta}} - f_0||^2_{n,\boldsymbol{\beta^*}} + 2(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} - \boldsymbol{\beta})^T h_n(\boldsymbol{\beta_0}) + 2\,\mathrm{pen}(\boldsymbol{\beta}) - 2\,\mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}).$$

We have to control the process $h_n(\boldsymbol{\beta_0})$.

For $j = 1, ..., M$, let introduce the following process:

$$h_{n,j}(\boldsymbol{\beta_0}) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau ((E_n(\boldsymbol{\beta_0}, t))_j - f_j(\boldsymbol{Z_i}))dM_i(t).$$

Its predictable variation is then defined as

$$H_{n,j}(\boldsymbol{\beta_0}) = n < h_{n,j}(\boldsymbol{\beta_0}) >= \frac{1}{n}\sum_{i=1}^n \int_0^\tau ((E_n(\boldsymbol{\beta_0}, t))_j - f_j(\boldsymbol{Z_i}))^2 \alpha_0(t) e^{f_0(\boldsymbol{Z_i})} Y_i(t)dt$$

and the optional variation as

$$\hat{H}_{n,j}(\boldsymbol{\beta_0}) = n[h_{n,j}(\boldsymbol{\beta_0})] = \frac{1}{n}\sum_{i=1}^n \int_0^\tau ((E_n(\boldsymbol{\beta_0}, t))_j - f_j(\boldsymbol{Z_i}))^2 dN_i(t).$$

Remark 2.9. Neither the predictable process nor the optional process are observable. However, we can bound the optional process as follows,

$$|\hat{H}_{n,j}(\boldsymbol{\beta_0})| = \left| \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left(\frac{\sum_{k=1}^n (f_j(\boldsymbol{Z_k}) - f_j(\boldsymbol{Z_i}))Y_k(t)e^{f_0(\boldsymbol{Z_k})}}{\sum_{k=1}^n Y_k(t)e^{f_0(\boldsymbol{Z_k})}} \right)^2 dN_i(t) \right|$$
$$\leq \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \sup_k |f_j(\boldsymbol{Z_k}) - f_j(\boldsymbol{Z_i})|^2 dN_i(t) =: \hat{K}_{n,j}. \tag{2.88}$$

Then, as in Section 2.5, we can apply an empirical Bernstein inequality for martingales with jumps to the process $h_{n,j}(\boldsymbol{\beta_0})$. We follow the proof of Theorem 2.7 in Appendix B to establish an empirical Bernstein inequality for $h_{n,j}(\boldsymbol{\beta_0})$.

Theorem 2.8. *For any numerical constant $c_\ell > 1$, $\varepsilon > 0$ and $c_0 > 0$ such that $ec_0 > 2(4/3+\varepsilon)c_\ell$, the following holds for any $x > 0$:*

$$\mathbb{P}\left[|h_{n,j}(\boldsymbol{\beta_0})| \geq c_{1,\varepsilon} \sqrt{\frac{\hat{H}_{n,j}(\boldsymbol{\beta_0})(x + \hat{\ell}_{n,x}(\boldsymbol{\beta_0}))}{n}} + c_{2,\varepsilon} \frac{x + 1 + \hat{\ell}_{n,x}(\boldsymbol{\beta_0})}{n} ||f_j||_{n,\infty} \right] \leq c_{3,\varepsilon,c_\ell} e^{-x},$$

with

$$\hat{\ell}_{n,x}(\boldsymbol{\beta_0}) = c_\ell \log\log \left(\frac{2en\hat{H}_{n,j}(\boldsymbol{\beta_0}) + 2e(4/3+\varepsilon)x||f_j||_{n,\infty}^2}{(ec_0 - 2(4/3+\varepsilon)c_\ell)||f_j||_{n,\infty}} \vee e \right),$$

$$c_{1,\varepsilon} = 2\sqrt{1+\varepsilon}, \quad c_{2,\varepsilon} = 2\sqrt{(1+\varepsilon)(4/3+\varepsilon)} + 1/3$$

and

$$c_{3,\varepsilon,c_\ell} = \sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon)) + 1/3}.$$

From Bound (2.88), we can replace the optional process, which is not observable, by $\hat{K}_{n,j}$.

Corollary 2.2. *For any numerical constant $c_\ell > 1$, $\varepsilon > 0$ and $c_0 > 0$ such that $ec_0 > 2(4/3+\varepsilon)c_\ell$, the following holds for any $x > 0$:*

$$\mathbb{P}\left[|h_{n,j}(\boldsymbol{\beta_0})| \geq c_{1,\varepsilon} \sqrt{\frac{\hat{K}_{n,j}(\boldsymbol{\beta_0})(x + \hat{\ell}'_{n,x})}{n}} + c_{2,\varepsilon} \frac{x + 1 + \hat{\ell}'_{n,x}}{n} ||f_j||_{n,\infty} \right] \leq c_{3,\varepsilon,c_\ell} e^{-x}, \tag{2.89}$$

with

$$\hat{\ell}'_{n,x}(\boldsymbol{\beta_0}) = c_\ell \log\log \left(\frac{2en\hat{K}_{n,j}(\boldsymbol{\beta_0}) + 2e(4/3+\varepsilon)x||f_j||_{n,\infty}^2}{(ec_0 - 2(4/3+\varepsilon)c_\ell)||f_j||_{n,\infty}} \vee e \right).$$

$$c_{1,\varepsilon} = 2\sqrt{1+\varepsilon}, \quad c_{2,\varepsilon} = 2\sqrt{(1+\varepsilon)(4/3+\varepsilon)} + 1/3$$

and

$$c_{3,\varepsilon,c_\ell} = \sqrt{2\max(c_0, 2(1+\varepsilon)(4/3+\varepsilon)) + 1/3}.$$

From Inequality (2.89), we deduce the following data-driven weights:

$$\omega_j = c_{1,\varepsilon}\sqrt{\hat{K}_{n,j}(\boldsymbol{\beta_0})\frac{(x+\log M+\hat{\ell}'_{n,x})}{n}} + c_{2,\varepsilon}\frac{x+\log M+1+\hat{\ell}'_{n,x}}{n}||f_j||_{n,\infty}. \quad (2.90)$$

To state the slow non-asymptotic oracle inequality, it suffices to follows the proof of Proposition 2.2.

Proposition 2.6. *Consider Model (2.81). Let $x > 0$ be fixed, ω_j be defined by (2.90) and for $\boldsymbol{\beta} \in \mathbb{R}^M$,*

$$\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j|.$$

Let $A_\varepsilon(x)$ be some numerical positive constant depending only on ε and x. Under Assumption 2.9, with a probability larger than $1 - A_\varepsilon(x)e^{-x}$, then

$$||f_{\hat{\beta}_L} - f_0||^2_{n,\beta^*} \leq \inf_{\boldsymbol{\beta}\in\mathbb{R}^M}\{||f_{\boldsymbol{\beta}} - f_0||^2_{n,\beta^*} + 2\,\mathrm{pen}(\boldsymbol{\beta})\}.$$

Towards practical applications

The advantage of this approach is that it is based on the partial log-likelihood. Indeed, as we have already pointed out in Remark 2.4, existing algorithms for the Lasso in the Cox model are implemented for the partial log-likelihood (e.g. in the R function glmnet). We could not directly use these algorithms to compare unweighted to weighted procedures in the context of Chapter 2, since our procedure is based on the total empirical log-likelihood (see Section 2.2). In the specific framework of Appendix C, we consider the partial log-likelihood and we propose a weighted procedure. An interesting perspective would be to compare in this specific case, unweighted to weighted procedures in terms of selection, estimation or prediction accuracies.

D Detailed comparison with the additive regression model

Dans cette annexe, nous allons faire une comparaison détaillée de l'étude du Lasso en régression additive et dans notre modèle de processus de comptage. Pour le modèle de régression additive, nous nous sommes inspirés de la démarche de Bickel et al. (2009) dans le cas gaussien et nous l'avons adapté au cas de bruits sous-exponentiels.

Analogies et différences entre les modèles :

1. **Modèle de régression additive**

Soient $(\boldsymbol{Z_1}, Y_1), ..., (\boldsymbol{Z_n}, Y_n)$ un échantillon de n couples aléatoires indépendants $(\boldsymbol{Z_i}, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ tels que

$$Y_i = f_0(\boldsymbol{Z_i}) + W_i, \quad i = 1, ..., n, \tag{2.91}$$

où $f_0 : \mathbb{R}^p \to \mathbb{R}$ est une fonction de régression à estimer, les $\boldsymbol{Z_i}$ sont déterministes et les erreurs de régression W_i sont centrées et de variance σ^2. Dans Bickel et al. (2009), les erreurs de régression considérées sont gaussiennes. Nous les considérerons sous-exponentielles pour faciliter la comparaison avec le modèle à intensité multiplicative d'Aalen.

2. **Modèle à intensité multiplicative d'Aalen**

   Nous observons n copies indépendantes $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$, où $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ est le vecteur de covariables, N_i un processus de comptage marqué, Y_i un processus aléatoire à valeur dans $[0, 1]$ et $[0, \tau]$ un intervalle de temps. La filtration $(\mathcal{F}_t)_{t \geq 0}$ est définie par

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

   D'après la décomposition de Doob-Meyer, on peut écrire, pour tout $i \in \{1, ..., n\}$, l'heuristique suivante :

$$\mathrm{d}N_i(t) = \lambda_0(t, \boldsymbol{Z_i}) Y_i(t) \mathrm{d}t + \mathrm{d}M_i(t), .$$

   où M_i est une martingale par rapport à $(\mathcal{F}_t)_{t \geq 0}$. Si l'on compare au modèle de régression additif (2.91), $\mathrm{d}M_i(.)$ peut être interprété comme un terme de bruit sous-exponentiel.

Procédures d'estimation :

1. **Modèle de régression additive**

   Soit $\mathbb{F}_M = \{f_1, ..., f_M\}$ un dictionnaire de fonctions. Les fonctions candidates pour estimer f_0 sont définies par $f_{\boldsymbol{\beta}} = \sum_{j=1}^M \beta_j f_j$, où $\boldsymbol{\beta} \in \mathbb{R}^M$. L'estimateur Lasso est alors défini par

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^M}{\arg\min} \{C_n(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \quad \text{avec } \mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^M \omega_j |\beta_j|,$$

   où

$$C_n(f_{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^n (Y_i - f_{\boldsymbol{\beta}}(\boldsymbol{Z_i}))^2 \tag{2.92}$$

   est le critère des moindres carrés et les $(\omega_j)_{j \in \{1,...,M\}}$ sont des poids construits à partir des données.

Remarque 2.1. Bickel et al. (2009) ont considéré des poids $\omega_j = \Gamma_n ||f_j||_n$ où Γ_n est d'ordre $\sqrt{\log M/n}$.

2. **Modèle à intensité multiplicative d'Aalen**

Soient $\mathbb{F}_M = \{f_1, ..., f_M\}$, où $f_j : \mathbb{R}^p \to \mathbb{R}$ et $\mathbb{G}_N = \{\theta_1, ..., \theta_N\}$, où $\theta_k : \mathbb{R}_+^* \to \mathbb{R}$ deux dictionnaires de fonctions. Les fonctions candidates pour estimer λ_0 sont définies par $\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z_i}) = \alpha_{\boldsymbol{\gamma}}(t)\mathrm{e}^{f_{\boldsymbol{\beta}}(\boldsymbol{Z_i})}$ avec

$$\log \alpha_{\boldsymbol{\gamma}} = \sum_{k=1}^{N} \gamma_k \theta_k \quad \text{et} \quad f_{\boldsymbol{\beta}} = \sum_{j=1}^{M} \beta_j f_j.$$

La procédure Lasso simultanée pour estimer les deux paramètres $\boldsymbol{\beta} \in \mathbb{R}^M$ et $\boldsymbol{\gamma} \in \mathbb{R}^N$ est définie par

$$(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})\in\mathbb{R}^M\times\mathbb{R}^N}{\arg\min} \{C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\gamma})\},$$

avec

$$\mathrm{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{M} \omega_j |\beta_j| \text{ et } \mathrm{pen}(\boldsymbol{\gamma}) = \sum_{k=1}^{N} \delta_k |\gamma_k|,$$

où $(\omega_j)_{j\in\{1,...,M\}}$ et $(\delta_k)_{k\in\{1,...,N\}}$ sont des poids positifs construits à partir des données.

Démarches pour obtenir une inégalité oracle lente :

1. **Modèle de régression additive**

Par définition de l'estimateur Lasso, pour tout $\boldsymbol{\beta} \in \mathbb{R}^M$,

$$C_n(f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) \leq C_n(f_{\boldsymbol{\beta}}) + \mathrm{pen}(\boldsymbol{\beta}).$$

Par définition (2.92) du critère des moindres carrés, on en déduit que

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq ||f_0 - f_{\boldsymbol{\beta}}||_n^2 + \frac{2}{n}\sum_{i=1}^{n}(f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_{\boldsymbol{\beta}})(\boldsymbol{Z_i})W_i + \mathrm{pen}(\boldsymbol{\beta}) + \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}). \tag{2.93}$$

Pour tout $\boldsymbol{\beta} \in \mathbb{R}^M$, $f_{\boldsymbol{\beta}} = \sum_{j=1}^{M} \beta_j f_j$. Donc en remplaçant $f_{\boldsymbol{\beta}}$ et $f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}}$ par leur expression dans (2.93), on obtient finalement

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq ||f_{\boldsymbol{\beta}} - f_0||_n^2 + \sum_{j=1}^{M}(\hat{\beta}_L - \beta)_j \eta_n(f_j) + \mathrm{pen}(\boldsymbol{\beta}) - \mathrm{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}), \tag{2.94}$$

avec $\eta_n(f_j) = \frac{1}{n}\sum_{i=1}^n f_j(\boldsymbol{Z_i})W_i$. Dans le cas sous-exponentiel, nous pouvons appliquer l'*inégalité de Bernstein* standard (cf. Massart (2007)), définie à l'Annexe A.3.1, au processus centré $V_n(f_j)$. Ainsi,

$$\mathbb{P}(|\eta_n(f_j)| \geq \sqrt{2vx} + cx) \leq 2e^{-x} \tag{2.95}$$

avec

$$v = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[(f_j(\boldsymbol{Z_i})W_i)^2] \quad \text{et} \quad c = \frac{1}{3n}\max_{1\leq i\leq n}|f_j(\boldsymbol{Z_i})W_i|.$$

Considérons alors des poids ω_j de la forme

$$\omega_j = c_{j,1}\sqrt{\frac{x+\log M}{n}} + c_{j,2}\frac{x+\log M}{n} \tag{2.96}$$

où $c_{j,1} = \sqrt{2\mathbb{E}[(f_j(\boldsymbol{Z_1})W_1)^2]}$ et $c_{j,2} = \frac{1}{3}\max_{1\leq i\leq n}|f_j(\boldsymbol{Z_i})W_i|$. On note $\mathcal{A}$ l'ensemble défini par

$$\mathcal{A} = \bigcap_{j=1}^M \{|\eta_n(f_j)| \leq \omega_j\}.$$

D'après (2.94), sur $\mathcal{A}$, nous avons donc

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq ||f_{\boldsymbol{\beta}} - f_0||_n^2 + \sum_{j=1}^M (\hat{\beta}_L - \beta)_j\omega_j + \sum_{j=1}^M \omega_j|\beta_j| - \sum_{j=1}^M \omega_j|\hat{\beta}_{L,j}|.$$

Et donc, sur $\mathcal{A}$, en prenant l'infimum sur les $\boldsymbol{\beta}$, nous obtenons l'inégalité oracle lente pour le modèle de régression additif suivante :

$$||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||_n^2 \leq \inf_{\boldsymbol{\beta}\in\mathbb{R}^M}\{||f_{\boldsymbol{\beta}} - f_0||_n^2 + 2\,\mathrm{pen}(\boldsymbol{\beta})\}. \tag{2.97}$$

À ce stade, il ne reste plus qu'à calculer $\mathbb{P}(\mathcal{A})$. Pour cela, nous calculons la probabilité du complémentaire. D'après (2.96), en appliquant l'inégalité de Bernstein (2.95) en remplaçant x par $x + \log M$, on en déduit

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\Big(\bigcup_{j=1}^M |\eta_n(f_j)| \geq \omega_j\Big) \leq \sum_{j=1}^M \mathbb{P}(|\eta_n(f_j)| \geq \omega_j) \leq 2M\mathrm{e}^{-x-\log M} \leq 2\mathrm{e}^{-x}.$$

L'inégalité oracle lente (2.97) est donc vérifiée avec probabilité supérieure à $1 - 2\mathrm{e}^{-x}$. Le $\log M$ que nous avons fait apparaître dans les poids, permet simplement d'obtenir une probabilité plus simple. Il apparaît dans la définition des poids, et la pénalité qui apparaît dans l'inégalité oracle est donc de l'ordre de $\sqrt{\log M/n}$, vitesse caractéristique d'une inégalité oracle dite lente.

Remarque 2.2. Pour des erreurs de régression gaussiennes, on considère plutôt une inégalité de déviation propre aux variables aléatoires gaussiennes (voir Bickel et al. (2009)). Pour des erreurs de régression centrées et bornées, on peut appliquer une inégalité de Hoeffding.

2. **Modèle à intensité multiplicative d'Aalen**

En suivant la même démarche que celle du modèle de régression additif, par définition, pour tout $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in $\mathbb{R}^M \times \mathbb{R}^N$,

$$C_n(\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}) + \text{pen}(\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}) + \text{pen}(\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}) \leq C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \text{pen}(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\gamma}). \tag{2.98}$$

De façon analogue, l'inégalit'e (2.98) devient

$$\begin{aligned}\widetilde{K}_n(\lambda_0, \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) + \sum_{j=1}^{M}(\hat{\beta}_{L,j} - \beta_j)\eta_{n,\tau}(f_j) + \sum_{j=1}^{M}\omega_j(|\beta_j| - |\hat{\beta}_{L,j}|) \\ + \sum_{k=1}^{N}(\hat{\gamma}_{L,k} - \gamma_k)^T\nu_{n,\tau}(\theta_k) + \sum_{k=1}^{N}\delta_k(|\gamma_k| - |\hat{\gamma}_{L,k}|),\end{aligned}$$

avec

$$\eta_{n,t}(f_j) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t f_j(\boldsymbol{Z_i})\mathrm{d}M_i(s) \text{ et } \nu_{n,t}(\theta_k) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \theta_k(s)\mathrm{d}M_i(s). \tag{2.99}$$

Ces processus sont des martingales par rapport à la filtration $\mathcal{F}_t$. Ils requierent donc l'application d'inégalités de Bernstein pour les martingales comme celles présentées en Annexe A.4 (voir le théorème A.4). En appliquant le théorème A.4 au processus $\eta_{n,t}(f_j)$, on obtient

$$\mathbb{P}\Big[\eta_{n,t}(f_j) \geq \Big(\sqrt{\frac{2\omega x}{n}} + \frac{x}{3n}\Big)||f_j||_{n,\infty}, V_{n,t}(f_j) \leq \omega\Big] \leq \mathrm{e}^{-x} \tag{2.100}$$

où $V_{n,t}(f_j)$ est le processus prévisible de $\eta_{n,t}(f_j)$ défini par

$$V_{n,t}(f_j) = n < \eta_n(f_j) >_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (f_j(\boldsymbol{Z_i}))^2\lambda_0(t, \boldsymbol{Z_i})Y_i(s)\mathrm{d}s.$$

Ce processus prévisible dépend de la fonction inconnue λ_0. Il en est de même du processus prévisible de $\nu_{n,t}(\theta_k)$. La démonstration nécessite donc une étape supplémentaire de façon à remplacer les processus prévisibles par leurs versions optionnelles définies respectivement par

$$\hat{V}_{n,t}(f_j) = n[\eta_n(f_j)]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (f_j(\boldsymbol{Z_i}))^2\mathrm{d}N_i(s) \tag{2.101}$$

et

$$\hat{R}_{n,t}(\theta_k) = n[\nu_n(\theta_k)]_t = \frac{1}{n}\sum_{i=1}^{n}\int_0^t (\theta_k(t))^2\mathrm{d}N_i(s). \tag{2.102}$$

Nous avons ainsi établi les inégalités de Bernstein empiriques (2.28), (2.29), (2.69) et (2.70) pour chacun des deux processus à contrôler. Des stratégies équivalentes

dans différents modèles ont été considérées par Hansen et al. (2012) et Gaïffas & Guilloux (2012). Notre démarche de preuve s'inspire des preuves de ces derniers. Nous renvoyons au Chapitre 2 pour une discussion détaillée sur chacune des stratégies. Les poids ω_j et δ_k découlent de l'application des inégalités de Bernstein via la contrainte $|\eta_{n,\tau}(f_j)| \leq \omega_j$ (respectivement $|\nu_{n,\tau}(\theta_k)| \leq \delta_k$) avec grande probabilité. Les poids que nous obtenons ont une forme complexe dont nous pouvons extraire l'ordre de grandeur :

$$\omega_j \approx \sqrt{\frac{\log M}{n}\hat{V}_{n,\tau}(f_j)} \quad \text{et} \quad \delta_k \approx \sqrt{\frac{\log N}{n}\hat{R}_{n,\tau}(\theta_k)}.$$

Nous renvoyons à la démonstration du Theorem 2.3 pour la fin de la preuve permettant d'obtenir l'inégalité oracle lente (2.19) (la preuve est similaire à celle de Bickel et al. (2009) mais sur deux ensembles $\mathcal{A}$ et $\mathcal{B}$ pour chacun des deux processus à contrôler).

Hypothèses aux valeurs propres restreintes (hypothèses RE) :

Les inégalités oracles rapides nécessitent une hypothèse sur la matrice de Gram. En effet, lorsque $p > n$, la matrice de Gram n'est plus inversible (voir Section 1.2.1 au Chapitre 1). L'hypothèse des valeurs propres restreintes introduite par Bickel et al. (2009) dans le cas du modèle de régression additive est alors l'une des hypothèses les plus faibles sur la matrice de Gram, permettant d'obtenir des inégalités oracles rapides. Nous avons adapté cette hypothèse à notre cadre de travail.

1. **Modèle de régression additive**

   La matrice de design est définie par $\boldsymbol{X} = (f_j(\boldsymbol{Z_i}))_{i,j}$ pour $i \in \{1, ..., n\}$ et $j \in \{1, ..., M\}$ et la matrice de Gram associée par $\boldsymbol{\Psi_n} = \frac{1}{n}\boldsymbol{X^T X}$.

   Hypothèse 2.1 (Condition $\mathbf{RE}(s, c_0)$ pour le modèle de régression additif)**.** *Pour un entier $s \in \{1, ..., M\}$ et une constante $c_0 > 0$, on dit que la condition aux valeurs propres restreintes (Restricted Eigenvalue Condition)* $\mathbf{RE}(s, c_0)$ *est vérifiée si*

$$0 < \boldsymbol{\kappa}(s, c_0) = \min_{\substack{J \subset \{1,...,M\}, \\ |J| \leq s}} \min_{\substack{\boldsymbol{b} \in \mathbb{R}^M \setminus \{0\}, \\ |\boldsymbol{b}_{J^c}|_1 \leq c_0 |\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b^T \Psi_n b})^{1/2}}{|\boldsymbol{b}_J|_2}. \tag{2.103}$$

   Nous renvoyons à la Section 1.2.1 de l'introduction de la thèse pour une interprétation détaillée de cette hypothèse.

2. **Modèle à intensité multiplicative d'Aalen**

La matrice de design associée à notre cadre de travail est définie par

$$\tilde{\boldsymbol{X}}(t)=\left[\begin{array}{c|ccc} & \theta_1(t) & \dots & \theta_N(t) \\ \boldsymbol{X} & \vdots & & \vdots \\ & \theta_1(t) & \dots & \theta_N(t) \end{array}\right] \in \mathbb{R}^{n\times(M+N)}$$

où $\boldsymbol{X} = (f_j(\boldsymbol{Z_i}))_{1\leq j\leq M, 1\leq i\leq n}$. Cette matrice de design fait intervenir les fonctions des deux dictionnaires $\mathbb{F}_M$ et $\mathbb{G}_N$. La matrice de Gram associée est alors définie par

$$\tilde{\boldsymbol{G}}_{\boldsymbol{n}} = \frac{1}{n}\int_0^\tau \tilde{\boldsymbol{X}}(t)^T\tilde{\boldsymbol{C}}(t)\tilde{\boldsymbol{X}}(t)dt, \quad \text{avec } \tilde{\boldsymbol{C}} = (\mathrm{diag}(\lambda_0(t,\boldsymbol{Z_i})Y_i(t)))_{1\leq i\leq n}.$$

Cette matrice de Gram est aléatoire, même conditionnellement aux covariables. Si nous avions appliqué directement l'Hypothèse 2.1 à la matrice de Gram $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ nous aurions obtenus un $\boldsymbol{\kappa}(s, c_0)$ aléatoire. Pour contourner ce problème, nous avons appliqué une hypothèse aux valeurs propres restreintes à $\mathbb{E}[\tilde{\boldsymbol{G}}_{\boldsymbol{n}}]$.

Hypothèse 2.2. *Pour un entier $s \in \{1, ..., M + N\}$ et une constante $r_0 > 0$, $\mathbb{E}[\tilde{\boldsymbol{G}}_{\boldsymbol{n}}]$ satisfait la condition aux valeurs propres restreintes $\widetilde{\mathbf{RE}}(s, r_0)$ si :*

$$0 < \tilde{\boldsymbol{\kappa}}_{\mathbf{0}}(s, r_0) = \min_{\substack{J\subset\{1,\dots,M+N\},\\ |J|\leq s}} \min_{\substack{\boldsymbol{b}\in\mathbb{R}^{M+N}\setminus\{0\},\\ |\boldsymbol{b}_{J^c}|_1\leq r_0|\boldsymbol{b}_J|_1}} \frac{(\boldsymbol{b}^{\boldsymbol{T}}\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})\boldsymbol{b})^{1/2}}{|\boldsymbol{b}_J|_2}.$$

Cette hypothèse est plus faible que l'Hypothèse **RE** classique 2.1. Nous avons montré que si la condition **RE** est vérifiée pour $\mathbb{E}(\tilde{\boldsymbol{G}}_{\boldsymbol{n}})$, alors la version empirique de la condition **RE** appliquée à $\tilde{\boldsymbol{G}}_{\boldsymbol{n}}$ est satisfaite avec grande probabilité (voir Lemme 2.2).
Cette version modifiée de la condition aux valeurs propres restreintes est nouvelle, tant par sa forme que par son utilisation.

Remarque 2.3. Si on avait plutôt considéré deux matrices de design, une associée à chaque dictionnaire de fonctions au lieu de considérer la matrice par bloc $\tilde{\boldsymbol{X}}(t)$ en supposant deux hypothèses **RE**, une sur chacune des matrices de Gram associées à chaque dictionnaire, on aurait obtenu une inégalité oracle en

$$||\log\alpha_{\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}} - \log\alpha_0||^2_{n,\Lambda} + ||f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}} - f_0||^2_{n,\Lambda}.$$

Or ce n'est pas ce qui nous intéresse puisque pour prédire la durée de survie, on a besoin d'un résultat sur le risque complet $\lambda_{\hat{\boldsymbol{\beta}}_{\boldsymbol{L}},\hat{\boldsymbol{\gamma}}_{\boldsymbol{L}}}$.

Deuxième partie

Estimation of the baseline function in the Cox model with high-dimensional covariates

Introduction

In Part I, we have studied a one step procedure to estimate simultaneously both parameters via a Lasso procedure. In this part, we consider a two-step procedure to estimate both parameters in the Cox model. For the estimation of the regression parameter in a high-dimensional setting, we use the Cox partial log-likelihood combined with a classical Lasso procedure as introduced in Tibshirani (1997). While our first objective is to predict the survival time from the whole hazard rate estimation, the estimation of the baseline function has its own interest. We focus our attention on the estimation of the baseline function.

II.1 Framework

Let us clarify our framework. We consider the general setting of counting processes. For $i = 1, ..., n$, let N_i be a marked counting process and Y_i a random process with values in $[0, 1]$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_t)_{t\geq 0}$ be the filtration defined by

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

where $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ is the $\mathcal{F}_0$-measurable random vector of covariates of individual i. Let Λ_i be the compensator of the process N_i with respect to $(\mathcal{F}_t)_{t\geq 0}$, so that $M_i = N_i - \Lambda_i$ is a $(\mathcal{F}_t)_{t\geq 0}$-martingale. We assume that the process N_i satisfies the Aalen multiplicative intensity model: for all $t \geq 0$,

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i})Y_i(s)\mathrm{d}s, \tag{2.104}$$

where λ_0 is an unknown nonnegative function called intensity.

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).

Remark 2.10 (Censoring case)**.** In the specific case of right censoring, let $(T_i)_{i=1,...,n}$ be independent and identically distributed (i.i.d.) survival times of n individuals and $(C_i)_{i=1,...,n}$ their i.i.d. censoring times. We observe $\{(X_i, \boldsymbol{Z_i}, \delta_i)\}_{i=1,...,n}$, where, for all

$i = 1, ..., n$, $X_i = \min(T_i, C_i)$ is the event time, $\boldsymbol{Z_i}$ is the vector of covariates and $\delta_i = \mathbb{1}_{\{T_i \leq C_i\}}$ is the censoring indicator. For $i = 1, ..., n$, the survival times T_i are supposed to be conditionally independent of the censoring times C_i given $\boldsymbol{Z_i}$. 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.

In this framework, we consider the following Cox model for a vector of covariates $\boldsymbol{Z} = (Z_1, ..., Z_p)^T$.

$$\lambda_0(t, \boldsymbol{Z}) = \alpha_0(t) \exp(\boldsymbol{\beta_0^T Z}), \tag{2.105}$$

where $\boldsymbol{\beta_0} = (\beta_{0_1}, ..., \beta_{0_p})^T$ is the vector of regression coefficients and α_0 is the baseline hazard function. In Part I, we have studied a one-step procedure to estimate simultaneously both parameters via a Lasso procedure. In this part, we consider a two-step procedure to estimate both parameters in the Cox model. We know how to estimate the regression parameter $\boldsymbol{\beta_0}$ from the Cox partial log-likelihood using a classical Lasso procedure. We focus our attention on the estimation of the baseline function. While it is true that we need to estimate both parameters in the Cox model to predict the survival time, in practice, the estimation of the baseline function has its own interest. Let us describe a concrete example. In Loi et al. (2007) and Tian et al. (2012), the considered data relate 414 patients with breast cancer among whom 277 are Tamoxifen treated patients and the 137 others are untreated patients. The observed variables are the time of relapse free survival, that can be right-censored, clinical variables, as the age or the size of the tumor, and 44 928 levels of gene expression. The goal is to predict survival time from relapse adjusted on the covariates. The estimation of the baseline function allows on one hand to make a prognosis for each patient and on the other hand to compare the baseline functions of the patients receiving Tamoxifen and the one of the untreated patients in order to predict the survival time from relapse adjusted on the treatment. In case of proportionality of the hazard functions in the two groups, the treatment variable can be included as a covariate.

In the two following chapters, we propose two methods to obtain an adaptive estimator of the baseline hazard function. For these two strategies, we establish non-asymptotic oracle inequalities for the baseline function, in possibly high-dimensional setting for $\boldsymbol{\beta_0}$.

II.2 Previous results

II.2.1 Preliminary estimation of $\boldsymbol{\beta_0}$: previous results

The estimation of $\boldsymbol{\beta_0}$ has been widely studied for both small and high dimensions of p. The usual procedures to estimate the regression parameter $\boldsymbol{\beta_0}$ are based on the minimization of the Cox partial log-likelihood introduced by Cox (1972) and defined,

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

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\frac{\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\mathrm{d}N_i(t), \quad \text{where } S_n(t,\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}Y_i(t). \tag{2.106}$$

This criterion allows to estimate $\boldsymbol{\beta_0}$ without having to know α_0 (see Appendix A.2 for details). The estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta_0}$ is then defined by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\{-l_n^*(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \tag{2.107}$$

$$\text{with} \quad \mathrm{pen}(\boldsymbol{\beta}) = \begin{cases} 0 & \text{when } p < n \\ \Gamma_n|\boldsymbol{\beta}|_1 & \text{in high-dimension when } p \gg n, \end{cases} \tag{2.108}$$

where Γ_n is a positive regularization parameter and $|\boldsymbol{\beta}|_1 = \sum_{j=1}^p |\beta_j|$. In high-dimension, the procedure with the ℓ_1-penalization is a classical Lasso procedure for the Cox model.

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. In high-dimension, when $p \gg n$, Bradic et al. (2012) have proved asymptotic results for the Lasso estimator $\hat{\boldsymbol{\beta}}$. More recently, Bradic & Song (2012) and Kong & Nan (2012) have established non-asymptotic oracle inequalities and Huang et al. (2013) have obtained non-asymptotic estimation bounds for the Lasso estimator.

II.2.2 Estimation of α_0 : previous results

Let $\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta_0}$. In a second step, one plugs the estimator $\hat{\boldsymbol{\beta}}$ to obtain an estimator of the cumulative baseline function A_0 defined by $A_0(t) = \int_0^t \alpha_0(s)\mathrm{d}s$. This estimator is called the Breslow estimator and is defined, for $t \in [0,\tau]$, by

$$\hat{A}_0(t,\hat{\boldsymbol{\beta}}) = \int_0^t \frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}\bar{N}(u), \tag{2.109}$$

where $\bar{Y} = \sum_{i=1}^n Y_i$ and $\bar{N} = \sum_{i=1}^n N_i$ (see Ramlau-Hansen (1983b) and Andersen et al. (1993)). From $\hat{A}_0(.,\hat{\boldsymbol{\beta}})$, an estimator of the underlying baseline function α_0 can be obtained by kernel function smoothing. The kernel function estimator for α_0 is derived by smoothing the increments of the Breslow estimator. It is defined by

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{h}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\mathrm{d}\hat{A}(u,\hat{\boldsymbol{\beta}}), \tag{2.110}$$

with $K : \mathbb{R} \mapsto \mathbb{R}$ a kernel with integral 1, and h a positive parameter called the bandwidth. This estimator has been introduced and studied by Ramlau-Hansen (1983a;b) within the framework of the multiplicative intensity model for counting processes, thereby extending its use to censored survival data. Ramlau-Hansen (1983b) has proved consistency and asymptotic normality results for this kernel estimator

with fixed bandwidth. In (2.110), both the kernel K and the bandwidth h have to be chosen. The kernel function frequently used is the Epanechnikov kernel function $K(x) = 0.75(1 - x^2)\mathbb{1}_{\{|x|\leq 1\}}$ but other kernel functions as the Gaussian kernel or the uniform kernel exist and can be chosen.

The choice of the bandwidth is more crucial. Ramlau-Hansen (1981) has suggested the cross-validation method to select the bandwidth but without any theoretical guarantees. For randomly censored survival data, Marron & Padgett (1987) have shown that the cross-validation method gives the optimal bandwidth for estimating the density in the sense that the ratio between the integrated squared error obtained by the cross-validation bandwidth and the infimum of the integrated squared error obtained by any bandwidth converges to one almost surely. Grégoire (1993) has considered the cross-validated method suggested by Ramlau-Hansen (1981) for the adaptive estimation of the intensity of a counting process and has proved some consistency and asymptotic normality results for the cross-validated kernel estimator. However, all the results for the adaptive kernel estimator with a cross-validated bandwidth are asymptotic. No non-asymptotic oracle inequalities have to date been stated for the kernel estimator of the baseline function and more generally no theoretical adaptive results have been established for the baseline function. In addition, to our knowledge, the construction of $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ has not yet been considered for high-dimensional covariates.

II.3 Our contributions

In the two following chapters, we propose two strategies to obtain adaptive estimators of the baseline function α_0 in a high-dimensional setting for $\boldsymbol{\beta_0}$: one is based on model selection and the other on adaptive kernel estimation. Both are two-step procedures, with a common preliminary step that consists in the estimation of the regression parameter $\boldsymbol{\beta_0}$ and a second step devoted to the estimation of the baseline hazard function.

In the first preliminary step, the regression parameters $\boldsymbol{\beta_0}$ is estimated with a Lasso procedure applied to the Cox partial log-likelihood (2.106). The Lasso estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta_0}$ is then defined by (2.107) and (2.108). We rely on a non-asymptotic estimation bound from Huang et al. (2013) (see Appendix A for details) to obtain an estimation bound on $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}$ for the Lasso estimator.

Our contributions focus more on the second step, which is specific to both approaches described in the two following chapters. In Chapter 3, we consider a selection model procedure, based on the minimization of a penalized contrast, to estimate the baseline function α_0. We obtain a penalized contrast estimator. In Chapter 4, we consider the usual kernel estimator of the baseline function defined by (2.110), and we consider the Goldenshluger and Lepski procedure (see Goldenshluger & Lepski (2011)) to select

a data-driven bandwidth. We establish non-asymptotic oracle inequalities, warranting the theoretical performances of our two estimators. These oracle inequalities depend on the non-asymptotic control of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}$ for the ℓ_1-norm described just above. Lastly, in Chapter 5, we compare the practical properties of our estimators on simulated data and then, we apply them to the real data from Loi et al. (2007) on the breast cancer described above.

Chapitre 3

Adaptive estimation of the baseline hazard function in the Cox model by model selection

Sommaire

3.1 Introduction 115
3.2 Estimation procedure 116
3.2.1 Assumptions and framework 116
3.2.2 Preliminary estimation of $\boldsymbol{\beta_0}$ 117
3.2.3 Estimation of α_0 119
Definition of the estimation criterion 119
Model selection 120
Assumptions and examples of the models 121
3.3 Non-asymptotic oracle inequalities 123
3.3.1 Preliminary result : oracle inequality for the estimator $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}}$, when β_0 is known 123
3.3.2 Non-asymptotic oracle inequality for the baseline hazard function α_0 124
3.4 Proofs 125
3.4.1 Technical results 125
Results used in the proofs of Theorems 3.1 and 3.2 127
Technical lemmas for the proofs of Proposition 3.4 and 3.5 129
A classical inequality: the Bürkholder Inequality 130
3.4.2 Proofs of the main theorems 130
3.4.3 Proofs of the technical propositions and lemmas 136
Appendices 144
A Prediction result on the Lasso estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta_0}$ for unbounded counting processes 144

Abstract. The aim of this chapter is to propose an adaptive estimator of the baseline function in the Cox model in a high-dimensional setting. Towards this end, we consider a two-step procedure, but different from the usual used procedure: first, we estimate the regression parameter in the Cox model via a Lasso procedure in high-dimension using the partial log-likelihood, secondly, we plug this Lasso estimator into a least-squares type criterion and use model selection methods 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.

Résumé L'objectif de ce chapitre est de proposer un estimateur adaptatif du risque de base dans le modèle de Cox dans un cadre de grande dimension. Pour ce faire, nous considérons une procédure en deux étapes, mais différente de celle habituellement utilisée: d'abord nous estimons le paramètre de régression du modèle de Cox par une procédure Lasso en grande dimension en utilisant la pseudo-log-vraisemblance, ensuite dans un deuxième temps, nous injectons l'estimateur Lasso dans un critère du type moindres carrés et utilisons des méthodes de sélection de modèles pour obtenir un estimateur adaptatif par contraste pénalisé du risque de base. En utilisant des résultats en estimation établis pour l'estimateur Lasso du paramètre de régression, nous établissons une inégalité oracle non-asymptotique pour cet estimateur par contraste pénalisé.

3.1 Introduction

Let us consider the Cox model (2.105) and the framework of counting processes described in the general introduction of Part II. In this chapter, we apply a model selection procedure to estimate the baseline hazard function α_0. This procedure relies on a preliminary first step, which consists in estimating the regression parameter $\boldsymbol{\beta_0}$ using the Lasso procedure described by (2.107) and (2.108).

Then, in the second step, we propose a strategy based on model selection for the estimation of the baseline function in the Cox model. Having its origins in the works of Akaike (1973) and Mallows (1973), the theory of model selection has been formalized by Birgé & Massart (1997) and Barron et al. (1999) for the estimation of densities and regression functions. We can also refer to 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 & 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 (2.104) (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. We propose to estimate the baseline function via a model selection method in high-dimension. Our contribution is, on the one hand, to consider a high-dimensional setting, and on the other hand, to focus on the baseline function of the semi-parametric Cox model. Using this two-step procedure, we provide a penalized contrast estimator, for which we establish a non-asymptotic oracle inequality in a high-dimensional setting for the regression parameter $\boldsymbol{\beta_0}$. This inequality depends on a non-asymptotic control of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}$ in ℓ_1-norm, stated from an estimation inequality of Huang et al. (2013).

The chapter is organized as follows. In Section 3.2, we describe the estimation procedure. Section 3.3 provides non-asymptotic oracle inequalities on the estimator of the baseline hazard function α_0, in a high-dimensional setting for $\boldsymbol{\beta_0}$. Section 3.4 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, in Appendix A, we describe the non-asymptotic estimation inequality of Huang et al. (2013), and explain how we have adapted it to establish a non-asymptotic prediction bound with large probability and how we have extend this result to unbounded counting processes.

3.2 Estimation procedure

After introducing some assumptions on the framework, we briefly recall the first preliminary step of our procedure: the estimation of the regression parameter $\boldsymbol{\beta_0}$ in high-dimension. Then, we focus on the second step of the procedure: the model selection estimation of the baseline function. This estimation procedure involves the minimization of a contrast, which is tuned to our model. In the third part of this section, we explain the choice of the contrast and describe the estimation procedure of the baseline function.

3.2.1 Assumptions and framework

Before defining the estimation procedure, we need to introduce some assumptions on the framework defined in Subsection II.1 in the introduction of Part II. 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 $|b|_1 = \sum_{j=1}^p |b_j|$.

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 $[0,1]$.

Assumption 3.1.

(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 $\boldsymbol{Z}$ admits 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 for all $(t,\boldsymbol{z}) \in [0,\tau] \times A$,*

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

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

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

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

Remark 3.1.

(i) Assumption 3.1.(i) is a classical assumption in the Cox model to obtain non-asymptotic oracle inequalities (see Huang et al. (2013) and Bradic & Song (2012)). In addition, this assumption seems reasonable since in practice the covariates are bounded.

(ii) With the introduction of the p.d.f. $f_{\boldsymbol{Z}}$ in Assumption 3.1.(ii), we can rewrite the deterministic norm as

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

(iii) In the specific case of right censoring, for T the survival time and C the censoring time, we can write

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

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 $[0, \tilde{\tau}]$, where $\tilde{\tau} \leq \sup\{t \geq 0, (1 - F_{T|\boldsymbol{Z}}(t))(1 - G_{C|\boldsymbol{Z}}(t)) > 0\}$. Hence when we can take $\tau = \tilde{\tau}$ and when $f_{\boldsymbol{Z}}$ is bounded from below on A, there exists $f_0 > 0$, such that

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

so that Assumption 3.1.(iii) is verified in this specific case.

Assumption 3.1.(iii) allows to compare the natural norm of the baseline function induced by our contrast $||.||_{det}$ to the standard $\mathbb{L}^2$-norm (see Proposition 3.3).

(iv) Assumptions 3.1.(ii)-(v) are classic and required when we are interested in intensity functions of counting processes in presence of covariates. Such assumptions are considered for example in Comte et al. (2011).

(v) In density estimation by penalized least squares model selection, the infinity norm of the true density is also assumed to be finite and appears in the penalty term (see Massart (2007)).

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

Let us describe the first preliminary step of our procedure on the estimation of the regression parameter $\boldsymbol{\beta_0}$ in high-dimension. We consider a classical Lasso procedure defined in the Introduction of Part II by (2.107) and (2.108) to estimate $\boldsymbol{\beta_0}$.

We now describe assumptions on $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta_0}$, required to ensure the existence of the estimator of the baseline function defined in the second step.

Assumption 3.2.

(i) $\hat{\boldsymbol{\beta}} \in \mathcal{B}(0, R_1)$, *where* $\mathcal{B}(0, R_1)$ *is the ball defined by*

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

On this ball, the procedure (2.107) becomes

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{B}(0,R_1)}{\arg\min} \{-l_n^*(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\beta})\}, \quad \textit{with} \quad \text{pen}(\boldsymbol{\beta}) = \Gamma_n |\boldsymbol{\beta}|_1, \tag{3.1}$$

(ii) $|\boldsymbol{\beta_0}|_1 < R_2 < +\infty$.

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

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq 2R \quad \text{a.s.} \tag{3.2}$$

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

We establish a non-asymptotic estimation bound with large probability for the Lasso estimator. This result is obtained from an estimation inequality stated by Huang et al. (2013) on a set. We have extended their result, by calculating the probability of this set in the case of unbounded counting processes. See Appendix A for further details.

Proposition 3.1. *Let* $k > 0$, $c > 0$ *and* $s := \text{Card}\{j \in \{1, ..., p\} : \beta_{0_j} \neq 0\}$ *be the sparsity index of* $\boldsymbol{\beta_0}$. *Under Assumptions 3.1.(i) and 3.1.(v) and Assumptions 3.2.(i)-(ii), with probability larger than* $1 - cn^{-k}$, *we have*

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}}, \tag{3.3}$$

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

This proposition is required to establish a non-asymptotic oracle inequality for the baseline function. It is proved in Appendix A.

Assumption 3.3. *We assume that* $p = o(e^n)$.

Remark 3.2. Assumption 3.3 ensures that

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

This assumption seems reasonable, since for higher p, we achieve the ultra-high dimension defined by Verzelen (2012), for which the expected rates of convergence in the oracle inequalities are not reached anymore.

In the rest of the chapter, we assume that the conditions of Proposition 3.1 are fulfilled, so that $\hat{\boldsymbol{\beta}}$ satisfies Inequality (3.3) with high probability. This bound will be intensively used for the estimation of α_0.

3.2.3 Estimation of α_0

We now come to our main subject that is the estimation of the baseline function α_0 via a model selection procedure. This procedure is based on the minimization of a contrast. Before describing the model selection procedure, let us introduce our contrast and explain this choice.

Definition of the estimation criterion

We estimate the baseline function α_0 using a modified least-squares criterion adapted to our problem. 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

$$C_n(\alpha, \boldsymbol{\beta}) = -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \alpha(t)\mathrm{d}N_i(t) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha^2(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(t)\mathrm{d}t. \tag{3.4}$$

In survival analysis, other recent papers are based on least-squares model selection, such that Reynaud-Bouret (2006) or Comte et al. (2011). In particular, Comte et al. (2011) have considered a least-squares type adapted to their problem of estimating the whole intensity $\lambda_0(., \boldsymbol{Z})$, and their criterion is obtained by applying the least-sqares to the function $\lambda(., \boldsymbol{Z})$, which can be defined by $\lambda(t, \boldsymbol{Z}) = \alpha(t)\mathrm{e}^{\boldsymbol{\beta^T Z}}$ in the specific case of the Cox model. We consider here another form of criterion since we remove from their contrast a term in $\mathrm{e}^{\boldsymbol{\beta^T Z}}$. This does not modify its properties but allows to derive suitable results.

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

$$\begin{aligned} \langle \alpha_1, \alpha_2 \rangle_{det(\boldsymbol{\beta})} &= \int_0^\tau \alpha_1(t)\alpha_2(t)\mathbb{E}[e^{\boldsymbol{\beta^T Z}}Y(t)]\mathrm{d}t, \\ ||\alpha||^2_{det(\boldsymbol{\beta})} &= \int_0^\tau \alpha^2(t)\mathbb{E}[e^{\boldsymbol{\beta^T Z}}Y(t)]\mathrm{d}t. \end{aligned} \tag{3.5}$$

A similar deterministic norm has already been introduced by Reynaud-Bouret (2006) to take into account the scarcity of the observations at its right-hand end and then by Comte et al. (2011). The main difference between our norms and the one being defined in those works is that their deterministic norms are weighted by $\mathbb{E}[Y(.)]$, when our is weighted by $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta^T Z}}Y(.)]$.

Using the Doob-Meyer decomposition $N_i = M_i + \Lambda_i$ and according to the multiplicative Aalen model (2.104), 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.

Model selection

The idea of model selection is to estimate the baseline hazard function α_0 by minimizing the empirical least-squares criterions $C_n(\alpha, \hat{\boldsymbol{\beta}})$ over a finite-dimensional function space S_m and then by selecting the appropriate space by penalization. We now describe it formally in our context.

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 (3.1). For each $m \in \mathcal{M}_n$, we define the estimator

$$\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}} = \arg\min_{\alpha \in S_m} \{C_n(\alpha, \hat{\boldsymbol{\beta}})\}. \tag{3.6}$$

The projection of α_0 on S_m with respect to the deterministic scalar product satisfies for all $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\alpha_m^{\boldsymbol{\beta}} = \arg\min_{\alpha \in S_m} \mathbb{E}[C_n(\alpha, \boldsymbol{\beta})],$$

and we remark that

$$\alpha_m^{\boldsymbol{\beta_0}} = \arg\min_{\alpha \in S_m} \mathbb{E}[C_n(\alpha, \boldsymbol{\beta}_0)] = \arg\min_{\alpha \in S_m} ||\alpha - \alpha_0||_{det}^2. \tag{3.7}$$

Inequality for one model. To describe our procedure, we start by assuming that $\boldsymbol{\beta_0}$ is known and then we state an upper bound for the risk on each fixed model. To this aim, set $\hat{\alpha}_m^{\boldsymbol{\beta_0}} = \arg\min_{\alpha \in S_m} \{C_n(\alpha, \boldsymbol{\beta_0})\}$ for $m \in \mathcal{M}_n$.

Proposition 3.2. *Let Assumptions 3.1.(i)-(v), Assumption 3.2.(ii) and Assumptions 3.4.(i)-(iii) hold. For any $m \in \mathcal{M}_n$ fixed*

$$\mathbb{E}||\hat{\alpha}_m^{\boldsymbol{\beta_0}} - \alpha_0||_{det}^2 \leq K_1 ||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||_{det}^2 + \kappa(1 + ||\alpha_0||_\infty)\frac{D_m}{n} + \frac{K_2}{n}.$$

for any $n \geq n_0$, where n_0 is a constant coming from Assumption 3.1.(iv), K_1 is a numerical constant and K_2 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, $|\boldsymbol{\beta_0}|_1$, s and κ_b a constant from the Bürkholder Inequality (see Theorem 3.3).

In this inequality, we recognize the bias term defined by $||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||_{det}^2$, the variance term of order D_m/n and a residual term, which is negligeable in relation to the first two previous terms.

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

$$\hat{m}^{\hat{\boldsymbol{\beta}}} = \underset{m\in\mathcal{M}_n}{\arg\min}\{C_n(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}, \hat{\boldsymbol{\beta}}) + \mathrm{pen}(m)\}, \tag{3.8}$$

where $\mathrm{pen} : \mathcal{M}_n \to \mathbb{R}$ will be defined later. From Proposition 3.2, we can justify heuristically this procedure on the same principle as the Mallows' heuristic (see Mallows (1973)), assuming for the sake of simplicity that $\boldsymbol{\beta_0}$ is known. The idea is the following: we are looking for the model S_m, such that $||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||_{det}^2 + CD_m/n$ is minimum, where $C > 0$ is a constant. This is equivalent to look at the minimum of $-||\alpha_m^{\boldsymbol{\beta_0}}||_{det} + CD_m/n$. Since $-||\alpha_m^{\boldsymbol{\beta_0}}||_{det}$ is unknown, we heuristically replace it by $C_n(\hat{\alpha}_m^{\boldsymbol{\beta_0}}, \boldsymbol{\beta_0}) + CD_m/n$.

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

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

$$\boldsymbol{G_m^{\hat{\beta}}} = \Big(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varphi_j(t)\varphi_k(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}Y_i(t)\mathrm{d}t\Big)_{(j,k)\in J_m^2}. \tag{3.9}$$

By definition, the estimator $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}$ is the solution of the equation $\boldsymbol{G_m^{\hat{\beta}}}\boldsymbol{A_m^{\hat{\beta}}} = \boldsymbol{\Gamma_m}$, where

$$\boldsymbol{A_m^{\hat{\beta}}} = (\hat{a}_j^{\hat{\boldsymbol{\beta}}})_{j\in J_m} \quad \text{and} \quad \boldsymbol{\Gamma_m} = \Big(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \varphi_j(t)\mathrm{d}N_i(t)\Big)_{j\in J_m}. \tag{3.10}$$

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

$$\hat{\mathcal{H}}_m^{\hat{\boldsymbol{\beta}}} = \left\{ \min \mathrm{Sp}(\boldsymbol{G_m^{\hat{\beta}}}) \geq \max\left(\frac{\hat{f}_0\mathrm{e}^{-B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{-B|\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}|_1}}{6}, \frac{1}{\sqrt{n}}\right)\right\}, \tag{3.11}$$

where $\mathrm{Sp}(\boldsymbol{M})$ denotes the spectrum of matrix $\boldsymbol{M}$. From Assumptions 3.2.(i)-(ii), on the set $\hat{\mathcal{H}}_m^{\hat{\boldsymbol{\beta}}}$, the matrix $\boldsymbol{G_m^{\hat{\beta}}}$ is invertible and $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}$ is thus uniquely defined as

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

Assumptions and examples of the models

The following assumptions on the models $\{S_m : m \in \mathcal{M}_n\}$ are usual in model selection techniques (see Comte et al. (2011) for example) and rather weak. 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 Example 3.1 for a description of histogram and trigonometric models, and to Barron et al. (1999) and Brunel & Comte (2005) for the other examples.

Assumption 3.4.

(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 3.3. Assumption 3.4.(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 3.4.(ii) implies a useful connection between the standard $\mathbb{L}^2$-norm and the infinite norm. Assumption 3.4.(iii) implies that $\forall m, m' \in \mathcal{M}_n$, $S_m + S_{m'} \subset \mathcal{S}_n$. This condition is useful for the chaining argument used by Comte et al. (2011) to prove Proposition 3.6. In addition, we have from Assumption 3.4.(i) that $\mathcal{D}_n \leq \sqrt{n}/\log n$.

We consider as examples the histogram basis and the trigonometric basis for which Assumptions 3.4.(i)-(iii) are satisfied.

Example 3.1.

1. *Consider the 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 3.4.(ii) is satisfied for $\phi = 1/\tau$.

2. *Consider the trigonometric basis,*

$$\begin{cases} \varphi_1(t) & = \sqrt{2/\tau}, \\ \varphi_{2j}(t) & = \sqrt{2/\tau}\cos(2\pi jt/\tau), \quad \text{for } j \geq 1, \\ \varphi_{2j+1}(t) & = \sqrt{2/\tau}\sin(2\pi jt/\tau), \quad \text{for } j \geq 1. \end{cases}$$

For this model, $D_m = m$ and Assumption 3.4.(ii) is satisfied for $\phi = 2/\tau$.

3.3 Non-asymptotic oracle inequalities

In this subsection, we establish non-asymptotic oracle inequalities for the estimator of the baseline function in the Cox model. As a preliminary result, we first establish, a non-asymptotic oracle inequality for the baseline function in the easier case where the regression parameter $\boldsymbol{\beta_0}$ is assumed to be known. This oracle inequality is mainly stated for a better understanding of the main case. Then, we establish our main result: a non-asymptotic oracle inequality for the estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ of the baseline function of the Cox model.

3.3.1 Preliminary result : oracle inequality for the estimator $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}}$, when β_0 is known

In the specific case of known $\boldsymbol{\beta_0}$, the estimation criterion (3.4) becomes

$$C_n(\alpha, \boldsymbol{\beta_0}) = -\frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \alpha(t)\mathrm{d}N_i(t) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha^2(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t,$$

and then, for each m, we define

$$\hat{\alpha}^{\boldsymbol{\beta_0}}_m = \underset{\alpha\in S_m}{\arg\min}\{C_n(\alpha, \boldsymbol{\beta}_0)\}. \tag{3.12}$$

We recall that

$$\alpha^{\boldsymbol{\beta_0}}_m = \underset{\alpha\in S_m}{\arg\min}\,\mathbb{E}[C_n(\alpha, \boldsymbol{\beta}_0)].$$

The relevant space is automatically selected by using the following penalized criterion

$$\hat{m}^{\boldsymbol{\beta}_0} = \underset{m\in\mathcal{M}_n}{\arg\min}\{C_n(\hat{\alpha}^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}_0) + \mathrm{pen}(m)\}, \tag{3.13}$$

where $\mathrm{pen}(m)$ is defined in Theorem (3.1). The final estimator is $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta}_0}}$. We now state for $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta}_0}}$ the following non-asymptotic oracle inequality.

Theorem 3.1. *Let Assumptions 3.1.(i)-(v), Assumption 3.2.(ii) and Assumptions 3.4.(i)-(iii) hold. Let $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta}_0}}$ be defined by (3.12) and (3.13) with*

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

where K_0 is a numerical constant. We have for any $n \geq n_0$, with n_0 a constant defined in Assumption 3.1.(iv),

$$\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta}_0}} - \alpha_0||^2_{det}] \leq \kappa_0 \inf_{m\in\mathcal{M}_n}\{||\alpha_0 - \alpha^{\boldsymbol{\beta}_0}_m||^2_{det} + 2\,\mathrm{pen}(m)\} + \frac{C}{n}, \tag{3.14}$$

where κ_0 is a numerical constant and C a constant 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, $|\boldsymbol{\beta_0}|_1$, s and κ_b a constant coming from the Bürkholder Inequality (see Theorem 3.3).

This inequality reveals all the usual terms of a classical non-asymptotic oracle inequality. The bias term defined by $||\alpha_0 - \alpha_m^{\boldsymbol{\beta}_0}||_{det}^2$ decreases with the dimension D_m of the model S_m, and even faster than α_0 is regular. The variance term $\text{pen}(m)$ is of order D_m/n, so that it increases with D_m. The rate of convergence is then of the same order than the one obtained in Proposition 3.2 for one model: there is no loss when we select one model among a collection with the model selection procedure (3.13). Lastly the residual term of order $1/n$ is negligible compared with the two previous terms.

3.3.2 Non-asymptotic oracle inequality for the baseline hazard function α_0

Now, let us state the main theorem. The estimator $\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$ satisfies the following non-asymptotique oracle inequality.

Theorem 3.2. *Let Assumptions 3.1.(i)-(v), Assumptions 3.2.(i)-(ii), Assumption 3.3 and Assumptions 3.4.(i)-(iii) hold. Let $\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$ be defined by (3.6) and (3.8) with*

$$\text{pen}(m) := K_0(1 + ||\alpha_0||_{\infty,\tau})\frac{D_m}{n}, \tag{3.15}$$

where K_0 is a numerical constant. Then, for any $n \geq n_0$, with n_0 a constant defined in Assumption 3.1.(iv),

$$\mathbb{E}[||\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}} - \alpha_0||_{det}^2] \leq \kappa_0 \inf_{m\in\mathcal{M}_n}\{||\alpha_0 - \alpha_m^{\boldsymbol{\beta}_0}||_{det}^2 + 2\,\text{pen}(m)\} + \frac{C_1}{n} + C_2(s)\frac{\log(np)}{n}, \tag{3.16}$$

where κ_0 is a numerical constant, C_1 and $C_2(s)$ 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 coming from the Bürkholder Inequality (see Theorem 3.3).

We refer to Subsection 3.4.2 for precisions about C_1 and $C_2(s)$. We recognize the bias term, the variance term of order D_m/n and a residual term of order $1/n$. We obtain an additional term with regard to Theorem 3.1. This term is of order $\log(np)/n$. This term comes from preliminary estimation of the regression parameter $\boldsymbol{\beta_0}$ in high-dimension and from the non-asymptotic control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ given by Proposition 3.1. The bound obtained for this error is of order $\log(np)/n$, which explain the order of our remaining term. We should also have write Inequality (3.16) before using the bound of control (3.3) and get:

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

This inequality makes clearer the role of the first step of the procedure in the control of the estimator $\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$ of the baseline function. In Inequality (3.16), the residual term $\log(np)/n$ gives the order of the loss depending on the dimension of

p; the rate of convergence is slower when we reach the ultra-high dimension with $p \gg n$ (more precisely, from Verzelen (2012), the ultra-high dimension is reached when $s\log(p/s)/n \geq 1/2$, where s is the sparsity index of $\boldsymbol{\beta_0}$). The remaining term $log(np)/n$ is characteristic of the high-dimensional setting. Inequality (3.16) provides the first non-asymptotic oracle inequality for an estimator of the baseline function. This inequality warrants the performances of our estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$.

Corollary 3.1. *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 3.2,*

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

where $\tilde{C}$ and $C_2(s)$ 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 3.3).

Proof. From Proposition 3.3 and the proof of Corollary 1 in Comte et al. (2011), we deduce that

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

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. □

Remark 3.4. From Reynaud-Bouret (2006) for an intensity function without covariates, we know that the minimax is $n^{-\frac{2\gamma}{2\gamma+1}}$. We infer that this would also be the optimal rate in our case when the term $\log(np)/n$ is negligible. However, when the high-dimensional 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 $\boldsymbol{\beta_0}$.

3.4 Proofs

3.4.1 Technical results

In this section, we introduce some propositions and lemmas that are necessary to prove the theorems. Their proofs are postponed to Subsection 3.4.3.

Let us first introduce the random norm revealed from the contrast (3.4) and associated to the deterministic norm defined by (3.5), 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,

$$||\alpha||^2_{rand(\boldsymbol{\beta})} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha^2(t)\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}Y_i(t)\mathrm{d}t, \tag{3.17}$$

$$\langle \alpha_1, \alpha_2\rangle_{rand(\boldsymbol{\beta})} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha_1(t)\alpha_2(t)\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}Y_i(t)\mathrm{d}t,$$

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 $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$C_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}, \boldsymbol{\beta}) + \mathrm{pen}(\hat{m}^{\boldsymbol{\beta}}) \leq C_n(\hat{\alpha}^{\boldsymbol{\beta}}_m, \boldsymbol{\beta}) + \mathrm{pen}(m) \leq C_n(\alpha^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}) + \mathrm{pen}(m), \tag{3.18}$$

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

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

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

$$\nu_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha(t)\mathrm{d}M_i(t). \tag{3.19}$$

It follows that

$$\begin{aligned} C_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}}, \boldsymbol{\beta}) - C_n(\alpha^{\boldsymbol{\beta_0}}_m, \boldsymbol{\beta}) = &||\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand(\boldsymbol{\beta})} - 2\nu_n(\hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m) \\ &+ 2\langle \hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\boldsymbol{\beta})} - 2\langle \hat{\alpha}^{\boldsymbol{\beta}}_{\hat{m}^{\boldsymbol{\beta}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha_0\rangle_{rand}. \end{aligned} \tag{3.20}$$

Let us now introduce the following events :

$$\Delta_1 = \left\{\alpha \in \mathcal{S}_n : \left|\frac{||\alpha||^2_{rand}}{||\alpha||^2_{det}} - 1\right| \leq \frac{1}{2}\right\}, \quad \text{and} \quad \Omega = \left\{\left|\frac{\hat{f}_0}{f_0} - 1\right| \leq \frac{1}{2}\right\} \tag{3.21}$$

$$\Delta_2 = \left\{\alpha \in \mathcal{S}_n : \left|\frac{||\alpha||^2_{rand(\hat{\boldsymbol{\beta}})}}{||\alpha||^2_{rand}} - 1\right| \leq \frac{1}{2}\right\}. \tag{3.22}$$

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 (3.5) and the standard $\mathbb{L}^2$-norm:

Proposition 3.3 (Connections between the norms)**.** *From Assumptions 3.1.(i), (ii) and (iii), we deduce the following connection between the deterministic norm and the standard* $\mathbb{L}^2$*-norm:*

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

The proof of this proposition is immediate.

Results used in the proofs of Theorems 3.1 and 3.2

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

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

The two following lemmas ensure the existence of the estimators $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}}$ and $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ respectively on $\Delta_1 \cap \Omega$ and on $\Delta_1 \cap \Delta_2 \cap \Omega$.

Lemma 3.1. *Under Assumptions 3.1.(i)-(v), Assumption 3.2.(ii) and Assumptions 3.4.(i)-(iii), for* $n \geq 16/(f_0 \mathrm{e}^{-B|\boldsymbol{\beta_0}|_1})^2$*, the following embedding holds:*

$$\Delta_1 \cap \Omega \subset \hat{\mathcal{H}} \cap \Omega, \qquad \textit{where } \hat{\mathcal{H}} := \bigcap_{m \in \mathcal{M}_n} \hat{\mathcal{H}}_m^{\boldsymbol{\beta_0}}.$$

Lemma 3.2. *Under Assumptions 3.1.(i)-(v), Assumptions 3.2.(i)-(ii) and Assumptions 3.4.(i)-(iii), for* $n \geq 16/(f_0 \mathrm{e}^{-3BR})^2$*, the following embedding holds:*

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

From these lemmas, for all $m \in \mathcal{M}_n$, the matrices $\boldsymbol{G_m^{\beta_0}}$ and $\boldsymbol{G_m^{\hat{\beta}}}$ are invertible respectively on $\Delta_1 \cap \Omega$ and $\Delta_1 \cap \Delta_2 \cap \Omega$, and thus the estimators of α_0 are well defined. Proofs of Lemma 3.1 and 3.2 are available in Subsection 3.4.3.

The two following propositions bound the quadratic difference between $\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta}_0}}$ or $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ and $\alpha_m^{\boldsymbol{\beta_0}}$ for $m \in \mathcal{M}_n$, respectively on the complements of $\Delta_1 \cap \Omega$ and 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

$$\Omega_H^k = \left\{ |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s) \sqrt{\frac{\log(pn^k)}{n}} \right\}, \tag{3.23}$$

for a constant $C(s)$ depending on the sparsity index of $\boldsymbol{\beta_0}$. From Proposition 3.1, $\mathbb{P}(\Omega_H^k) \geq 1 - cn^{-k}$ for a constant $c > 0$. Now, let us state the two following propositions.

Proposition 3.4. *Under Assumptions 3.1.(i)-(v), Assumption 3.2.(ii) and Assumptions 3.4.(i)-(iii),*

$$\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{(\Delta_1\cap\Omega)^c}] \leq c_1/n, \tag{3.24}$$

where 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, $|\boldsymbol{\beta_0}|_1$ and κ_b a constant that comes from the Bürkholder Inequality (see Theorem 3.3).

Proposition 3.5. *Under Assumptions 3.1.(i)-(v), Assumptions 3.2.(i)-(ii) and Assumptions 3.4.(i)-(iii),*

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{\aleph_k^c}] \leq \tilde{c}_1/n, \tag{3.25}$$

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 3.3).

The proofs of Propositions 3.4 and 3.5 are similar. We refer to Subsection 3.4.3 for the proof of Proposition 3.5. A detailed proof of Proposition 3.4 is available in Comte et al. (2011). These two propositions are directly used in the proofs of Theorems 3.1 and 3.2 in Subsection 3.4.2.

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 (3.19), 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 3.6. *Let $m, m' \in \mathcal{M}_n$. Define*

$$\mathcal{B}^{det}_{m,m'}(0,1) = \{\alpha \in S_m + S'_m : ||\alpha||_{\text{det}} \leq 1\}. \tag{3.26}$$

Under the assumptions of Theorem 3.1, there exists $\kappa > 0$ such that for

$$p(m,m') = \frac{\kappa}{K_0}(\text{pen}(m) + \text{pen}(m')), \tag{3.27}$$

where the constant K_0 and $\text{pen}(m)$ are defined in (3.15), 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 Theorems 3.1 and 3.2. We admit the proof of this proposition and refer to Comte et al. (2011) for a detailed proof of this result.

We need Proposition 3.7 to prove Theorem 3.2: 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 3.2, when we control the difference between the scalar products $\langle ., . \rangle_{rand} - \langle ., . \rangle_{rand(\hat{\boldsymbol{\beta}})}$ (see Subsection 3.4.2). Proposition 3.7 allows to control this process.

Proposition 3.7. *Let introduce the ball* $\mathcal{B}_n^{det}(0,1) \subset \mathcal{S}_n$ *defined by*

$$\mathcal{B}_n^{det}(0,1) = \{\alpha \in \mathcal{S}_n : ||\alpha||_{det} \leq 1\}. \tag{3.28}$$

Under Assumptions 3.1.(i)-(v) and Assumption 3.2.(ii), 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 3.7 is proved in Subsection 3.4.3.

Technical lemmas for the proofs of Proposition 3.4 and 3.5

In order to prove Propositions 3.4 and 3.5, we need three lemmas:

Lemma 3.3. *Under Assumptions 3.1.(i)-(v), Assumptions 3.2.(i)-(ii) and Assumptions 3.4.(i)-(iii), we have*

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

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 3.3) and on the choice of the basis.*

Lemma 3.4. *Under Assumptions 3.1.(i)-(v) and Assumptions 3.4.(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 3.5. *Under Assumptions 3.1.(i)-(v), Assumptions 3.2.(i)-(ii) 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* $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta}_0^T \boldsymbol{Z}}]$.

The first two lemmas are required to prove Proposition 3.4 and we need to add the third one to prove Proposition 3.5. The three lemmas are proved in Subsection 3.4.3.

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 & Shiryayev (1989) p.75, for the proof of this result.

Theorem 3.3 (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 3.3 and in the oracle inequalities of Theorem 3.1 and 3.2, the constants depend on κ_b.

3.4.2 Proofs of the main theorems

Proofs of Proposition 3.2 and Theorem 3.1

The proof of Proposition 3.2 is very similar to this of Theorem 3.1. So, we only prove Theorem 3.1 and it suffices to consider only one model in the following proof to establish Proposition 3.2.

We have the following decomposition

$$\begin{aligned}\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_0||^2_{det}] \leq 2||\alpha_0 - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} &+ 2\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} \mathbb{1}_{(\Delta_1 \cap \Omega)}] \\ &+ 2\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} \mathbb{1}_{(\Delta_1 \cap \Omega)^c}], \end{aligned} \tag{3.29}$$

where Δ_1 and Ω are defined by (3.21). The first term in (3.29) is usual and is referred to as the bias term. The last term is controlled via Proposition 3.4. We now focus on the second term $\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{det} \mathbb{1}_{(\Delta_1 \cap \Omega)}]$. According to Lemma 3.1 as soon as $n \geq 16/(f_0 \mathrm{e}^{-B|\boldsymbol{\beta_0}|_1})^2$, the estimator $\hat{\alpha}^{\boldsymbol{\beta_0}}_{m^{\boldsymbol{\beta_0}}}$ of α_0 is well defined. From (3.18) and (3.20), with $\boldsymbol{\beta} = \boldsymbol{\beta_0}$, we have for all $m \in \mathcal{M}_n$,

$$||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_m^{\boldsymbol{\beta_0}}||^2_{rand} \leq 2\langle \hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_m^{\boldsymbol{\beta_0}}, \alpha_0 - \alpha_m^{\boldsymbol{\beta_0}} \rangle_{rand} + 2\nu_n(\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha_m^{\boldsymbol{\beta_0}}) + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}^{\boldsymbol{\beta_0}}), \tag{3.30}$$

where $\nu_n(.)$ is the empirical process defined by Equation (3.19). From (3.30), using the classical inequality $2xy \leq bx^2 + y^2/b$ with $b > 0$, we obtain

$$||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} \leq \frac{1}{4}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} + 4||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} + \text{pen}(m) - \text{pen}(\hat{m}^{\boldsymbol{\beta_0}})$$
$$+\frac{1}{4}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + 4 \sup_{\alpha \in \mathcal{B}^{det}_{m,\hat{m}^{\boldsymbol{\beta_0}}}(0,1)} \nu_n^2(\alpha),$$

also rewritten as

$$\frac{3}{4}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} \leq 4||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} + \frac{1}{4}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + \text{pen}(m) - \text{pen}(\hat{m}^{\boldsymbol{\beta_0}})$$
$$+ 4 \sum_{m' \in \mathcal{M}_n} \Big(\sup_{\alpha \in \mathcal{B}^{det}_{m,m'}(0,1)} \nu_n^2(\alpha) - p(m,m') \Big)_+ + 4p(m, \hat{m}^{\boldsymbol{\beta_0}})$$

where $\mathcal{B}^{det}_{m,m'}(0,1)$ is defined by (3.26) and $p(m,m')$ by (3.27). Using the relation between the random norm and the deterministic norm on $\Delta_1 \cap \Omega$, we get

$$\frac{3}{8}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} \leq 4||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} + \text{pen}(m) + \frac{1}{4}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}$$
$$+ 4 \sum_{m' \in \mathcal{M}_n} \Big(\sup_{\alpha \in \mathcal{B}^{det}_{m,m'}(0,1)} \nu_n^2(\alpha) - p(m,m') \Big)_+$$
$$+ 4p(m, \hat{m}^{\boldsymbol{\beta_0}}) - \text{pen}(\hat{m}^{\boldsymbol{\beta_0}}).$$

We fix $K_0 \geq 4\kappa$ such that for all m, m', $4p(m,m') \leq \text{pen}(m) + \text{pen}(m')$ and obtain that on $\Delta_1 \cap \Omega$, we have

$$\frac{1}{8}||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} \leq 4||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{rand} + 2\,\text{pen}(m)$$
$$+ 4 \sum_{m' \in \mathcal{M}_n} \Big(\sup_{\alpha \in \mathcal{B}^{det}_{m,m'}(0,1)} \nu_n^2(\alpha) - p(m,m') \Big)_+ . \tag{3.31}$$

The control of the process

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

is given in expectation by Proposition 3.6.

Taking the expectation in (3.31) and applying Proposition 3.6, we conclude that

$$\frac{1}{8}\mathbb{E}[||\hat{\alpha}^{\boldsymbol{\beta_0}}_{\hat{m}^{\boldsymbol{\beta_0}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} \mathbb{1}_{\{\Delta_1 \cap \Omega\}}] \leq 4||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + 2\,\text{pen}(m) + \frac{C_3}{n}. \tag{3.32}$$

Finally, combining (3.29), (3.32) and Proposition 3.4, we establish Inequality (3.14) in Theorem 3.1. □

Proof of Theorem 3.2

In the following, we consider the sets Δ_1, Δ_2 and Ω defined by (3.21) and (3.22) and the set Ω_H^k defined by (3.23). 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:

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

The first term is the usual bias term. From Proposition 3.5, we deduce that the last term is bounded by $\tilde{c}_1/n$. We now focus on the term $\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{\aleph_k}]$. Once again, for all $m \in \mathcal{M}_n$, the matrices $\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}$ 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 $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ of α_0 is well defined. From (3.18) and (3.20), with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, we have for all $m \in \mathcal{M}_n$,

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

where the empirical process $\nu_n(.)$ is defined by Equation (3.19) and the random norm by (3.17). For $\mathcal{B}^{det}_{m,m'}(0,1)$ defined by (3.26), using the classical inequality $2xy \leq bx^2 + y^2/b$ with $b > 0$, we obtain

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

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

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

also be rewritten for $p(m,m')$ defined by (3.27) for all $m' \in \mathcal{M}_n$, as

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

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

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

that is

$$\frac{3}{32}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det}\mathbb{1}_{\aleph_k}] \leq 16||\alpha_0 - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} + 2\,\text{pen}(m) + A(m) + \mathbb{E}[B(m,\hat{m}^{\hat{\boldsymbol{\beta}}})\mathbb{1}_{\aleph_k}] \tag{3.33}$$

where

$$A(m) = 16\sum_{m'\in\mathcal{M}_n}\mathbb{E}\left(\left(\sup_{\alpha\in\mathcal{B}^{det}_{m,m'}(0,1)}\nu_n^2(\alpha) - p(m,m')\right)_+\mathbb{1}_{\aleph_k}\right), \tag{3.34}$$

$$B(m,\hat{m}^{\hat{\boldsymbol{\beta}}}) = 2\Big(\langle\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand} - \langle\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\hat{\boldsymbol{\beta}})}\Big). \tag{3.35}$$

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

Study of (3.34). According to Proposition 3.6, 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 (3.27) 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

$$A(m) \leq \frac{C_3'}{n}. \tag{3.36}$$

Study of (3.35). Using again the classical inequality $2xy \leq bx^2 + y^2/b$ with $b > 0$, we obtain

$$\begin{aligned}\langle\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand} - \langle\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m, \alpha^{\boldsymbol{\beta_0}}_m\rangle_{rand(\hat{\boldsymbol{\beta}})} &\leq \frac{1}{32}||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}} - \alpha^{\boldsymbol{\beta_0}}_m||^2_{det} \\ + 32\sup_{\alpha\in\mathcal{B}^{det}_{m,\hat{m}^{\hat{\boldsymbol{\beta}}}}(0,1)}\Big(\frac{1}{n}\sum_{i=1}^n\int_0^\tau\alpha(t)\alpha^{\boldsymbol{\beta_0}}_m(t)(\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} - \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}})Y_i(t)\mathrm{d}t\Big)^2.\end{aligned} \tag{3.37}$$

Now, from Assumption 3.4.(iii) and by definition (3.28) of $\mathcal{B}_n^{det}(0,1)$, we write that

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

is less than

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

We have

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

Using the fact that $|\mathrm{e}^x-\mathrm{e}^y|\leq|x-y|\mathrm{e}^{x\vee y}$ for all $(x,y)\in\mathbb{R}^2$ and applying Assumptions 3.1.(i) and Assumptions 3.2.(i)-(ii), we obtain that

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

Now, write

$$\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\left(\frac{1}{n}\sum_{i=1}^n\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}(1-\mathrm{e}^{\boldsymbol{\hat\beta^T Z_i}-\boldsymbol{\beta_0^T Z_i}})Y_i(t)\mathrm{d}t\right)^2$$
$$\leq B^2\mathrm{e}^{4BR}|\boldsymbol{\hat\beta}-\boldsymbol{\beta_0}|_1^2\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\frac{1}{n}\sum_{i=1}^n\left(\int_0^\tau \alpha(t)\alpha_m^{\boldsymbol{\beta_0}}(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right)^2$$
$$\leq B^2\mathrm{e}^{4BR}|\boldsymbol{\hat\beta}-\boldsymbol{\beta_0}|_1^2\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\{\eta_n(\alpha,\alpha_m^{\boldsymbol{\beta_0}})+D_n(\alpha,\alpha_m^{\boldsymbol{\beta_0}})\} \tag{3.38}$$

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 (3.37) and (3.38, we obtain that

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

So, taking the expectation and applying Proposition 3.7 to control

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

we get

$$\begin{aligned}\mathbb{E}[B(m,\hat{m}^{\hat{\boldsymbol{\beta}}})\mathbb{1}_{\aleph_k}]&\leq\frac{1}{16}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}-\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{1}_{\aleph_k}]\\&+64B^2\mathrm{e}^{4BR}\left\{\mathbb{E}^{1/2}[|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|_1^4\mathbb{1}_{\aleph_k}]\mathbb{E}^{1/2}\left[\sup_{\alpha\in\mathcal{B}_n^{det}(0,1)}\{\eta_n^2(\alpha,\alpha_m^{\boldsymbol{\beta_0}})\}\right]+||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\mathbb{E}[|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|_1^2\mathbb{1}_{\aleph_k}]\right\}.\end{aligned}\tag{3.39}$$

Finally, combining (3.33), (3.36) and (3.39) we conclude that

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

On $\Omega\cap\Omega_H^k$, using that, from definition (3.7) and Proposition 3.3, $||\alpha_m^{\boldsymbol{\beta_0}}||_{det}^2\leq 2||\alpha_0||_{det}\leq\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0}^T Z}]\tau||\alpha_0||_{\infty,\tau}$, 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 $C(s,B,R,\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}],||\alpha_0||_{\infty,\tau},\tau)$ and $\tilde{C}(s,B,|\boldsymbol{\beta_0}|_1,R,\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}],\mathbb{E}[\mathrm{e}^{4\boldsymbol{\beta_0^T Z}}],||\alpha_0||_{\infty,\tau},\tau,f_0)$ are constants depending on the

elements in brackets. Combining the previous bounds with Proposition 3.5, we conclude that Theorem 3.2 is proved since

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

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.

□

3.4.3 Proofs of the technical propositions and lemmas

Proofs of Lemma 3.1 and Lemma 3.2

Proofs of Lemma 3.1 and Lemma 3.2 rely on the same principle. We only detail the one that establishes Lemma 3.2 and we refer to Comte et al. (2011) for a similar proof of Lemma 3.1.

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 $\boldsymbol{A}^{\boldsymbol{T}}_{\boldsymbol{m}}\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}\boldsymbol{A_m} = v\boldsymbol{A}^{\boldsymbol{T}}_{\boldsymbol{m}}\boldsymbol{A_m}$. Now, take $h := \sum_j a_j\varphi_j \in S_m$. We have $||h||^2_{rand(\hat{\boldsymbol{\beta}})} = \boldsymbol{A}^{\boldsymbol{T}}_{\boldsymbol{m}}\boldsymbol{G}^{\hat{\boldsymbol{\beta}}}_{\boldsymbol{m}}\boldsymbol{A}^{\boldsymbol{T}}_{\boldsymbol{m}}$ and $||h||^2_2 = \boldsymbol{A}^{\boldsymbol{T}}_{\boldsymbol{m}}\boldsymbol{A_m}$. Thus, on $\Delta_1 \cap \Delta_2$ defined in (3.21) and (3.22) and from Proposition 3.3:

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

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 $n \geq 16/(f_0\mathrm{e}^{-3BR})^2$. □

Proof of Propositions 3.4 and 3.5

We only prove here Proposition 3.5 because Proposition 3.4 is proved in Comte et al. (2011).

We have the following decomposition :

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

We deduce that

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

From definition (3.7) of $\alpha_m^{\boldsymbol{\beta_0}}$ and Proposition 3.3, we have $||\alpha_m^{\boldsymbol{\beta_0}} - \alpha_0||_{det}^2 \leq ||\alpha_0||_{det}^2 \leq \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]||\alpha_0||_2^2$. From this relation and using Cauchy-Schwarz Inequality, we have

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

From Assumption 3.1.(iv), Proposition 3.1, Lemmas 3.3, 3.4 and 3.5 with $k = 6$, we conclude that

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

which ends the proof of Proposition 3.5. □

Proof of Proposition 3.7

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 3.4.(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

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

Applying the Cauchy-Schwarz Inequality, we get

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

From Proposition 3.3, we have

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

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

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

Proof of Lemma 3.3

From Assumption 3.2, we recall that $|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|_1 \leq 2R$. On $\hat{\mathcal{H}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$, we have

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

So we have

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

where $\mathcal{K}_n$ is a set of indices of the global nesting space $\mathcal{S}_n$, defined in Assumption 3.4.(iii), and $\dim \mathcal{S}_n = \mathcal{D}_n = |\mathcal{K}_n|$. Thus, we deduce that

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

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 & 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 3.4.(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

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

Eventually, under Assumption 3.4.(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. □

Proof of Lemma 3.4

The event Δ_1 defined by (3.21) can be rewritten as

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

and consider

$$\vartheta_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\Big(\alpha(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t) - \mathbb{E}[\alpha(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)]\Big)\mathrm{d}t = ||\sqrt{\alpha}||^2_{rand} - ||\sqrt{\alpha}||^2_{det}. \tag{3.40}$$

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

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

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

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

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

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

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 (3.28). 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 3.3, 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 $|U_i^{(j,k)}| \leq \mathrm{e}^{B|\boldsymbol{\beta}_0|_1}$ 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

$$\mathbb{P}\Big(|\vartheta_n(\varphi_j\varphi_k)| \geq \mathrm{e}^{B|\boldsymbol{\beta}_0|_1} x + \sqrt{2\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} x}\Big) \leq 2\mathrm{e}^{-nx}. \tag{3.41}$$

Let introduce

$$\Theta := \{\forall j,k, |\vartheta_n(\varphi_j\varphi_k)| \leq \mathrm{e}^{B|\boldsymbol{\beta_0}|_1} x + \mathrm{e}^{B|\boldsymbol{\beta_0}|_1}\sqrt{2x}\} \quad \text{and} \quad x := \frac{f_0^2 \mathrm{e}^{-2B|\boldsymbol{\beta}_0|_1}}{16\mathcal{D}_n^2 \mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}}.$$

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

$$\begin{aligned}
&\frac{1}{f_0 \mathrm{e}^{-B|\boldsymbol{\beta}_0|_1}} \sup_{(a_j),\sum_{j\in\mathcal{K}_n} a_j^2\leq 1} \sum_{j,k} |a_j a_k| (\mathrm{e}^{B|\boldsymbol{\beta}_0|_1} x + \mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\sqrt{2x}) \\
\leq &\frac{1}{f_0 \mathrm{e}^{-B|\boldsymbol{\beta}_0|_1}} \sup_{(a_j),\sum_{j\in\mathcal{K}_n} a_j^2\leq 1} \Big(\sum_j |a_j|\Big)^2 (\mathrm{e}^{B|\boldsymbol{\beta}_0|_1} x + \mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\sqrt{2x}),
\end{aligned}$$

which is less than

$$\begin{aligned}
&\leq \frac{1}{f_0 \mathrm{e}^{-B|\boldsymbol{\beta}_0|_1}} D_m \left(\frac{\mathrm{e}^{B|\boldsymbol{\beta}_0|_1} f_0^2 \mathrm{e}^{-2B|\boldsymbol{\beta}_0|_1}}{16 \mathcal{D}_n^2 \mathrm{e}^{2B|\boldsymbol{\beta}_0|_1}} + \frac{\mathrm{e}^{B|\boldsymbol{\beta}_0|_1} \sqrt{2} f_0 \mathrm{e}^{-B|\boldsymbol{\beta}_0|_1}}{4 \mathcal{D}_n \mathrm{e}^{B|\boldsymbol{\beta}_0|_1}} \right) \\
&\leq \frac{1}{2} \left(\frac{1}{8} \frac{f_0}{\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} \mathcal{D}_n} + \frac{1}{\sqrt{2}} \right) \\
&\leq \frac{1}{2} \left(\frac{1}{4} + \frac{1}{\sqrt{2}} \right) \\
&\leq \frac{1}{2}. \qquad (3.42)
\end{aligned}$$

From Inequality (3.42), we deduce that $\mathbb{P}((\Delta_1)^c) \leq \mathbb{P}(\Theta^c)$. So using Inequality (3.41), we can conclude that

$$\begin{aligned}
\mathbb{P}((\Delta_1)^c) &\leq \sum_{j,k} \mathbb{P}\Big(|\vartheta_n(\varphi_j \varphi_k)| > \mathrm{e}^{B|\boldsymbol{\beta}_0|_1} x + \mathrm{e}^{B|\boldsymbol{\beta}_0|_1} \sqrt{2x} \Big) \\
&\leq 2\mathcal{D}_n^2 \exp\left(-\frac{n f_0^2 \mathrm{e}^{-2B|\boldsymbol{\beta}_0|_1}}{16 \mathcal{D}_n^2 \mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}} \right) \\
&\leq 2n \exp\left(-\frac{f_0^2}{16 \mathrm{e}^{4B|\boldsymbol{\beta}_0|_1}} \frac{n}{\mathcal{D}_n^2} \right) \\
&\leq 2n \exp\left(-\frac{f_0^2}{16 \mathrm{e}^{4B|\boldsymbol{\beta}_0|_1}} \log n \right) \\
&\leq \frac{C_k^{\Delta_1}}{n^k}, \quad \forall k \geq 1,
\end{aligned}$$

as $\mathcal{D}_n \leq \sqrt{n}/\log n$ from Assumption 3.4.(iii), which ends the proof of Lemma 3.4 with $C_k^{\Delta_1}$ a constant depending on ρ_1, f_0, B and $|\boldsymbol{\beta}_0|_1$. □

Proof of Lemma 3.5

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

$$\tilde{\vartheta}_n(\alpha) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau (\alpha(t) \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{Z_i}} Y_i(t) - \alpha(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} Y_i(t)) \mathrm{d}t = ||\sqrt{\alpha}||^2_{rand(\hat{\boldsymbol{\beta}})} - ||\sqrt{\alpha}||^2_{rand}.$$

Following the same approach as in the proof of Lemma 3.4, we have

$$\mathbb{P}((\Delta_2^{\rho_2})^c) \leq \mathbb{P}\left(\sup_{\alpha \in \mathcal{B}_n^{det}(0,1)} |\tilde{\vartheta}_n(\alpha^2)| > 1 - \frac{1}{\rho_2} \right), \qquad (3.43)$$

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

$$|\tilde{\vartheta}_n(\alpha^2)| \leq B\mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1||\alpha||_2^2 \leq |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \frac{B\mathrm{e}^{B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}}{f_0\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}}||\alpha||_{det}^2.$$

So we get

$$\sup_{\alpha \in \mathcal{B}_{\mathcal{S}_n}^{det}(0,1)} |\tilde{\vartheta}_n(\alpha^2)| \leq |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \frac{B\mathrm{e}^{2B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}}{f_0}.$$

From Proposition 3.1, we have with probability larger than $1 - cn^{-k}$

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

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

$$\sup_{\alpha \in \mathcal{B}_{\mathcal{S}_n}^{det}(0,1)} |\tilde{\vartheta}_n(\alpha^2)| \leq C(s)\sqrt{\frac{\log(pn^k)}{n}}\frac{B\mathrm{e}^{2B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}}{f_0}.$$

Thus, by taking $1 - 1/\rho_2 = C(s)\sqrt{\frac{\log(pn^k)}{n}}\frac{B\mathrm{e}^{2B|\boldsymbol{\beta}_0|_1}\mathrm{e}^{2BR}}{f_0}$ in (3.43), 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 (3.22) verifies $\mathbb{P}((\Delta_2)^c) \leq \mathbb{P}((\Delta_2^{\rho_2})^c) \leq C_k^{(\Delta_2)}n^{-k}$, with $C_k^{(\Delta_2)} = c > 0$. We have shown in Appendix A that the constant c depends on τ, $||\alpha_0||_{\infty,\tau}$ and $\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}]$. □

Appendix

A Prediction result on the Lasso estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta_0}$ 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 (2.107) and (2.108) 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 (2.106) 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 $\mathcal{O} = \{j : \beta_{0_j} \neq 0\}$, $\mathcal{O}^c = \{j : \beta_{0_j} = 0\}$ 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 & 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 3.4 (Huang et al. (2013))**.** *Let $k > 0$ and $\nu = B(\xi + 1)s\Gamma_{n,k}/\{2\kappa^2(\xi, \mathcal{O})\}$. Suppose Assumption 3.1.(i) holds and $\nu \leq 1/\mathrm{e}$. Then, on the event*

$$\widetilde{\Omega}^k_H = \left\{ |\dot{\boldsymbol{l}}^*_{\boldsymbol{n}}(\boldsymbol{\beta_0})|_\infty \leq \frac{\xi - 1}{\xi + 1}\Gamma_{n,k} \right\}, \quad \textit{with} \quad \Gamma_{n,k} = C_0 B \frac{\xi + 1}{\xi - 1}\sqrt{2\frac{\log(pn^k)}{n}}, \tag{3.44}$$

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 3.4. 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 3.6. *Let consider, for $k > 0$, the event $\widetilde{\Omega}_H^k$ defined by (3.44). Then, under Assumptions (i) and 3.1.(v), 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 3.4 as:

Corollary 3.2. *Let $\nu = B(\xi+1)s\Gamma_{n,k}/\{2\kappa^2(\xi,\mathcal{O})\}$, $k > 0$ and $c > 0$. Suppose Assumptions (i) and 3.1.(v) 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 3.2 and Assumption (i), we deduce a prediction inequality given by the following proposition.

Proposition 3.8. *Let $k > 0$ and $c > 0$. Under Assumptions (i) and 3.1.(v), with probability larger than $1 - cn^{-k}$, we have*

$$|\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z} - \boldsymbol{\beta_0^T Z}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}}, \tag{3.45}$$

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

Remark 3.5. From Proposition 3.8 and Definition (3.23) of Ω_H^k, we deduce that $\widetilde{\Omega}_H^k \subset \Omega_H^k$.

Proof of Lemma 3.6. To prove Lemma 3.6, we start from Lemma 3.3. p.10 in the paper of Huang et al. (2013), that we enounce below.

Lemma 3.7 (Lemma 3.3 from Huang et al. (2013))**.** *Suppose that Assumption 3.1.(i) is verified. Let $\dot{\boldsymbol{l}}_{\boldsymbol{n}}^*(\boldsymbol{\beta})$ be the gradient of the $l_n^*(\boldsymbol{\beta})$ defined by (2.106). Then, for all $C_0 > 0$,*

$$\mathbb{P}\left(|\dot{\boldsymbol{l}}_{\boldsymbol{n}}^*(\boldsymbol{\beta_0})|_\infty > C_0 Bx, \sum_{i=1}^n \int_0^\tau Y_i(t)\mathrm{d}N_i(t) \leq C_0^2 n\right) \leq 2p\mathrm{e}^{-nx^2/2}. \tag{3.46}$$

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

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

Lemma 3.8 (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 3.7, to prove Lemma 3.6, 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 3.8 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

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

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 3.6.

□

Chapitre 4

Adaptive kernel estimation of the baseline function in the Cox model

Sommaire

4.1 Introduction 149
4.2 Estimation procedure 150
4.2.1 Preliminary estimation of $\boldsymbol{\beta_0}$ 150
4.2.2 Estimation of α_0 151
Kernel estimator 151
Functional and kernel assumptions 152
Collection of estimators 153
Adaptive selection of the bandwidth 154
4.3 Non-asymptotic bounds for kernel estimators 155
4.3.1 Bound for the kernel estimator of α_0 with a fixed bandwidth 155
4.3.2 Oracle inequality for the adaptive kernel estimator of α_0 156
4.4 Proofs 158
4.4.1 Intermediate Results to prove Theorem 4.1 from a pseudo-estimator . 158
4.4.2 Proof of the oracle inequality in Theorem 4.2 160
Appendices 172
A Technical lemma 172
B Unbounded case 173

Abstract. The aim of this chapter is to propose an estimation of the baseline function in the Cox model, based on a kernel estimator, in a high-dimensional setting. Towards this end, we proceed in two steps. We first estimate the regression parameter in the Cox model via a Lasso procedure in high-dimension. Then, we plug this estimator into the classical kernel estimator of the baseline function, obtained by smoothing the so-called Breslow estimator of the cumulative baseline function. The problem with kernel function estimators is the choice of the bandwidth. To resolve this problem, we begin to study a kernel function estimator with a fixed bandwidth, for which we provide a bound for the Mean Integrated Squared Errors (MISE); then, we describe and study an adaptive procedure of bandwidth selection, in the spirit of Goldenshluger & Lepski (2011), and we state a non-asymptotic oracle inequality for the final estimator.

Résumé L'objectif de ce chapitre est de proposer une estimation adaptative du risque de base dans le modèle de Cox, basée sur un estimateur à noyau, dans un cadre de grande dimension. Pour ce faire, nous procédons en deux étapes. Nous estimons d'abord le paramètre de régression dans le modèle de Cox à l'aide d'une procédure Lasso en grande dimension. Puis, nous injectons cet estimateur dans l'estimateur à noyau usuel du risque de base, obtenu en lissant l'estimateur de Breslow du risque de base cumulé. Le problème des estimateurs à noyau est le choix de la fenêtre. Pour résoudre ce problème, nous commençons par étudier l'estimateur à noyau avec une fenêtre fixée, pour lequel nous obtenons une borne pour l'erreur quadratique moyenne intégrée (MISE); ensuite, nous décrivons et étudions une procédure adaptative pour sélectionner la fenêtre, dans l'esprit de Goldenshluger & Lepski (2011), et nous établissons une inégalité oracle non-asymptotique pour l'estimateur final.

4.1 Introduction

In this chapter, we consider the Cox model (2.105) with high-dimensional covariates and we aim at estimating the function of time $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}$, via kernel estimation. More precisely, in the framework described in the introduction of Part II, the idea is to consider the usual kernel estimator (2.110) of the baseline hazard function, in which we plug the Lasso estimator (2.107) of the regression parameter $\boldsymbol{\beta_0}$, and then use a data-driven procedure to select a data-driven bandwidth.

The choice of the bandwidth in kernel estimation is crucial, in particular when one is interested in establishing non-asymptotic adaptive inequalities. State-of-the-art methods are based on cross-validation, however, as recalled in the introduction of Part II, no non-asymptotic result has been yet established for the kernel estimator with a bandwidth selected by cross-validation. More recently, Goldenshluger & Lepski (2011) have proposed a procedure that addresses the problem of bandwidth selection in kernel density estimation and provides an adaptive estimator, which satisfies non-asymptotic minimax bounds. This method was then considered by Doumic et al. (2012) to estimate the division rate of a size-structured population in a non-parametric setting, by Bouaziz et al. (2013) to estimate the intensity function of a recurrent event process and by Chagny (2014) for the estimation of a real function via a warped kernel strategy.

This method has never been applied to date to the kernel estimator (2.110) and we consider it to obtain an adaptive kernel estimator of the baseline function with a data-driven bandwidth. We establish the first adaptive and non-asymptotic oracle inequality, which warrants the theoretical performances of this kernel estimator. The oracle inequality depends on the non-asymptotic control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ deduced from an estimation inequality stated by Huang et al. (2013) and extended to the case of unbounded counting processes (see Appendix A in Chapter 3 for details).

The chapter is organized as follows. In Section 4.2, we describe the two-step procedure to estimate the baseline function: first, we describe the estimation of $\boldsymbol{\beta_0}$ as a preliminary step and give the bound for $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$, then, we focus on the kernel estimation of α_0 and describe the adaptive estimation procedure of Goldenshluger and Lepski to select a data-driven bandwidth. In Section 4.3, we establish non-asymptotic bounds, first for the kernel estimator with a fixed bandwidth and then, for the adaptive kernel estimator. Proofs are gathered in Section 4.4. Lastly, Appendix A provides some technical results needed in the proofs.

Notations

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

$$||f||_q = \Big(\int |f(x)|^q \mathrm{d}x \Big)^{1/q} \text{ and } ||f||_\infty = \sup_{x \in \mathbb{R}} |f(x)|.$$

The integrals and the supremum are restricted to the support of f and for τ a positive real number, we set $||f||_{\infty,\tau} = \sup_{x\in[0,\tau]} |f(x)|$. For h a positive real number, we define $f_h(.) = f(./h)/h$. For square-integrable functions f and g from $\mathbb{R}$ to $\mathbb{R}$, we denote the convolution product of f and g by $f * g$.

For a real $0 < a < \tau/2$, we also introduce the norm $||.||_{2,a}$ defined for a function $\alpha : [0,\tau] \longmapsto \mathbb{R}$ by

$$||\alpha||_{2,a} = \Big(\int_a^{\tau-a} \alpha^2(t)\mathrm{d}t \Big)^{1/2},$$

where τ is defined in the introduction of Part II as the time at the end the study. We denote by $\langle ., . \rangle_{2,a}$ its associated scalar product. For a vector $\boldsymbol{b} \in \mathbb{R}^p$ and a real $q \geq 1$, we denote $|\boldsymbol{b}|_q = (\sum_{j=1}^p |b_j|^q)^{1/q}$.

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

Finally, 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 $\boldsymbol{Z}$.

4.2 Estimation procedure

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

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

As recalled in the introduction, the estimation of the baseline function relies on the preliminary estimation of the regression parameter $\boldsymbol{\beta_0}$ in high-dimension. This preliminary first step is the same as the one of Chapter 3. For the sake of completeness, we recall here briefly this first step.

As in the previous chapter, the regression parameter $\boldsymbol{\beta_0}$ is estimated via a classical Lasso procedure described by (2.107) and (2.108). We also need the two following assumptions on $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta_0}$:

Assumption 4.1.

(i) *$\hat{\boldsymbol{\beta}} \in \mathcal{B}(0,R)$, where $\mathcal{B}(0,R)$ is the ball defined by*

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

On this ball, the procedure (2.107) becomes

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{B}(0,R)}{\arg\min}\{-l_n^*(\boldsymbol{\beta}) + \mathrm{pen}(\boldsymbol{\beta})\}, \quad \textit{with } \mathrm{pen}(\boldsymbol{\beta}) = \Gamma_n|\boldsymbol{\beta}|_1, \tag{4.1}$$

(ii) $|\boldsymbol{\beta_0}|_1 < +\infty$.

These assumptions are reasonable since in practice we always assume that $\boldsymbol{\beta_0}$ is bounded and $\hat{\boldsymbol{\beta}}$ is in a neighborhood of the true parameter. Such conditions have already been considered by van de Geer (2008) or Kong & Nan (2012). Roughly speaking, it means that we can restrict our attention to a ball in a neighborhood of $\boldsymbol{\beta_0}$ for finding a good estimator of $\boldsymbol{\beta_0}$. Assumptions 4.1.(i) and (ii) are required to control the kernel estimator of the baseline function $\boldsymbol{\beta_0}$. Concerning the covariates, we introduce the following assumption:

Assumption 4.2. *There exists a positive constant B such that for all $j \in \{1, ..., p\}$,*

$$|Z_j| \leq B.$$

Assumption 4.2 is a classical assumption in the Cox model to obtain non-asymptotic oracle inequalities (see Huang et al. (2013) and Bradic & Song (2012)). In addition, this assumption seems reasonable since in practice the covariates are bounded.

From Theorem 3.1 in Huang et al. (2013), we deduce an estimation inequality given by the following proposition. We refer to Chapter 3, Appendix A for a proof of Proposition 4.1.

Proposition 4.1. *Let $k > 0$, $c > 0$ and $s := \text{Card}\{j \in \{1, ..., p\} : \beta_{0_j} \neq 0\}$ be the sparsity index of $\boldsymbol{\beta_0}$. Assume that $||\alpha_0||_{\infty,\tau} < \infty$. Then, under Assumptions 4.1.(i)-(ii) and Assumption 4.2, with probability larger than $1 - cn^{-k}$, we have*

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}} \tag{4.2}$$

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

In the rest of the chapter, the conditions of Proposition 4.1 will be fulfilled, so that $\hat{\boldsymbol{\beta}}$ satisfies Inequality (4.2).

4.2.2 Estimation of α_0

Now, let us focus on the adaptive kernel estimation of the baseline hazard function α_0. In this subsection, we define the kernel estimator on which our procedure relies, we state some functional and kernel assumptions, and we describe the Goldenshluger and Lepski procedure to select a data-driven bandwidth.

Kernel estimator

The kernel function method has been first developed to estimate a probability density function. Ramlau-Hansen (1983b) has generalized this method to estimate counting process intensities using kernel functions to smooth the increments of the non-parametric Breslow estimator (2.109) for the cumulative intensity and provide with

this method an estimator of the baseline function. We describe this kernel function estimator in the following.

Let define $K : \mathbb{R} \to \mathbb{R}$ a kernel, namely K is a function such that $\int_{\mathbb{R}} K(x)\mathrm{d}x = 1$. From (2.110), the usual kernel function estimator introduced by Ramlau-Hansen (1983b) to estimate α_0 is then defined by

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}N_i(u), \tag{4.3}$$

with

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

The parameter $h > 0$ is called the bandwidth. In kernel function estimation, the bandwidth has to be chosen by the user. The cross-validation method is a classical data-driven procedure to select the bandwidths in the kernel function methods. Grégoire (1993) has defined a procedure for selecting the bandwidth for the smooth estimate of intensity in the Aalen counting process, based on the cross-validation. To our knowledge, all theoretical results for the kernel function estimator (4.3) with a bandwidth selected by cross-validation are asymptotic and no non-asymptotic results has yet been established for this estimator. In practice, in medical surveys, there is only a few patients that are observed and the asymptotic framework is far from being reached. This explains our interest in providing a data-driven method to select automatically the bandwidth and obtain a kernel function estimator, for which we can warrant some non-asymptotic properties.

In what follows, we denote by $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ the kernel estimator (4.3), in which the Lasso estimator (4.1) has been plugged.

Functional and kernel assumptions

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

Assumption 4.3.

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

(ii) *For $S(t, \boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} Y_i(t)]$, there exists a positive constant c_S such that,*

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

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

(iv) *There exists $c_\tau > 0$, such that $N_i(t) \leq c_\tau$ almost surely for every $t \in [0, \tau]$ and $i \in \{1, ..., n\}$.*

Assumption 4.3.(i) is satisfied for all the examples quoted in the general introduction of Part II. In fact, this assumption is needed to ensure that the random process Y_i has a lower bound when it is nonzero. We could also have considered a modified estimator of $S_n(u, \boldsymbol{\beta})$, defined by (2.106), as it is usually done in the censoring case without covariates. Assumption 4.3.(ii) is common in the context of estimation with censored observations (see Andersen et al. (1993)). Assumption 4.3.(iii) is required to obtain Theorems 4.1 and 4.2 below. Nevertheless, the value $||\alpha_0||_{\infty,\tau}$ is not needed to compute the estimator (see Chapter 5, Section 5.3). Assumption 4.3.(iv) can be found e.g. in Dauxois & Sencey (2009) or in Bouaziz et al. (2013).

Remark 4.1. We have also studied the unbounded case, when Assumption 4.3.(iv) is not assumed. We work in this chapter in the bounded case because the proofs are easier and the constants more readable, but we refer to Appendix B for details on the proofs of some results when the counting process N_i is unbounded.

The following assumptions are fulfilled by many standard kernel functions and are standard in kernel function estimation.

Assumption 4.4.

(i) *K has a compact support* $[-1, 1]$,
$$\int_{-1}^{1} K(u)\mathrm{d}u = 1 \text{ and } ||K||_2^2 = \int_{-1}^{1} K^2(u)\mathrm{d}u < \infty.$$

(ii) $||K||_\infty = \sup_{u\in[-1,1]} |K(u)| < \infty$.

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

For practical purposes, Assumption 4.4.(i) is not a real restriction. This assumption can also be found in Ramlau-Hansen (1983b). Assumptions 4.4.(ii) and 4.4.(iii) are rather standard in kernel density estimation (see Goldenshluger & Lepski (2011)) and has also been considered in the kernel intensity estimation by Bouaziz et al. (2013).

As the kernel has bounded support on $[-1, 1]$, the integrands in the Breslow estimator (4.3) and in the pseudo-estimator (4.12) vanishes outside $[t-h, t+h]$. Therefore, given a bandwidth h, we only discuss estimation of α_0 for t such that $t \pm h \in [0, \tau]$ (see Andersen et al. (1993) p.230 and Bouaziz et al. (2013)).

Collection of estimators

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

Assumption 4.5.

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

(ii) *For some* $a \geq 0$, $\sum_{h\in\mathcal{H}_n} \dfrac{1}{nh} \lesssim \log^a(n)$.

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

Assumptions 4.5.(i)-(iii) mean that the bandwidth collection should not be too large. Let us give an example of grid $\mathcal{H}_n$ that satisfies the three previous assumptions.

Example 4.1 (Example of $\mathcal{H}_n$)**.** *The following grid is considered in the simulations in Chapter 5*

$$\mathcal{H}_n = \Big\{ h_j = \frac{\varepsilon}{2^{j-1}}, j = 1, ..., \lfloor \log(n)/\log(2) \rfloor \Big\},$$

where $\varepsilon \in [0, \tau/2]$ is a small constant chosen arbitrarily as close as possible to 0. For this grid, all the assumptions required on the bandwidths are verified. Indeed, $\mathrm{Card}(\mathcal{H}_n) \leq \log(n)/\log(2) \leq n$ *and* $\forall k = 1, ..., \lfloor \log(n)/\log(2) \rfloor$*, we have $h_j \in [n^{-1}, 1]$. Moreover, Assumption 4.5.(ii) holds true since*

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

Lastly,

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

and Assumption 4.5.(iii) is verified.

On the grid $\mathcal{H}_n$, we obtain from the definition (4.3), a set of kernel estimators $\mathcal{F}(\mathcal{H}_n) = \{\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}, h \in \mathcal{H}_n\}$ of the baseline function α_0.

Adaptive selection of the bandwidth

We wish to automatically select a relevant bandwidth $h \in \mathcal{H}_n$, in such a way to then be able to select a kernel estimator among the set $\mathcal{F}(\mathcal{H}_n)$. As usual, we must choose a bandwidth h which realizes the best compromise between the squared-bias and the variance terms. The choice should be data-driven. For this, we use a quite recent method introduced by Goldenshluger & Lepski (2011) for the problem of density estimation. The "Goldenshluger and Lepski method" has only been considered in two different settings: Bouaziz et al. (2013) has applied this method to provide an adaptive kernel function estimator of the intensity function of a recurrent event process and Chagny (2014) has used it to estimate a real valued function from a sample of random couples (see Chagny (2014)). Lastly, Chagny (2013) has also proposed a "mixed strategy", which consists in applying the "Goldenshluger and Lepski method" to select the relevant model in model selection methods for real valued function in regression models. We consider this method to obtain an adaptive kernel function estimator of the baseline function, for which we establish a non-asymptotic oracle inequality.

Let us begin to describe the method. We can explain the idea of the method of Goldenshluger & Lepski (2011) from an heuristic proposed by Chagny (2013). We want to define $\hat{\alpha}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$ so that the risk is as close as possible as

$$\min_{h \in \mathcal{H}_n} \{ ||\alpha_0 - K_h * \alpha_0||_{2,\varepsilon}^2 + V(h) \}, \text{ with } V(h) = \kappa \frac{c_\tau \tau ||\alpha_0||_{\infty,\tau}}{c_S} \frac{||K||_2^2}{nh}$$

for a constant $\kappa > 0$ and $\varepsilon = \max\{h : h \in \mathcal{H}_n\}$. In order to get closer from the bias term $||\alpha_0 - K_h * \alpha_0||^2_{2,\varepsilon}$, we replace α_0 with an estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}$, with a fixed bandwidth h', so that we obtain $||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'} - K_h * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}||^2_{2,\varepsilon}$. However, unlike the bias term, this quantity is random and so contains some variability. We need to correct this variability by retiring the part of the variance $V(h')$. Lastly, since there are no reason to choose one bandwidth $h' \in \mathcal{H}_n$ rather than an other one, we consider the entire collection and take the maximum over this collection.

Formally, we define for $h \in \mathcal{H}_n$ and $\varepsilon = \max\{h : h \in \mathcal{H}_n\}$,

$$A^{\hat{\boldsymbol{\beta}}}(h) = \sup_{h' \in \mathcal{H}_n} \left\{ ||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}||^2_{2,\varepsilon} - V(h') \right\}_+ \tag{4.4}$$

where

$$V(h) = \kappa \frac{c_\tau \tau ||\alpha_0||_{\infty,\tau}}{c_S} \frac{||K||_2^2}{nh}, \tag{4.5}$$

for some numerical constant $\kappa > 0$ to be specified later (see Chapter 5, Appendix B.2), and

$$\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(t) = K_{h'} * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h(t), \tag{4.6}$$

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

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

$$\hat{h}^{\hat{\boldsymbol{\beta}}} = \underset{h \in \mathcal{H}_n}{\arg\min} \{A^{\hat{\boldsymbol{\beta}}}(h) + V(h)\}. \tag{4.7}$$

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

4.3 Non-asymptotic bounds for kernel estimators

In this section, we provide inequalities for kernel function estimators. To assist the reading, we first focus on kernel function estimators with a fixed bandwidth. For these estimators $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h$, with h fixed, we obtain a non-asymptotic global bound on Mean Integrated Squared Error (MISE). Then, in the second subsection, we consider the Goldenshluger and Lepski method to automatically select a bandwidth $\hat{h}^{\hat{\boldsymbol{\beta}}}$. We establish a non-asymptotic oracle inequality for the resulting estimator.

4.3.1 Bound for the kernel estimator of α_0 with a fixed bandwidth

In this subsection, we consider the kernel function estimators $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h$ for any $h \in \mathcal{H}_n$. We study the global properties of these estimators. Specifically, we bound the quadratic risk of the kernel function estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h$, with a fixed bandwidth.

Theorem 4.1. *Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(ii)-(iv) and 4.4.(i)-(iii), for a fixed $h > 0$ and n large enough*

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_{2,h}^2] \leq 2||\alpha_h - \alpha_0||_{2,h}^2 + 2\frac{c_\tau \tau ||\alpha_0||_{\infty,\tau}}{c_S}\frac{||K||_2^2}{nh} + 2c(s)\frac{\log np}{n}, \tag{4.8}$$

where $c(s)$ is a constant depending on B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_1$, $||K||_2$, c_τ and on the sparsity index s of $\boldsymbol{\beta_0}$.

The inequality stated in Theorem 4.1 is non-asymptotic. It provides a bound that contains a squared bias term which decreases with $1/h$, a variance term of order $(1/nh)$, which increases with $1/h$ and a residual term of order $\log(pn)/n$. The residual term comes from the control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ given by Proposition 4.1 and is characteristic of the high-dimensional setting.

4.3.2 Oracle inequality for the adaptive kernel estimator of α_0

Now, let us state the main theorem of the chapter, which provides the first non-asymptotic oracle inequality for the adaptive kernel baseline estimator in high-dimension.

Theorem 4.2. *Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(i)-(iv) and 4.4.(i)-(iii), if $\mathcal{H}_n$ is a finite discrete set of bandwidths such that 4.5.(i)-(iii) are satisfied and $\varepsilon = \max\{h : h \in \mathcal{H}_n\}$, then there exists a constant κ such that $\hat{\alpha}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$ defined by (4.5), (4.4) and (4.7) satisfies for n large enough:*

$$\mathbb{E}[||\hat{\alpha}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}} - \alpha_0||_{2,\varepsilon}^2] \leq C \inf_{h\in\mathcal{H}_n}\left\{||\alpha_h - \alpha_0||_{2,\varepsilon}^2 + V(h)\right\} + C'(s)\frac{\log^a(n)\log(np)}{n} \tag{4.9}$$

$$\leq C \inf_{h\in\mathcal{H}_n}\left\{||\alpha_h - \alpha_0||_{2,\varepsilon}^2 + \frac{L}{nh}\right\} + C'(s)\frac{\log^a(n)\log(np)}{n} \tag{4.10}$$

where C is a numerical constant, $C'(s)$ a constant depending on τ, c_τ, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$, $||K||_2$ and on the sparsity index s of $\boldsymbol{\beta_0}$ and $L = \kappa c_\tau \tau ||\alpha_0||_{\infty,\tau}||K||_2^2/c_S$.

This inequality ensures that the adaptive kernel estimator $\hat{\alpha}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$ automatically makes the squared-bias/variance compromise. The selected bandwidth $\hat{h}^{\hat{\boldsymbol{\beta}}}$ is performing as well as the unknown oracle:

$$h_{oracle}^{\hat{\boldsymbol{\beta}}} := \arg\min_{h\in\mathcal{H}_n}\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_{2,\varepsilon}^2],$$

up to the multiplicative constant C and up to a remaining term of order $\log^a(n)\log(np)/n$, which is negligible. In Inequality (4.9), the infimum term is classic in such oracle inequalities for kernel estimators: the bias term $||\alpha_h - \alpha_0||_{2,\varepsilon}^2$ decreases

when h decreases and the variance term $V(h)$ increases when h decreases. The remaining term is of order $\log^a(n)\log(np)/n$. Chagny (2014), in the context of an additive regression model, has established an oracle inequality for the kernel estimator of the real-value regression function with a remaining term of order $1/n$. Bouaziz et al. (2013) have a logarithm term which appears in their oracle inequality with a remaining term of order $\log^{1+a}(n)/n$ instead of the expected $1/n$. This logarithm term comes from the control in $\log(n)/n$ between the distribution function G and its modified Kaplan-Meier estimator $\hat{G}$, which appears in the kernel intensity estimator. The exponent a in the remaining term arises from Assumption 4.5.(ii), which is needed for the control of the difference between the kernel intensity estimator involving $\hat{G}$ and a pseudo-estimator that does not depend of $\hat{G}$. As well as Bouaziz et al. (2013), our kernel estimator depends on an other estimator, so that we need Assumption 4.5.(ii) in order to control the difference between the kernel estimator (4.3) and the pseudo-estimator (4.12). If our kernel estimator had not involved an other estimator, we would have considered instead of Assumption 4.5.(ii), condition $\sum 1/h \leq k_0 n_0^a$, as in Chagny (2014). The term in $\log(np)/n$ in the remaining term comes from the control of $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ given by Proposition 4.1. This term is typical of the estimation of the regression parameter $\boldsymbol{\beta_0}$ when the number of covariates is large. There is no hope to capture up to usual rates in this high-dimensional setting, but the loss in the variance term is only of order $\log p/n$.

Remark 4.2. We compare Inequality (4.9) with this obtained for the penalized contrast estimator of the baseline function in Chapter 3 (Inequality (3.16)). For this, from Proposition 3.3 in Chapter 3, we deduce a similar oracle inequality to (4.9) in deterministic norm (see Definition (3.5)) restricted to the integration interval $[\varepsilon, \tau-\varepsilon]$, we denote this norm $||.||_{det,\varepsilon}$. Let Assumptions 3.1.(i)-(iii) be satisfied. Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(ii)-(iv) and 4.4.(i)-(iii), if $\mathcal{H}_n$ is a finite discrete set of bandwidths such that Assumptions 4.5.(i)-(iii) hold, then there exists a constant κ such that $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ defined by (4.5), (4.4) and (4.7) satisfies :

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{det,\varepsilon}] \leq \tilde{C} \inf_{h\in\mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||^2_{det,\varepsilon} + V(h) \right\} + \tilde{C}'(s)\frac{\log^a(n)\log(np)}{n}, \quad (4.11)$$

where $\tilde{C}$ and $\tilde{C}'(s)$ are constants depending on f_0, τ, c_τ, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$, $||K||_2$ and on the sparsity index s of $\boldsymbol{\beta_0}$.

Inequality (4.11) can be compared to Inequality (3.16) in Chapter 3. These two inequalities make appear the infimum over the models in model selection and over the bandwidths in kernel function estimators of a bias term. The penalty term pen(m) defined by (3.15) in model selection can be viewed as the mirror image of $V(h)$ defined by (4.5) in kernel function estimation. Both inequalities reveal a term that contains $\log(np)/n$. This term arises from the first step of each procedure, which is the same: the estimation of the regression parameter of the Cox model in high-dimension via a

Lasso procedure. The ℓ_1-norm $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ is controlled by a term of order $\log(np)/n$. Both inequalities highlight the link between the estimation of the baseline function as a function of the time and the estimation of the regression parameter in the Cox model in high-dimension.

Corollary 4.1. *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 4.2,*

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

where $\tilde{L}$ and $C_2'(s)$ are constants depending on τ, c_τ, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$, $||K||_2$ and on the sparsity index s of $\boldsymbol{\beta_0}$.

The proof of this corollary is similar to the one of Corollary 3.1 in model selection, by replacing D_m by $1/h$ and we refer to Remark 3.4 in Chapter 3 for comments on this corollary.

4.4 Proofs

Steps of the proofs are inspired from Bouaziz et al. (2013).

4.4.1 Intermediate Results to prove Theorem 4.1 from a pseudo-estimator

To link the kernel estimator to the true baseline function α_0, the trick is to introduce a pseudo-estimator, which does not depend on $\hat{\boldsymbol{\beta}}$. Consider for $h > 0$ the pseudo-estimator

$$\bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau K\Big(\frac{t-u}{h}\Big)\frac{1}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u), \tag{4.12}$$

which corresponds to the kernel estimator of α_0 when $S(u,\boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(u)]$ is known. To justify the choice of the pseudo-estimator, let us calculate its expectation:

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

which is a unit approximation of α_0, so that $K_h * \alpha_0 \underset{h\to 0}{\longrightarrow} \alpha_0$ under mild conditions (see Bochner Lemma).

In the following, we define for all $t \in [0, \tau]$

$$\alpha_h(t) := \mathbb{E}[\bar{\alpha}_h(t)] = K_h * \alpha_0(t). \tag{4.13}$$

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

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_{2,h}^2] \leq 2\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2] + 2\mathbb{E}[||\bar{\alpha}_h - \alpha_0||_{2,h}^2]. \tag{4.14}$$

Since the pseudo-estimator (4.12) does not depend on the estimator $\hat{\boldsymbol{\beta}}$, the error $\mathbb{E}[||\bar{\alpha}_h - \alpha_0||_{2,h}^2]$ is easier to bound than directly the error $\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_{2,h}^2]$. The study of the error of $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0$ is then divided into two parts: the error of $\bar{\alpha}_h - \alpha_0$ and the one of $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h$.

The following lemma provides the classical bias/variance inequality for the pseudo-estimator (4.12).

Lemma 4.1. *Under Assumptions 4.3.(ii)-(iv), 4.4.(i)-(iii), for $h > 0$ fixed*

$$\mathbb{E}[||\bar{\alpha}_h - \alpha_0||_{2,h}^2] \leq ||\alpha_h - \alpha_0||_{2,h}^2 + \frac{c_\tau \tau ||\alpha_0||_{\infty,\tau}}{c_S} \frac{||K||_2^2}{nh}. \tag{4.15}$$

The next lemma controls the quadratic error between $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ and $\bar{\alpha}_h$. The term to be controlled in this difference is in fact the difference between the regression parameter $\boldsymbol{\beta_0}$ and its Lasso estimator $\hat{\boldsymbol{\beta}}$. The ℓ_1-norm of this difference is bounded from Proposition 4.1 by a term of order $\log(np)/n$. This explains the obtained bound in the following lemma.

Lemma 4.2. *Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(i)-(iv) and 4.4.(i)-(iii), for a fixed $h > 0$,*

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2] \leq c(s) \frac{\log(np)}{n},$$

where $c(s)$ is a constant depending on B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_1$, $||K||_2$, c_τ and on the sparsity index s of $\boldsymbol{\beta_0}$.

From Equation (4.14), gathering Lemmas 4.1 and 4.2 provide directly Theorem 4.1.

Remark 4.3. For all $h \in \mathcal{H}_n$, by definition of ε as the maximal bandwidth in $\mathcal{H}_n$, we have for all integrable function $\alpha : [0, \tau] \to \mathbb{R}$,

$$||\alpha||_{2,\varepsilon} \leq ||\alpha||_{2,h}.$$

4.4.2 Proof of the oracle inequality in Theorem 4.2

For all $h \in \mathcal{H}_n$, let $\varepsilon = \max\{h : h \in \mathcal{H}_n\}$. Then, $A^{\hat{\boldsymbol{\beta}}}(h)$ is defined by (4.4) and we can apply this definition for $h = \hat{h}^{\hat{\boldsymbol{\beta}}}$. We deduce from this, using the definition (4.7) of $\hat{h}^{\hat{\boldsymbol{\beta}}}$, that for all $h \in \mathcal{H}_n$

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

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

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{2,\varepsilon}] \leq 6\mathbb{E}[A^{\hat{\boldsymbol{\beta}}}(h)] + 6V(h) + 3\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h} - \alpha_0||^2_{2,\varepsilon}]. \tag{4.16}$$

By definition of ε, for all $h \in \mathcal{H}_n$,

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h} - \alpha_0||^2_{2,\varepsilon}] \leq \mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h} - \alpha_0||^2_{2,h}]$$

and Theorem 4.1 gives a bound of $\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h} - \alpha_0||^2_{2,h}]$, which reveals the bias term, the variance term of order $1/nh$ and a remaining term of order $\log(np)/n$, and $V(h)$ is of the expected order $1/nh$. $\mathbb{E}[A^{\hat{\boldsymbol{\beta}}}(h)]$ is bounded in the following proposition.

Proposition 4.2. *Let $h \in \mathcal{H}_n$ be fixed. Under the assumptions of Theorem 4.2, there exist constants C_1, C_2, $C_3(s)$ such that,*

$$\mathbb{E}[A^{\hat{\boldsymbol{\beta}}}(h)] \leq C_1||\alpha_h - \alpha_0||^2_{2,\varepsilon} + C_2(s)\frac{\log^a(n)\log(pn)}{n} + C_3(s)\frac{\log(np)}{n}, \tag{4.17}$$

where the constant C_1 only depends on $||K||_1$, $C_2(s)$ and $C_3(s)$ depends, among others, on the sparsity index s of $\boldsymbol{\beta_0}$.

Applying Inequalities (4.8) and (4.17) in Equation (4.16) implies Inequality (4.9) by taking the infimum over $h \in \mathcal{H}_n$. This ends the proof of Theorem 4.2. □

4.4.3 Proof of Proposition 4.2

We introduce several additional notations, for $t \in [0, \tau]$, $\bar{\alpha}_{h,h'}(t) = K_{h'} * \bar{\alpha}_h(t)$, $\alpha_h(t) = \mathbb{E}[\bar{\alpha}_h(t)]$, $\alpha_{h,h'}(t) = \mathbb{E}[\bar{\alpha}_{h,h'}(t)]$ and write

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

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

$$||\bar{\alpha}_h - \alpha_h||_{2,h}^2 = \sup_{f \in \mathbb{L}^2([h,\tau-h]), ||f||_{2,h}=1} \langle \bar{\alpha}_h - \alpha_h, f \rangle_{2,h}^2. \tag{4.18}$$

We introduce the centered empirical process $\nu_{n,h}(f) = \langle \bar{\alpha}_h - \alpha_h, f \rangle_{2,h}$, which is equal to

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

As $f \longmapsto \nu_{n,h}(f)$ is continuous, the supremum in (4.18) can be taken over a countable dense subset of $\{f \in \mathbb{L}^2([h,\tau-h]), ||f||_{2,h} = 1\}$, which we denote by $\mathcal{B}_\tau(h)$. Therefore, we can write

$$\begin{aligned}
\mathbb{E}[T_1] &\leq \mathbb{E}\Big[\Big\{ \sup_{h' \in \mathcal{H}_n} ||\bar{\alpha}_{h'} - \alpha_{h'}||_{2,h'}^2 - V(h')/10 \Big\}_+ \Big] \\
&\leq \sum_{j, h_j \in \mathcal{H}_n} \mathbb{E}\Big[\Big\{ ||\bar{\alpha}_{h_j} - \alpha_{h_j}||_{2,h_j}^2 - V(h_j)/10 \Big\}_+ \Big] \\
&\leq \sum_{j, h_j \in \mathcal{H}_n} \mathbb{E}\left[\left\{ \sup_{f \in \mathcal{B}_\tau(h_j)} \nu_{n,h_j}^2(f) - V(h_j)/10 \right\}_+ \right]. \qquad (4.19)
\end{aligned}$$

Let introduce a key lemma, which allows to bound (4.19).

Lemma 4.3. *Let us introduced the centered process $\nu_{n,h}(f) = \langle \bar{\alpha}_h - \alpha_h, f \rangle_{2,h}$, for any $h \in \mathcal{H}_n$ and $f \in \mathbb{L}^2([h,\tau-h])$ and $\mathcal{B}_\tau(h) = \{f \in \mathbb{L}^2([h,\tau-h]), ||f||_2 = 1\}$. Under the assumptions of Theorem 4.2, with $V(h')$ defined by (4.5) for any $h' \in \mathcal{H}_n$, there exists a constant c_6 depending on the bound c_τ of the counting process, τ, $||\alpha_0||_{\infty,\tau}$, the bound c_S of $S(t, \boldsymbol{\beta_0})$, $||K||_1$ and $||K||_2$, such that*

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

So, from Lemma 4.3, there exists a constant $c_6 > 0$ such that

$$\mathbb{E}[T_1] \leq \frac{c_6}{n}, \tag{4.20}$$

where c_6 depends on c_τ, τ, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$ and $||K||_2$.

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

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

From Lemma 4.3 (see the remark at the end of the proof of Lemma 4.3), there exists a constant $c_7 > 0$ such that

$$\mathbb{E}[T_2] \leq \frac{c_7}{n}, \tag{4.21}$$

where c_7 depends on c_τ, τ, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$ and $||K||_2$.

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

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

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

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) - \bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int K\left(\frac{t-u}{h}\right)\frac{S(u,\boldsymbol{\beta_0}) - S_n(u,\hat{\boldsymbol{\beta}})}{S_n(u,\hat{\boldsymbol{\beta}})S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{\bar{Y}(u)>0\}}\mathrm{d}N_i(u). \quad (4.22)$$

Consider the following sets:

$$\Omega_{H,k} = \Big\{\omega : \forall u \in [0,\tau], |S_n(u,\hat{\boldsymbol{\beta}}) - S(u,\boldsymbol{\beta_0})| \leq 2C(s)B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}\sqrt{\frac{\log(pn^k)}{n}}\Big\}, \quad (4.23)$$

$$\Omega_{S_n} = \Big\{\omega : \forall u \in [0,\tau], S_n(u,\hat{\boldsymbol{\beta}}) - S(u,\boldsymbol{\beta_0}) \geq -\frac{c_S}{2}\Big\}, \quad (4.24)$$

$$\Omega_k = \Omega_{H,k} \cap \Omega_{S_n}, \quad (4.25)$$

where the constant $C(s)$ and c_S are respectively defined in Proposition 4.2 and Assumption 4.3.(ii). We decompose T_3 on Ω_k and on its complement. On Ω_k^c, let introduce the following lemma:

Lemma 4.4. *Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(i)-(iv) and 4.4.(i)-(iii), for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2\mathbb{1}(\Omega_k^c)] \leq c_3 n^{3-k/2},$$

where c_3 is a constant depending on B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_1$, $||K||_2$. Choosing $k \geq 8$ yields $\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2\mathbb{1}(\Omega_k^c)] \leq c_3/n$.

From Lemma 4.4,

$$\begin{aligned}\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}||\hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_{h'}||_{2,\varepsilon}^2\mathbb{1}(\Omega_k^c)\right] &\leq \sum_{j,h_j\in\mathcal{H}_n}\mathbb{E}[||\hat{\alpha}_{h_j}^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_{h_j}||_{2,h_j}^2\mathbb{1}(\Omega_k^c)] \\ &\leq \sum_{j,h_j\in\mathcal{H}_n} c_3 n^{3-k/2},\end{aligned}$$

which is of order $1/n$ as long as $k \geq 10$ from Assumption 4.5.(i). On the other hand, from (4.22) on Ω_k, we have

$$\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_{h'}^{\tau-h'}(\hat{\alpha}_{h'}^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_{h'})^2(t)\mathbb{1}(\Omega_k)\mathrm{d}t\right]$$

$$\leq \frac{16C(s)^2B^2\mathrm{e}^{2BR}\mathrm{e}^{4B|\boldsymbol{\beta_0}|_1}}{c_S^2}\frac{\log(pn^k)}{n}\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_{h'}^{\tau-h'}\left(\int\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}N_i(u)\right)\right)^2\right].$$

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

$$\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_{h'}^{\tau-h'}\left\{\int\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}N_i(u)\Big)\right\}^2\mathrm{d}t\right]$$
$$\leq 2\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_{h'}^{\tau-h'}\left\{\int\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}N_i(u)-\alpha_0(u)S(u,\boldsymbol{\beta_0})\mathrm{d}u\Big)\right\}^2\mathrm{d}t\right] \quad (4.26)$$
$$+\,2\sup_{h'\in\mathcal{H}_n}\int_{h'}^{\tau-h'}\left\{\int|K_{h'}(t-u)|\alpha_0(u)\mathrm{d}u\right\}^2\mathrm{d}t. \quad (4.27)$$

The term (4.27) is bounded by $2||K||_2^2\tau||\alpha_0||_{\infty,\tau}^2$. Let us bound the term (4.26),

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

It remains to bound the variance term.

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

We apply Lemma 4.7 and use Assumption 4.3.(iv) to get

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\int\frac{|K_{h_j}(t-u)|}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u)\right]&\leq\frac{1}{n}\mathbb{E}\left[N_1(\tau)\int\frac{K_{h_j}^2(t-u)}{S^2(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right]\\&\leq\frac{c_\tau}{n}\int\frac{K_{h_j}^2(t-u)}{S(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{d}u\\&\leq\frac{c_\tau}{nc_S}\int K_{h_j}^2(t-u)\alpha_0(u)\mathrm{d}u. \quad (4.28)\end{aligned}$$

A change of variables gives the integrated variance term:

$$\int_{h_j}^{\tau-h_j}\int_{t-h_j}^{t+h_j}K_{h_j}^2(t-u)\alpha_0(u)\mathrm{d}u\mathrm{d}t=\frac{||K||_2^2}{h_j}\int_0^\tau\alpha_0(t)\mathrm{d}t\leq\frac{||K||_2^2}{h_j}\tau||\alpha_0||_{\infty,\tau}. \quad (4.29)$$

Therefore, we get

$$\mathbb{E}\left[\sup_{h'\in\mathcal{H}_n}\int_{h'}^{\tau-h'}\Big(\int\frac{|K_{h'}(t-u)|}{S(u,\boldsymbol{\beta_0})}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}N_i(s)\Big)\Big)^2\right]$$
$$\leq 2\sum_{j,h_j\in\mathcal{H}_n}\frac{\tau||\alpha_0||_{\infty,\tau}c_\tau}{c_S}\frac{||K||_2^2}{nh_j}+2||K||_2^2\tau||\alpha_0||_{\infty,\tau}.$$

From Condition 4.4.(ii), there exists a constant c_5 such that

$$\mathbb{E}[T_3] \leq c_5(s)\frac{\log^a(n)\log(pn)}{n}, \tag{4.30}$$

where $c_5(s)$ depends on c_S, τ, $||\alpha_0||_{\infty,\tau}$, B, R, $|\boldsymbol{\beta_0}|_1$, c_τ, $||K||_2$ and on the sparsity index s of $\boldsymbol{\beta_0}$.

• Study of $\mathbb{E}[T_4]$: Since

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

we have from Young Inequality (Lemma 4.8) with $p = 1, q = 2, r = 2$,

$$\mathbb{E}[T_4] \leq ||K||_1^2\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2] \leq c(s)||K||_1^2\frac{\log(np)}{n}, \tag{4.31}$$

where the last inequality is obtained from Lemma 4.2.

• Study of $\mathbb{E}[T_5]$: From Young Inequality (Lemma 4.8) with $p = 1, q = 2, r = 2$, we obtain that

$$||\alpha_{h'} - \alpha_{h,h'}||_{2,\varepsilon}^2 = ||K_{h'} * (\alpha_0 - K_h * \alpha_0)||_{2,\varepsilon}^2 \leq ||K||_1^2||\alpha_0 - K_h * \alpha_0||_{2,\varepsilon}^2$$

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

$$\mathbb{E}[T_5] \leq ||K||_1^2||\alpha_0 - \alpha_h||_{2,\varepsilon}^2, \tag{4.32}$$

which corresponds to a bias term.

Finally, gathering the bounds of the five terms (4.20), (4.21), (4.30), (4.31) and (4.32), gives the result of Proposition 4.2. □

4.4.4 Proof of Lemma 4.3

The proof of Lemma 4.3 is based on the Talagrand Inequality given in Theorem 4.3. For all j such that $h_j \in \mathcal{H}_n$ and $f \in \mathcal{B}_\tau(h_j) = \{g \in \mathbb{L}^2([h_j, \tau - h_j]), ||g||_{2,h_j} = 1\}$, we have

$$\nu_{n,h_j}(f) = \frac{1}{n}\sum_{i=1}^n \int_{h_j}^{\tau-h_j} f(t)\left(\int \frac{K_{h_j}(t-u)}{S(u,\boldsymbol{\beta_0})}\left(\mathrm{d}N_i(u) - \alpha_0(u)S(u,\boldsymbol{\beta_0})\mathrm{d}u\right)\right)\mathrm{d}t. \tag{4.33}$$

To apply this concentration inequality, we need to determine the bounds H, M, W and the constant μ (see Theorem 4.3 in Appendix A for the notations).

• Determination of the constant M:

$$\Big| \int_{h_j}^{\tau-h_j} f(t) \int K_{h_j}(t-u) \frac{\mathrm{d}N_1(u)}{S(u, \boldsymbol{\beta_0})} \mathrm{d}t \Big| \leq ||f||_{2,h_j} \int \left(\int_{h_j}^{\tau-h_j} K_{h_j}^2(t-u) \mathrm{d}t \right)^{1/2} \frac{\mathrm{d}N_1(u)}{S(u, \boldsymbol{\beta_0})}$$
$$\leq \frac{c_\tau ||K||_2}{c_S} \frac{1}{\sqrt{h_j}} := M.$$

• Determination of the constant H:
From Equation (4.18), we have that $\sup_{f \in \mathcal{B}_\tau(h_j)} \nu_{n,h_j}^2(f) = ||\bar{\alpha}_{h_j} - \alpha_{h_j}||_{2,h_j}^2$ and from (4.13), we have that

$$\mathbb{E}\left[\sup_{f \in \mathcal{B}_\tau(h_j)} \nu_{n,h_j}^2(f) \right] \leq \mathbb{E}[||\bar{\alpha}_{h_j} - \alpha_{h_j}||_{2,h_j}^2] \leq \int_{h_j}^{\tau-h_j} \mathrm{Var}[\bar{\alpha}_{h_j}] \mathrm{d}t.$$

Combining (4.28) and (4.29), we can bound the variance term to get

$$\mathbb{E}\left[\sup_{f \in \mathcal{B}_\tau(h_j)} \nu_{n,h_j}^2(f) \right] \leq \int_{h_j}^{\tau-h_j} \mathrm{Var}[\bar{\alpha}_{h_j}] \mathrm{d}t \leq \frac{c_\tau \tau ||\alpha_0||_{\infty,\tau}}{c_S} \frac{||K||_2^2}{n h_j} := H^2.$$

We have $H^2 = V(h_j)/\kappa$. Then we set $\mu^2 = 1/2$ and $\kappa = 40$ in order to have $2(1+2\mu^2)H^2 = V(h_j)/10 = O(\frac{1}{nh_j})$.

• Determination of the constant W:
We have the following inequality

$$\mathrm{Var}\left[\int_{h_j}^{\tau-h_j} f(t) \int K_{h_j}(t-u) \frac{\mathrm{d}N_i(u)}{S(u, \boldsymbol{\beta_0})} \mathrm{d}t \right] \leq \mathbb{E}\left[\left(\int \int_{h_j}^{\tau-h_j} K_{h_j}(t-u) f(t) \mathrm{d}t \frac{\mathrm{d}N_i(u)}{S(u, \boldsymbol{\beta_0})} \right)^2 \right].$$

We introduce $K_{h_j}^-(u) = K_{h_j}(-u)$. From Lemmas 4.7 (Cauchy-Schwarz) and 4.8 (Young Inequality) with $p = 1, q = 2, r = 2$, we obtain

$$\mathrm{Var}\left[\int_{h_j}^{\tau-h_j} f(t) \int K_{h_j}(t-u) \frac{\mathrm{d}N_i(u)}{S(u, \boldsymbol{\beta_0})} \mathrm{d}t \right] \leq \mathbb{E}\left[\left(\int K_{h_j}^- * f(u) \frac{\mathrm{d}N_i(u)}{S(u, \boldsymbol{\beta_0})} \right)^2 \right]$$
$$\leq c_\tau \left(\int \frac{(K_{h_j}^- * f)^2(u)}{S(u, \boldsymbol{\beta_0})} \alpha_0(u) \mathrm{d}u \right)$$
$$\leq \frac{c_\tau ||\alpha_0||_{\infty,\tau}}{c_S} ||K_{h_j}^- * f||_2^2$$
$$\leq \frac{c_\tau ||\alpha_0||_{\infty,\tau}}{c_S} ||K_{h_j}^-||_1^2 := W.$$

We apply the Talagrand Inequality with the bounds M, H, W and the constant $\mu^2 = 1/2$ to obtain

$$\mathbb{E}\Big[\Big\{ \sup_{f \in \mathcal{B}_\tau(h_j)} \nu_{n,h_j}^2(f) - V(h_j)/10 \Big\}_+ \Big] \leq \frac{\vartheta_1}{n} \Big(\exp\Big(-\frac{\vartheta_2}{h_j} \Big) + \frac{1}{n h_j} \exp(-\vartheta_3 \sqrt{n}) \Big)$$

for some positive constants ϑ_1, ϑ_2 and ϑ_3. Then, from Assumptions 4.5.(ii)-(iii), we deduce that

$$\sum_{j,h_j\in\mathcal{H}_n}\mathbb{E}\left[\left\{\sup_{f\in\mathcal{B}_\tau(h_j)}\nu^2_{n,h_j}(f)-V(h_j)/10\right\}_+\right]\leq\frac{\vartheta_1}{n}\sum_{j,h_j\in\mathcal{H}_n}\left(\mathrm{e}^{-\frac{\vartheta_2}{h_j}}+\frac{1}{nh_j}\mathrm{e}^{-\vartheta_3\sqrt{n}}\right)\lesssim\frac{1}{n}.$$

So, there exists a constant $c_6>0$ such that

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

where c_6 depends on c_τ, τ, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$ and $||K||_2$. □

Remark: A similar lemma can be obtained for the centered process $\langle\bar{\alpha}_{h,h_j}-\alpha_{h,h_j},f\rangle_{2,h_j}$, where $\bar{\alpha}_{h,h'}=K_{h'}*\bar{\alpha}_h$. Just take

$$M=\frac{c_\tau||K||_1||K||_2}{c_S\sqrt{h_j}},\quad H^2=\frac{V(h_j)||K||_1^2}{\kappa}\quad\text{and}\quad W=\frac{c_\tau||\alpha_0||_{\infty,\tau}}{c_S}||K||_1^4$$

in the Talagrand inequality, since we have from Lemma 4.8,

$$||K_h*K_{h_j}||_2\leq||K||_1\frac{||K||_2}{\sqrt{h_j}}\quad\text{and}\quad||K_h*K^-_{h_j}||_1\leq||K||_1^2.$$

4.4.5 Proof of Lemma 4.1

From Definition (4.13), we have $\alpha_h(t)=K_h*\alpha_0(t)$ for all $t\in[0,\tau]$ so that

$$\mathbb{E}[||\bar{\alpha}_h-\alpha_0||^2_{2,h}]=\mathbb{E}[||\bar{\alpha}_h-\alpha_h||^2_{2,h}]+||\alpha_h-\alpha_0||^2_{2,h}.$$

The first term of the right part of this equality can be rewritten as

$$\mathbb{E}[||\bar{\alpha}_h-\alpha_h||^2_{2,h}]=\int_h^{\tau-h}\mathrm{Var}[\bar{\alpha}_h(t)]\mathrm{d}t.$$

It remains to bound $\mathrm{Var}[\bar{\alpha}_h(t)]$:

$$\begin{aligned}\mathrm{Var}[\bar{\alpha}_h(t)]&=\frac{1}{n}\,\mathrm{Var}\left[\int K_h(t-u)\frac{1}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right]\\&\leq\frac{1}{n}\mathbb{E}\left[\left(\int K_h(t-u)\frac{1}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right)^2\right].\end{aligned}$$

We apply Lemma 4.7 (Cauchy-Schwarz) and use Assumption 4.3.(iv) to get

$$\begin{aligned}\mathrm{Var}[\bar{\alpha}_h(t)] &\leq \frac{c_\tau}{n}\mathbb{E}\left[\int K_h^2(t-u)\frac{1}{S^2(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right] \\ &\leq \frac{c_\tau}{n}\int K_h^2(t-u)\frac{1}{S(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{d}u \\ &\leq \frac{c_\tau}{nc_S}\int K_h^2(t-u)\alpha_0(u)\mathrm{d}u. \end{aligned} \tag{4.34}$$

A change of variables gives the integrated variance term:

$$\int_h^{\tau-h}\int_{t-h}^{t+h} K_h^2(t-u)\alpha_0(u)\mathrm{d}u\mathrm{d}t = \frac{||K||_2^2}{h}\int_0^\tau \alpha_0(t)\mathrm{d}t \leq \frac{||K||_2^2}{h}\tau||\alpha_0||_{\infty,\tau}.$$

Gathering the bias term $||\alpha_h - \alpha_0||_{2,h}^2$ and the bound on variance term gives Inequality (4.15) in Lemma 4.1. □

4.4.6 Proof of Lemma 4.2

The proof of Lemma 4.2 relies on an additional lemma. First, write

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

We study the difference process $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h$ on Ω_k, defined by (4.25) and on its complement. From Lemma 4.4, the process $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h$ is controled on Ω_k^c. The following lemma allows to bound the difference process on Ω_k.

Lemma 4.5. *Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(ii)-(iv) and 4.4.(i)-(iii), for any $k \in \mathbb{N}$, we have*

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2 \mathbb{1}(\Omega_k)] \leq c_4(s)\frac{\log(pn^k)}{n},$$

where $c_4(s)$ is a constant depending on c_τ, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, τ, $||K||_1$, $||K||_2$ and on the sparsity index s of $\boldsymbol{\beta_0}$.

Gathering Lemmas 4.4 and 4.5, we finally get that, for a fixed k

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2] \leq c(s)\frac{\log(pn)}{n},$$

with $c(s)$ a constant depending on $B, |\boldsymbol{\beta_0}|_1, R, ||\alpha_0||_{\infty,\tau}, c_S, ||K||_1, ||K||_2, \tau$, c_τ and on the sparsity index s of $\boldsymbol{\beta_0}$, and Lemma 4.2 is then proved. Let now prove the Lemmas 4.4 and 4.5.

4.4.7 Proof of Lemma 4.4 :

We have to bound $\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2 \mathbb{1}(\Omega_k^c)]$, which is equal to

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2 \mathbb{1}(\Omega_k^c)] = \int_h^{\tau-h} \mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k^c)]\mathrm{d}t.$$

First, let us focus on $\mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k^c)]$ defined by

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \int K_h(t-u)\mathbb{1}_{\{\bar{Y}(u)>0\}} \frac{S(u,\boldsymbol{\beta_0}) - S_n(u,\hat{\boldsymbol{\beta}})}{S(u,\boldsymbol{\beta_0})S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}N_i(t)\right)^2 \mathbb{1}(\Omega_k^c)\right].$$

From Assumptions 4.1.(i)-(ii), $\hat{\boldsymbol{\beta}}$ belongs to a ball $\mathcal{B}(0,R)$ and $|\boldsymbol{\beta_0}|_1 < \infty$, so we have the following bound

$$S_n(u,\hat{\boldsymbol{\beta}}) - S(u,\boldsymbol{\beta_0}) \leq \frac{1}{n}\sum_{i=1}^n \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} \mathrm{e}^{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0})^{\boldsymbol{T}}\boldsymbol{Z_i}} \leq \mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}\mathrm{e}^{BR}. \tag{4.35}$$

For sake of simplicity, let us denote $C(B,R,|\boldsymbol{\beta_0}|_1)$ the bound in (4.35). From $\mathbb{1}_{\{\bar{Y}(u)>0\}}$ in the definition of $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$, there exists $i_0 \in \{1,...,n\}$ such that $Y_{i_0} \neq 0$, so that from Assumption 4.3.(i)

$$S_n(u,\hat{\boldsymbol{\beta}}) \geq \frac{1}{n}\mathrm{e}^{-B|\boldsymbol{\beta_0}|_1}\mathrm{e}^{-B|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|_1} \geq \frac{1}{n}e^{-2B|\boldsymbol{\beta_0}|_1}\mathrm{e}^{-BR}. \tag{4.36}$$

Combining (4.35) and (4.36), for $\tilde{C}(B,R,|\boldsymbol{\beta_0}|_1) = \mathrm{e}^{8B|\boldsymbol{\beta_0}|_1}\mathrm{e}^{4BR}$, we obtain the following bound

$$\mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k^c)] \leq \tilde{C}(B,R,|\boldsymbol{\beta_0}|_1)\frac{n^2}{c_S^2}\mathbb{E}\Big[\Big(\int K_h(t-u)\mathrm{d}N_1(u)\Big)^2\mathbb{1}(\Omega_k^c)\Big]. \tag{4.37}$$

We then apply the Cauchy-Schwarz Inequality (see Lemma 4.7) to get

$$\begin{aligned}\mathbb{E}\Big[\Big(\int K_h(t-u)\mathrm{d}N_1(u)\Big)^2\mathbb{1}(\Omega_k^c)\Big] &\leq \mathbb{E}\Big[N_1(\tau)\int K_h^2(t-u)\mathrm{d}N_1(u)\mathbb{1}(\Omega_k^c)\Big] \\ &\leq c_\tau \mathbb{E}\Big[\int K_h^2(t-u)\mathrm{d}N_1(u)\mathbb{1}(\Omega_k^c)\Big].\end{aligned} \tag{4.38}$$

Now we focus on $\displaystyle\int_h^{\tau-h} \mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k^c)]\mathrm{d}t$. From the two bounds (4.37) and (4.38) obtained above, we have

$$\int_h^{\tau-h} \mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k^c)]\mathrm{d}t \leq \tilde{C}(B,R,|\boldsymbol{\beta_0}|_1)\frac{n^2}{c_S^2}\Big(c_\tau \frac{||K||_2^2}{h}\mathbb{E}^{1/2}[N(\tau)^2]\sqrt{\mathbb{P}(\Omega_k^c)}\Big).$$

Let introduce the following lemma that gives a bound for $\mathbb{P}(\Omega_k^c)$.

Lemma 4.6. *Under Assumptions 4.1.(i)-(ii) and 4.2, for all $k \in \mathbb{N}$, there exists $n_0 \in \mathbb{N}$, such that for $n > n_0$ we have*

$$\mathbb{P}[\Omega_k^c] \leq c_2 n^{-k}, \tag{4.39}$$

where c_2 is a constant depending on B, $|\boldsymbol{\beta_0}|_1$ and s.

From Lemma 4.6 and the fact that $h^{-1} \leq n$ from Assumption 4.4.(iii), we get that

$$\int_h^{\tau-h} \mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k^c)]\mathrm{d}t \leq C(B, |\boldsymbol{\beta_0}|_1, R, ||\alpha_0||_{\infty,\tau}, c_S, \tau, ||K||_1, ||K||_2) n^{3-k/2},$$

where $C(B, |\boldsymbol{\beta_0}|_1, R, ||\alpha_0||_{\infty,\tau}, c_S, \tau, ||K||_1, ||K||_2)$ is a constant depending on elements in brackets. Finally, we obtain

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2 \mathbb{1}(\Omega_k^c)] \leq C(B, |\boldsymbol{\beta_0}|_1, R, ||\alpha_0||_{\infty,\tau}, c_S, \tau, ||K||_1, ||K||_2) n^{3-k/2},$$

which ends the proof of Lemma 4.4. □

4.4.8 Proof of Lemma 4.5 :

On Ω_k, we have

$$\mathbb{E}[(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h)^2(t)\mathbb{1}(\Omega_k)] \leq \frac{4B^2 \mathrm{e}^{2BR} \mathrm{e}^{4B|\boldsymbol{\beta_0}|_1}}{c_S^2} C^2(s) \frac{\log(pn^k)}{n} \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \int \frac{|K_h(t-u)|}{S(u, \boldsymbol{\beta_0})} \mathrm{d}N_i(u)\right)^2\right].$$

Then, from a change of variables and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\int_h^{\tau-h} \left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \int \frac{|K_h(t-u)|}{S(u, \boldsymbol{\beta_0})} \mathrm{d}N_i(u)\right]\right)^2 \mathrm{d}t &= \int_h^{\tau-h} \Big(\int |K_h(t-u)|\alpha_0(u)\mathrm{d}u\Big)^2 \mathrm{d}t \\
&\leq \Big(\int_{-1}^1 K^2(v)\mathrm{d}v\Big) \int_h^{\tau-h} \int_{-1}^1 \alpha_0^2(t-vh)\mathrm{d}v\mathrm{d}t \\
&\leq 2||K||_2^2 \int_0^\tau \alpha_0^2(t)\mathrm{d}t, \qquad (4.40)
\end{aligned}$$

where the last inequality is obtained with another change of variables and is bounded by $2||K||_2^2 \tau ||\alpha_0||_{\infty,\tau}^2$. From arguments as in the proof of Lemma 4.1, we have

$$\int_h^{\tau-h} \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n \int \frac{|K_h(t-u)|}{S(u, \boldsymbol{\beta_0})} \mathrm{d}N_i(u)\right] \leq \frac{\tau ||\alpha_0||_{\infty,\tau} c_\tau}{c_S} \frac{||K||_2^2}{nh}. \tag{4.41}$$

Combining the bounds (4.40) and (4.41) and with the fact that $h^{-1} \leq n$, we obtain that

$$\mathbb{E}[||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_{2,h}^2 \mathbb{1}(\Omega_k)] \leq C(\tau, c_\tau, c_S, ||\alpha_0||_{\infty,\tau}, |\boldsymbol{\beta_0}|_1, R, B, ||K||_1, ||K||_2, s) \frac{\log(pn^k)}{n}. \tag{4.42}$$

□

4.4.9 Proof of Lemma 4.6

In order to calculate $\mathbb{P}(\Omega_k^c)$, let us begin by study the set $\Omega_{H,k}$ defined by (4.23). Let us introduce the two following sets

$$\Omega_1 := \left\{ \omega : \forall u \in [0,\tau], |S_n(u,\hat{\boldsymbol{\beta}}) - S_n(u,\boldsymbol{\beta_0})| \leq B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} C(s)\sqrt{\frac{\log(pn^k)}{n}} \right\},$$
$$\Omega_2 := \left\{ \omega : \forall u \in [0,\tau], |S_n(u,\boldsymbol{\beta_0}) - S(u,\boldsymbol{\beta_0})| \leq B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} C(s)\sqrt{\frac{\log(pn^k)}{n}} \right\}.$$

We have $\Omega_{H,k} \supset \Omega_1 \cap \Omega_2$. We begin to calculate $\mathbb{P}(\Omega_1^c)$. By definition, we have

$$\begin{aligned} |S_n(u,\hat{\boldsymbol{\beta}}) - S_n(u,\boldsymbol{\beta_0})| &= \Big|\frac{1}{n}\sum_{i=1}^n (\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}} - \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}) Y_i(u)\Big| \\ &\leq \mathrm{e}^{B|\boldsymbol{\beta_0}|_1} |\mathrm{e}^{B|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|_1} - 1| \end{aligned}$$

Under Assumptions 4.1 and 4.2, from Proposition 4.1, there exists a constant $c > 0$ such that, with probability larger than $1 - cn^{-k}$,

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

So, with probability larger than $1 - cn^{-k}$, using that $|\mathrm{e}^x - \mathrm{e}^y| \leq |x-y|\mathrm{e}^{x\vee y}$ for all x, y, we have

$$\begin{aligned} |S_n(u,\hat{\boldsymbol{\beta}}) - S_n(u,\boldsymbol{\beta_0})| &\leq \mathrm{e}^{B|\boldsymbol{\beta_0}|_1} B|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \mathrm{e}^{B|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|_1} \\ &\leq B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1} C(s)\sqrt{\frac{\log(pn^k)}{n}}. \end{aligned}$$

We deduce that

$$\mathbb{P}(\Omega_1^c) \leq cn^{-k}. \tag{4.43}$$

To calculate $\mathbb{P}(\Omega_2^c)$, we remark that

$$n(S_n(u,\boldsymbol{\beta_0}) - S(u,\boldsymbol{\beta_0})) = \sum_{i=1}^n \Big(\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} Y_i(u) - \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} Y_i(u)]\Big).$$

As $0 \leq \mathrm{e}^{\boldsymbol{\beta_0^T Z_1}} Y_1(u) \leq \mathrm{e}^{B|\boldsymbol{\beta_0}|_1}$, we apply a Hoeffding inequality:

$$\mathbb{P}\left(|S_n(u,\boldsymbol{\beta_0}) - S(u,\boldsymbol{\beta_0})| \geq \frac{y}{n}\right) \leq 2\exp\left(-\frac{2y^2}{n\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}}\right),$$

and with $y = B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}C(s)\sqrt{n\log(pn^k)}/2$, we finally get

$$\begin{aligned}\mathbb{P}\left(|S_n(u,\boldsymbol{\beta_0}) - S(u,\boldsymbol{\beta_0})| \geq C(s)B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}\sqrt{\frac{\log(pn^k)}{n}}\right) \\ \leq 2\exp\left(-\frac{2B^2\mathrm{e}^{2BR}\mathrm{e}^{4B|\boldsymbol{\beta_0}|_1}C^2(s)\log(pn^k)}{\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}}\right) \\ \leq \frac{2}{pn^k}.\end{aligned}$$

We conclude that there exists a constant $c_7 > 0$ such that

$$\mathbb{P}(\Omega_2^c) \leq c_7 n^{-k}. \tag{4.44}$$

Gathering (4.43) and (4.44), we obtain

$$\mathbb{P}(\Omega_{H,k}^c) \leq \mathbb{P}(\Omega_1^c) + \mathbb{P}(\Omega_2^c) \leq \tilde{c}n^{-k}, \tag{4.45}$$

where $\tilde{c} > 0$ is a constant. It remains to calculate $\mathbb{P}(\Omega_{S_n}^c)$, with $\Omega_{S_n}^c$ defined by (4.24), to obtain $\mathbb{P}(\Omega_k^c)$. We decompose

$$S_n(u,\hat{\boldsymbol{\beta}}) - S(u,\boldsymbol{\beta_0}) = S_n(u,\hat{\boldsymbol{\beta}}) - S_n(u,\boldsymbol{\beta_0}) + S_n(u,\boldsymbol{\beta_0}) - S(u,\boldsymbol{\beta_0}).$$

On $\Omega_1 \cap \Omega_2$,

$$S_n(u,\hat{\boldsymbol{\beta}}) - S_n(u,\boldsymbol{\beta_0}) \geq -2B\mathrm{e}^{BR}\mathrm{e}^{2B|\boldsymbol{\beta_0}|_1}C(s)\sqrt{\frac{\log(pn^k)}{n}} \in (-\infty, 0)$$

So for n large enough, we have that $S_n(u,\hat{\boldsymbol{\beta}}) - S_n(u,\boldsymbol{\beta_0}) \geq -c_S/2$. For n large enough, $\Omega_1 \cap \Omega_2 \subset \Omega_{S_n}$, and

$$\mathbb{P}(\Omega_{S_n}^c) \leq \mathbb{P}(\Omega_1^c) + \mathbb{P}(\Omega_2^c) \leq \tilde{c}n^{-k}. \tag{4.46}$$

Gathering (4.45) and (4.46), we finally obtain for n large enough that $\mathbb{P}(\Omega_k^c) \leq c_2 n^{-k}$, where c_2 is a constant depending on B, $|\boldsymbol{\beta_0}|_1$ and s. □

Appendix

A Technical lemma

In this appendix, some classical technical lemmas and a theorem needed for the proofs of the two main theorems of the chapter, are listed. We do not give the proofs of these well-known results but we give the references where to find their proofs.

A.1 A Cauchy-Schwarz Inequality

The following lemma gives a useful inequality concerning integrals with respect to the counting process N.

Lemma 4.7 (Cauchy-Schwarz)**.** *For all function g bounded on $[0,\tau]$,*

$$N(\tau)\int_{\tau_1}^{\tau_2} g^2(s)\mathrm{d}N(s) \geq \Big(\int_{\tau_1}^{\tau_2} g(s)\mathrm{d}N(s)\Big)^2, \quad 0 \leq \tau_1 \leq \tau_2 \leq \tau$$

We refer to Bouaziz et al. (2013) for the proof of this lemma.

A.2 Young Inequality

The following lemma provides an inequality that bounds a norm of the convolution product of two functions by a product of norms of each function.

Lemma 4.8 (Young Inequality)**.** *Let $p,q \in [1,+\infty)$ such that $1/p + 1/q \geq 1$. If $s \in \mathbb{L}^p(\mathbb{R})$ and $t \in \mathbb{L}^q(\mathbb{R})$, then s and t are convolable. Moreover, if $1/r = 1/p+1/q-1$, then $f * g \in \mathbb{L}^r(\mathbb{R})$ and*

$$||s * t||_r \leq ||s||_p ||t||_q$$

This convolution inequality is proved in Hirsch & Lacombe (1999) (Theorem 3.4 p.149).

A.3 Talagrand Inequality

The following Talagrand Inequality is a concentration inequality that allows to control the supremum of an empirical process.

Theorem 4.3 (Talagrand Inequality)**.** *Let $\xi_1, ..., \xi_n$ be independent random values, and let*

$$\nu_{n,\xi}(f) = \frac{1}{n}\sum_{i=1}^{n}\{f(\xi_i) - \mathbb{E}[f(\xi_i)]\}.$$

Then, for a countable class of functions $\mathcal{F}$ uniformly bounded and $\mu > 0$, we have

$$\mathbb{E}\left[\left\{\sup_{f\in\mathcal{F}}\nu^2_{n,\xi}(f) - 2(1+2\mu^2)H^2\right\}_+\right] \leq \frac{4}{d}\left(\frac{W}{n}\mathrm{e}^{-d\mu^2\frac{nH^2}{W}} + \frac{98M^2}{dn^2\varphi^2(\mu)}\mathrm{e}^{-\frac{2d\varphi(\mu)\mu}{\tau\sqrt{2}}\frac{nH}{M}}\right),$$

with $\varphi(\mu) = \sqrt{1+\mu^2} - 1$, $d = 1/6$ and

$$\sup_{f\in\mathcal{F}}||f||_\infty \leq M, \quad \mathbb{E}\Big[\sup_{f\in\mathcal{F}}|\nu_{n,\xi}(f)|\Big] \leq H, \quad \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}[f(\xi)] \leq W.$$

This theorem is a useful corollary from the classical Talagrand established by Talagrand (1996). The proof of Theorem 4.3 can be found in Comte et al. (2008) (Lemma 6.1). The proof of the theorem follows from a concentration Inequality in Klein & Rio (2005) and arguments that can be found in Birgé & Massart (1998).

B Unbounded case

We do not assume anymore that $N_i(t)$ is bounded for every $t \in [0,\tau]$ and $i \in \{1,...,n\}$ (see Assumption 4.3.(iv)).

Theorem 4.4. *Under Assumptions 4.1.(i)-(ii), 4.2, 4.3.(i)-(iii) and 4.4.(i)-(iii), if $\mathcal{H}_n$ is a finite discrete set of bandwidths such that 4.5.(i)-(iii) are satisfied and $\varepsilon = \max\{h : h \in \mathcal{H}_n\}$, then there exists a constant κ such that $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ defined by (4.4) and (4.7) with*

$$V(h) = \kappa\frac{\tau||\alpha||_{\infty,\tau}}{c_S}\frac{||K||_2^2}{nh},$$

satisfies :

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||^2_{2,\varepsilon}] \leq C\inf_{h\in\mathcal{H}_n}\Big\{||\alpha_h - \alpha_0||^2_{2,\varepsilon} + V(h)\Big\} + C'(s)\frac{\log^a(n)\log(np)}{n} \qquad (4.47)$$

where C is a numerical constant, and $C'(s)$ a constant depending on τ, B, $|\boldsymbol{\beta_0}|_1$, R, $||\alpha_0||_{\infty,\tau}$, c_S, $||K||_1$, $||K||_2$, s the sparsity index of $\boldsymbol{\beta_0}$ and κ_b a constant from the Bürkholder Inequality (see Theorem A.2 in Appendix A.1).

Let us detail the important changes in the proofs of Theorem 4.4 in the unbounded case. The steps of the proof of Theorem 4.2 are similar. The only differences lie in the way we bound the variance of the pseudo-estimator and in the proof of Lemma 4.3.

B.1 Bound on the variance of the pseudo-estimator in the unbounded case:

Let us detail how we bound the variance of the pseudo-estimator in the unbounded case.

$$\begin{aligned}\mathrm{Var}[\bar{\alpha}_h(t)] &= \frac{1}{n}\,\mathrm{Var}\left[\int K_h(t-u)\frac{1}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right] \\ &\leq \frac{1}{n}\mathbb{E}\left[\left(\int K_h(t-u)\frac{1}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\right)^2\right].\end{aligned}$$

Instead of using the Cauchy-Schwarz Inequality (see Lemma 4.7), we apply the following Doob-Meyer decomposition $N_1 = M_1 + \Lambda_1$:

$$\begin{aligned}\mathrm{Var}[\bar{\alpha}_h(t)] \leq\ & \frac{2}{n}\mathbb{E}\left[\left(\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}M_1(u)\right)^2\right] \\ &+ \frac{2}{n}\mathbb{E}\left[\left(\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0^T Z_1}}Y_1(u)\mathrm{d}u\right)^2\right]\end{aligned}$$

Then, we have

$$\mathbb{E}\left[\left(\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}M_1\right)^2\right] \leq \mathbb{E}\left[\int \frac{K_h^2(t-u)}{S^2(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0^T Z_1}}Y_1(u)\mathrm{d}u\right],$$

and using the standard Cauchy-Schwarz Inequality, we finally get that

$$\mathbb{E}\left[\left(\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}M\right)^2\right] \leq \frac{\mathbb{E}[\mathrm{e}^{2\boldsymbol{\beta_0^T Z}}]||\alpha_0||^2_{\infty,\tau}\tau}{c_S^2}\frac{||K||_2^2}{h}.$$

Thus, in the unbounded case, the variance term of the pseudo-estimator is bounded by

$$\mathrm{Var}[\bar{\alpha}_h(t)] \leq \frac{2||\alpha_0||_{\infty,\tau}}{c_S^2}(\mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z}}] + ||\alpha_0||_{\infty,\tau}\mathbb{E}[\mathrm{e}^{2\boldsymbol{\beta_0^T Z}}]\tau)\frac{||K||_2^2}{nh},$$

which gives a bound of the same order $1/nh$, than the one obtained in the bounded case (see Inequality (4.34)). □

B.2 Proof of Lemma 4.3 in the unbounded case:

We have to control the supremum of $\nu_{n,h}(f)$ defined by (4.33) over the ball $\mathcal{B}_\tau(h) = \{f \in \mathbb{L}^2([h,\tau-h]), ||f||_{2,h} = 1\}$. In the unbounded case, we can not directly apply the Talagrand Inequality: we have to introduce a truncation (see Chagny (2014) for a close approach). Let us define for a constant c,

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

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

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

where

$$\nu_{n,h}^{(1)}(f) = \frac{1}{n}\sum_{i=1}^{n}\int_h^{\tau-h} f(t)\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)\leq\kappa_n\}}\mathrm{d}M_i(u)\mathrm{d}t$$
$$- \frac{1}{n}\sum_{i=1}^{n}\int_h^{\tau-h} f(t)\int \mathbb{E}\left[\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_i(\tau)\leq\kappa_n\}}\mathrm{d}M_i(u)\right]\mathrm{d}t$$

and

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

We follow the lines and the notations of the proof of Lemma 4.3 (see Subsection 4.4.4).

• Control of $\nu_{n,h}^{(1)}(f)$:

As in the proof of Lemma 4.3, we can apply a Talagrand Inequality to $\nu_{n,h}^{(1)}(f)$, which is bounded, with the following bounds:

— Determination of the constant M:
Using the Doob-Meyer decomposition and the Cauchy-Schwarz Inequality, we have for $f \in \mathcal{B}_\tau(h)$

$$\left|\int_h^{\tau-h} f(t)\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}M_1(u)\mathrm{d}t\right|$$
$$\leq \left|\int_h^{\tau-h} f(t)\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}N_1(u)\mathrm{d}t\right|$$
$$+\left|\int_h^{\tau-h} f(t)\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0^T Z_1}}Y_1(u)\mathrm{d}u\mathrm{d}t\right|$$
$$\leq ||f||_{2,h}\left|\int\left(\int_h^{\tau-h} K_h^2(t-u)\frac{\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_1(u)\mathrm{d}t\right|\right.$$
$$+||f||_{2,h}\left|\int\left(\int_h^{\tau-h} K_h^2(t-u)\frac{\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}}{S(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0^T Z_1}}Y_1(u)\mathrm{d}u\mathrm{d}t\right|\right.$$
$$\leq \frac{||K||_2^2}{\sqrt{h}}\frac{|N_1(\tau)\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}|}{c_S} + \frac{||K||_2^2}{\sqrt{h}}\frac{||\alpha_0||_{\infty,\tau}\mathrm{e}^{B|\boldsymbol{\beta_0}|_1\tau}}{c_S}$$
$$\leq \frac{||K||_2^2}{c_S\sqrt{h}}\left(\kappa_n + \frac{||\alpha_0||_{\infty,\tau}\mathrm{e}^{B|\boldsymbol{\beta_0}|_1\tau}}{c_S}\right) := M \sim \frac{\sqrt{n}}{\log n\sqrt{h}}.$$

— Determination of the constant H:

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{B}_\tau(h)}(\nu^{(1)}_{n,h}(f))^2\right] &\leq \frac{1}{n}\int_h^{\tau-h}\mathrm{Var}\left[\int\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}M_1(u)\right]\mathrm{d}t\\ &\leq \frac{1}{n}\int_h^{\tau-h}\mathbb{E}\left[\left(\int\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}M_1(u)\right)^2\right]\mathrm{d}t\\ &\leq \frac{\tau||\alpha||_{\infty,\tau}}{c_S}\frac{||K||_2^2}{nh} := H^2\end{aligned}$$

We have $H^2 = V(h_j)/\kappa$. Then we set $\mu^2 = 1/2$ and $\kappa = 40$ in order to have $2(1+2\mu^2)H^2 = V(h_j)/10 = O(\frac{1}{nh_j})$.

— Determination of the constant W:
From Young Lemma 4.8, we have

$$\begin{aligned}&\mathrm{Var}\left[\int_h^{\tau-h}f(t)\int\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}M_1(u)\mathrm{d}t\right]\\ &\quad\leq\mathbb{E}\left[\left(\int_h^{\tau-h}f(t)\int\frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\mathrm{d}M_1(u)\mathrm{d}t\right)^2\right]\\ &\quad\leq\mathbb{E}\left[\mathbb{1}_{\{N_1(\tau)\leq\kappa_n\}}\left(\int\frac{(K_h*f)(u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}M_1(u)\right)^2\right]\\ &\quad\leq\mathbb{E}\left[\int\frac{(K_h*f)^2(u)}{S^2(u,\boldsymbol{\beta_0})}\alpha_0(u)\mathrm{e}^{\boldsymbol{\beta_0^T Z_1}}Y_1(u)\mathrm{d}u\right]\\ &\quad\leq\frac{||K_h*f||_2^2}{c_S}||\alpha_0||_{\infty,\tau}\leq\frac{||\alpha_0||_{\infty,\tau}}{c_S}||K||_1^2 := W.\end{aligned}$$

Then, from Assumptions 4.5.(i)-(iii), we deduce that

$$\begin{aligned}\sum_{j,h_j\in\mathcal{H}_n}\mathbb{E}\left[\left\{\sup_{f\in\mathcal{B}_\tau(h_j)}\nu^2_{n,h_j}(f)-V(h_j)/10\right\}_+\right]&\leq\frac{\vartheta_1}{n}\sum_{j,h_j\in\mathcal{H}_n}\left(\frac{1}{n}\mathrm{e}^{-\frac{\vartheta_2}{h_j}}+\frac{1}{n\log(n)h_j}\mathrm{e}^{-\vartheta_3\log n}\right)\\ &\lesssim\frac{1}{n},\end{aligned}$$

with

$$V(h)=\kappa\frac{\tau||\alpha||_{\infty,\tau}}{c_S}\frac{||K||_2^2}{nh}.$$

From Assumptions 4.5.(ii) and (iii), we obtain for a constant $c>0$, that

$$\sum_{j,h_j\in\mathcal{H}_n}\mathbb{E}\left[\left\{\sup_{f\in\mathcal{B}_\tau(h_j)}(\nu^{(1)}_{n,h_j}(f))^2-V(h_j)/10\right\}_+\right]\leq\frac{c}{n}.$$

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{B}_\tau(h)}(\nu_{n,h}^{(2)}(f))^2\right] &\leq \mathbb{E}\left[\int_h^{\tau-h}\Psi^2(t)\mathrm{d}t\right] \\
&\leq \frac{1}{n}\int_h^{\tau-h}\mathrm{Var}\left[\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}\mathrm{d}M_1(u)\right]\mathrm{d}t \\
&\leq \frac{1}{n}\int_h^{\tau-h}\mathbb{E}\left[\left(\int \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}\mathrm{d}M_1(u)\right)^2\right]\mathrm{d}t
\end{aligned}$$

Applying the Bürkholder Inequality (see Theorem A.2 in Appendix A.1, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{B}_\tau(h)}(\nu_{n,h}^{(2)}(f))^2\right] &\leq \frac{\kappa_b}{c_S^2}\frac{||K||_2^2}{nh}\mathbb{E}[N_1(\tau)\mathbb{1}_{\{N_1(\tau)>\kappa_n\}}] \\
&\leq \frac{\kappa_b}{c_S^2}\frac{||K||_2^2}{nh}\frac{\mathbb{E}[N_1(\tau)^{p+1}]}{\kappa_n^p}.
\end{aligned}$$

From Assumption 4.5.(ii), we deduce that for p large enough

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

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

$$\sum_{j,h_j\in\mathcal{H}_n}\mathbb{E}\left[\sup_{f\in\mathcal{B}_\tau(h)}(\nu_{n,h}^{(2)}(f))^2\right] \leq \frac{C}{n},$$

for a constant $C > 0$, which ends the proof. □

Chapitre 5

Simulations

Sommaire

5.1 Simulated data 181
5.2 Step 1: Estimation of the regression parameter β_0 184
5.3 Step 2: Estimation of the baseline function α_0 187
5.3.1 Penalized contrast estimator and model selection 187
5.3.2 Kernel estimators and bandwidth selection 191
5.4 Measuring the quality of the estimators 193
5.5 Results in the case of simulated data 194
5.5.1 Results for the regression parameter estimation 194
5.5.2 Results for the baseline function estimation 198
5.6 Application to a real dataset on breast cancer 205
5.6.1 Variable selection 206
5.6.2 Estimation of the baseline hazard function 206
Appendices 210
A The rectangle method 210
B Calibration of the constants 210
C Description of the real data 212

Abstract. The aim of this chapter is to study the practical performances of the estimators proposed in Chapters 3 and 4. With the model selection approach of Chapter 3, we obtain a penalized contrast estimator. The approach of Chapter 4 provides an adaptive kernel estimator. In addition to these two estimators, we also consider the usual kernel estimator of the baseline function, with a bandwidth selected by cross-validation, which we refer to as "cross-validated kernel estimator". We describe how we implement the three estimators by decomposing the estimation procedures. We then apply these procedures to simulated data to illustrate the behavior of these three estimators. We then compare theirs respective performances by calculating different Mean Integrated Squared Errors (MISE). Lastly, we apply the three procedures to a real dataset on breast cancer, to compare the baseline functions for two groups of patients, the treated and the untreated, and to draw informations on the lifetime in these two groups.

Résumé L'objectif de ce chapitre est d'étudier les performances pratiques des estimateurs obtenus dans les chapitres 3 et 4. À partir de l'approche par sélection de modèles du chapitre 3, nous obtenons un estimateur par contraste pénalisé. L'approche du chapitre 4 nous a permis d'obtenir un estimateur à noyau adaptatif. En plus de ces deux estimateurs, nous considérons aussi l'estimateur à noyau usuel du risque de base avec une fenêtre choisie par validation croisée, que nous appelons « estimateur à noyau cross-validé ». Nous décrivons comment nous implémentons ces trois estimateurs en décomposant les procédures d'estimation. Nous appliquons ensuite ces procédures à des données simulées pour illustrer le comportement de ces trois estimateurs par rapport au vrai risque de base simulé et nous les comparons les uns aux autres en calculant des erreurs quadratiques moyennes intégrées (MISE) et en traçant leurs courbes sur un même graphe. Enfin, nous appliquons les trois procédures à un jeu de données sur le cancer du sein pour comparer les risques de base pour deux types de patients, les patients traités et les patients non-traités, et pour obtenir des informations sur la durée de survie dans ces deux groupes.

In this chapter, we describe the implementation of the estimation procedures described in Chapters 3 and 4 to estimate the two parameters of the Cox model.

The chapter is organized as follows. In Section 5.1, we describe how we generate simulated data in the context of right censoring. In Section 5.2, we propose four different estimation procedures for the regression parameter $\boldsymbol{\beta_0}$. In Section 5.3, we describe the implementation of the different procedures to estimate the baseline function α_0: first, we explain how we compute the model selection procedure in an histogram basis and in a trigonometric basis, and then we specify the implementation of the kernel estimator and the selection of the bandwidth with cross-validation and with the Goldenshluger and Lepski method. Section 5.4 is devoted to the definitions of three different Integrated Squared Errors (ISE) that we consider to measure the quality of the estimators. In Section 5.5, we present the results for the estimations of the regression parameter and for the baseline function. Section 5.6 ends the study with the application of the implemented procedures to a real dataset on breast cancer. Appendix A describes the Rectangle method, which allows to approximate integrals, Appendix B explains how we calibrate some constants in our procedures and Appendix C gives some details on the real dataset on breast cancer.

This simulation study has been implemented with the software **R**. The programs are available on request.

5.1 Simulated data

In this section, we describe data generation process in the right censoring framework (as described in the introduction of Part II, see Remark 2.10). Recall the Cox model:

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \text{for } i = 1, ..., n \tag{5.1}$$

where $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T$ is the vector of covariates of individual i.

Sample size and number of covariates. 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 $n = 50$, $n = 200$ and $n = 500$ and p varies as follows:

— $p = 5 < n$, referred to as the small dimension case,

— $p = \sqrt{n}$, being 7, 15 and 22 respectively and referred to as the intermediate case,

— $p \geq n$, being 50, 100, 200, 500 and 1000 respectively and referred to as the high-dimension case.

Regression parameter in the Cox model. The true regression parameter $\boldsymbol{\beta_0}$ is chosen as a vector of dimension p, defined by $\boldsymbol{\beta_0} = (0.1, 0.3, 0.5, 0, ..., 0)^T \in \mathbb{R}^p$, for various $p \geq 3$.

Matrix design. 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 $[-1, 1]$.

Survival times and baseline function. We consider two forms for the true baseline function α_0 depending on the distribution of the survival time.

(i) If the survival times are distributed according to a Weibull distribution $\mathcal{W}(a, \lambda)$, the baseline function is of the form $\alpha_0(t) = a\lambda^a t^{a-1}$, where a and λ are two positive parameters. From the relation between the conditional survival function and the conditional hazard rate function, we have for $i = 1, ..., n$,

$$S(t, \boldsymbol{Z_i}) = \exp\Big[- \int_0^t \lambda_0(s, \boldsymbol{Z_i}) \mathrm{d}s \Big] = \exp(-\lambda^a t^a \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}).$$

We simulate n uniform random variables $Y_i \sim \mathcal{U}([0, 1])$, for $i = 1, ..., n$. Inverting the conditional survival function, one gets an explicit expression of the survival times T_i:

$$T_i = \frac{1}{\lambda}\Big(- \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}} \log(1 - Y_i)\Big)^{1/a}.$$

We implement T_i from this expression. We simulate three Weibull distributions for the survival times: $\mathcal{W}(1.5, 1)$, $\mathcal{W}(0.5, 2)$, $\mathcal{W}(3, 4)$.

(ii) If the survival times are distributed according to a log-normal distribution $\ln\mathcal{N}(\lambda, a)$, the baseline function is of the form

$$\alpha_0(t) = \frac{\frac{1}{a\sqrt{2\pi t}} \exp\Big(- \frac{(\log t - \lambda)^2}{2a^2}\Big)}{1 - \Phi\Big(\frac{\log t - \lambda}{a}\Big)},$$

where Φ is the cumulative distribution function of the standard normal distribution. The associated survival time T_i is defined for $i = 1, ..., n$, by

$$T_i = \exp\left(a\Phi^{-1}\left(1 - \exp\left(\frac{\log(1 - Y_i)}{\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}}\right)\right) + \lambda\right), \quad \text{where } Y_i \sim \mathcal{U}([0, 1]).$$

We simulate three log-normal distributions for the survival times: $\ln\mathcal{N}(0.25, 0)$, $\ln\mathcal{N}(1, 0)$, $\ln\mathcal{N}(10, 0)$.

From the different distributions considered, we obtain several different forms for the baseline hazard function α_0 (see Figure 1.0.1).

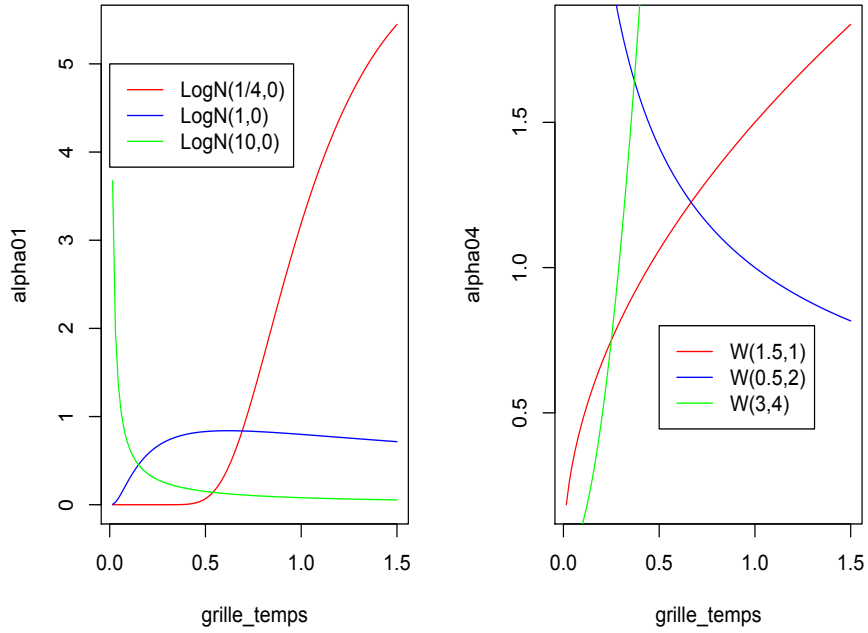


FIGURE 1.0.1 – *Plots of the baseline hazard function for different parameters of a log-normal distribution (left) and of a Weibull distribution (right).*

Censoring times. We set a rate of censoring of the form $1/\gamma\mathbb{E}[T_1]$, where $\gamma > 0$ is a constant to be adjusted to the rate of censorship. We consider two different rates of censoring:

— a large rate of censoring of 50% by taking $\gamma = 1.2$,

— a moderate rate of 20% of censorship by taking $\gamma = 4.5$.

The censoring times $(\tilde{C}_i)_{1\leq i\leq n}$ are simulated independently from the survival times via an exponential distribution $\mathcal{E}(1/\gamma\mathbb{E}[T_1])$, with rate parameter $1/\gamma\mathbb{E}[T_1]$.

Time interval of the study. The time τ of the end of the study is taken as the quantile at 90% of $(T_i \wedge \tilde{C}_i)_{i=1,\dots,n}$. We consider in the following that τ is non-random. A theoretical study with a random τ is beyond the scope of the present study.

Observed times and censoring indicator. For $i = 1, ..., n$, we compute the observed times $X_i = \min(T_i, C_i)$, where $C_i = \tilde{C}_i \wedge \tau$, and the censoring indicators $\delta_i = \mathbb{1}_{T_i \leq C_i}$. The definition of 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 (we consider

$N_e = 500$ for some experiments but it takes too much time to rise N_e to 500 in each one).

5.2 Step 1: Estimation of the regression parameter $\boldsymbol{\beta_0}$

We implement four procedures to estimate the regression parameter $\boldsymbol{\beta_0}$: the classical minimization of the opposite of the partial log-likelihood, the classical Lasso procedure, a procedure that performs a Lasso procedure and a minimization of the partial log-likelihood on the covariates selected from the Lasso, and lastly the adaptive Lasso procedure. These algorithms are all based on the Cox partial likelihood (2.106), which can be rewritten in the censoring case by

$$l_n^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \log \frac{\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}}}{S_n(X_i, \boldsymbol{\beta})}, \quad \text{where } S_n(\boldsymbol{\beta}, X_i) = \sum_{k:X_k \geq X_i} \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_k}}. \tag{5.2}$$

Let us describe these four different strategies.

1. **The maximum Cox partial log-likelihood estimation:** According to (5.2), the maximum Cox partial log-likelihood estimator is defined by

$$\hat{\boldsymbol{\beta}}^{ML} = \arg\max_{\boldsymbol{\beta}} \{l_n^*(\boldsymbol{\beta}))\}.$$

   This procedure is implemented in R with the *coxph* function, located in the *survival* library in R (see Andersen & Gill (1982) and Therneau & Grambsch (2000) for theoretical and practical details about this procedure). The obtained estimator is consistent (see Tsiatis (1981) and Andersen & Gill (1982)) and asymptotically efficient. The maximum Cox partial log-likelihood procedure works well when the number of covariates p is not too large compared with n. When p increase, the estimations become wrong and when p is really too large, the procedure does not converge anymore. In theory, the high-dimension is reached when p is of order n, but in practice the *coxph* algorithm can no longer be used as soon as p exceeds $\sqrt{n}$. Thus, we consider the *coxph* algorithm only in the cases when $p \leq \sqrt{n}$.

2. **The Lasso:** Recall that the Lasso estimator, introduced by Tibshirani (1997), is defined by

$$\hat{\boldsymbol{\beta}}_{\mu}^{Lasso} = \arg\min_{\boldsymbol{\beta}} \{-l_n^*(\boldsymbol{\beta}) + \mu|\boldsymbol{\beta}|_1\}, \tag{5.3}$$

   where $\mu > 0$ is a regularization parameter.

The problem (5.3) is convex (see Huang et al. (2013) for example) and implementable. The Lasso procedure for the Cox model has been implemented in R by Simon et al. (2011) with the *glmnet* function (located in the *glmnet* library in R). This algorithm is based on a cyclical coordinate descent. The idea is the following: one minimizes on one coordinate a quadratic approximation of the Cox partial log-likelihood. This amounts to consider in the hessian only one non-zero coordinate on the diagonal. We obtain a gradient descent, to which we apply a soft-threshholding operator (associated with the ℓ_1 penalty). One does this procedure in every directions. As proposed by Tibshirani (1996) and implemented in the *glmnet* procedure, the regularization parameter μ is chosen via cross-validation. The Lasso procedure (5.3) is defined for all p, in particular it is adapted to the case $p \geq n$. The resulting estimator is sparse and thus easily interpretable.

One limitation of the Lasso procedure is that it is in general biased (see Zhang & Huang (2008a)) and in some case inconsistent for variable selection (see Zou (2006)). To overcome these problems, we consider two other procedures (see Paragraph 3 for a procedure that corrects the bias of the Lasso and Paragraph 4 for a procedure that provides a consistent estimator).

3. **A Lasso associated to a maximum Cox partial log-likelihood estimation:** In order to correct the bias of the Lasso procedure, we propose a two-step estimation procedure, which combines the two previous procedures. First, we select the non-zero coordinates of $\boldsymbol{\beta_0}$ with the *glmnet* procedure, we obtain the Lasso estimator $\tilde{\boldsymbol{\beta}}$. Second, to the Cox model with those selected variables, we apply a *coxph* procedure. To be more precise, let define the sparsity set of $\tilde{\boldsymbol{\beta}}$ by $J(\tilde{\boldsymbol{\beta}}) = \{j \in \{1, ..., p\} : \tilde{\beta}_j \neq 0\}$ and its sparsity index by $|J(\tilde{\boldsymbol{\beta}})| = \text{Card}\, J(\tilde{\boldsymbol{\beta}})$. Let also consider $\tilde{\boldsymbol{Z}}$ the vector of covariates such that $\tilde{Z}_j = Z_j$ when $j \in J(\tilde{\boldsymbol{\beta}})$ and $\tilde{Z}_j = 0$ otherwise. The resulting estimator is then defined by

$$\hat{\boldsymbol{\beta}}^{Lasso+ML} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{|J(\tilde{\boldsymbol{\beta}})|}}{\arg\min} \{-\tilde{l}_n^*(\boldsymbol{\beta})\},$$

where

$$\tilde{l}_n^*(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \frac{\mathrm{e}^{\boldsymbol{\beta}^T \tilde{\boldsymbol{Z}}_i}}{\tilde{S}_n(X_i, \boldsymbol{\beta})} \quad \text{and} \quad \tilde{S}_n(X_i, \boldsymbol{\beta}) = \sum_{k: X_k \geq X_i} \mathrm{e}^{\boldsymbol{\beta}^T \tilde{\boldsymbol{Z}}_k}.$$

This procedure is not always defined. Indeed, when the number of selected covariates with the Lasso procedure is still too large (i.e. larger than $\sqrt{n}$), the minimization of the opposite of the Cox partial log-likelihood is not possible. In this case, we only consider a Lasso procedure, without adding the second step with the *coxph* procedure.

4. **The adaptive Lasso:** The Lasso problem, interesting for its stability, convexity and sparse properties, has inspired lots of works, which address the problem of providing interpretable estimators with selection consistency (i.e. estimators that select covariates with nonzero coefficients with probability converging to one). From the practical point of view, this problem deals with improving the selective and predictive performances of the Lasso. The adaptive Lasso, introduced by Zou (2006) in the linear regression, provides an estimator both sparse and consistent. It has been extended by Zhang & Lu (2007) to the Cox model. The adaptive Lasso is a two-step method. First, one calculates a preliminary estimator $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$, which can be the maximum Cox partial log-likelihood estimator when p is not too large or any other estimator. Then, the preliminary estimator $\tilde{\boldsymbol{\beta}}$ is used to fit the penalty imposed on each regression parameter of the adaptive Lasso as follows

$$\hat{\boldsymbol{\beta}}^{AdapL} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left\{ - l_n^*(\boldsymbol{\beta}) + \Gamma_n \sum_{j=1}^{p} \frac{1}{|\tilde{\beta}_j|}|\beta_j| \right\}. \tag{5.4}$$

In this work, we propose to use the Lasso estimator as the preliminary estimator $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{Lasso}$. This leads to a penalty of the form

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{Lasso}| + \varepsilon}, \tag{5.5}$$

where ε is the minimum of the $|\hat{\beta}_j^{Lasso}|$ that are non-zeros and the procedure is then

$$\hat{\boldsymbol{\beta}}^{AdapL} = \underset{\boldsymbol{\beta}}{\arg\min}\{l_n^*(\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\beta})\}.$$

This procedure can be obtained by using the function *glmnet* in R. Indeed, *glmnet* admits optional parameters, that allows to consider other forms of penalties such as this defines by (5.5). The resulting estimator $\hat{\boldsymbol{\beta}}^{AdapL}$ has the advantage to be consistent in selection under ad hoc conditions, but is more biased than the Lasso (see Zou (2006)).

The maximum Cox partial log-likelihood procedure and the Lasso associated to a maximum Cox partial log-likelihood procedure do not work when the number of covariates is too large compared to n. Thus, we consider the following different regimes:

— the classical setting, also called the small-dimension case (generally when $p \leq \sqrt{n}$), in which all the estimators described above are well defined,

— the moderate high-dimension setting, when the classical maximum Cox partial log-likelihood procedure to estimate $\boldsymbol{\beta_0}$ does not work anymore (generally from $\sqrt{n}$) and when we consider either the Lasso associated to a maximum Cox partial log-likelihood procedure, the Lasso or the adaptive Lasso,

— the high-dimensional setting, when $p \geq n$, for which only the Lasso and the adaptive Lasso work.

Even without reaching the ultra high-dimension, as defined by Verzelen (2012), the algorithms, such as the Lasso or the adaptive Lasso, are performing poorly when $p \gg n$. In this case, one proceeds to a preliminary screening to reduce the number of covariates to be entered in the multivariate model (see Section 5.6 for details on the screening and application to a real dataset on breast cancer). Generally, one attempts to obtain a p of order n.

5.3 Step 2: Estimation of the baseline function α_0

We now describe the implementation of the two strategies to estimate the baseline hazard function: the model selection and the kernel estimation. We implement the model selection in two different bases of functions and for the kernel estimation, we implement two procedures to select the bandwidth. The resulting estimators are the following:

1. The penalized contrast estimator $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ obtained by model selection, in
   - **(i)** Histogram basis
   - **(ii)** Trigonometric basis
2. The kernel estimator with a bandwidth selected by
   - **(i)** Cross-validation,
   - **(ii)** The Goldenshluger and Lepski method,

   the resulting estimators are denoted by $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}}$ and $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ respectively.

Note that all the three estimators of the baseline function depends on the estimator $\hat{\boldsymbol{\beta}}$ implemented in the first step. Let us precise how we implement all the procedures.

5.3.1 Penalized contrast estimator and model selection

We refer to Chapter 3 for a detailed description of the procedure and briefly recall it.

The model selection method is based on the minimization of a contrast. This contrast in our context is defined for a function $\alpha \in (\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$ by

$$C_n(\alpha, \hat{\boldsymbol{\beta}}) = -\frac{2}{n}\sum_{i=1}^{n} \alpha(X_i)\delta_i + \frac{1}{n}\sum_{i=1}^{n}\int_0^{X_i} \alpha^2(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}\mathrm{d}t. \tag{5.6}$$

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

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

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. $|J_m| = D_m$. From Assumption 3.1.(i), one must have $D_m \leq \sqrt{n}/\log n$ for all $m \in \mathcal{M}_n$. Associated to the collection of models, we define the sequence of minimum contrast estimators $(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}})_{m\in\mathcal{M}_n}$ with

$$\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}} = \underset{\alpha\in S_m}{\arg\min}\{C_n(\alpha, \hat{\boldsymbol{\beta}})\}.$$

The relevant space is automatically selected by using following penalized criterion

$$\hat{m}^{\hat{\boldsymbol{\beta}}} = \underset{m\in\mathcal{M}_n}{\arg\min}\{C_n(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}, \hat{\boldsymbol{\beta}}) + \text{pen}(m)\}, \tag{5.7}$$

with

$$\text{pen}(m) := \kappa(1 + ||\alpha_0||_{\infty,\tau})\frac{D_m}{n}, \tag{5.8}$$

The penalized contrast estimator is then $\hat{\alpha}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$.

Implementation of the minimum contrast estimators $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}$

By definition, for $m \in \mathcal{M}_n$, $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}$ belongs to the model S_m. Thus, there exists $\hat{\boldsymbol{A}}_{\boldsymbol{m}}^{\hat{\boldsymbol{\beta}}} = (\hat{a}_j^{\hat{\boldsymbol{\beta}}})_{j\in J_m}$ such that, for $(\varphi_j^m)_{j\in J_m}$ an orthonormal basis of $(\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$ and

$$\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}(t) = \sum_{j\in J_m} \hat{a}_j^{\hat{\boldsymbol{\beta}}}\varphi_j^m(t), \quad \forall t \in [0,\tau] \tag{5.9}$$

is the solution of the equation

$$\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\boldsymbol{\beta}}}\hat{\boldsymbol{A}}_{\boldsymbol{m}}^{\hat{\boldsymbol{\beta}}} = \boldsymbol{\Gamma}_{\boldsymbol{m}}, \tag{5.10}$$

where $\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\boldsymbol{\beta}}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{m}}$ are respectively defined by

$$\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\boldsymbol{\beta}}} = \Big(\frac{1}{n}\sum_{i=1}^n \int_0^{X_i} \varphi_j(t)\varphi_k(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}Y_i(t)\mathrm{d}t\Big)_{(j,k)\in J_m^2}, \tag{5.11}$$

$$\boldsymbol{\Gamma}_{\boldsymbol{m}} = \Big(\frac{1}{n}\sum_{i=1}^n \delta_i\varphi_j(X_i)\Big)_{j\in J_m}. \tag{5.12}$$

We implement $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}$ in two different bases of $(\mathbb{L}^2 \cap \mathbb{L}^\infty)([0,\tau])$: the histogram basis and the trigonometric basis, both described in Chapter 3, Example 3.1.

(i) Histogram basis (see Chapter 3, Example 3.1.1.). In this case,

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

such that $D_m = 2^m \leq \sqrt{n}/\log n$, according to Assumption 3.1.(i). We take in the program $m = 0, ..., \lfloor \log(n/\log(n))/\log(2) \rfloor$. The solution of Equation (5.10) satisfies

$$\hat{a}_j^{\hat{\boldsymbol{\beta}}} = \frac{\tau}{2^m} \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \left(\left(\min\left(X_i, \frac{j\tau}{2^m}\right) - \frac{(j-1)\tau}{2^m} \right) \vee 0 \right)} \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{2^{m/2}}{\sqrt{\tau}} \mathbb{1}_{\left[\frac{(j-1)\tau}{2^m}, \frac{j\tau}{2^m}\right)}(X_i). \tag{5.13}$$

For $m = 0, ..., \lfloor \log(n/\log(n))/\log(2) \rfloor$, we compute the candidates $\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}$ for the estimation of α_0 using relation (5.9).

(ii) Trigonometric basis (see Chapter 3, Example 3.1.2.). Recall that

$$\begin{cases} \varphi_1(t) & = \sqrt{2/\tau}, \\ \varphi_{2j}(t) & = \sqrt{2/\tau}\cos(2\pi j t/\tau), \quad \text{for } j \geq 1, \\ \varphi_{2j+1}(t) & = \sqrt{2/\tau}\sin(2\pi j t/\tau), \quad \text{for } j \geq 1. \end{cases}$$

For this model, $D_m = m \leq \sqrt{n}/\log n$ according to Assumption 3.1.(i). In the program, we add two models by taking $m = 1, ..., \lfloor \sqrt{n}/\log(n) + 2 \rfloor$. In practice, it is classic to consider more models than what the theory authorizes (see Brunel et al. (2010) for example). Then, to solve Equation (5.10), we compute the Gram matrix $\boldsymbol{G}_{\boldsymbol{m}}^{\hat{\boldsymbol{\beta}}}$ and the vector $\boldsymbol{\Gamma_m}$ by calculating the integrals defined in (5.11) and (5.12) for all possibilities: let $q \neq r$ in $\mathbb{N}^*$,

$$\begin{aligned}
(G_m^{\hat{\beta}})_{1,1} &= \frac{1}{n\tau}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} X_i, \\
(G_m^{\hat{\beta}})_{1,2q} &= \frac{\sqrt{2}}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \frac{\sin(2\pi q X_i/\tau)}{2\pi q}, \\
(G_m^{\hat{\beta}})_{1,2q+1} &= \frac{\sqrt{2}}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \left(\frac{1}{2\pi q} - \frac{\cos(2\pi q X_i/\tau)}{2\pi q} \right), \\
(G_m^{\hat{\beta}})_{2q,2q} &= \frac{2}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \left(\frac{\cos(2\pi q X_i/\tau)\sin(2\pi q X_i/\tau)}{4\pi q} + \frac{X_i}{2\tau} \right), \\
(G_m^{\hat{\beta}})_{2q+1,2q+1} &= \frac{2}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \left(\frac{X_i}{2\tau} - \frac{\cos(2\pi q X_i/\tau)\sin(2\pi q X_i/\tau)}{4\pi q} \right), \\
(G_m^{\hat{\beta}})_{2q+1,2r+1} &= \frac{2}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \left(\frac{\sin(2\pi(q-r)X_i/\tau)}{4\pi(q-r)} - \frac{\sin(2\pi(q+r)X_i/\tau)}{4\pi(q+r)} \right), \\
(G_m^{\hat{\beta}})_{2q,2r} &= \frac{2}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} \left(\frac{\sin(2\pi(q+r)X_i/\tau)}{4\pi(q+r)} + \frac{\sin(2\pi(q-r)X_i/\tau)}{4\pi(q-r)} \right),
\end{aligned}$$

$$(G_m^{\hat\beta})_{2q+1,2r} = \frac{2}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol\beta}^{\boldsymbol T}\boldsymbol{Z_i}} \left(\frac{\sin^2(\pi(q+r)X_i/\tau)}{2\pi(q+r)} + \frac{\sin^2(\pi(q-r)X_i/\tau)}{2\pi(q-r)} \right),$$

$$(G_m^{\hat\beta})_{2q+1,2q} = \frac{2}{n}\sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol\beta}^{\boldsymbol T}\boldsymbol{Z_i}} \frac{\sin^2(2\pi q X_i/\tau)}{4\pi q},$$

and

$$(\Gamma_m)_1 = \frac{1}{n\sqrt{\tau}}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\le\tau,\delta_i=1\}},$$

$$(\Gamma_m)_{2q} = \frac{1}{n}\sqrt{\frac{2}{\tau}}\sum_{i=1}^{n}\delta_i\cos(2\pi q X_i), \quad (\Gamma_m)_{2q+1} = \frac{1}{n}\sqrt{\frac{2}{\tau}}\sum_{i=1}^{n}\delta_i\sin(2\pi q X_i).$$

From these relations, we get the Gram matrix $\boldsymbol{G_m^{\hat\beta}}$ and the vector $\boldsymbol{\Gamma_m}$, so that Equation (5.10) can be solve by inverting $\boldsymbol{G_m^{\hat\beta}}$. Thus, for each $m \in \{0,...,\lfloor\sqrt{n}/\log(n)+2\rfloor\}$, we obtain a solution $\hat{\boldsymbol{A}}_{\boldsymbol m}^{\hat{\boldsymbol\beta}}$ and the candidates $\hat\alpha_m^{\hat{\boldsymbol\beta}}$ for the estimation of α_0 are implemented by using relation (5.9).

Implementation of the model selection procedure

The estimator of the baseline function α_0 is obtained via the model selection procedure (5.7), where $\mathcal{M}_n = \{1,...,\lfloor\log(n/\log(n))/\log(2)\rfloor\}$ in the histogram case and $\mathcal{M}_n = \{1,...,\lfloor\sqrt{n}/\log(n)+2\rfloor\}$ in the trigonometric case. The contrast is implemented from definition (5.6). Applying the rectangle method (see Appendix A for details on this method), the contrast (5.6) is approximated by

$$C_n(\hat\alpha_m^{\hat{\boldsymbol\beta}},\hat{\boldsymbol\beta}) \approx -\frac{2}{n}\sum_{i=1}^{n}\hat\alpha_m^{\hat{\boldsymbol\beta}}(X_i)\delta_i + \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\hat{\boldsymbol\beta}^{\boldsymbol T}\boldsymbol{Z_i}}\sum_{j:u_j\le X_i}(\hat\alpha_m^{\hat{\boldsymbol\beta}}(u_j))^2\frac{\tau}{N},$$

where $N = 100$, and $(u_j)_{j=1,...,N}$ are grid points evenly distributed across $[0,\tau]$.

The penalty defined in Equation (5.8) involves the unknown quantity $||\alpha_0||_{\infty,\tau}$. This problem occurs occasionally in penalization procedures, see for instance Comte (2001), Lacour (2007) or Comte et al. (2011). The solution is to replace it by an estimator:

$$\widehat{\mathrm{pen}}(m) = \kappa(1+||\hat\alpha_{\max(m)}^{\hat{\boldsymbol\beta}}||_{\infty,\tau})\frac{D_m}{n}, \tag{5.14}$$

where κ is a numerical constant and $\hat\alpha_{\max(m)}^{\hat{\boldsymbol\beta}}$ is an estimator of α_0 computed on the arbitrary larger space $S_{\max(m)}$. The constant κ in the penalty term is a universal numerical constant. From several tests, the constant κ is tuned to $\kappa = 5$. We refer to Appendix B for the details on the procedure to choose κ.

We compute $\hat m^{\hat{\boldsymbol\beta}}$ from (5.7) and the resulting estimator is $\hat\alpha_{\hat m^{\hat{\boldsymbol\beta}}}^{\hat{\boldsymbol\beta}}$, $\hat{\boldsymbol\beta}$ being alternately $\hat{\boldsymbol\beta}^{ML}$, $\hat{\boldsymbol\beta}^{Lasso}$, $\hat{\boldsymbol\beta}^{Lasso+ML}$, $\hat{\boldsymbol\beta}^{AdapL}$.

5.3.2 Kernel estimators and bandwidth selection

Implementation of the kernel estimator $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$

Let recall the definition of the usual kernel estimator of the baseline function, introduced by Ramlau-Hansen (1983b), in the specific case of right censoring: for a kernel function $K : \mathbb{R} \mapsto \mathbb{R}$ with integral 1 and a bandwidth $h > 0$,

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{h}\sum_{i=1}^{n} \frac{\delta_i}{\sum_{k=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_k}} \mathbb{1}_{\{X_k \geq X_i\}}} K\left(\frac{t - X_i}{h}\right). \tag{5.15}$$

Among the classical kernel function used in kernel estimation, we choose to work with the Epanechnikov kernel, defined by

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

Estimator (5.15) has been calculated for $\hat{\boldsymbol{\beta}}$ being alternately $\hat{\boldsymbol{\beta}}^{ML}$, $\hat{\boldsymbol{\beta}}^{Lasso}$, $\hat{\boldsymbol{\beta}}^{Lasso+ML}$, $\hat{\boldsymbol{\beta}}^{AdapL}$.

Implementation of the bandwidth selection

We propose two data-driven procedures to select the bandwidth: the cross-validation and the Goldenshluger and Lepski method.

(i) Implementation of the cross-validation

Cross-validation is a popular and readily implemented heuristic for selecting the bandwidth in kernel estimation. The cross-validation method for the multiplicative intensity model for counting processes has been suggested by Ramlau-Hansen (1981). The cross-validated bandwidth is chosen to minimize an estimate of the Mean Integrated Squared Error (MISE) as a function of h, and defined by

$$\mathrm{MISE}(h) = \mathbb{E}\int_0^\tau (\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t))^2 \mathrm{d}t - 2\mathbb{E}\int_0^\tau \hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t)\alpha_0(t)\mathrm{d}t + \int_0^\tau \alpha_0^2(t)\mathrm{d}t$$

The last term does not depend on h and the first term is estimated according to (5.15). Ramlau-Hansen (1981) showed that an approximatively unbiased estimator of the second term is the "cross-validation" estimate

$$-2\sum_{i \neq j} \frac{1}{h} K\Big(\frac{X_i - X_j}{h}\Big) \frac{\Delta N(X_i)}{\bar{Y}(X_i)} \frac{\Delta N(X_j)}{\bar{Y}(X_j)},$$

where the sum is over i and j such that $i \neq j$ and $0 \leq X_i \leq \tau$ and $\bar{Y} = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \geq t\}}$. In this case, the bandwidth selected by cross-validation is

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

(ii) Implementation of the Goldenshluger and Lepski method

In this paragraph, we describe the implementation of the adaptive Goldenshluger and Lepski method (described in Chapter 4) to select an adaptive bandwidth for the kernel estimator (5.15).

Let us recall few outlines of the Goldenshluger and Lepski method. For $\mathcal{H}_n$ a discrete set of bandwidths specified in the following, we have for $h > 0$

$$A^{\hat{\boldsymbol{\beta}}}(h) = \sup_{h' \in \mathcal{H}_n} \left\{ ||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}||_2^2 - V(h') \right\}_+,$$

where

$$V(h) = \kappa' \frac{||\alpha_0||_{\infty,\tau} ||K||_2^2}{nh}, \tag{5.16}$$

for some numerical constant $\kappa' > 0$ to be specified later, and

$$\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(t) = K_{h'} * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h}(t), \tag{5.17}$$

for any $t \geq 0$ and h, h' two positive real numbers. From these definitions, the selected bandwidth is:

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

and the resulting kernel estimator is denoted $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$.

In order to implement the adaptive bandwidth selection method, we consider a grid of bandwidths $\mathcal{H}_n$ defined by

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

This grid satisfies the theoretical conditions required to establish the non-asymptotic oracle inequality (see Example 4.1 in Chapter 4). The auxiliary estimates $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}$ defined by (5.17) are easily computed by the rectangle method (see Appendix A for details about this approximation method):

$$\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(t) \approx \sum_{j=1}^{N} K_{h'}(t - u_j) \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h}(u_j) \frac{\tau}{N},$$

and the standard $\mathbb{L}^2$-norm in $A^{\hat{\boldsymbol{\beta}}}(h)$ is then approximate by

$$||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}||_2^2 \approx \sum_{kj=1}^{N} (\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}(u_j) - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'}(u_j))^2 \frac{\tau}{N},$$

where $N = 100$, and $(u_j)_j$ are grid points evenly distributed across $[0, \tau]$.

The variance term $V(h)$ defined by (5.16) involves the unknown quantity $||\alpha_0||_{\infty,\tau}$. Following Bouaziz et al. (2013), we propose to replace the true unknown function by an estimator and to consider rather:

$$\hat{V}^{\hat{\boldsymbol{\beta}}}(h) = \kappa' \frac{||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(h)}||_{\infty,\tau}||K||_2^2}{nh}, \tag{5.18}$$

where κ' is a numerical constant and $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(h)}$ an estimator computed for the largest bandwidth h in the grid $\mathcal{H}_n$. The constant κ' is a universal constant that we tuned from different tests detailed in Appendix B.We take $\kappa' = 1$.
Finally, we select $\hat{h}^{\hat{\boldsymbol{\beta}}}$ such that

$$\hat{h}^{\hat{\boldsymbol{\beta}}} = \underset{h\in\mathcal{H}_n}{\arg\min}\{A^{\hat{\boldsymbol{\beta}}}(h) + \hat{V}^{\hat{\boldsymbol{\beta}}}(h)\}. \tag{5.19}$$

We compute $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h^{\hat{\boldsymbol{\beta}}}}$ from (5.15) and (5.19).

5.4 Measuring the quality of the estimators

The performances of the different estimators of the baseline hazard function and of the complete hazard function are evaluated via Mean Integrated Squared errors (MISEs). Let first define the three Integrated Squared Errors (ISEs) considered for some function $\alpha \in \mathbb{L}^2([0,\tau])$ and some parameter $\boldsymbol{\beta} \in \mathbb{R}^p$: the standard ISE, the random ISE more adapted to the Cox model and the total ISE are respectively defined by

$$\begin{aligned}
\text{ISEstand}(\alpha) &= \int_0^\tau (\alpha(t) - \alpha_0(t))^2 \mathrm{d}t, \\
\text{ISErand}(\alpha, \boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n \int_0^{X_i} (\alpha(t) - \alpha_0(t))^2 \mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}} \mathrm{d}t, \\
\text{ISEtotal}(\alpha, \boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau (\alpha(t)\mathrm{e}^{\boldsymbol{\beta}^T \boldsymbol{Z_i}} - \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta}_0^T \boldsymbol{Z_i}})^2 \mathrm{d}t.
\end{aligned} \tag{5.20}$$

Remark 5.1. In Chapter 3, we have introduced a random norm defined by (3.17) in Section 3.2. This explain why we introduce the ISErand which is inspired by the random norm. In addition the ISErand is more adapted to Cox model than the standard ISE. We also introduce the ISEtotal to measure the quadratic error of the estimator $\hat{\lambda}(t, \boldsymbol{Z_i}) = \hat{\alpha}^{\hat{\boldsymbol{\beta}}}(t)\mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}}$ of the whole intensity λ_0 in the Cox model, for $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}$ being either $\alpha^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}}$, $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ or $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$. It is close to the quadratic norm considered in Chapter 2.

All these ISEs are implemented via the rectangle method (see Appendix). With this approximation method, we compute:

$$\begin{aligned}
\text{ISEstand}(\alpha) &\approx \sum_{j=1}^{N}(\alpha(u_j)-\alpha_0(u_j))^2\frac{\tau}{N},\\
\text{ISErand}(\alpha,\boldsymbol{\beta}) &\approx \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\boldsymbol{\beta^T Z_i}}\sum_{j:u_j\le X_i}(\alpha(u_j)-\alpha_0(u_j))^2\frac{\tau}{N},\\
\text{ISEtotal}(\alpha,\boldsymbol{\beta}) &\approx \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N}(\alpha(u_j)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}-\alpha_0(u_j)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}})^2\frac{\tau}{N},
\end{aligned} \tag{5.21}$$

where $N=100$, and $(u_j)_{j=1,\dots,N}$ are grid points evenly distributed across $[0,\tau]$.

The associated Mean Integrated Squared Errors are defined by $\text{MISEg}(\alpha) = \mathbb{E}[\text{ISEg}(\alpha)]$, for g=stand, rand or total, where the expectation is taken on $(T_i, C_i, \boldsymbol{Z_i})$ (for sake of simplicity, we write $\text{MISEg}(\alpha)$ even if the MISE depends on $\boldsymbol{\beta}$). We obtain an estimation of the different MISEs by taking the empirical mean for the N_e replications ($N_e = 100$ or 500 in some cases).

$$\begin{aligned}
\text{MISE}^{g,k}_{emp}(\alpha) &= \frac{1}{N_e}\sum_{j=1}^{N_e}(\text{ISEg}(\alpha_j)),\\
\text{Var}_{emp}(\text{MISE}^{g,k}(\alpha)) &= \frac{1}{N_e-1}\sum_{j=1}^{N_e}(\text{ISEg}(\alpha_j)-\text{ISE}^{g,k}_{emp}(\alpha))^2
\end{aligned}$$

with g=stand, rand or total, $k = ML, Lasso, Lasso+ML$ or $AdapL$, $\alpha = \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}}$, $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$ or $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}}$ and $\alpha_j = (\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}_{CV}})_j$, $(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}})_j$ or $(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}})_j$ corresponds to the estimator obtained at the jth experiment.

We refer to Section 5.5 for the detailed results and the comparison of the MISEs of all estimators.

5.5 Results in the case of simulated data

5.5.1 Results for the regression parameter estimation

We compare the estimation errors $|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta_0}|$ and the prediction errors $|\boldsymbol{Z}\hat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta_0}|$, where $\boldsymbol{Z}=(Z_{i,j})$ is the design matrix for $i=1,\dots,n$ and $j=1,\dots,p$, obtained from each procedures (see Tables 5.1.1 and 5.1.2). We have plotted the first and the third quartiles and the median of the ℓ_1-norm and ℓ_2-norm of errors between the true regression parameter $\boldsymbol{\beta_0}$ and its estimators $\hat{\boldsymbol{\beta}}^{AdapL}$ (see Figure 5.1.1) for different sizes

of n and p. All the regimes are represented : the small dimension for $p = 5$, the intermediate case for $p = \sqrt{n}$ and the high-dimension for $p = n$ or $p = 2n$.

			ML	Lasso	AdapL
$n = 200$	$p = 5$	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_1$	0.586	0.502	0.508
		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_2$	0.115	0.103	0.116
$n = 500$	$p = 5$	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_1$	0.359	0.324	0.306
		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_2$	0.043	0.043	0.042

			ML	Lasso	AdapL
$n = 200$	$p = 15$	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_1$	1.750	0.690	0.818
		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_2$	0.353	0.149	0.216
$n = 500$	$p = 22$	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_1$	1.67	0.556	0.692
		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_2$	0.203	0.072	0.112

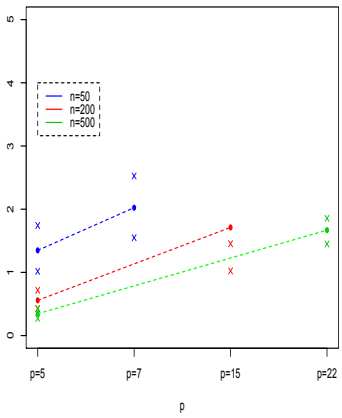
			ML	Lasso	AdapL
$n = 200$	$p = 200$	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_1$	NA	0.984	1.62
		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_2$	NA	0.276	0.541
$n = 500$	$p = 500$	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_1$	NA	0.808	1.67
		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}\rvert_2$	NA	0.161	0.361

TABLE 5.1.1 – *ℓ_1-norm and ℓ_2 norm of estimation errors of the estimators $\hat{\boldsymbol{\beta}}^{ML}$, $\hat{\boldsymbol{\beta}}^{Lasso}$ and $\hat{\boldsymbol{\beta}}^{AdapL}$ for $n = 200$ and $p = 5, 15, 200$, and $n = 500$ and $p = 5, 22, 500$.*

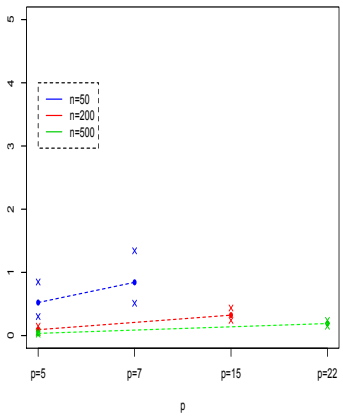
Comments on Table 5.1.1, Table 5.1.2 and Figure 5.1.1.

On Figure 5.1.1, we plot the three quartiles (the first and the third quartiles represented by a cross ($\times$) and the median represented by a bullet ($\bullet$)) of the ℓ_1-norm (on the left) and the ℓ_2-norm (on the right) of the estimation errors for the maximum Cox partial log-likelihood estimator (see 5.1.1a and 5.1.1b), for the Lasso estimator (see 5.1.1c and 5.1.1d) and for the adaptive Lasso estimator (see 5.1.1e and 5.1.1e) of the regression parameter $\boldsymbol{\beta_0}$ for different values of n and p. In blue, we plot the quartiles of the estimation errors of the three estimators for $n = 50$ and $p = 5$, $p = 7$, $p = 100$; the red color corresponds to the quartiles of the estimation errors of the three estimators for $n = 200$ and $p = 5$, $p = 15$, $p = 500$; the green color corresponds to the quartiles of the estimation errors of the three estimators for $n = 500$ and $p = 5$, $p = 22$, $p = 1000$. The dotted lines just help in interpreting the trends in the graphs.

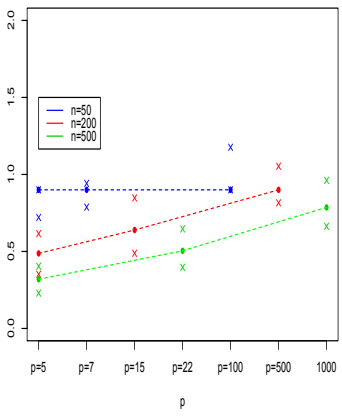
As expected, we observe that the larger the sample size n is, the smaller the ℓ_1-norm and ℓ_2-norm of the estimation and prediction errors of the estimators of $\boldsymbol{\beta_0}$ are; and inversely the higher the number of covariates is, the greater the estimation and prediction errors of the estimators of $\boldsymbol{\beta_0}$ are. The maximum Cox partial log-likelihood estimator $\hat{\boldsymbol{\beta}}^{ML}$ is not defined for $p > \sqrt{n}$. From Tables 5.1.1 and 5.1.2, it seems that



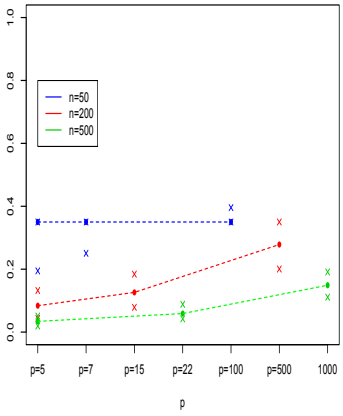
(a) *Quartiles of* $|\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta_0}|_1$



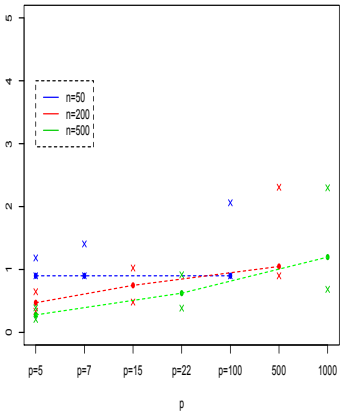
(b) *Quartiles of* $|\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta_0}|_2$



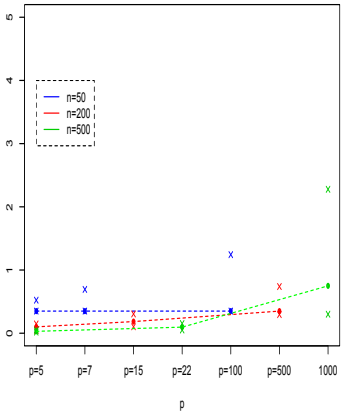
(c) *Quartiles of* $|\hat{\boldsymbol{\beta}}^{Lasso} - \boldsymbol{\beta_0}|_1$



(d) *Quartiles of* $|\hat{\boldsymbol{\beta}}^{Lasso} - \boldsymbol{\beta_0}|_2$



(e) *Quartiles of* $|\hat{\boldsymbol{\beta}}^{AdapL} - \boldsymbol{\beta_0}|_1$



(f) *Quartiles of* $|\hat{\boldsymbol{\beta}}^{AdapL} - \boldsymbol{\beta_0}|_2$

FIGURE 5.1.1 – *Median, first and third quartiles of* $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1$ *(on the left) and for* $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_2$ *(on the right),* $\hat{\boldsymbol{\beta}}$ *being either* $\hat{\boldsymbol{\beta}}^{ML}$*, either* $\hat{\boldsymbol{\beta}}^{Lasso}$ *or* $\hat{\boldsymbol{\beta}}^{AdapL}$*.*

			ML	Lasso	AdapL
$n = 200$	$p = 5$	$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_1$	0.148	0.139	0.146
		$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_2$	0.036	0.033	0.036
$n = 500$	$p = 5$	$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_1$	0.090	0.088	0.087
		$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_2$	0.013	0.013	0.013

			ML	Lasso	AdapL
$n = 200$	$p = 15$	$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_1$	0.259	0.170	0.200
		$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_2$	0.108	0.047	0.069
$n = 500$	$p = 22$	$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_1$	0.200	0.118	0.144
		$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_2$	0.063	0.022	0.036

			ML	Lasso	AdapL
$n = 200$	$p = 200$	$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_1$	NA	0.231	0.315
		$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_2$	NA	0.083	0.181
$n = 500$	$p = 500$	$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_1$	NA	0.170	0.255
		$\lvert \boldsymbol{Z}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta_0}\rvert_2$	NA	0.046	0.118

TABLE 5.1.2 – *ℓ_1-norm and ℓ_2 norm of prediction errors of the estimators $\hat{\boldsymbol{\beta}}^{ML}$, $\hat{\boldsymbol{\beta}}^{Lasso}$ and $\hat{\boldsymbol{\beta}}^{AdapL}$ for $n = 200$ and $p = 5, 15, 200$, and $n = 500$ and $p = 5, 22, 500$.*

the Lasso performs better than the Adaptive Lasso in estimation and in prediction. The reason for this is that the adaptive Lasso estimator is more biased than the Lasso estimator. However, the adaptive Lasso performs better in selection since it sets more coefficients to zero. In order to explain this effect more precisely, we compute specificities (SPEC) and sensitivities (SENS) for each method for $n = 200$ and $p = 200$. For an estimation $\hat{\boldsymbol{\beta}}_m$, the false (FP) and true (TP) positives are defined as

$$\begin{aligned} \mathrm{FP}(\hat{\boldsymbol{\beta}}_m) &= \mathrm{Card}\left(j \in \{1, ..., p\} : \hat{\beta}_{m_j} \neq 0 \text{ and } \beta_{0_j} = 0\right) \\ \mathrm{TP}(\hat{\boldsymbol{\beta}}_m) &= \mathrm{Card}\left(j \in \{1, ..., p\} : \hat{\beta}_{m_j} \neq 0 \text{ and } \beta_{0_j} \neq 0\right) \end{aligned}$$

and false (FN) and true (TN) negatives are defined by exchanging $=$ and $\neq$ in the above definitions. The specificity and sensibility of a method are defined over the $N_e = 50$ replications as

$$\begin{aligned} \mathrm{SPEC} &= \frac{1}{N_e}\sum_{m=1}^{N_e} \frac{\mathrm{TN}(\hat{\boldsymbol{\beta}}_m)}{\mathrm{TN}(\hat{\boldsymbol{\beta}}_m) + \mathrm{FP}(\hat{\boldsymbol{\beta}}_m)} \\ \mathrm{SENS} &= \frac{1}{N_e}\sum_{m=1}^{N_e} \frac{\mathrm{TP}(\hat{\boldsymbol{\beta}}_m)}{\mathrm{TP}(\hat{\boldsymbol{\beta}}_m) + \mathrm{FN}(\hat{\boldsymbol{\beta}}_m)}. \end{aligned}$$

From Table 5.1.3, the adaptive Lasso has the best specificity, but has a lower sensitivity than the Lasso. This is a consequence of the fact that the adaptive Lasso sets more coefficients to zero.

	SPEC	SENS
Lasso	0.97	0.36
AdapL	0.98	0.35

TABLE 5.1.3 – *Specificities (SPEC) and sensitivities (SENS) for the Lasso and the adaptive Lasso (AdapL) for* $n = 200$ *and* $p = 200$.

We would have also expected that the maximum Cox partial likelihood estimator performs better than Lasso estimator when p is small, but it does not. This can been explained from the calculation of the ℓ_1-norm of the estimation error on the support of $\boldsymbol{\beta_0}$, defined by $J(\boldsymbol{\beta_0}) = \{j \in \{1, ..., p\} : \beta_{0_j} \neq 0\}$, and on its complementary $J(\boldsymbol{\beta_0})^c$.

	$\lvert(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})_{J(\boldsymbol{\beta_0})}\rvert_1$	$\lvert(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})_{J(\boldsymbol{\beta_0})^c}\rvert_1$
ML	0.36	1.37
Lasso	0.53	0.21

TABLE 5.1.4 – ℓ_1*-norm of the estimation errors on the support of* $\boldsymbol{\beta_0}$ *and its complementary for the maximum Cox partial likelihood and the Lasso estimators for* $n = 200$ *and* $p = 15$.

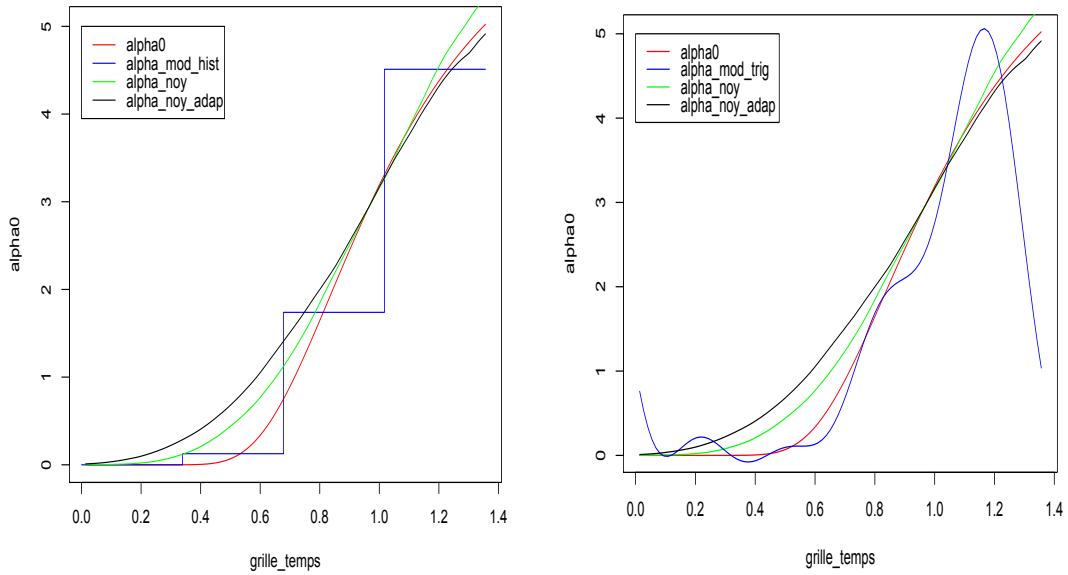
From Table 5.1.4, it follows that the maximum Cox partial likelihood estimator is better than the Lasso estimator on the support of $\boldsymbol{\beta_0}$, but since the maximum Cox partial likelihood estimator sets no coefficient to zero, its estimation error on the complementary of the support of $\boldsymbol{\beta_0}$ is much higher than the one of the Lasso estimator. This explains why the estimation errors in Table 5.1.1 are better for the Lasso estimator than for the maximum Cox partial likelihood estimator.

5.5.2 Results for the baseline function estimation

Comments on Figure 5.2.1.

Figure 5.2.1 represents in the case of a log-normal distribution $\ln \mathcal{N}(1/4, 0)$, the plots of the true baseline function α_0, the kernel estimator with a bandwidth selected by cross-validation, the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method and the penalized contrast estimator in a histogram basis (see Figure 5.2.1a) and in a trigonometric basis (see Figure 5.2.1b).

First of all, Figure 5.2.1 shows that all procedures seems to estimate suitably the true function (in red). The curves of the kernel estimators with a bandwidth selected by cross-validation and by Goldenshluger and Lepski are close to each other. The penalized contrast estimator seems to have comparable performances to those of the two other estimators for small t, but performs less well after $t = 0.8$. This is probably related to the chosen bases and form of the true function. Indeed, the penalized contrast estimator in a histogram basis is a piecewise function whereas the true function to be estimated is continuous and the form of the penalized contrast estimator in a trigonometric basis does not fit exactly the true function α_0 because of the periodicity of the basis. We confirm these first visual impressions by comparing the different MISEs of these estimators.



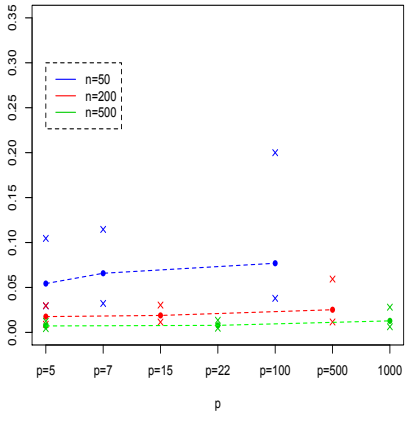
(a) *Histogram basis.* (b) *Trigonometric basis.*

FIGURE 5.2.1 – *Plots of the true baseline function (in red), the cross-validated kernel estimator (in green), the penalized contrast estimator (in blue) and the adaptive kernel estimator (in black). The plots have been obtain for:* $n = 500$, $p = \lfloor \sqrt{n} \rfloor$, $\alpha_0 \sim \ln \mathcal{N}(1/4, 0)$, $d = 4.5$

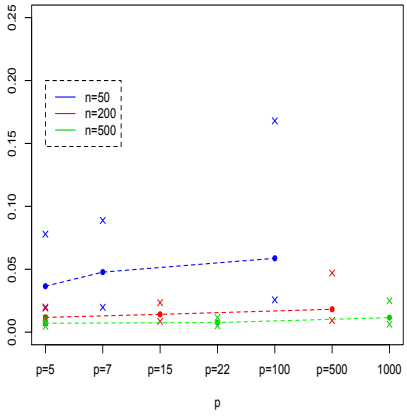
Comments on Figure 5.2.2.

On Figure 5.2.2, we plot the three quartiles (the first and the third quartiles represented by a cross ($\times$) and the median represented by a bullet ($\bullet$)) of the empirical random MISEs of the kernel estimators with a bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method, and of the penalized contrast estimator in a histogram basis and in a trigonometric basis of the true baseline function in the case of a Weibull distribution $\mathcal{W}(1.5, 1)$ for different values of n and p. In blue, we plot the quartiles of the MISErands of the four estimators for $n = 50$ and $p = 5$, $p = 7$ and $p = 100$; the red color corresponds to the quartiles of the MISErands of the four estimators for $n = 200$ and $p = 5$, $p = 15$ and $p = 500$; the green color corresponds to the quartiles of the MISErands of the four estimators for $n = 500$ and $p = 5$, $p = 22$ and $p = 1000$. The dotted lines just help in interpreting the graphs.

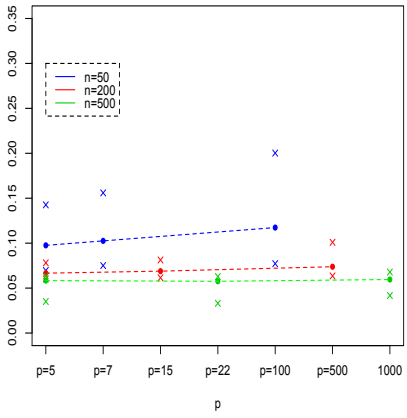
As expected, the random empirical MISEs are smaller, when the sample size increases, and are degraded when the number of covariates p increases. In additions, it seems that the two kernel estimators perform better than the penalized contrast estimators. The kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method seems to have the smallest MISEs for all n and p, and the penalized contrast estimator in a trigonometric basis the worst.



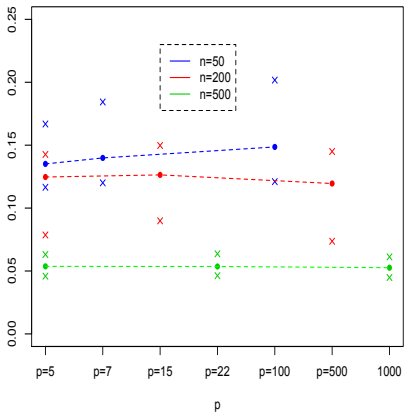
(a) MISErand $\left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}^{AdapL}}_{\hat{h}^{\hat{\boldsymbol{\beta}}^{AdapL}}_{CV}}\right)$.



(b) MISErand $\left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}^{AdapL}}_{\hat{h}^{\hat{\boldsymbol{\beta}}^{AdapL}}}\right)$.



(c) MISErand $\left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}^{AdapL}}_{\hat{m}^{\hat{\boldsymbol{\beta}}^{AdapL}}}\right)$ *in the case of histograms.*



(d) MISErand $\left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}^{AdapL}}_{\hat{m}^{\hat{\boldsymbol{\beta}}^{AdapL}}}\right)$ *in the trigonometric case.*

FIGURE 5.2.2 – *Median, first and third quartiles of the random* MISE*s of the kernel estimator with a bandwidth selected by cross-validation (5.2.2a), of the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method (5.2.2b), of the penalized contrast estimator in the case of histograms (5.2.2c) and of the penalized contrast estimator in the trigonometric case (5.2.2d), with the adaptive Lasso estimator of the regression parameter in the case of a Weibull distribution* $\mathcal{W}(1.5, 1)$.

Comments on Table 5.2.1.

Table 5.2.1 gives the three empirical MISEs of the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method and of the penalized contrast estimator in a histogram basis for an adaptive Lasso estimator of the regression parameter and survival times that are distributed from $\mathcal{W}(1.5, 1)$, in different censoring situation. We consider the results for two rates of censoring: a usual rate of 20% of censoring and large rate of 50% of censoring.

For a fixed censure, the better results (i.e the smallest values) are obtained for the random MISE and the worst results for the total MISE, which seems consistent, since the total MISE measures the performances of the complete intensity estimators, whereas the two other MISEs measures the performances of the estimators of the baseline function.

Intuitively, one would expect that the MISEs would be degraded when the censoring rate increases. This is not what we observe on Table 5.2.1. The reason for this is that the time interval of the study $[0, \tau]$ is smaller for a large rate of censoring. Indeed, the time τ is the quantile to 90% of the observed times $(X_i)_{i=1,\dots,n}$. We calculate its value on one experiment, we obtain $\tau_1 = 1.48$ for 20% of censoring and $\tau_2 = 1.19$ for 50% of censoring. More precisely, if we take the example of the standard ISE, we have

$$\begin{aligned}
\mathrm{ISE}_{20\%}(\alpha) &= \int_0^{\tau_1} (\alpha(t) - \alpha_0(t))^2 \mathrm{d}t \quad \text{for 20\% of censoring,} \\
\mathrm{ISE}_{50\%}(\alpha) &= \int_0^{\tau_2} (\alpha(t) - \alpha_0(t))^2 \mathrm{d}t \quad \text{for 50\% of censoring.}
\end{aligned}$$

Thus, the ISE for 20% of censoring is obtained from the one of 50% by adding

$$\int_{\tau_2}^{\tau_1} (\alpha(t) - \alpha_0(t))^2 \mathrm{d}t.$$

This explains why the MISEs obtained for a higher rate of censoring seem better. In order to compare the MISEs in different censoring situations, we could have calculate the MISEs on the same time interval.

In most cases, the results are better for the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method than for the penalized contrast estimator in a histogram basis, unless for the total empirical MISEs in a high-dimensional setting ($n = 200$ and $p = 500$), i.e. when we consider the estimator of the complete intensity $\hat{\lambda}(t, \boldsymbol{Z}) = \hat{\alpha}^{\hat{\boldsymbol{\beta}}}(t) \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{Z}}$.

Dimensions \ MISEs		20%			50%		
		MISEstand	MISErand	MISEtot	MISEstand	MISErand	MISEtot
$n = 200$	$p = 15$	0.069	0.018	0.442	0.05	0.01	0.45
	$p = 500$	0.44	0.07	7.35	0.33	0.053	9.01
$n = 500$	$p = 22$	0.03	0.009	0.20	0.022	0.006	0.16
	$p = 1000$	0.13	0.02	1.38	0.02	0.004	0.10

(a) MISE*s of the kernel estimator of the baseline function with a bandwidth selected by the Goldenshluger and Lepski method and an adaptive Lasso estimator of the regression parameter.*

Dimensions \ MISEs		20%			50%		
		MISEstand	MISErand	MISEtot	MISEstand	MISErand	MISEtot
$n = 200$	$p = 15$	0.16	0.07	0.49	0.10	0.04	0.45
	$p = 500$	0.36	0.11	5.48	0.16	0.05	3.88
$n = 500$	$p = 22$	0.10	0.05	0.28	0.07	0.03	0.21
	$p = 1000$	0.17	0.06	1.27	0.04	0.03	0.15

(b) MISE*s of the penalized contrast estimator in a histogram basis of the baseline function with an adaptive Lasso estimator of the regression parameter.*

TABLE 5.2.1 – *Random empirical MISE of the kernel estimator with a bandwidth selected by the Goldenshluger et Lepski method (5.2.1a) and of the penalized contrast estimator in a histogram basis (5.2.1b) obtained from an adaptive Lasso estimator of the regression parameter given two rate of censoring: 20% and 50% of censoring.*

Comments on Table 5.2.2.

Table 5.2.2 allows to compare the random MISE of the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method and of the penalized contrast estimator in a histogram basis, with an adaptive Lasso estimator of the regression parameter, for different distributions of the survival times.

In both tables, the results are the worst for a Weibull distribution with parameters $a = 3$ and $\lambda = 4$. This can easily be explained from Figure 1.0.1: the baseline hazard function associated to a $\mathcal{W}(3, 4)$ has the most complicated form to be estimated because it increases steeply. Otherwise, the previous comments on the influences of n and p on the MISEs are still the same: the results are better for a large n and a small p. Lastly, the MISEs are better in the case of the kernel estimator with a bandwidth selected by cross-validation than in the case of the penalized contrast estimator in a histogram basis for the Weibull distribution $\mathcal{W}(1.5, 1)$, but the penalized contrast estimator in a histogram basis performs better for the Weibull distributions $\mathcal{W}(0.5, 2)$ and $\mathcal{W}(3, 4)$, for which the true baseline function is harder to estimate.

Dimensions \ Distributions		$\mathcal{W}(1.5, 1)$	$\mathcal{W}(0.5, 2)$	$\mathcal{W}(3, 4)$
$n = 200$	$p = 15$	0.018	0.83	7.38
	$p = 200$	0.04	0.79	10.41
$n = 500$	$p = 22$	0.009	0.56	7.21
	$p = 500$	0.02	0.69	10.23

(a) *MISEs for the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method with an adaptive Lasso estimator of the regression parameter.*

Dimensions \ Distributions		$\mathcal{W}(1.5, 1)$	$\mathcal{W}(0.5, 2)$	$\mathcal{W}(3, 4)$
$n = 200$	$p = 15$	0.07	0.65	5.40
	$p = 200$	0.09	0.66	6.51
$n = 500$	$p = 22$	0.05	0.41	5.32
	$p = 500$	0.06	0.42	6.15

(b) *MISEs for the penalized contrast estimator in a histogram basis, with an adaptive Lasso estimator of the regression parameter.*

TABLE 5.2.2 – *Random empirical MISE for the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method (5.2.2a) and for the penalized contrast estimator in a histogram basis (5.2.2b), with an adaptive Lasso estimator of the regression parameter, for three different Weibull distributions of the survival times.*

Comments on Table 5.2.3.

Lastly, we compare the random MISEs of the four implemented estimators of the baseline function, in the case of an adaptive Lasso estimator of the regression parameter and for a Weibull distribution $\mathcal{W}(1.5, 1)$.

As observed on Figure 5.2.1, the best estimator of the baseline function is the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method. The kernel estimator with a bandwidth selected by cross-validation performs a little bit less, then comes the penalized contrast estimator in a histogram basis and lastly the penalized contrast estimator in a trigonometric basis gives the poorer MISEs.

Estimators \ Dimensions	$n = 200$		$n = 500$	
	$p = 15$	$p = 200$	$p = 22$	$p = 500$
KernelCV	0.023	0.045	0.010	0.023
KernelGL	0.017	0.044	0.009	0.022
MShist	0.073	0.089	0.051	0.059
MStrig	0.178	0.110	0.059	0.06

TABLE 5.2.3 – *Random empirical MISE of the kernel estimators with a bandwidth selected by cross-validation (KernelCV) and by the Goldenshluger and Lepski method (KernelGL) and of the penalized contrast estimators in a histogram basis (MShist) and in a trigonometric basis (MStrigo) for an adaptive Lasso estimator of the regression parameter and a Weibull distributions* $\mathcal{W}(1.5, 1)$ *for the survival times.*

Conclusion

This study reveals known classical results: the estimations of both parameters are degraded for large number of covariates p or when the sample size is too small. We have not been able to highlight the impact of the rate of censoring on the estimation of the baseline function, but we have provided an explanation, which leads us to believe that the estimation of the baseline function performs less for a large rate of censoring. We observe also that the random MISEs allows to obtain the better results for all estimators and that the form of the true baseline function influences its estimation. Lastly, from a practical point of view, it seems that for a very regular true baseline function (in the case of a Weibull distribution $\mathcal{W}(1.5, 1)$ for example) the best procedure to estimate the baseline hazard function is the kernel estimation with a bandwidth selected with the Goldenshluger and Lepski method. In the case of baseline functions with brutal changes in their derivatives, we have seen that the contrast penalized estimator in a histogram basis performs better than the kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method.

5.6 Application to a real dataset on breast cancer

In this section, we apply the proposed method to study the relapse free survival (RFS) from breast cancer adjusted on high-dimensional covariates in two groups of patients. We consider a Cox model (2.105) to link the RFS to the covariates. We aim at answering the two questions of the introduction concerning the biomarkers that influence the RFS and the prediction of the RFS for each individual. The dataset is available on the website `www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532`.

The dataset consists of 414 patients in the cohort GSE6532 collected by Loi et al. (2007) for the purpose of characterizing Estrogen Receptor (ER)-positive subtypes with gene expression profiles. Estrogen receptors are a group of proteins found inside cells, which is activate by the hormone estrogen. There are different forms of the estrogen receptors, referred to as subtypes of estrogen receptors. When they are over expressed, they are referred to as ER-positive. The dataset has been studied from a survival analysis point of view in Tian et al. (2012). Following them, we apply our procedures to the same survival time of interest (the RFS). Excluding patients with incomplete informations, as it is done by Loi et al. (2007), there are 142 patients receiving Tamoxifen and 104 untreated patients. It should be underlined that we should do better to handle the missing data, but in this study we also exclude the patients with missing data. In addition to clinical informations such as the age or the size of the tumor (the complete clinical informations are listed in Table (C.0.1) in Appendix C), we have 44 928 gene expression measurements for each of the 246 patients. Two different survival times are available in this study: the time of relapse free survival and the time of distant metastasis free survival. We are interested in this study in the time of relapse free survival, which subjects to right censoring due to incomplete follow-up. There are 60% of censorship in the group of the untreated patients and 66% in the group of patients receiving Tamoxifen. Our goal is to compare the baseline functions in the two groups of patients: the patients receiving Tamoxifen and the untreated patients.

We start by a preliminary variable selection among the 44928 levels of gene expression. This corresponds to a screening step (see Fan et al. (2010)). This preliminary variable selection is based on the score statistics of each Cox model considered for each variable separately. We only keep the variables which score statistics are superior to a threshold. The difference from the procedure proposed by Fan et al. (2010) is that we fix the number of covariates we want to keep and then we tune a threshold to select this number of covariates. We define four different thresholds. The first one is defined as the 95^{th} percentile of a Chi-squared distribution with 1 degree of freedom, so that 996 probesets have been selected and with the clinical covariates we have $p = 1000$. We

then take 3 other thresholds such that we can work alternatively with $p = 6$, $p = 10$ and $p = 100$ covariates.

We refer to Appendix C for a detailed description of the data contained in R workspace.

5.6.1 Variable selection

Comments on Table 6.1.1.

Table 6.1.1 recapitulates the variables that have been selected by the one step Lasso, the one step Lasso associated to a maximum Cox partial log-likelihood procedure and the adaptive Lasso associated to a maximum Cox partial log-likelihood procedure procedure for different values of p in the two groups of patients.

p / Group of patients	$p = 6$	$p = 10$	$p = 100$	$p = 1000$
Untreated	size	size	size+2 g.e. size+3g.e.	size+15 g.e. adap: size+18 g.e.
Tamoxifen	grade	grade	10 g.e. adap: 9 g.e.	24 g.e. adap: 38 g.e.

TABLE 6.1.1 – *Selected variables in the two groups of patients. We precise the name of the clinical selected variables and only give the number of selected genes (e.g.=gene expression).*

In the group of untreated patients, the covariate "size", which corresponds to the tumor size in centimeter, is the only clinical variable selected for all considered p. When $p > \sqrt{n}$ the procedures select also some genes that explain the time of relapse free survival: 2 probesets when $p = 100$ and 15 probesets when p=1000 (18, for the adaptive Lasso procedure).

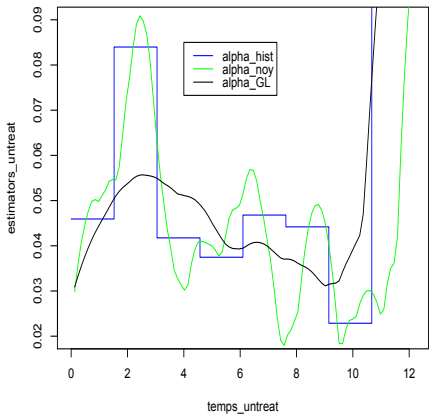
Concerning the patients receiving Tamoxifen, when $p \leq \sqrt{n}$, the clinical variable "grade", which corresponds to histologic grade (Elston-Ellis) category, is the only variable that seems explain the time of relapse free survival and when $p > n$ some probesets are also selected.

We have considered two different models for each treatment group, we now investigate if one model could be fitted for all the cohort with the treatment as a covariate.

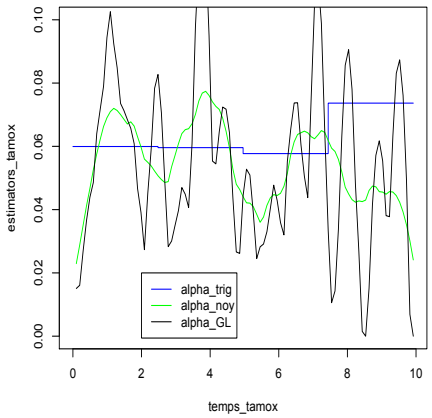
5.6.2 Estimation of the baseline hazard function

Comments on Figures 6.2.1 and 6.2.2.

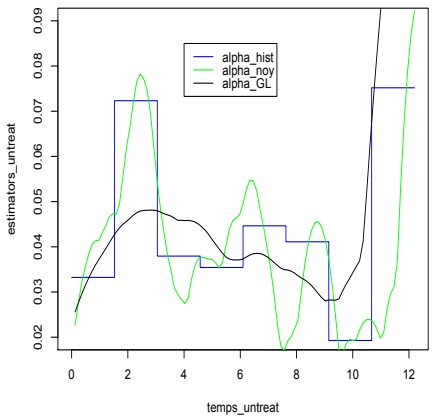
Figures 6.2.1 and 6.2.2 show the graphs of the estimators of the baseline function via the kernel estimator with a bandwidth selected by cross-validation, the kernel estimator with a bandwidth select with the Goldenshluger and Lepski method and the penalized contrast estimator in histogram and trigonometric bases respectively, in the two groups of patients for $p = 10, 100, 1000$.



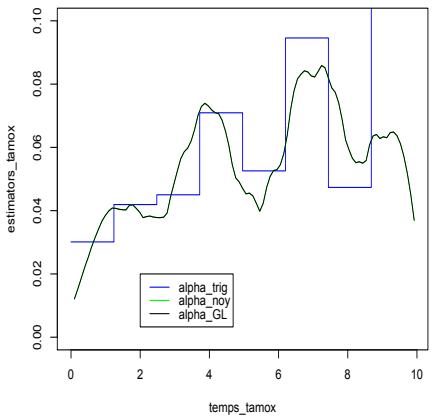
(a) *Untreated patients (p=10).*



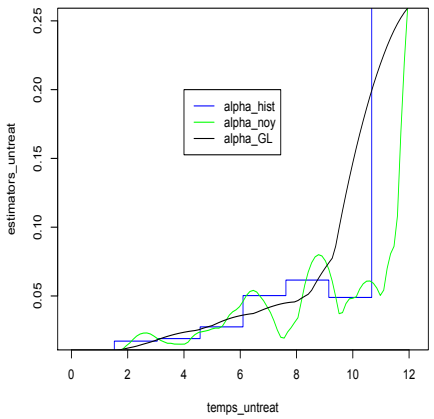
(b) *Tamoxifen patients (p=10).*



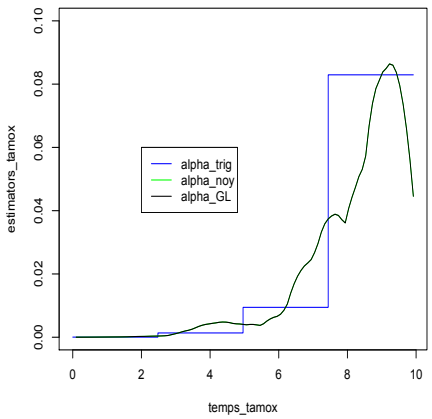
(c) *Untreated patients (p=100).*



(d) *Tamoxifen patients (p=100).*

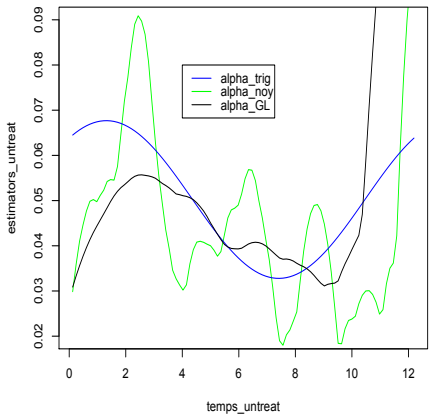


(e) *Untreated patients (p=1000).*

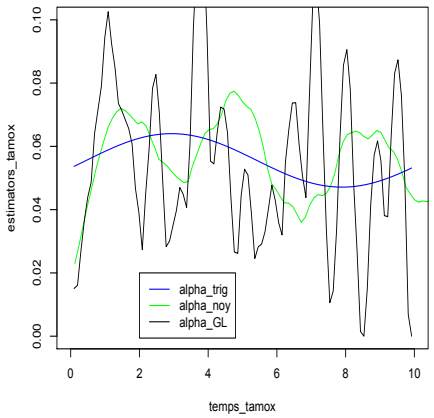


(f) *Tamoxifen patients (p=1000).*

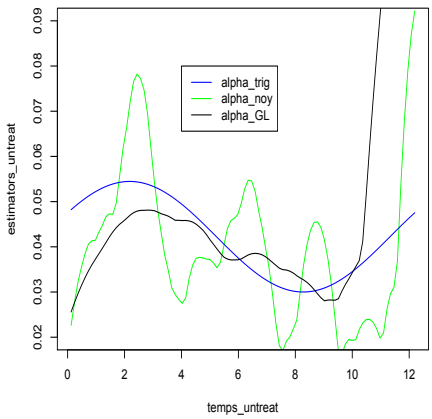
FIGURE 6.2.1 – *Kernel estimators with a bandwidth selected either by cross-validation (in green) or by the Goldenshluger and Lepski method (in black) and by model selection estimator in the histogram basis (in blue). The first column is associated to the group of untreated patients and the second column corresponds to the group of Tamoxifen patients for* $p = 10$ *(first line),* $p = 100$ *(second line) and* $p = 1000$ *(third line).*



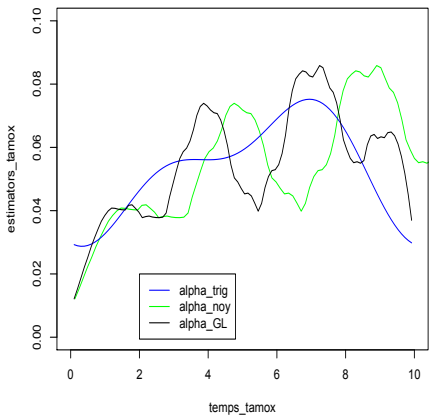
(a) *Untreated patients (p=10).*



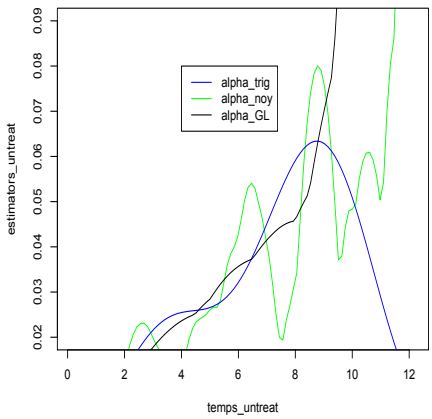
(b) *Tamoxifen patients (p=10).*



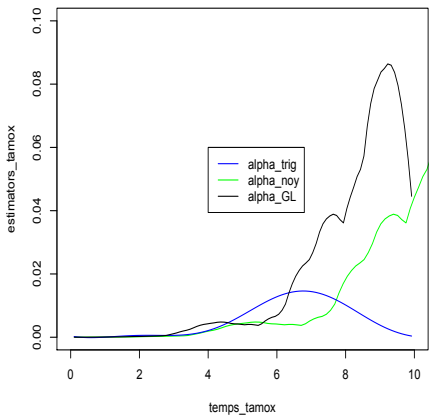
(c) *Untreated patients (p=100).*



(d) *Tamoxifen patients (p=100).*



(e) *Untreated patients (p=1000).*



(f) *Tamoxifen patients (p=1000).*

FIGURE 6.2.2 – *Kernel estimator with a bandwidth selected either by cross-validation (in green) or by the Goldenshluger and Lepski method (in black) and model selection estimator in the trigonometric basis (in blue). The first column is associated to the group of untreated patients and the second column corresponds to the group of Tamoxifen patients for* $p = 10$ *(first line),* $p = 100$ *(second line) and* $p = 1000$ *(third line).*

A first remark is that on Figures 6.2.1d and 6.2.1f, the plots of the kernel estimators with a bandwidth selected either by cross-validation or by the Goldenshluger and Lepski method are overlaid. This means that the two estimators give the same estimations of the baseline hazard ratio. Then, we observe that the more the time goes by, the more the baseline function is high for the untreated patient compared with those receiving tamoxifen. Indeed at time 10 the estimators of the baseline function for the untreated patients are exploding whereas those of the Tamoxifen patient are bounded and tend to decrease. This leads us to believe that the treatment has a positive influence on the survival time. Then, for $p = 10$ and $p = 100$, the baseline function in the two groups of patients are not proportional and thus, we can not consider a Cox model with a mix of the two groups of patients. But, for $p = 1000$, it seems that the baseline hazard functions are proportional. In this case, we could therefore consider a Cox model by grouping all the patients and adding the treatment to the model as a covariate to study the effect of the treatment on the survival time. As the sets of selected variables in the two groups of patients are different (in particular the selected probesets are not the same), a more interesting work would be to consider an interaction model to investigate potential crossed effects of the treatment and some genes.

Appendix

A The rectangle method

The rectangle method computes an approximation of a definite integral of a function f, made by covering the area under the graph of f with a collection of rectangles whose heights are determined by the values of the function. Specifically, the interval (a, b) over which the function f is to be integrated is divided into N equal subintervals of length $h = (b - a)/N$. The rectangles are then drawn so that either their left or right corners, or the middle of their top line lies on the graph of the function, with bases running along the x-axis. The approximation to the integral is then calculated by adding up the areas of the N rectangles, giving the formula:

$$\int_a^b f(x)\mathrm{d}x \approx h \sum_{n=1}^{N} f(x_n)$$

where $h = (b - a)/N$ and $x_n = a + nh$.

The formula for x_n above gives x_n for the top-right corner approximation.

As N gets larger, this approximation gets more accurate. In fact, this computation is the spirit of the definition of the Riemann integral and the limit of this approximation as n $\to \infty$ is defined and equal to the integral of f on (a,b) if this Riemann integral is defined. Note that this is true regardless of which top line is used, however the midpoint approximation tends to be more accurate for finite n.

B Calibration of the constants

B.1 In model selection

The penalty term in model selection is defined by (5.14), where κ is a universal constant. To determine this constant, we consider a grid of constants $\kappa = (\kappa_1, ..., \kappa_K)$. For $k = 1, ..., K$, we determine $\hat{m}^{\hat{\boldsymbol{\beta}}}(\kappa_k) = \arg\min_m \{C_n(\hat{\alpha}_m^{\hat{\boldsymbol{\beta}}}, \hat{\boldsymbol{\beta}}) + \widehat{\mathrm{pen}}(m, \kappa_k)\}$, where

$$\widehat{\mathrm{pen}}(m, \kappa_k) = \kappa_k(1 + ||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(m)}||_{\infty,\tau})\frac{D_m}{n}, \quad \text{for } k = 1, ..., K,$$

and we obtain K estimators $(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}(\kappa_k)})_{k=1,...,K}$. Then, we compute, for each estimator, the random Integrated Squared Error (ISErand) defined and approximated via the

rectangle method by

$$\begin{aligned}\mathrm{ISErand}(\kappa_k) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}(\kappa_k)}(t)-\alpha_0(t)\right)^2 \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}} Y_i(t)\mathrm{d}t, \quad \text{for } k=1,..,K.\\ &\approx \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}\sum_{j=1}^{N}\left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}(\kappa_k)}(u_j)-\alpha_0(u_j)\right)^2 \mathbb{1}_{\{X_i\geq u_j\}}\frac{\tau}{N},\end{aligned}$$

where $N = 100$, and $(u_j)_j$ are grid points evenly distributed across $[0, \tau]$. Lastly, we conduct a new Monte Carlo study with $N_e = 500$ replications and approximate the empirical random Mean Integrated Squared Errors (MISErand) for each κ in the grid: for $k = 1, ..., K$,

$$\mathrm{MISErand}^{emp}(\kappa_k) = \frac{1}{N_e}\sum_{j=1}^{N_e}(\mathrm{ISErand}(\kappa_k))_j. \tag{5.22}$$

We finally choose the constant κ that minimizes the empirical $\mathrm{MISErand}^{emp}$ (5.22). Since the constant κ is universal, we should obtain the same κ for different distributions of the survival time. In order to observe this, we iterate all the procedure for three Weibull distributions with different parameters ($\mathcal{W}(0.5, 1)$, $\mathcal{W}(1, 1)$ and $\mathcal{W}(3, 1)$). We consider the grid $\kappa = (0.1, 0.5, 1, 2, 5, 8, 10)$ in the case of histograms and $\kappa = (0.1, 0.5, 1, 2, 5, 8, 10, 15, 20)$ in the trigonometric case.

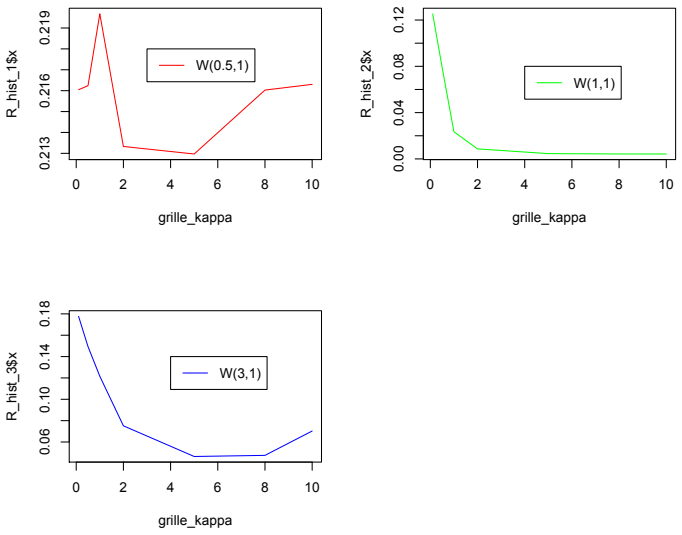


FIGURE B.1.1 – *Plots of the empirical random* MISE*s of* $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}(\kappa_k)}$ *as functions of* κ *in the case of histograms.*

Figures B.1.1 represents graphically the empirical random MISEs of $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}^{\hat{\boldsymbol{\beta}}}(\kappa_k)}$ as functions of κ in the case of histograms (we proceed similarly in the case of a trigonometric basis). On Figure B.1.1, we observe clearly the U-form in the cases of $\mathcal{W}(0.5, 1)$

and $\mathcal{W}(3,1)$ distributions, it is less visible for the $\mathcal{W}(1,1)$ distribution. In this case, we could have considered a largest grid $\mathcal{H}_n$ with greatest values, but from the constraint of the two other cases, we choose for the three distributions the fifth value, which seems to be the more suitable intersection. This corresponds to the choice of $\kappa = 5$.

B.2 In Goldenshluger and Lepski method

The variance term $V(h)$ in the Goldenshluger and Lepski method is defined by (5.18), where κ' is a universal constant. The principle to determine this constant is close to the one in model selection to determine the universal constant in the penalty term, so that we explain the procedure more briefly in this case.

We consider a grid of constants $\kappa' = (10^{-4}, 0.1, 1, 10, 1000)$. We calculate for each constant κ'_k, $k = 1, ..., K'$,

$$\hat{h}^{\hat{\boldsymbol{\beta}}}(\kappa'_k) = \underset{h\in\mathcal{M}_n}{\arg\min}\{A^{\hat{\boldsymbol{\beta}}}(h) + \hat{V}^{\hat{\boldsymbol{\beta}}}(h,\kappa'_k)\}, \quad \text{where } \hat{V}^{\hat{\boldsymbol{\beta}}}(h,\kappa'_k) = \kappa'_k \frac{||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\max(h)}||_{\infty,\tau}||K||_2^2}{nh}.$$

Then, we compute the random Integrated Squared Error (ISErand) defined for $k = 1, ..., K'$ by

$$\begin{aligned}\text{ISErand}(\kappa'_k) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}(\kappa'_k)}(t) - \alpha_0(t)\right)^2 \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}} Y_i(t)\mathrm{d}(t) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{Z_i}}\sum_{j=1}^{N}\left(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}(\kappa'_k)}(u_j) - \alpha_0(u_j)\right)^2 \mathbb{1}_{\{X_i\geq u_j\}}\frac{\tau}{N},\end{aligned}$$

where $N = 100$, and $(u_j)_j$ are grid points evenly distributed across $[0,\tau]$ and the approximation comes from the rectangle method (see Appendix A). Lastly, we conduct a new Monte Carlo study with $N_e = 500$ replications: for $k = 1, ..., K'$,

$$\text{MISErand}^{emp}(\kappa'_k) = \frac{1}{N_e}\sum_{j=1}^{N_e}(\text{MISErand}(\kappa'_k))_j. \tag{5.23}$$

We finally choose the constant κ' that minimizes (5.23) for three Weibull distributions with different parameters ($\mathcal{W}(0.5,1)$, $\mathcal{W}(1,1)$ and $\mathcal{W}(2,1)$). This study leads to choose $\kappa' = 1$.

C Description of the real data

Demographics file containing the clinical information for the untreated patients (137) and the tamoxifen treated patients (277).

The clinical information are listed in Table (C.0.1).

Notations	Definitions
samplename	unique id for the patient
id	original id for the patients (id given by the institution, useful to look for intersection between multiple datasets)
series	arbitrary name to specify the institution
age	age in years
grade	histologic grade (Elston-Ellis) category
size	tumor size in cm
er	estrogen receptor status
pgr	progesterone receptor status
node	nodal status
t.rfs	time of relapse free survival
e.rfs	event of relaps free survival
t.dmfs	time of distant metastasis free survival
e.dmfs	event of distant metastasis free survival
treatment	type of treatment (none/tamoxifen)

TABLE C.0.1 – *Table of the clinical information*

R workspace containing the following objects:

data.untreated: matrix of gene expressions with the untreated patients in rows (137) and the affy probesets in columns (44928). The probesets are all the probsets from the chip hgu133a and the chip hgu133b (with some probesets being renamed with "_2" because of same name than in hgu133a chip). RMA normalization was performed separately for each population ("series" in demo) and all the probesets were median centered.

demo.untreated: matrix with clinical information concerning the untreated patients (137).

annot.untreated: annotations of all the probesets (44928).

data.tam: matrix of gene expressions with the tamoxifen treated patients in rows (277) and the affy probesets in columns (44928). The probesets are all the probsets from the chip hgu133a and the chip hgu133b (with some probesets being renamed with "_2" because of same name than in hgu133a chip). The chip hgu133plus2 was used for "GUYT" series and only the common probesets with hgu133a/b were kept for further analysis. RMA normalization was performed separately for each population ("series" in demo) and all the probesets were median centered.

demo.tam: matrix with clinical information concerning the tamoxifen treated patients (277).

annot.tam: annotations of all the probesets (44928).

Conclusion

Les travaux présentés dans cette thèse portent sur l'estimation de l'intensité d'un processus de comptage dans un modèle à intensité multiplicative d'Aalen en présence d'un grand nombre de covariables. Nous avons envisagé plusieurs approches pour différents modèles. Dans la Partie I, nous avons considéré l'intensité comme fonction non-paramétrique du temps et des covariables et nous l'avons estimée par une fonction satisfaisant un modèle de Cox non-paramétrique. Nous avons proposé une procédure Lasso spécifique à la grande dimension pour estimer simultanément les deux paramètres du modèle de Cox. Dans la Partie II, nous avons supposé que l'intensité vérifiait un modèle de Cox. Pour estimer les deux paramètres du modèle, nous avons alors opté pour des procédures en deux étapes, permettant d'estimer le paramètre de régression associé aux covariables à l'aide d'une procédure Lasso, spécifique à la grande dimension, et de proposer des procédures d'estimation du risque de base, fonction du temps, par des procédures usuellement utilisées en estimation non-paramétrique, mais non spécifiques à la grande dimension. Nous avons établi des inégalités oracles non-asymptotiques pour l'estimateur de l'intensité complète (Partie I) et pour les estimateurs du risque de base obtenus par sélection de modèles et par l'estimation à noyaux avec un choix de fenêtre par la méthode de Goldenshluger et Lepski (Partie II). Ces résultats sont les premiers résultats non-asymptotiques obtenus dans un cadre de grande dimension pour les covariables. Nous avons de plus montré que les procédures pour estimer le risque de base à la Partie II sont pertinentes en pratique et sont au moins aussi performantes que les procédures existantes. Ces travaux, considérés dans le cadre général des processus de comptage, ont été motivés par la question pratique de la prédiction de la durée de survie d'un individu pour une maladie donnée. En cela, les procédures proposées, ainsi que les résultats théoriques et pratiques permettent de répondre à cette question. Il reste tout de même différentes pistes de recherche à envisager afin de poursuivre l'étude que nous avons menée et de l'étendre à d'autres types de problèmes.

Perspectives de recherche

Dans les deux parties de cette thèse, nous avons établi des inégalités oracles pour les estimateurs obtenus. Les vitesses de convergence obtenues pour ces estimateurs impliquent la taille de l'échantillon, comme c'est classiquement le cas, mais aussi un

terme de l'ordre du logarithme du nombre de covariables, spécifique au cadre de la grande dimension étudié. Nous ne pouvons pour l'instant nous permettre d'affirmer l'optimalité de cette vitesse au sens minimax. Une perspective immédiate consiste donc à étudier la borne inférieure pour le risque minimax associé sur des boules d'espaces de Besov.

Dans la Partie I, nous avons considéré le critère de la vraisemblance empirique totale de manière à estimer simultanément les deux paramètres du modèle de Cox non-paramétrique. Ce critère n'est pas le critère usuellement considéré dans le modèle de Cox, puisque la procédure habituelle consiste d'abord à estimer le paramètre de régression à l'aide de la vraisemblance partielle de Cox et ensuite à considérer un estimateur à noyau du risque de base. À notre connaissance, les algorithmes existants à ce jour pour estimer le paramètre de régression du modèle de Cox en grande dimension, tels que les algorithmes *glmnet* ou *penalized* dans R, sont programmés à partir de la vraisemblance partielle de Cox. Cependant, pour une fonction

$$\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z_i}) = \exp\Big(\sum_{k=1}^{N} \gamma_k \theta_k(t) + \boldsymbol{Z_i}\boldsymbol{\beta}\Big) \tag{5.24}$$

avec les notations de la Partie I, la log-vraisemblance empirique définie par

$$C_n(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \log(\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z_i}))\mathrm{d}N_i(t) - \int_0^\tau \lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z_i})Y_i(t)\mathrm{d}t\right\},$$

est de la forme de celle associée à la régression poissonnienne. En effet, si $\nu_1, ..., \nu_n$ sont n variables aléatoires de loi de Poisson de paramètre λ, alors la log-vraisemblance associée à cet échantillon est définie par

$$-\frac{1}{n}\log \mathcal{L}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}(\nu_i \log\lambda - \lambda).$$

Or la régression poissonnienne est implémentée dans les packages *glmnet* (pour des sauts positifs) et *penalized* de R. On pourrait donc utiliser ces algorithmes basés sur une descente de coordonnées cycliques pour implémenter la procédure d'estimation de la Partie I. L'algorithme SPAMS implémenté sous MATLAB est un algorithme de type FISTA permettant aussi de considérer une procédure Lasso pour une régression poissonnienne. Cependant, le gradient de la vraisemblance n'étant pas Lipschitz, on ne connaît pas de résultat théorique garantissant la convergence de ces trois algorithmes. Cela donne néanmoins une possibilité de mise en pratique de notre procédure et dans un futur proche, nous nous attacherons à l'implémenter à l'aide d'un de ces algorithmes. Cela permettra d'une part de vérifier les performances pratiques de l'estimateur Lasso de l'intensité que nous proposons, et d'autre part d'introduire les poids que nous obtenons et de comparer les procédures avec et sans les poids. De plus,

une fois la procédure de la Partie I implémentée, nous pourrons effectuer une étude comparative des procédures proposées dans les Parties I et II de cette thèse.

Ensuite, une perspective naturelle, sera de s'intéresser à une des généralisations classique et très utilisée en pratique qui consiste à considérer des covariables et un coefficient de régression qui dépendent du temps. Ainsi, nous avons par exemple récemment amorcé un travail, en collaboration avec Mokhtar Alaya et Agathe Guilloux, dans lequel nous considérons un modèle de Cox avec un paramètre de régression qui dépend du temps : pour tout $t \geq 0$,

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

où $\boldsymbol{Z} = (Z_1, ..., Z_p)^T$ est un vecteur de covariables p-dimensionnel et $\boldsymbol{\beta_0}(t) = (\beta_{0_1}(t), ..., \beta_{0_p}(t))^T$ est un vecteur de fonctions du temps p-dimensionnel. L'idée est alors d'approximer chaque coordonnée du paramètre de régression $\boldsymbol{\beta_0(t)}$ par des histogrammes. Plus précisément, nous considérons des fonctions candidates de la forme de (5.24), en supposant que les fonctions θ_k sont des histogrammes, i.e. $\theta_k = \mathbb{1}_{I_k}$ pour $k = 1, ..., N$, où I_k est un intervalle, et en approximant $\boldsymbol{\beta_0}(t)$ par une combinaison linéaire de ces mêmes histogrammes. Nous obtenons donc des fonctions candidates de la forme

$$\lambda_{\boldsymbol{\beta},\boldsymbol{\gamma}}(t, \boldsymbol{Z}) = \exp\left(\sum_{k=1}^{N} \gamma_k \mathbb{1}_{I_k}(t) + \sum_{j=1}^{p} \sum_{k=1}^{N} \beta_{j,k} \mathbb{1}_{I_k}(t) Z_j \right).$$

Nous étudierons alors les aspects algorithmiques et théoriques des procédures qui estiment des histogrammes pénalisés par leurs variations totales.

Dans la partie II, nous avons mené une étude sur une base de données sur le cancer du sein. Nous avons ainsi montré que pour 1000 covariables les courbes du risque de base semblent avoir la même forme dans les deux groupes de patients considérés. Le modèle de Cox étant un modèle à risque proportionnel, si les risques de base estimés dans les deux groupes de patients ont la même forme, nous pouvons regrouper les individus et considérer un nouveau modèle de Cox en ajoutant la variable traitement. De plus, les ensembles des variables sélectionnées dans les deux groupes de patients sont différents. Il se peut donc que certaines variables sélectionnées aient un effet en intéraction avec le traitement. Il serait donc intéressant dans notre étude d'étudier un modèle de Cox en considérant la variable traitement comme une variable explicative et en ajoutant des termes d'interaction pour mettre en évidence l'effet combiné du traitement et de certaines variables, lesquelles sont alors appelées chimio-sensibilisantes. Ce type d'étude mettant en évidence la relation d'interaction entre des facteurs génétiques et le traitement fait partie des grandes questions statistiques en analyse de survie pour

la médecine personnalisée (voir Collura et al. (2014) pour un exemple de mise en évidence d'une interaction entre une anomalie au niveau du génome et le traitement). De plus, des travaux récents portent sur le Lasso dans des modèles à interaction (voir Bien et al. (2013) par exemple). Il serait donc intéressant d'approfondir notre étude en considérant ces modèles.

A

Appendices

Sommaire

A.1 Un peu de théorie sur les processus de comptage 220
A.2 Construction de la vraisemblance de Cox 224
A.3 Quelques inégalités de concentration 227
A.3.1 Inégalité de Bernstein standard 227
A.3.2 Inégalité de Bernstein pour les martingales 227
A.3.3 Inégalité de Talagrand 227

A.1 Un peu de théorie sur les processus de comptage [1] :

Processus stochastiques et notions de base

Tous les phénomènes étudiés dans ce travail se déroulent en temps continu. Fixons, pour la suite, un intervalle $\mathcal{T} = [0, \tau[$ de $\mathbb{R}_+$ qui indexera le temps. Soient $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé, où $\mathcal{F}$ une σ-algèbre et $\mathbb{P}$ une mesure de probabilité définie sur $\mathcal{F}$. On appelle *filtration* une famille $(\mathcal{F}_t)_{t\in\mathcal{T}}$ croissante et continue à droite (càd) de sous-tribus, c'est-à-dire telle que :

(i) $\mathcal{F}_s \subseteq \mathcal{F}_t$ pour tout $s \leq t$ dans $\mathcal{T}$ (croissante)

(ii) $\mathcal{F}_s = \bigcap_{t>s} \mathcal{F}_t$ pour tout $s \in \mathcal{T}$ (càd).

Un processus stochastique W est alors une famille de variables aléatoires (v.a.) $\{W(t), t \in \mathcal{T}\}$ indexées par le temps. On dira que W est *adapté* à la filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ si pour tout t dans $\mathcal{T}$, la v.a. $W(t)$ est $\mathcal{F}_t$-mesurable. Pour un ω dans Ω, on appellera trajectoire de W la fonction $t \mapsto W(t, \omega)$. Le processus W est alors dit *càdlàg* si $\mathbb{P}$-presque toutes ses trajectoires appartiennent à l'espace $\mathbb{D}(\mathcal{T})$ des fonctions càdlàg (continues à droite, limites à gauche) sur $\mathcal{T}$, appelé espace de Skorohod dans la théorie de la convergence faible (voir Billingsley (1968)). Il est dit *prévisible* si, en tant que fonction de $(t, \omega) \in \mathcal{T} \times \Omega$, il est mesurable par rapport à la tribu sur $\mathcal{T} \times \Omega$ engendrée par les processus càg et adaptés.

À tout processus W càdlàg, on peut, grâce au résultat de Courrège & Priouret (1965), associer une filtration « historique » $(\mathcal{F}_t)_{t\in\mathcal{T}}$ telle que, pour tout t dans T, la sous-tribu $\mathcal{F}_t = \sigma\{W(s), s \leq t \in \mathcal{T}\}$ est engendrée par le passé à t du processus. Cette famille peut être complétée en ajoutant les négligeables à la tribu $\mathcal{F}_0$ sans qu'elle perde sa propriété de filtration. Par construction, le processus W est adapté à $(\mathcal{F}_t)_{t\in\mathcal{T}}$.

On appellera *temps d'arrêt* une v.a. T à valeurs dans $\bar{\mathcal{T}} = [0, \tau]$ telle que les événements $\{T \leq t\}$ sont $\mathcal{F}_t$-mesurables pour tout t dans $\mathcal{T}$. Une suite *localisante* de temps d'arrêt est une suite (T_n) croissante de temps d'arrêt telle que :

$$\mathbb{P}(T_n > t) \underset{n\to\infty}{\to} 1,$$

pour tout t dans $\mathcal{T}$. On dira qu'un processus vérifie *localement* une propriété s'il existe une suite localisante (T_n) telle que le processus $I(\{T_n > 0\})X^T$ la vérifie, où X^T est le processus arrêté à T tel que

$$X^T(.) = X(. \wedge T).$$

Martingales à temps continu :

Une martingale M par rapport à la filtration $(\mathcal{F}_t)$ est un processus stochastique continu à droite avec limite à gauche tel que :

1. Nous nous sommes inspirés de la thèse de Guilloux (2004) pour écrire cette annexe

(i) M est $\mathcal{F}_t$-adapté

(ii) $\mathbb{E}|M(t)| < \infty$ for all $s \leq t$,

(iii) $\mathbb{E}(M(t)|\mathcal{F}_s) = M(s)$ for all $s \leq t$

Une martingale M est dite de *carré intégrable* si :

$$\sup_{t\in\mathcal{T}} \mathbb{E}\left(M(t)^2\right) < \infty.$$

Un processus M est une *martingale locale* s'il existe une suite localisante (T_n) de temps d'arrêt telle que, pour tout n, le processus $I(\{T_n > 0\})M^{T_n}$ soit une martingale. Une martingale locale est un processus càdlàg et adapté. Si, de plus, la suite (T_n) peut être choisie de telle sorte que $I(\{T_n > 0\})M^{T_n}$ soit de carré intégrable, alors M est une *martingale locale de carré intégrable.*

Nous donnons maintenant une version du célèbre Théorème de décomposition de Doob-Meyer.

Définition A.1 (Décomposition de Doob-Meyer)**.** *Pour un processus W càdlàg et adapté, on définit, s'il existe, son* compensateur *comme l'unique processus $\widetilde{W}$ prévisible, càdlàg, à variation finie et tel que le processus $W - \widetilde{W}$ soit une martingale locale nulle en zéro.*

En particulier, si M est une martingale locale de carré intégrable, il existe un unique processus prévisible, càdlàg, à variation finie et nul en zéro, noté $\langle M, M\rangle$ ou $\langle M\rangle$, tel que $M^2 - \langle M, M\rangle$ soit une martingale locale. On appelle $\langle M\rangle$ la *variation prévisible* de M. Si la variation prévisible est une fonction déterministe, on parlera alors de *fonction de variance* de M. On introduit aussi le *processus de variation optionnel* de M, noté $[M, M]$ ou $[M]$ et défini pour $t \in \mathcal{T}$, par $[M]_t = \sum_{s\leq t}(\Delta M(s))^2$, avec $\Delta M = M(t) - M(t^-)$.

Théorème A.1. *Soient M une martingale locale de carré intégrable et à variation finie et H un processus prévisible localement borné. L'intégrale stochastique $\int H dM$ définie pour tout t dans $\mathcal{T}$ et presque tout w dans Ω par :*

$$\int_0^t H(w, s)dM(w, s)$$

est alors une martingale locale de carré intégrable et à variation finie. On a de plus :

$$\left\langle \int H dM \right\rangle = \int H^2 d\langle M\rangle \quad \textit{and} \quad \left[\int H \mathrm{d}M\right] = \int H^2 \mathrm{d}[M].$$

Processus de comptage :

En notant $(\Omega, \mathcal{F}, \mathbb{P})$ l'espace probabilisé filtré par $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathcal{T}}$ sur lequel on travaille, on dira que $N = (N_1, ..., N_n)$ est un processus de comptage s'il est adapté et tel que :

(i) les applications $t \mapsto N_i(t)$, pour $i = 1, ..., n$, sont càd, étagées, nulles en zéro et admettent un nombre dénombrable de sauts de longueur $+1$;

(ii) deux processus N_i et N_j, pour $i \neq j$, ne peuvent sauter en même temps.

On suppose de plus que, pour tout $i = 1, ..., n$, la v.a. $N_i(t)$ est *p.s.* finie pour tout t positif. Pour le processus N, les instants de saut $T_k = \inf\{t : \sum_{i=1}^{n} N_i(t) \geq k\}$ sont des $(\mathcal{F}_t)$-temps d'arrêt. On pose $T_k = \tau$ si $k > \sum_{i=1}^{n} N_i(\tau)$. Les composantes N_i, pour $i = 1, ..., n$, sont des processus adaptés, càdlàg, croissants et localement bornés $(0 \leq N_i^{T_k}(t) \leq n)$. Ils admettent des compensateurs notés Λ_i croissants tels que, pour tout $i = 1, ..., n$, le processus :

$$M_i = N_i - \Lambda_i$$

soit une martingale locale de carré intégrable dont le processus de variation prévisible est donné par $\langle M_i \rangle = \Lambda_i - \int \Delta\Lambda_i d\Lambda_i$. De plus, on a $\langle M_i, M_{i'} \rangle = -\int \Delta\Lambda_i d\Lambda_{i'}$ pour $i \neq i'$. On a pour tout $t \in \mathcal{T}$ et $i = 1, ..., n$, $\mathbb{E}[N_i(t)] = \mathbb{E}[\Lambda_i(t)]$. De plus, dans le cas continu, il existe, pour $i = 1, ..., n$, des processus prévisibles f_i, appelés *intensités*, tels que

$$\Lambda_i(t) = \int_0^t f_i(s)\mathrm{d}s.$$

On dit que les processus N_i sont des processus de comptage à intensité.

Par définition, le processus de variation optionnel de $M_i = N_i - \Lambda_i$ est défini par $[M_i] = N_i$ et par unicité du compensateur, le processus de variation prévisible est défini par $\langle M_i \rangle = \Lambda_i$.

Cas général : les processus de comptage marqué

Les processus de comptage marqués sont une généralisation des processus de comptage. L'idée est la suivante : au lieu de ne considérer que les instants T_k auquel se produit un évènement spécifique comme c'est le cas pour les processus de comptage, nous observons aussi une variable supplémentaire Z_k à chaque instant T_k. Plus précisément, fixons un espace mesurable $(E, \mathcal{E})$ appelé espace marqué, et supposont que

(i) $(Z_k, k \geq 1)$ est une suite de variables aléatoires dans E,

(ii) la suite $(T_k, k \geq 1)$ constitue un processus de comptage

$$N(t) = \sum_k \mathbb{1}_{\{T_k \leq t\}}.$$

La suite de couple (T_k, Z_k) est appelée processus ponctuel marqué, où les Z_k sont les marques. Pour tout $A \in \mathcal{E}$, nous pouvons définir le processus de comptage

$$N_t(A) = \sum_k \mathbb{1}_{\{T_k \leq t\}} \mathbb{1}_{\{Z_k \in A\}},$$

qui compte le nombre de sauts avant t dont les marques sont dans A.

Une inégalité classique : l'inégalité de Bürkholder

L'inégalité de Bürkholder permet de relier un martingle à son processus optionel. Nous renvoyons à Liptser & Shiryayev (1989) p.75, pour la démonstration de ce résultat.

Théorème A.2 (Inégalité de Bürkholder)**.** *Si M est une $(\mathcal{F}_t)$-martingale, alors pour tout $b > 1$, il existe des constantes universelles γ_b et κ_b (indépendentes of M) telles que pour tout $t \geq 0$*

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

où $[M]_t$ est la variation quadratique de M_t. Les constantes γ_b et κ_b peuvent être définie par

$$\gamma_b = [18b^{3/2}/(b-1)]^{-1}, \quad \kappa_b = 18b^{3/2}/(b-1)^{1/2}.$$

A.2 Construction de la vraisemblance de Cox

Soient $N_1, ..., N_n$ des processus de comptage i.i.d., vérifiant le modèle à intensité multiplicative d'Aalen (1.3) et pour $i = 1, ..., n$, la fonction de risque $\lambda_0(., \boldsymbol{Z_i})$ est définie selon le modèle de Cox (1.1). On observe $\mathcal{O}_n = (\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$ l'ensemble des observations. Pour tout i et tout $t \in [0, \tau]$, d'après la décomposition de Doob-Meyer, nous avons

$$\mathrm{d}N_i(t) = \alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t + \mathrm{d}M_i(t),$$

où M_i est une martingale locale de carré intégrable. Pour $i = 1, ..., n$, comme les N_i, sont des processus de comptage indépendants, le processus agrégé $\bar{N} = \sum_{i=1}^n N_i$ est encore un processus de comptage qui vérifie

$$\mathrm{d}\bar{N}(t) = \alpha_0(t)\frac{1}{n}\sum_{i=1}^{n}\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t + \mathrm{d}\bar{M}(t),$$

avec $\bar{M} = \sum_{i=1}^n M_i$. Pour tout $\boldsymbol{\beta} \in \mathbb{R}^p$, notons

$$nS_n(t, \boldsymbol{\beta}) = \sum_{i=1}^{n}\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t),$$

et définissons $\gamma_0(t) = \alpha_0(t)S_n(t, \boldsymbol{\beta_0})$. Avec ces notations nous avons

$$\mathrm{d}\bar{N}(t) = \gamma_0(t)\mathrm{d}t + \mathrm{d}\bar{M}(t).$$

D'après la formule de Jacod (voir Andersen et al. (1993)), la *log-vraisemblance pour les processus de comptage* basée sur les observations $\bar{\mathcal{O}} = (\bar{N}(t), \bar{Y}(t), 0 \leq t \leq \tau)$ vérifie :

$$\mathcal{L}(\gamma; \bar{\mathcal{O}}) = \int_0^\tau \log(\gamma(t))\mathrm{d}\bar{N}(t) - \int_0^\tau \gamma(t)\mathrm{d}t \tag{A.1}$$

L'estimateur du maximum de vraisemblance de γ_0 est alors défini formellement par

$$\hat{\gamma} = \arg\max_{\gamma}\left\{\int_0^\tau \log(\gamma(t))\mathrm{d}\bar{N}(t) - \int_0^\tau \gamma(t)\mathrm{d}t\right\}.$$

D'après (A.1), la log-vraisemblance pour estimer les deux paramètres du modèle de Cox vérifie

$$\mathcal{L}(\alpha, \boldsymbol{\beta}; \mathcal{O}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau \log(\alpha(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}})\mathrm{d}N_i(t) - \int_0^\tau \alpha(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}Y_i(t)\mathrm{d}t\right\},$$

que l'on peut décomposer en

$$\begin{aligned}\mathcal{L}(\alpha,\boldsymbol{\beta};\mathcal{O}_n) =& \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\alpha(t)S_n(t,\boldsymbol{\beta})\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}N_i(t) - \int_0^\tau \alpha(t)S_n(t,\boldsymbol{\beta})\mathrm{d}t \\ =& \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}N_i(t) \qquad (A.2) \\ &+ \int_0^\tau \log(\alpha(t)S_n(t,\boldsymbol{\beta}))\mathrm{d}\bar{N}(t) - \int_0^\tau \alpha(t)S_n(t,\boldsymbol{\beta})\mathrm{d}t. \qquad (A.3)\end{aligned}$$

Le terme (A.3) correspond à la vraisemblance (A.1) pour $\gamma(t) = \alpha(t)S_n(t,\boldsymbol{\beta})$. On définit alors le terme (A.2) comme la vraisemblance partielle de Cox :

$$\begin{aligned}l_n^*(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}N_i(t) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0}^T\boldsymbol{Z_i}}Y_i(t)\mathrm{d}t + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}M_i(t).\end{aligned}$$

où le deuxième terme est centré.

Nous pouvons construire un estimateur de type Breslow. Toujours, d'après la décomposition de Doob-Meyer, nous avons

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathrm{d}N_i(t)}{S_n(t,\boldsymbol{\beta})} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \alpha_0(t)\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}Y_i(t)}{S_n(t,\boldsymbol{\beta})}\mathrm{d}t + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathrm{d}M_i(t)}{S_n(t,\boldsymbol{\beta})}.$$

Le deuxième terme de droite en $\mathrm{d}M_i$ est centré peut être interprété comme un bruit aléatoire. En $\boldsymbol{\beta_0}$, on peut écrire l'approximation suivante

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathrm{d}N_i(t)}{S_n(t,\boldsymbol{\beta_0})} \approx \int_0^\tau \alpha_0(t)\mathrm{d}t,$$

on dit que $\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}N_i(t)/S_n(t,\boldsymbol{\beta_0})$ « estime » $\alpha_0(t)\mathrm{d}t$. Dans le cas particulier de la censure aléatoire à droite, cela revient à dire que $\frac{1}{n}\delta_i/S_n(X_i,\boldsymbol{\beta_0})$ « estime » $\alpha_0(X_i)$. On retrouve une justification de l'estimateur de Breslow et de Ramlau-Hansen (1983b).

Dans le cas de la censure aléatoire à droite, la vraisemblance partielle de Cox se réécrit

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}N_i(t) = \frac{1}{n}\sum_{i=1}^{n}\delta_i \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(X_i,\boldsymbol{\beta})}\right),$$

avec $(X_i,\delta_i)_{i\in\{1,\dots,n\}}$ i.i.d. Si on identifie $\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}/S_n(t,\boldsymbol{\beta})$ à un estimateur de l'intensité de N_i définie par $\mathrm{e}^{\boldsymbol{\beta_0}^T\boldsymbol{Z_i}}\alpha_0(t)$, on devrait donc avoir

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}N_i(t) - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{\mathrm{e}^{\boldsymbol{\beta}^T\boldsymbol{Z_i}}}{S_n(t,\boldsymbol{\beta})}Y_i(t)\mathrm{d}t,$$

où le deuxième terme est nul. On est bien en train d'estimer $\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}$ par maximum de vraisemblance. En effet, la fonction de perte naturelle associée à la vraisemblance partielle de Cox est définie par :

$$\ell(\boldsymbol{\beta}, \boldsymbol{\beta_0}) = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta^T Z_i}}}{S_n(t,\boldsymbol{\beta})}\right)\mathrm{d}N_i(t) - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}}{S_n(t,\boldsymbol{\beta_0})}\right)\mathrm{d}N_i(t)\right]$$

qui se décompose par Doob-Meyer en

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta^T Z_i}}}{S_n(t,\boldsymbol{\beta})}\frac{S_n(t,\boldsymbol{\beta_0}}{\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}}\right)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta^T Z_i}}}{S_n(t,\boldsymbol{\beta})}\frac{S_n(t,\boldsymbol{\beta_0}}{\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}}\right)\mathrm{d}M_i(t)\right]
\end{aligned}$$

où le deuxième terme est nul. La fonction de perte s'écrit donc

$$\ell(\boldsymbol{\beta}, \boldsymbol{\beta_0}) = -\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\int_0^\tau \log\left(\frac{\mathrm{e}^{\boldsymbol{\beta^T Z_i}}}{S_n(t,\boldsymbol{\beta})}\frac{S_n(t,\boldsymbol{\beta_0}}{\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}}\right)\alpha_0(t)\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}Y_i(t)\mathrm{d}t\right].$$

On retrouve la fonction de perte en kullback introduite par Castellan et Letué (voir Letué (2000)).

A.3 Quelques inégalités de concentration

A.3.1 Inégalité de Bernstein standard

Nous rappelons l'inégalité de Bernstein standard (voir Proposition 2.9 dans le livre de Massart (2007)).

Théorème A.3. *Soient $\zeta_1, ..., \zeta_n$, n variables aléatoires réelles indépendantes. Supposons qu'il existe des constantes positives v et c telles que pour tout entier $m \geq 2$*

$$\sum_{i=1}^{n} \mathbb{E}[|\zeta_i|^m] \leq m! v c^{m-2}. \tag{A.4}$$

Pour tout réel positif x, nous avons

$$\mathbb{P}(\sum_{i}^{n} (\zeta_i - \mathbb{E}(\zeta_i)) \geq x) \leq \exp\Big(- \frac{x^2}{2(v + cx)}. \Big) \tag{A.5}$$

Lorsque les variables ζ_i sont bornées, $|\zeta_i| \leq b$ pour tout $i \in \{1, ..., n\}$, alors l'hypothèse (A.4) est satisfaite avec

$$v = \sum_{i=1}^{n} \mathbb{E}[\zeta_i^2] \quad \text{et} \quad c = b/3.$$

A.3.2 Inégalité de Bernstein pour les martingales

Rappelons l'inégalité de Bernstein pour les martingales (voir van de Geer (1995), Lemme 2.1).

Théorème A.4. *Soit $\{M_t\}_{t \geq 0}$ une martingale localement de carré intégrable par rapport à la filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Définissons le processus prévisible de $\{M_t\}$ par $V_t = \langle M, M \rangle_t$, $t \geq 0$, et ses sauts par $\Delta M_t = M_t - M_{t^-}$. Supposons que $|\Delta M_t| \leq K$ pour tout $t > 0$ et un certain K tel que $0 \leq K < \infty$. Alors, pour tout $a > 0$, $b > 0$,*

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

A.3.3 Inégalité de Talagrand

L'inégalité de Talagrand est une inégalité de concentration qui permet de contrôler le supremum d'un processus empirique sur une classe de fonctions. Plusieurs formes de cette inégalité existent. Nous donnons ici celle qui nous a servi dans cette thèse. On peut trouver une preuve de ce théorème chez Comte et al. (2008) (Lemme 6.1).

Théorème A.5. *Soit $\mathcal{F}$ une classe dénombrable de fonctions uniformément bornées. Soient $\xi_1, ..., \xi_n$ des variables aléatoires indépendantes, et pour $f \in \mathcal{F}$, on définit*

$$\nu_{n,\xi}(f) = \frac{1}{n}\sum_{i=1}^{n}\{f(\xi_i) - \mathbb{E}[f(\xi_i)]\}.$$

Alors, pour $\varepsilon > 0$, on a

$$\mathbb{E}\left[\left\{\sup_{f\in\mathcal{F}}\nu_{n,\xi}^2(f) - 2(1+2\varepsilon^2)H^2\right\}_+\right] \leq \frac{4}{d}\left(\frac{W}{n}\mathrm{e}^{-d\varepsilon^2\frac{nH^2}{W}} + \frac{98M^2}{dn^2\varphi^2(\varepsilon)}\mathrm{e}^{-\frac{2d\varphi(\varepsilon)\varepsilon}{\tau\sqrt{2}}\frac{nH}{M}}\right),$$

avec $\varphi(\varepsilon) = \sqrt{1+\varepsilon^2} - 1$, $d = 1/6$ et

$$\sup_{f\in\mathcal{F}}||f||_\infty \leq M, \quad \mathbb{E}\Big[\sup_{f\in\mathcal{F}}|\nu_{n,\xi}(f)|\Big] \leq H, \quad \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[f(\xi)] \leq W.$$

Bibliographie

O. Aalen. A model for nonparametric regression analysis of counting processes. Dans *Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisła, 1978)*, volume 2 de *Lecture Notes in Statist.*, pages 1–25. Springer, New York, 1980. (Cité pages 5, 54 et 107.)

O.O Aalen. *Statistical inference for a family of counting processes*. PhD thesis, University of California, Berkeley, 1975. (Cité page 7.)

O.O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978. (Cité page 7.)

H. Akaike. Information theory and an extension of the maximum likelihood principle. Dans *In Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973. (Cité pages 9, 18 et 115.)

B. Altshuler. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences*, 6:1–11, 1970. (Cité page 7.)

P. K. Andersen, Ø. Borgan, R. D. Gill, & Niels Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0. (Cité pages 5, 29, 33, 54, 56, 93, 107, 109, 117, 153 et 224.)

P. K. Andersen & R. D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):pp. 1100–1120, 1982. ISSN 00905364. URL `http://www.jstor.org/stable/2240714`. (Cité pages 6, 7 et 184.)

A. Antoniadis, P. Fryzlewicz, & F. Letué. The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):pp. 531–552, 2010. ISSN 1467-9469. URL `http://dx.doi.org/10.1111/j.1467-9469.2009.00685.x`. (Cité pages 29, 55 et 59.)

F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:pp. 384–414, 2010. (Cité pages 56, 61, 84 et 86.)

Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117(4):467–493, 2000. (Cité page 18.)

Y. Baraud. Model selection for regression on a random design. *ESAIM: Probability and Statistics*, 6:127–146, 2002. (Cité pages 18 et 22.)

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. (Cité page 128.)

A. Barron, L. Birgé, & P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999. (Cité pages 18, 22, 115 et 121.)

P. L. Bartlett, S. Mendelson, & J. Neeman. l1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields*, 154(1-2):193–224, 2012. (Cité page 54.)

K. Bertin, C. Lacour, & V. Rivoirard. Adaptive estimation of conditional density function. *arXiv preprint arXiv:1312.7402*, 2013. (Cité pages 26 et 27.)

K. Bertin, E. Le Pennec, & V. Rivoirard. Adaptive Dantzig density estimation. *Annales de l'IHP, Probabilités et Statistiques*, 47(1):pp. 43–74, 2011. (Cité pages 11, 39, 61 et 66.)

P. J. Bickel, B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, & A. van der Vaart. Regularization in statistics. *Test*, 15(2):271–344, 2006. (Cité page 9.)

P. J. Bickel, Y. Ritov, & 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`. (Cité pages 11, 14, 15, 17, 18, 29, 54, 55, 56, 59, 61, 62, 97, 98, 99, 100 et 102.)

J. Bien, J. Taylor, & R. Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013. (Cité page 218.)

P. Billingsley. Convergence of probability measures. *John Wiley, New York*, 228:229, 1968. (Cité page 220.)

L. Birgé. Model selection for density estimation with l2-loss. *arXiv preprint arXiv:0808.1416*, 2008. (Cité page 18.)

L. Birgé & P. Massart. *From model selection to adaptive estimation*. Springer, 1997. (Cité pages 18, 21 et 115.)

L. Birgé & P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375, 1998. (Cité pages 18, 22 et 173.)

L. Birgé & P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001. (Cité page 18.)

O. Bouaziz, F. Comte, & A. Guilloux. Nonparametric estimation of the intensity function of a recurrent event process. *Statistica Sinica*, 23(2):635–665, 2013. (Cité pages 31, 149, 153, 154, 157, 158, 172 et 193.)

A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984. (Cité page 25.)

J. Bradic, Fan, J., & J. Jiang. Regularization for Cox's proportional hazards model with NP-dimensionality. *The Annals of Statistics*, 39(6):pp. 3092–3120, 2012. (Cité pages 29, 55 et 109.)

J. Bradic & R. Song. Gaussian Oracle Inequalities for Structured Selection in Non-Parametric Cox Model. *arXiv preprint arXiv:1207.4510*, 2012. (Cité pages 29, 55, 68, 92, 109, 117 et 151.)

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995. (Cité page 9.)

E. Brunel & F. Comte. Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā: The Indian Journal of Statistics*, pages 441–475, 2005. (Cité pages 22, 30, 115 et 121.)

E Brunel, F Comte, & A. Guilloux. Nonparametric density estimation in presence of bias and censoring. *test*, 18(1):166–194, 2009. (Cité pages 30 et 115.)

E. Brunel, F. Comte, & C. Lacour. Minimax estimation of the conditional cumulative distribution function. *Sankhya A*, 72(2):293–330, 2010. (Cité pages 30, 115, 137, 138 et 189.)

P. Bühlmann & 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`. (Cité pages 14, 17, 62 et 144.)

F. Bunea. Consistent selection via the lasso for high dimensional approximating regression models. Dans *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 122–137. Institute of Mathematical Statistics, 2008. (Cité page 15.)

F. Bunea, A. Tsybakov, & M. Wegkamp. Aggregation for regression learning. *arXiv preprint math/0410214*, 2004. (Cité page 11.)

F. Bunea, A. B. Tsybakov, & 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. (Cité pages 11, 15 et 54.)

F. Bunea, A. B. Tsybakov, & M. H. Wegkamp. Aggregation and sparsity via l1 penalized least squares. Dans *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. (Cité pages 11 et 54.)

F. Bunea, A.B. Tsybakov, & M.H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007b. (Cité pages 11, 13 et 15.)

F. Bunea, A.B. Tsybakov, & M.H. Wegkamp. Sparse density estimation with l1 penalties. Dans *Learning theory*, pages 530–543. Springer, 2007c. (Cité page 11.)

F. Bunea, A.B. Tsybakov, M.H. Wegkamp, & A. Barbu. Spades and mixture models. *The Annals of Statistics*, 38(4):pp. 2525–2558, 2010. (Cité page 61.)

G. Chagny. *Estimation adaptative avec des données transformées ou incomplètes. Application à des modèles de survie.* PhD thesis, Université René Descartes-Paris V, 2013. (Cité pages 29, 42 et 154.)

G. Chagny. Adaptive warped kernel estimators. *accepted for publication in Scandinavian Journal of statistics*, 2014. (Cité pages 26, 149, 154, 157 et 174.)

A. Collura, A. Lagrange, M. Svrcek, L. Marisa, O. Buhard, A. Guilloux, K. Wanherdrick, C. Dorard, A. Taieb, & A. Saget. Patients with colorectal tumors with microsatellite instability and large deletions in hsp110 t17 have improved response to 5-fluorouracil–based chemotherapy. *Gastroenterology*, 146(2):401–411, 2014. (Cité page 218.)

F. Comte. Adaptive estimation of the spectrum of a stationary gaussian sequence. *Bernoulli*, 7(2):267–298, 2001. (Cité page 190.)

F. Comte, J. Dedecker, & M.L. Taupin. Adaptive density deconvolution with dependent inputs. *Mathematical Methods of Statistics*, 17(2):87–112, 2008. (Cité pages 173 et 227.)

F. Comte, S. Gaïffas, & 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. (Cité pages 30, 33, 55, 115, 117, 119, 121, 122, 125, 128, 129, 136 et 190.)

P. Courrège & P. Priouret. Temps d?arrêt d?une fonction aléatoire: relations d?équivalence associées et propriétés de décomposition. *Publications de l?Institut de Statistique de l?Université de Paris*, 14:245–274, 1965. (Cité page 220.)

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. (Cité pages 3, 6, 53, 93 et 108.)

J.Y. Dauxois & S. Sencey. Non-parametric tests for recurrent events under competing risks. *Scandinavian Journal of Statistics*, 36(4):649–670, 2009. (Cité page 153.)

S. S. Dave, G. Wright, B. Tan, A. Rosenwald, R. D. Gascoyne, W. C. Chan, R. I. Fisher, R. M. Braziel, L. M. Rimsza, T. M. Grogan, T. P. Miller, M. LeBlanc, T. C. Greiner, D. D. Weisenburger, J. C. Lynch, J. Vose, J. O. Armitage, E. B. Smeland, S. Kvaloy, H. Holte, J. Delabie, J. M. Connors, P. M. Lansdorp, Q. Ouyang, T. A. Lister, A. J. Davies, A. J. Norton, H. K. Muller-Hermelink, G. Ott, E. Campo, E. Montserrat, W. H. Wilson, E. S. Jaffe, R. Simon, L. Yang, J. Powell, H. Zhao, N. Goldschmidt, M. Chiorazzi, & L. M. Staudt. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, 351(21):pp. 2159–2169, 2004. URL `http://www.nejm.org/doi/full/10.1056/NEJMoa041869`. (Cité pages 53 et 55.)

D.L. Donoho, M. Elad, & V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006. (Cité page 11.)

M. Doumic, M. Hoffmann, P. Reynaud-Bouret, & V. Rivoirard. Nonparametric estimation of the division rate of a size-structured population. *SIAM Journal on Numerical Analysis*, 50(2):925–950, 2012. (Cité pages 26 et 149.)

B. Efron, T. Hastie, I. Johnstone, & R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. (Cité pages 11 et 13.)

J. Fan, Y. Feng, & Y. Wu. High-dimensional variable selection for Cox?s proportional hazards model. Dans *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics, 2010. (Cité page 205.)

J. Fan & R. Li. Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, 30(1):74–99, 2002. ISSN 0090-5364. URL `http://dx.doi.org/10.1214/aos/1015362185`. (Cité page 55.)

J. Friedman, T. Hastie, H. Höfling, & R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. (Cité page 13.)

S. Gaïffas & A. Guilloux. High-dimensional additive hazard models and the Lasso. *Electronic Journal of Statistics*, 6:pp. 522–546, 2012. ISSN 1935-7524. URL http://dx.doi.org/10.1214/12-EJS68. (Cité pages 29, 38, 55, 56, 69, 70, 71, 87, 88 et 102.)

R. Gill. Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics*, pages 49–58, 1983. (Cité page 57.)

A. Goldenshluger & O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39 (3):1608–1632, 2011. (Cité pages 24, 26, 27, 28, 110, 148, 149, 153 et 154.)

M. L. Gourlay, J. P. Fine, J. S. Preisser, R. C. May, C. Li, LY. Lui, D. F. Ransohoff, J. A. Cauley, & K. E. Ensrud. Bone-density testing interval and transition to osteoporosis in older women. *New England Journal of Medicine*, 366(3):pp. 225–233, 2012. URL http://www.nejm.org/doi/full/10.1056/NEJMoa1107142. (Cité page 53.)

E. Greenshtein & Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004. (Cité page 11.)

G. Grégoire. Least squares cross-validation for counting process intensities. *Scandinavian journal of statistics*, pages 343–360, 1993. (Cité pages 30, 110 et 152.)

A. Guilloux. *Inférence non paramétrique en statistique des durées de vie sous biais de sélection*. PhD thesis, Rennes 1, 2004. (Cité page 220.)

P. Hall & J.S. Marron. Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6(2):109–115, 1987. (Cité page 25.)

N. R. Hansen, P. Reynaud-Bouret, & V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *arXiv preprint arXiv1208.0570*, 2012. (Cité pages 55, 56, 60, 70, 71, 81 et 102.)

T. Hastie & R. Tibshirani. Generalized additive models. *Statistical science*, pages 297–310, 1986. (Cité page 6.)

M. Hebiri. *Quelques questions de sélection de variables autour de l'estimateur LASSO*. PhD thesis, Université Paris-Diderot-Paris VII, 2009. (Cité page 13.)

F. Hirsch & G. Lacombe. *Elements of functional analysis*, volume 192. Springer, 1999. (Cité page 172.)

J. Huang, T. Sun, Z. Ying, Y. Yu, & C.H. Zhang. Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, 41(3):1142–1165, 2013. (Cité pages 29, 32, 40, 109, 110, 115, 117, 118, 127, 144, 145, 149, 151 et 185.)

J. Jacod. On the stochastic intensity of a random point process over the half-line. *Princeton Univ., Dept. Statist., Techn. Rep*, 51, 1973. (Cité page 33.)

J. Jacod. Multivariate point processes: predictable projection, radon-nikodym derivatives, representation of martingales. *Probability Theory and Related Fields*, 31(3): 235–253, 1975. (Cité page 33.)

A. Juditsky & A. Nemirovski. Functional aggregation for nonparametric regression. *Annals of Statistics*, pages 681–712, 2000. (Cité page 11.)

M. J. Kearns, R. E. Schapire, & L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994. (Cité page 55.)

T. Klein & E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005. (Cité page 173.)

K. Knight & W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000. (Cité page 11.)

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011. (Cité page 54.)

V. Komornik. *Précis d'analyse réelle: Analyse fonctionnelle, intégrale de Lebesgue, espaces fonctionnels*. Ellipses, 2002. (Cité page 24.)

S. Kong & B. Nan. Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Arxiv preprint arXiv:1204.1992*, 2012. (Cité pages 29, 55, 61, 68, 109, 118 et 151.)

C. Lacour. *Estimation non paramétrique adaptative pour les chaînes de Markov et les chaînes de Markov cachées*. PhD thesis, Université René Descartes-Paris V, 2007. (Cité page 190.)

E. Le Pennec & S.X. Cohen. Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, eFirst, 3 2013. ISSN 1262-3318. URL `http://www.esaim-ps.org/article_S1292810012000171`. (Cité page 57.)

S. Lemler. Oracle inequalities for the lasso in the high-dimensional multiplicative Aalen intensity model. *To appear in Les Annales de l'Institut Henri Poincaré, arXiv preprint arXiv:1206.5628*, 2012. (Cité pages 39 et 71.)

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. (Cité pages 6, 30, 55, 115 et 226.)

R. Sh. Liptser & A. N. Shiryayev. *Theory of martingales*, volume 49 de *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]. (Cité pages 130, 139 et 223.)

S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J.A. Foekens, et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. *Journal of clinical oncology*, 25(10):1239–1246, 2007. (Cité pages 2, 10, 46, 108, 111 et 205.)

C.L. Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973. (Cité pages 18, 20, 115 et 121.)

J.S. Marron & W.J. Padgett. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics*, pages 1520–1535, 1987. (Cité pages 25 et 110.)

T. Martinussen & T. H. Scheike. Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, 36(4):602–619, 2009. ISSN 0303-6898. URL http://dx.doi.org/10.1111/j.1467-9469.2009.00650.x. (Cité pages 29 et 55.)

P. Massart. *Concentration inequalities and model selection*, volume 1896 de *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. (Cité pages 18, 20, 22, 23, 72, 89, 100, 115, 117, 128, 141 et 227.)

P. Massart & C. Meynet. The Lasso as an l1-ball model selection procedure. *Electronic Journal of Statistics*, 5:pp. 669–687, 2011. ISSN 1935-7524. URL http://dx.doi.org/10.1214/11-EJS623. (Cité pages 14, 15, 54 et 59.)

L. Meier, S. Van De Geer, & P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. (Cité page 13.)

N. Meinshausen & P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):pp. 1436–1462, 2006. (Cité pages 11 et 54.)

N. Meinshausen & B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pages 246–270, 2009. (Cité pages 11 et 15.)

W. Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1), 1969. (Cité page 7.)

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972. (Cité page 7.)

A. Nemirovski. Topics in nonparametric statistics. *Ecole d'Ete de Probabilites de Saint-Flour XXVIII, 1998*, 28:85, 2000. (Cité page 11.)

E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076, 1962. (Cité page 23.)

H. Ramlau-Hansen. *Udglatning med kernefunktioner i forbindelse med tælleprocesser: Del 1.* Forsikringsmatematisk Laboratorium, Københavns universitet, 1981. (Cité pages 30, 110 et 191.)

H. Ramlau-Hansen. The choice of a kernel function in the graduation of counting process intensities. *Scandinavian Actuarial Journal*, 1983(3):165–182, 1983a. (Cité page 109.)

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466, 1983b. (Cité pages 7, 30, 42, 109, 151, 152, 153, 191 et 225.)

P. Reynaud-Bouret. Penalized projection estimators of the aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006. (Cité pages 30, 33, 115, 119 et 125.)

J. Rice & M. Rosenblatt. Estimation of the log survivor function and hazard function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 60–78, 1976. (Cité page 30.)

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956. (Cité pages 23 et 24.)

M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78, 1982. (Cité page 24.)

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978. (Cité page 9.)

R. Senoussi. Problème d'identification dans le modèle de Cox. *Annales de l'Institut Henri Poincaré*, 26:pp. 45–64, 1988. (Cité pages 36, 57 et 73.)

N. Simon, J. Friedman, T. Hastie, & R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. (Cité page 185.)

E. W. Steyerberg, M. Y. V. Homs, A. Stokvis, ML. Essink-Bot, & P. D. Siersema. Stent placement or brachytherapy for palliation of dysphagia from esophageal cancer: a prognostic model to guide treatment selection. *Gastrointestinal Endoscopy*, 62(3): pp. 333–340, 2005. ISSN 0016-5107. URL http://www.sciencedirect.com/science/article/pii/S0016510705015877. (Cité page 53.)

C. J. Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, 22(1):118–184, 1994. ISSN 0090-5364. URL http://dx.doi.org/10.1214/aos/1176325361. With discussion by Andreas Buja and Trevor Hastie and a rejoinder by the author. (Cité page 57.)

M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996. (Cité page 173.)

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. (Cité page 128.)

T.M. Therneau & P.M. Grambsch. *Modeling survival data: extending the Cox model*. Springer, 2000. (Cité page 184.)

L. Tian, A. Alizadeh, A. Gentles, & R. Tibshirani. A simple method for detecting interactions between a treatment and a large number of covariates. *arXiv preprint arXiv:1212.2995*, 2012. (Cité pages 108 et 205.)

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. (Cité pages 11, 14, 54 et 185.)

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. (Cité pages 29, 54, 107 et 184.)

A.A. Tsiatis. A large sample study of cox's regression model. *The Annals of Statistics*, pages 93–108, 1981. (Cité page 184.)

A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 de *Mathématiques & Applications (Berlin)*. Springer-Verlag, Berlin, 2004. ISBN 3-540-40592-5. (Cité page 28.)

A.B. Tsybakov & V. Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009. (Cité page 23.)

J. V. Uspensky. *Introduction to mathematical probability.* New York: McGraw-Hill , 1937. (Cité page 89.)

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`. (Cité pages 36, 57, 69, 82, 89, 128, 146 et 227.)

S. van de Geer. The deterministic lasso. Rapport technique, ETH Zürich, Switzerland, Available at http://stat.ethz.ch/research/publ archive/2007/140., 2007. (Cité page 144.)

S. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):pp. 614–645, 2008. (Cité pages 59, 61, 118 et 151.)

N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012. (Cité pages 10, 118, 125 et 187.)

G.S. Watson & M.R. Leadbetter. Hazard analysis i. *Biometrika*, pages 175–184, 1964a. (Cité page 30.)

G.S. Watson & M.R. Leadbetter. Hazard analysis ii. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 101–116, 1964b. (Cité page 30.)

M. Wegkamp. Model selection in nonparametric regression. *Annals of Statistics*, pages 252–273, 2003. (Cité page 18.)

Y. Yang. Model selection for nonparametric regression. *Statistica Sinica*, 9(2):475–499, 1999. (Cité page 18.)

C.H. Zhang & J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):pp. 1567–1594, 2008a. (Cité pages 11, 15, 54 et 185.)

C.H. Zhang & J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, pages 1567–1594, 2008b. (Cité page 15.)

H.H. Zhang & W. Lu. Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007. (Cité pages 29, 55, 59 et 186.)

T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *The Journal of Machine Learning Research*, 11:pp. 1081–1107, 2010. (Cité page 59.)

P. Zhao & B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006. (Cité pages 11 et 15.)

P. Zhao & B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(2):pp. 2541, 2007. (Cité page 54.)

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):pp. 1418–1429, 2006. (Cité pages 59, 185 et 186.)

H. Zou. A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, 95(1):241–247, 2008. ISSN 0006-3444. URL `http://dx.doi.org/10.1093/biomet/asm083`. (Cité pages 29 et 55.)

Ce document a été préparé à l'aide de l'éditeur de texte GNU Emacs et du logiciel de composition typographique LaTeX 2_ε.

Title **Estimation for counting processes with high-dimensional covariates**

Abstract We consider the problem of estimating the intensity of a counting process adjusted on high-dimensional covariates. We propose two different approaches. First, we consider a non-parametric intensity function and estimate it by the best Cox proportional hazards model given two dictionaries of functions. The first dictionary is used to construct an approximation of the logarithm of the baseline hazard function and the second to approximate the relative risk. In this high-dimensional setting, we consider the Lasso procedure to estimate simultaneously the unknown parameters of the best Cox model approximating the intensity. We provide non-asymptotic oracle inequalities for the resulting Lasso estimator. In a second part, we consider an intensity that rely on the Cox model. We propose two two-step procedures to estimate the unknown parameters of the Cox model. Both procedures rely on a first step which consists in estimating the regression parameter in high-dimension via a Lasso procedure. The baseline function is then estimated either via model selection or by a kernel estimator with a bandwidth selected by the Goldenshluger and Lepski method. We establish non-asymptotic oracle inequalities for the two resulting estimators of the baseline function. We conduct a comparative study of these estimators on simulated data, and finally, we apply the implemented procedure to a real dataset on breast cancer.

Keywords Survival analysis, Cox model, Aalen multiplicative intensity model, counting process, high-dimension, Lasso, non-asymptotic oracle inequalitie, Bernstein inequalities, model selection, kernel estimator, Goldenshluger and Lepski method.

Titre **Estimation pour les processus de comptage avec beaucoup de covariables**

Résumé Nous cherchons à estimer l'intensité de sauts d'un processus de comptage en présence d'un grand nombre de covariables. Nous proposons deux approches. D'abord, nous considérons une intensité non-paramétrique et nous l'estimons par le meilleur modèle de Cox étant donné deux dictionnaires de fonctions. Le premier dictionnaire est utilisé pour construire une approximation du logarithme du risque de base et le second pour approximer le risque relatif. Nous considérons une procédure Lasso, spécifique à la grande dimension, pour estimer simultanément les deux paramètres inconnus du meilleur modèle de Cox approximant l'intensité. Nous prouvons des inégalités oracles non-asymptotiques pour l'estimateur Lasso obtenu. Dans une seconde partie, nous supposons que l'intensité satisfait un modèle de Cox. Nous proposons deux procédures en deux étapes pour estimer les paramètres inconnus du modèle de Cox. La première étape est commune aux deux procédures, il s'agit d'estimer le paramètre de régression en grande dimension via une procédure Lasso. Le risque de base est ensuite estimé soit par sélection de modèles, soit par un estimateur à noyau avec une fenêtre choisie par la méthode de Goldenshluger et Lepski. Nous établissons des inégalités oracles non-asymptotiques pour les deux estimateurs du risque de base ainsi obtenus. Nous menons une étude comparative de ces estimateurs sur des données simulées, et enfin, nous appliquons les procédures implémentées à une base de données sur le cancer du sein.

Mots-clés Analyse de survie, modèle de Cox, modèle à intensité multiplicative d'Aalen, processus de comptage, grande dimension, procédure Lasso, inégalité oracles non-asymptotique, inégalités de Bernstein, sélection de modèles, estimateur à noyaux, méthode de Goldenshluger et Lepski.