Modèle de Cox : estimation par sélection de modèle et modèle de chocs bivarié

Frédérique Letué

ORSAY
N° D'ORDRE:

UNIVERSITÉ DE PARIS SUD
U.F.R SCIENTIFIQUE D'ORSAY

THÈSE

présentée

pour obtenir

Le TITRE de DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY

SPÉCIALITÉ: MATHÉMATIQUES

par

Frédérique LETUÉ

Sujet: **MODÈLE DE COX: ESTIMATION PAR SÉLECTION DE MODÈLE ET MODÈLE DE CHOCS BIVARIÉ.**

Rapporteurs: M. Mark van der LAAN
Mme Shulamith T. GROSS

Soutenue le 21 décembre 2000 devant le jury composé de:

M.	Jean	BRETAGNOLLE	Directeur de Thèse
M.	Gérard	GRÉGOIRE	Examinateur
Mme	Éva	LELIÈVRE	Examinatrice
M.	Pascal	MASSART	Examinateur
Mme	Odile	PONS	Examinatrice

A ma grand-mère Marguerite.

Remerciements

Procédons par ordre chronologique :

Je remercie Jean Bretagnolle d'avoir accepté d'encadrer cette thèse. Pendant ces trois années, il a toujours été présent pour répondre avec patience à mes questions et me faire découvrir les secrets de l'analyse de survie.

Je remercie Éva Lelièvre et Daniel Courgeau, qui ont participé activement à la naissance de cette thèse, en me proposant le sujet du deuxième chapitre. Je remercie également Sylvie Huet, qui m'a utilement conseillée pour les simulations sur cette partie.

Je remercie Gwénaëlle Castellan pour notre travail commun, exposé dans le premier chapitre de cette thèse. Celui-ci fut long et difficile, mais nous avons finalement réussi à le mener à bien. Je remercie Pascal Massart de nous avoir proposé ce sujet et de m'avoir initiée aux techniques de sélection de modèle. Je remercie aussi Marie-Laure Martin pour notre collaboration concernant les simulations sur cette partie. Je remercie enfin le "Cox club", groupe de travail à l'INRA, qui nous a permis, à Gwénaëlle et à moi, de faire nos premiers exposés sur ce sujet.

Je remercie Shulamith T. Gross et Mark van der Laan d'avoir accepté de rapporter cette thèse. Mes échanges avec Shulamith T. Gross m'ont permis d'enrichir mon manuscrit grâce à ses remarques intéressantes.

Je remercie Odile Pons, Gérard Grégoire, et à nouveau Éva Lelièvre, et Pascal Massart d'avoir accepté de faire partie du jury de ma thèse. Odile Pons a suivi avec attention et intérêt l'évolution de mes travaux. Ses conseils m'ont toujours été très précieux.

Je remercie l'équipe de "Probabilités, Statistiques et Modélisation" de l'Université Paris-Sud, qui m'a très bien accueillie à Orsay pendant trois ans. Je remercie en particulier Sabine Hoarau et Catherine Rivalain pour leur aide administrative, et Yves Misiti et Patrick Jakubowicz pour leur aide informatique. Je remercie enfin l'équipe de Probabilités du LATP, à l'Université de Provence de m'avoir accueillie en tant qu'ATER, alors que je terminais cette thèse.

Je remercie bien sûr les locataires passés et présents du bureau 110. Maintes fois ils ont dû subir mes bavardages, maintes fois ils m'ont soutenue dans les moments difficiles. Je remercie aussi les occupants réguliers ou occasionnels de la salle café du premier étage du bâtiment 430, avec lesquels j'ai passé de bien bons moments.

Je remercie Jérôme et Marie-Luce, Antoine et Julie, Catherine et Jimmy, Fida, qui m'ont accueillie et soutenue pendant la fin de rédaction de cette thèse. Je remercie aussi Marie-Luce, Marguerite et Sabine d'avoir pris soin de ma santé physique en tentant de faire de moi une sportive ...

Je remercie Nicolas pour ses patientes relectures de ma thèse.

COX MODEL: ESTIMATION VIA MODEL SELECTION AND BIVARIATE SHOCK MODEL

Abstract:

This thesis contains two independent parts, both based on the Cox model.

In the first chapter which presents a joint work with Gwénaëlle Castellan, the Cox model is considered when the regression function of the covariates is not necessarily linear. To estimate this regression function, we devise a nonparametric estimation by model selection. A model is defined as a L_∞-ball of some finite-dimensional linear space of functions. In each model, the regression function is estimated by maximizing the Cox partial log-likelihood. We define then some penalized maximum partial log-likelihood estimator, from this collection of estimators. We give a risk bound for our estimator, in comparison to the smallest risk bound over the considered estimators collection.

In the second chapter, we propose a semiparametric shock model in order to model situations in demography where the biographies of a pair of individuals cannot be considered as independent. For that purpose, we construct two dependent counting processes representing these biographies in such a way that, whenever either one of both counting processes jumps, the hazard rate of the other one is instantaneously multiplied by a constant, called a shock parameter. Moreover, these counting processes may be censored. In such a context, assuming a Cox model, we propose maximum partial log-likelihood estimators for the shock parameters and for the Cox regression parameters, from a sample of independent and identically distributed, possibly censored pairs. Consistency and asymptotic normality of these estimators are established. We illustrate our results with simulations.

Keywords and phrases: Bivariate censored data, Bivariate survival analysis, Cox model, Kullback-Leibler Information, Model selection, Nonparametric estimation, Penalization.

AMS Classification: 62G05, 62G07, 62M09, 62F10, 62J15, 62J15, 62P10, 62P25

Table des matières

1. Ce chapitre présente un travail en collaboration avec G. Castellan.

2. Ce chapitre présente un travail en collaboration avec M.-L. Martin et G. Castellan.

Introduction

Cette thèse est constituée de deux parties indépendantes: la première est consacrée à l'estimation de la fonction de régression dans le modèle de Cox non paramétrique par sélection de modèle, alors que la seconde consiste en la modélisation d'un modèle de Cox bivarié. Bien entendu, le point commun entre les deux parties est le modèle que l'on considère: le modèle de Cox. Dans cette introduction, nous présenterons l'analyse de survie en général et le modèle de Cox en particulier, ainsi que quelques unes des nombreuses contributions sur le sujet. Puis, nous exposerons nos résultats pour chacune des deux parties de la thèse.

Analyse de survie, données censurées

L'analyse de survie consiste en l'étude de variables aléatoires positives qui modélisent des durées. C'est pourquoi elle apparaît dans de nombreux champs d'application. En effet, on pourra s'intéresser: en médecine à la durée entre l'apparition du symptôme d'une maladie chez un patient et le décès de cet individu suite à cette maladie, en démographie à la durée entre la mise en couple de deux individus (mariage ou concubinage) et l'arrivée d'un enfant dans ce foyer, en économétrie à la durée d'une période de chômage entre deux périodes de travail, en fiabilité à la durée de bon fonctionnement d'un composant, etc ... Alors que classiquement en statistiques, on cherche à estimer la fonction de répartition d'une variable aléatoire ou sa densité, en analyse de survie, on s'intéressera plutôt à la fonction de survie S ou au taux de risque α de la variable aléatoire X modélisant la durée, qui sont définis respectivement, pour tout $t \geq 0$, par

$$\begin{aligned} S(t) &= \mathbb{P}(X > t) \\ \text{et } \alpha(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(t < X \leq t + \Delta t \mid T > t) = -\frac{S'(t)}{S(t)}, \end{aligned}$$

où S' est la dérivée de S. Le taux de risque α caractérise à lui seul la loi de X puisque pour tout $t \geq 0$,

$$S(t) = e^{-\int_0^t \alpha(s)ds}.$$

Pour estimer ces grandeurs, on va collecter des données en choisissant par exemple un échantillon représentatif d'une population et en relevant pour chaque individu dans l'échantillon, la durée qui nous intéresse. Mais il arrive souvent que l'on ne puisse observer pour toutes les personnes la durée réelle du phénomène d'intérêt, soit parce que cette durée ne parvient pas à expiration avant la fin de la période d'étude, soit parce qu'un événement est intervenu entre-temps, interrompant l'étude avant son terme. On dit alors que les données sont censurées. Prenons un exemple: supposons que l'on observe pendant un laps de temps donné (par exemple deux ans) des patients soumis à une maladie et que l'on s'intéresse à la durée entre l'apparition de symptômes et le décès du patient dû à la maladie. Il se peut que certains patients ne soient pas décédés à la fin des deux ans, ou bien que certains patients soient décédés d'une autre maladie ou d'un accident. La mise à l'écart pure et simple des données de ces patients n'est pas possible, d'une part car elle

conduit à une réduction parfois trop importante de la taille de l'échantillon d'étude (si trop de patients sont dans ce cas), et d'autre part à un biais dans l'estimation. Il faut donc garder tous les patients dans l'échantillon et trouver une méthode d'estimation qui tienne compte du fait que certaines des données sont censurées.

D'un point de vue mathématique, ceci se traduit de la façon suivante : on introduit une nouvelle variable aléatoire U qui représente aussi une durée dite de censure (dans l'exemple précédent, il s'agira par exemple, de la durée entre le début d'étude et l'accident). Dans le cas d'une censure à droite, au lieu d'observer la variable aléatoire d'intérêt X, on observe en réalité $T = \min(X,U)$ et $\delta = \mathbb{1}_{\{X \leq U\}}$. Il s'agit alors d'estimer les fonctions S et/ou α à partir d'un n-échantillon de variables $(T_i,\delta_i)_{1 \leq i \leq n}$.

La littérature sur l'estimation non paramétrique de la fonction de survie et du taux de risque cumulé est très vaste. Nous nous contentons ici de citer quelques papiers historiques incontournables : Kaplan et Meier [39], Nelson [45, 46], Breslow [17], Aalen [1, 2], ...

Une approche possible pour aborder ce problème consiste à utiliser la théorie des processus de comptage et des martingales. L'idée est la suivante : pour une observation (T,δ), on introduit le processus de comptage $N(t) = \delta\, \mathbb{1}_{\{T \leq t\}}$, qui admet la décomposition suivante

$$N(t) = \Lambda(t) + M(t),$$

où Λ est un processus croissant et prévisible par rapport à la filtration $(\mathcal{N}_t)_{t \geq 0}$ définie par

$$\mathcal{N}_t = \sigma\{\delta, \mathbb{1}_{\{T \leq u\}}\,, 0 \leq u \leq t\},$$

et M est une martingale par rapport à la même filtration. Le processus Λ s'appelle le compensateur prévisible de N, il est unique et s'écrit sous la forme

$$\Lambda(t) = \int_0^t \alpha(s)\, \mathbb{1}_{\{T \geq s\}}\, ds.$$

On définit aussi le processus intensité λ par la relation

$$\Lambda(t) = \int_0^t \lambda(s) ds,$$

et donc ici, $\lambda(s) = \alpha(s)\, \mathbb{1}_{\{T \geq s\}}$, pour tout $s \geq 0$.

En interprétant la martingale comme un terme de bruit, il paraît alors assez naturel de s'inspirer des processus de comptage pour estimer α. On peut en particulier calculer la vraisemblance du processus de comptage pour en déduire des estimateurs du maximum de vraisemblance. Jacod [35, 36] donne la vraisemblance d'un processus de comptage multivarié à partir de son compensateur prévisible. Si on observe un processus de comptage multivarié $N = (N_h)_{1 \leq h \leq k}$, (c'est-à-dire que chaque coordonnée N_h est un processus de comptage et ces processus n'ont pas de sauts communs), sur un intervalle donné $[0,\tau]$, dont la loi P_{θ_0} appartient à une certaine famille $\{P_\theta, \theta \in \Theta\}$ (Θ ouvert de $\mathbb{R}^d$) de probabilités dominées par une même probabilité Q, la vraisemblance comme fonction de θ est

proportionnelle à

$$L(\theta) = \prod_{1\leq h\leq k} \prod_{0\leq t\leq \tau} \lambda_h^\theta(t)^{\Delta N_h(t)} \exp(-\Lambda_h^\theta(\tau)),$$

où Λ^θ est le compensateur prévisible de N et λ^θ son processus intensité sous la loi P_θ. Le logarithme de cette quantité peut s'écrire

$$l(\theta) = \ln L(\theta) = \sum_{h=1}^{k} \int_0^\tau \ln \lambda_h^\theta(s) dN_h(s) - \bar{\Lambda}^\theta(\tau),$$

où $\bar{\Lambda}^\theta(t) = \sum_{h=1}^k \int_0^t \lambda_h^\theta(s)ds$. Ce cadre permet notamment de construire des estimateurs du maximum de vraisemblance partielle (voir les bibliographies détaillées de Gross et Clark [31] (Chap. 3 et 4), Kalbfleisch et Prentice [38] (Chap.3), Lawless [41] (Chap 3 à 5), Andersen *et al.*[5] (Chap. VI), ...).

Outre l'estimation non paramétrique et paramétrique, on peut encore tenter d'estimer les fonctions S et α dans un cadre semi-paramétrique. C'est le contexte du modèle de Cox.

Le modèle de Cox

Le modèle de Cox est un modèle semi-paramétrique de régression pour des données censurées. On suppose désormais que la loi de la durée à laquelle nous nous intéressons dépend d'un certain nombre de caractéristiques. Dans notre exemple médical, il est en effet raisonnable de penser que la probabilité de survie à une maladie dépend de l'âge, du sexe, des antécédents médicaux ... du patient. De même, la date d'arrivée du premier enfant dans un foyer a des chances de dépendre de l'activité professionelle des parents, la durée d'une période de chômage du dernier diplôme obtenu par le chômeur, la durée de fonctionnement d'un composant du régime imposé à la machine, etc ...

Pour modéliser ceci, on introduit une variable aléatoire Z à valeur dans $\mathbb{R}^d$, appelé vecteur de covariables et on définit cette fois le taux de risque conditionnel α_Z par

$$\alpha_Z(t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} \mathbb{P}(t < X \leq t + \Delta t \mid T > t, Z) = \frac{f_Z(t)}{\mathbb{P}(X > t \mid Z)},$$

où f_Z est la densité conditionelle de X sachant Z. Dans le modèle de Cox, on suppose que ce taux de risque s'écrit sous la forme

$$\alpha_Z(t) = e^{\theta_0^T Z} \alpha_0(t),$$

où θ_0 est un paramètre de $\mathbb{R}^d$ à estimer, et α_0 est le taux de risque de base, qu'on ne cherche pas à modéliser et qui jouera le rôle de paramètre de nuisance pour l'estimation de θ_0. Cette écriture peut s'interpréter de la manière suivante: le taux de risque de base α_0 est le taux de risque d'un individu standard dont toutes les covariables sont nulles ($Z = 0$), et les autres individus voient leur taux de risque individuel multliplié par un facteur $e^{\theta_0^T Z}$ par rapport à cet individu standard. Si les coordonnées du vecteur Z sont positives, un

coefficient positif $\theta_{0,k}$ traduira un effet accélérateur pour l'événement étudié, alors qu'un coefficient négatif traduira un ralentissement par rapport à un individu standard.

Pour les mêmes raisons que précédemment, les données dont on dispose pour estimer le paramètre θ_0 auront des chances d'être censurées. On cherchera donc à estimer θ_0 à partir d'un n-échantillon de variables $(Z_i, \mathbb{1}_{\{T_i \leq t\}}, 0 \leq t \leq \tau, \delta_i)_{1 \leq i \leq n}$ avec les notations $T_i = \min(X_i, U_i)$ et $\delta_i = \mathbb{1}_{\{X_i \leq U_i\}}$ pour $1 \leq i \leq n$.

Comme nous l'avons fait dans le cas non paramétrique, nous pouvons définir les processus de comptage $N_i(t) = \delta_i \mathbb{1}_{\{T_i \leq t\}}$ pour $t \geq 0$ et $1 \leq i \leq n$, et calculer leurs compensateurs prévisibles par rapport à la filtration $(\mathcal{F}_t)_{0 \leq t \leq \tau}$, définie cette fois par

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

Ces compensateurs s'écrivent

$$\Lambda_i(t) = \int_0^t e^{\theta_0^T Z_i} \alpha_0(s) ds.$$

En suivant le même raisonnement que plus haut et en appliquant la formule de Jacod pour la vraisemblance d'un processus de comptage, on peut écrire la vraisemblance partielle du modèle de Cox

$$L_n(\theta, \alpha) = \prod_{1 \leq i \leq n} \prod_{0 \leq t \leq \tau} (e^{\theta_0^T Z_i} \alpha(s))^{\Delta N_i(t)} \exp(-\Lambda_i(\tau)). \tag{0.0.1}$$

Ce critère dépend bien sûr à la fois de θ et de α. Pour estimer θ, il faut donc trouver un critère qui ne dépende que de θ et plus de α. On en obtient un en maximisant le critère (0.0.1) en α (voir Johansen [37]). Ceci constitue une construction possible du critère proposé par Cox dans [25]

$$l_n(\theta) = \sum_{i=1}^n \ln \frac{e^{\theta^T Z_i}}{S_n(\theta, T_i)} N_i(\tau) = \sum_{i=1}^n \int_0^\tau \ln \frac{e^{\theta^T Z_i}}{S_n(\theta, s)} dN_i(s),$$

où $N_i(s) = \delta_i \mathbb{1}_{\{T_i \leq s\}}$, pour $0 \leq s \leq \tau$, et $1 \leq i \leq n$, et

$$S_n(\theta, s) = \frac{1}{n} \sum_{j=1}^n e^{\theta^T Z_j} \mathbb{1}_{\{T_j \geq s\}}, \text{ pour } 0 \leq s \leq \tau.$$

L'estimateur de Cox est alors défini par

$$\hat{\theta}_n = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}}\, l_n(\theta).$$

Dans [6], Andersen et Gill ont prouvé, en utilisant la théorie des processus de comptage, la consistance et la normalité asymptotique de cet estimateur, et ont donné un estimateur de sa matrice de covariance asymptotique. D'autres preuves ont également été proposées

par Tsiatis [55], Næs [44], et Bailey [7].

Dans la discussion qui suit l'article original de Cox [25] ainsi que dans [16], Breslow propose un estimateur $\hat{A}_n$ du taux de risque de base cumulé A_0 défini par

$$A_0(t) = \int_0^t \alpha_0(s)ds.$$

Cet estimateur, calculé à partir de l'estimateur $\hat{\theta}_n$, a pour expression avec les notations précédentes

$$\hat{A}_n(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{S_n(\hat{\theta}_n,s)} dN_i(s),$$

et maximise le critère (0.0.1) au point $\hat{\theta}_n$

$$\hat{A}_n = \underset{A}{\operatorname{argmax}}\, L(\hat{\theta}_n,dA).$$

Cette propriété permet notamment de prouver l'efficacité asymptotique de ces estimateurs (voir Bickel *et al.*[10], van der Vaart [57, 58]).

Le modèle de Cox non paramétrique

Dans la première partie de cette thèse, qui est un travail commun avec Gwénaëlle Castellan, nous considérons un modèle de Cox purement non paramétrique. Si nous désignons par W la covariable dans un espace E (en pratique [0,1] ou $\mathbb{R}^d$), le taux de risque conditionnel est supposé être de la forme

$$\alpha_W(t) = e^{f(W)}\alpha_0(t),$$

où f est une fonction de régression à estimer et α_0 est toujours le taux de risque de base. On remarque que pour toute constante c, le modèle $(f + c,\ e^{-c}\alpha_0)$ est le même que le modèle $(f,\ \alpha_0)$. C'est pourquoi nous sommes amenés à lever cette indétermination en imposant la contrainte $\mathbb{E}(f(W)) = 0$. On cherche alors à estimer la fonction f à partir d'un n-échantillon de variables $(W_i, \mathbb{1}_{\{T_i\leq t\}},0 \leq t \leq \tau,\delta_i)_{1\leq i\leq n}$, avec les mêmes notations que précédemment.

Nous avons en réalité deux exemples en tête :

Sélection de covariables :

Comme dans le modèle de Cox classique, la fonction f est linéaire en la covariable, qui est un vecteur de $\mathbb{R}^d$, d'espérance nulle : $f(W) = \theta_0^T W$, où le vecteur θ_0 de $\mathbb{R}^d$ peut avoir plusieurs coordonnées nulles. Dans ce cas, cela signifie que les coordonnées correspondantes de W n'interviennent en réalité pas dans le phénomène que l'on étudie. Dans un tel cas, si on savait quelles coordonnées de θ_0 étaient nulles, on aurait intérêt à chercher à estimer θ_0 non pas dans $\mathbb{R}^d$ tout entier, mais dans le sous-espace vectoriel de $\mathbb{R}^d$ engendré par les coordonnées non nulles de θ_0. En effet, en procédant de la sorte, on diminue le nombre de

coefficients à estimer et par conséquent, on améliore l'estimation de ces coefficients. On cherchera donc à estimer la fonction f par une fonction $\hat{f}$ de la forme

$$\hat{f} = \sum_{\lambda \in \Lambda} \hat{\beta}_\lambda W_\lambda,$$

où Λ est un sous-ensemble d'indices de $\{1,\ldots,d\}$, et les $W_\lambda, 1 \leq \lambda \leq d$ sont les coordonnées de W. La question que l'on se pose alors est: quel est le "bon" sous-ensemble Λ à considérer?

Le problème de la sélection de covariables dans le modèle de Cox a déjà été étudié par Senoussi [53]. Il propose un estimateur du maximum de log-vraisemblance partielle de Cox pénalisé, en utilisant la méthode de vraisemblance d'Akaike, qui consiste à maximiser la log-vraisemblance partielle tout en minimisant le nombre de coordonnées non-nulles. Il prouve que son estimateur est consistant et présente des simulations qui montrent quelles formes de pénalité sont acceptables.

Estimation par histogrammes:

Cette fois, E est l'intervalle [0,1] et f une fonction de [0,1] dans $\mathbb{R}$ telle que $\mathbb{E}(f(W)) = 0$. On cherche à estimer f par histogrammes. Pour cela, on se donne une partition Λ de [0,1], dont les intervalles sont notés $(I_\lambda, \lambda \in \Lambda)$, et on cherche à estimer f par une fonction $\hat{f}$ de la forme

$$\hat{f} = \sum_{\lambda \in \Lambda} \hat{\beta}_\lambda \, \mathbb{1}_{\{I_\lambda\}},$$

avec la contrainte $\sum_{\lambda \in \Lambda} \hat{\beta}_\lambda \mu(I_\lambda) = 0$. Si la partition est dyadique, on pourra chercher $\hat{f}$ sous la forme

$$\hat{f} = \sum_{\lambda \in \Lambda} \hat{\beta}_\lambda \varphi_\lambda(W),$$

où $(\varphi_\lambda, \lambda \in \Lambda)$ est la base de Haar.

Là aussi, nous sommes amenés à nous demander quelle est la "bonne" partition à choisir pour bien estimer f. En effet, pour bien approcher la fonction, on aura envie de prendre une partition à beaucoup de morceaux, mais en augmentant le nombre d'intervalles, on augmente le nombre de coefficients à estimer et on dégrade donc la qualité de l'estimation. Il faut donc faire un compromis entre ces deux tendances inverses.

Dans le cadre de l'estimation non paramétrique de la fonction de régression f, O'Sullivan [49] propose un estimateur du maximum de vraisemblance partielle pénalisé, en suivant la méthode développée par Wahba [61]. Cet estimateur minimise la somme de deux termes, le premier étant l'opposé de la log-vraisemblance partielle de Cox et le second, proportionnel à un opérateur linéaire borné sur une classe donnée de fonctions de régression, qui est typiquement l'opérateur laplacien standard. Les vitesses de convergence obtenues correspondent aux vitesses usuelles dans les espaces de Hölder.

Le but de ce chapitre est d'adapter à l'estimation de la fonction de régression f, la méthode d'estimation non paramétrique par sélection de modèle, proposée par Barron *et al.*[8] dans le cadre de l'estimation de densité ou de régression. Cette méthode consiste

à choisir une collection de modèles (par exemple, des espaces linéaires de fonctions) supposés avoir de bonnes propriétés d'approximation pour une classe de fonctions donnée. On commence par estimer la fonction f sur chacun des modèles en minimisant un contraste, obtenant ainsi une collection d'estimateurs. On sélectionne alors un modèle dans la collection en minimisant sur tous les modèles, un critère issu uniquement des données, qui est la somme du contraste pris en l'estimateur et d'un terme de pénalité, proportionnel au nombre de paramètres à estimer divisé par le nombre d'observations. Il s'agit alors de trouver une forme de la fonction de pénalité telle que le modèle sélectionné réalise un bon compromis entre l'erreur d'approximation et l'erreur d'estimation. Quand ceci est possible, le risque de l'estimateur sélectionné est alors de l'ordre du plus petit risque des estimateurs de la collection.

On cherchera donc ici à estimer la fonction f dans un espace linéaire $\mathcal{G}_\Lambda$ indexé par un ensemble d'indices Λ, dont le cardinal est la dimension de l'espace $\mathcal{G}_\Lambda$. Si $(\varphi_\lambda)_{\lambda\in\Lambda}$ est une base de $\mathcal{G}_\Lambda$, l'ensemble $\mathcal{G}_\Lambda$ peut s'écrire

$$\mathcal{G}_\Lambda = \left\{ \sum_{\lambda\in\Lambda} \beta_\lambda \varphi_\lambda, \beta \in \mathbb{R}^{|\Lambda|} \right\}.$$

Pour des raisons techniques, nous aurons besoin de manipuler des fonctions bornées. C'est pourquoi nous introduisons des espaces $\mathcal{S}_m$ qui sont des boules L_∞ des espaces linéaires précédents. Un modèle m sera un couple $m = (\Lambda,B)$, où Λ est un ensemble d'indices comme précédemment et B est un entier supérieur ou égal à 2 jouant le rôle de borne. On introduit ensuite les espaces $\mathcal{S}_m$ définis pour $m = (\Lambda,B)$ par

$$\mathcal{S}_m = \{g \in \mathcal{G}_\Lambda, \|g\|_\infty \leq B\}.$$

Pour estimer f sur le modèle m, on peut utiliser le critère de Cox. Définissons pour toute fonction g telle que $\mathbb{E}(g(W)) = 0$, et tout $0 \leq s \leq \tau$,

$$\begin{aligned} N_i(s) &= \delta_i \, \mathbb{1}_{\{T_i \leq s\}}, \\ S_n(g,s) &= \frac{1}{n} \sum_{j=1}^{n} e^{g(W_j)} \, \mathbb{1}_{\{T_j \geq s\}}, \\ \gamma_n(g) &= -\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \ln \frac{e^{g(W_i)}}{S_n(g,s)} dN_i(s). \end{aligned}$$

Il est maintenant naturel d'estimer f en minimisant γ_n sur l'ensemble $\mathcal{S}_m$

$$\hat{f}_m = \operatorname*{argmin}_{g\in\mathcal{S}_m} \gamma_n(g).$$

En procédant de la sorte sur plusieurs espaces linéaires $(\mathcal{G}_\Lambda, \Lambda \in \Lambda_n)$ et pour plusieurs

bornes B, ($B \in \mathbb{N}$, $B \geq 2$), on obtient une collection d'estimateurs

$$\left\{\hat{f}_m,\, m = (\Lambda,B),\, \Lambda \in \Lambda_n,\, B \geq 2\right\}.$$

On se pose alors la question: quel est le "meilleur" estimateur de cette collection?

Pour pouvoir parler de "meilleur" estimateur, il faut pouvoir mesurer leur qualité d'estimation. Pour cela, nous introduisons une fonction de perte K entre la fonction f et toute fonction g telle que $\mathbb{E}\,(g(W)) = 0$

$$K(f,g) = \mathbb{E}\left(\int_0^\tau \left(\log \frac{A(f)}{A(g)}\right)(s,W)A(f)(s,W)S(f,s)\,\mathbb{1}_{\{T\geq s\}}\,\alpha_0(s)ds\right),$$

avec les notations $A(g)(s,W) = \frac{e^{g(W)}}{S(g,s)}$ et $S(g,s) = \mathbb{E}\,(S_n(g,s)) = \mathbb{E}\left(e^{g(W)}\,\mathbb{1}_{\{T\geq s\}}\right)$. Cette fonction de perte est analogue à l'Information de Kullback-Leibler dans le cadre des densités, elle en a en effet toutes les qualités (voir Senoussi [53]), et nous la désignerons d'ailleurs ainsi par la suite.

Supposons à partir de maintenant que la fonction de régression f est bornée par une constante connue M et que la dimension de chacun des modèles est inférieure à $N_n \leq n/\ln n$. Nous introduisons des poids $L_\Lambda \geq 1, \Lambda \in \Lambda_n$ qui dépendront de la complexité de la collection Λ_n et qui vérifient

$$\sum_{\Lambda\in\Lambda_n} e^{-L_\Lambda|\Lambda|} \leq \Sigma < +\infty. \tag{0.0.2}$$

Nous proposons une fonction de pénalité de la forme

$$\mathrm{pen}_n(m) = c\left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{B\log(B)L_\Lambda D\log(n)}{n},$$

où c est une constante assez grande et D est la dimension de l'ensemble $\mathcal{S}_m$. Nous prouvons alors l'inégalité suivante

$$\mathbb{E}\left(K(f,\tilde{f})\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C \inf_{m\in\mathcal{M}_n} \inf_{\substack{f_m^* \in \mathcal{S}_m,\\ \|f_m^*\|_\infty \leq M}} \left(K(f,f_m^*) + \mathbb{E}\left(\mathrm{pen}_n(m)\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right)\right) + \frac{C'\Sigma}{n},$$

où $\Omega_n[\varepsilon]$ est un ensemble de grande probabilité dépendant d'un réel $\varepsilon \in]0,1[$, C ne dépend que de c et de M, et C' ne dépend que de c. De plus, nous obtenons le contrôle suivant

$$\mathbb{E}\left(\mathrm{pen}_n(m)\,\mathbb{1}_{\{\Omega_n\}}[\varepsilon]\right) \leq C''\left(\frac{e^{2(B+M)}}{\mathbb{P}\,(T\geq\tau)}\right)^2 \frac{B\log(B)L_\Lambda D\log(n)}{n},$$

où C'' ne dépend que de c et ε.

Ces inégalités permettent en quelque sorte de comparer le risque de l'estimateur sélectionné $\hat{f}_{\hat{m}}$, avec la plus petite des bornes de risques des estimateurs $\hat{f}_m$ de la collection. Bien que ces bornes de risques soient sans doute trop grossières, la procédure permet de choisir dans certains cas la plus petite borne de risque à un terme $\ln n$ près: en effet, quand la collection de modèles n'est pas trop riche (par exemple, s'il n'y a qu'un modèle de dimension D donnée dans la collection), on peut choisir des poids L_Λ constants. En revanche, si la collection est trop riche (par exemple, s'il y a $\binom{N_n}{D}$ modèles de dimension D donnée dans la collection), il est nécessaire de prendre des poids L_Λ de l'ordre de $\ln n$ pour que la somme (0.0.2) reste bornée pour toute valeur de n. La plus petite borne de risques n'est alors atteinte qu'à un terme $\ln^2 n$ près.

Dans le cas de l'estimation de la fonction de régression f par histogrammes réguliers, nous donnons la vitesse de convergence de notre estimateur, quand la vraie fonction de régression f appartient à l'ensemble des fonctions höldériennes d'ordre α et de constante de Hölder H

$$\mathbb{E}\left(K(f,\tilde{f})\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C'''(M,\alpha)H^{\frac{2}{2\alpha+1}}\left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + \frac{C'\Sigma}{n},$$

où C''' dépend de M et de α.

Dans l'annexe A, nous donnons les tout premiers résultats d'une étude de simulations réalisée par Marie-Laure Martin, en collaboration avec Gwénaëlle Castellan et moi-même, qui tente d'appliquer la méthode de sélection de modèle en pratique. L'estimation se fait cette fois sur les espaces linéaires $\mathcal{G}_\Lambda$ entiers, et non plus sur des boules L_∞ de ces espaces. La fonction de pénalité est choisie proportionnelle à D/n, et nous présentons une heuristique due à Birgé et Massart [11, 14] pour trouver la constante de proportionalité à partir des données elles-mêmes.

Un modèle de Cox bivarié

Le modèle de Cox, bien que décrit ci-dessus avec des covariables Z indépendantes du temps, peut en fait s'adapter pour des covariables dépendant du temps. En effet, si l'on cherche à décrire comment la fréquence de crises d'asthme dépend de la pollution de l'air, on aura envie de prendre comme covariables le processus $(Z(t),t \in [0,\tau])$ donnant la concentration de tel polluant dans l'air. La log-vraisemblance partielle de Cox s'adapte alors en remplaçant les covariables fixes Z par le processus $(Z(t),t \in [0,\tau])$, à condition toutefois que les covariables soient externes. Kalbfleisch et Prentice [38] distinguent trois types de covariables externes: soit la covariable est fixe (indépendante du temps), soit elle est "définie", c'est-à-dire qu'elle varie au cours du temps mais son aléa est entièrement contenu dans sa valeur initiale, soit elle est "ancillaire" (*ancillary*), c'est-à-dire que c'est un processus stochastique extérieur à l'individu étudié (mathématiquement, sa loi marginale

ne dépend pas des paramètres de la loi de X, ici θ_0 et α_0).

Il existe des cas où l'on aimerait prendre en compte des covariables qui ne sont clairement pas externes. Prenons le cas d'un couple de chômeurs cherchant du travail: on peut penser qu'à partir du moment où l'un des deux aura trouvé du travail, l'autre verra ses chances d'en trouver à son tour modifiées. C'est pourquoi, si l'on s'intéresse au cas de la femme, on aimerait prendre pour covariable le processus "travail" du mari, valant 0 tant que celui-ci n'a pas de travail et 1 à partir du moment où il en trouve, mais ce processus dépend clairement du processus "travail" de la femme elle-même: le processus n'est pas externe.

Dans ce genre de situations, la solution peut être d'envisager des processus de comptage bivarié: on considère maintenant deux variables aléatoires X_1 et X_2, qui ne sont pas indépendantes, et on cherche à estimer par exemple leur fonction de survie bivariée

$$F(t_1,t_2) = \mathbb{P}(X_1 > t_1, X_2 > t_2).$$

La littérature sur l'analyse de survie bivariée est elle aussi abondante. Citons les premiers travaux sur ce sujet de Dabrowska [27, 28], Gill *et al.*[30], Keiding [40], McKeague et Utikal [43], Pons [50, 51], van der Laan [56], . Ces travaux montrent que l'estimation non paramétrique en dimension supérieure ou égale à 2 s'avère notablement plus difficile qu'en dimension 1. L'une des raisons en est notamment la difficulté d'établir une théorie pour des martingales en dimension 2.

Cependant, des modèles de survie multivariés ont été proposés. Une classe très utilisée de modèles de survie multivariés est la classe des "modèles vulnérables" (*frailty models*). Ils consistent à supposer que les taux de risques d'individus d'un même groupe (par exemple d'une même famille) sont proportionnels à une même variable aléatoire, qui n'est pas observée (voir par exemple Vaupel *et al.*[60], Clayton et Cuzick [23], Gross et Huber [32], Hougaard [34], Oakes [48], Nielsen *et al.*[47]). Une autre classe de modèles bivariés est la classe des "modèles de copule" (*copula models*), dans lesquels la fonction de répartition du couple (X_1,X_2) est une fonction des fonctions de répartition marginales. Un modèle particulier est le modèle de chocs proposé par Clayton, où l'on suppose la relation suivante entre les taux de risque conditionnel d'un fils (λ_s) et de son père (λ_f)

$$\frac{\lambda_s(s \mid T = t_0)}{\lambda_s(s \mid T > t_0)} = \frac{\lambda_f(t \mid S = s_0)}{\lambda_f(t \mid S > s_0)} = \theta, \text{ pour tous } s,t,s_0,t_0 > 0.$$

Dans le second chapitre de la thèse, nous proposons un modèle de chocs bivarié pour des données censurées. Il s'agit de modéliser la situation suivante, issue d'un problème de démographie (voir Lelièvre *et al.*[42]) : dans un couple d'individus, tous deux susceptibles de subir des événements à des dates $\tilde{X}_A$ et $\tilde{X}_B$, quand l'un des deux individus subit son événement, le taux de risque de l'autre est instantanément multiplié par une constante. Mathématiquement, ce modèle est défini de la façon suivante: soit (X_A,X_B) un couple de variables aléatoires indépendantes positives de taux de risque α_A et α_B.

Si $X_A < X_B$, nous posons $\tilde{X}_A = X_A$ et $\tilde{X}_B = X_A + X'_B$ où X'_B est une variable

aléatoire positive, indépendante de X_B, de taux de risque $e^{\rho_{B,0}}\alpha_B(X_A+s)$, où $\rho_{A,0}$ est un nombre réel.

Si $X_B < X_A$, nous posons $\tilde{X}_B = X_B$ et $\tilde{X}_A = X_B + X'_A$ où X'_A est une variable aléatoire positive, indépendante de X_A, de taux de risque $e^{\rho_{A,0}}\alpha_A(X_B+s)$, où $\rho_{B,0}$ est un nombre réel.

Pour tenir compte de covariables Z_A et Z_B, il suffit alors de remplacer les taux de risques α_A et α_B par les taux de risques conditionnels $\alpha_{Z_A} = e^{\theta_{A,0}{}^T Z_A}\alpha_{A,0}$ et $\alpha_{Z_B} = e^{\theta_{B,0}{}^T Z_B}\alpha_{B,0}$, où $\theta_{A,0}$ et $\theta_{B,0}$ sont des vecteurs de dimension D, appelés paramètres de régression de Cox.

On suppose de plus que les durées $\tilde{X}_A$ et $\tilde{X}_B$ peuvent être censurées : au lieu d'observer $\tilde{X}_A$ et $\tilde{X}_B$, on observe $T_h = \min(\tilde{X}_h, U_h)$ et $\delta_h = \mathbb{1}_{\{\tilde{X}_h \le U_h\}}$ pour $h = A,B$, où les U_h sont des variables aléatoires positives de censure. Il s'agit alors d'estimer le paramètre $\beta_0 = (\rho_{A,0},\theta_{A,0},\rho_{B,0},\theta_{B,0})$ à partir d'un n-échantillon de variables aléatoires

$$(Z_{A,i},Z_{B,i},\mathbb{1}_{\{T_{A,i}\le t\}},\mathbb{1}_{\{T_{B,i}\le t\}},t\in[0,\tau],\delta_{A,i},\delta_{B,i})_{1\le i\le n}$$

de même loi que $(Z_A,Z_B,\mathbb{1}_{\{T_A\le t\}},\mathbb{1}_{\{T_B\le t\}},t\in[0,\tau],\delta_A,\delta_B)$, en présence des taux de risque de base $\alpha_{A,0}$ et $\alpha_{B,0}$, considérés comme des paramètres de nuisance : le cadre est à nouveau semi-paramétrique.

Pour cela, nous définissons des processus de comptage

$$\begin{aligned} N_{A,i}(t) &= \delta_{A,i}\,\mathbb{1}_{\{T_{A,i}\le t\}}(\delta_{B,i}\,\mathbb{1}_{\{T_{B,i}<T_{A,i}\}} + \mathbb{1}_{\{T_{A,i}\le T_{B,i}\}}),\\ N_{B,i}(t) &= \delta_{B,i}\,\mathbb{1}_{\{T_{B,i}\le t\}}(\delta_{A,i}\,\mathbb{1}_{\{T_{A,i}<T_{B,i}\}} + \mathbb{1}_{\{T_{B,i}\le T_{A,i}\}}), \text{ pour tout } t>0, \end{aligned}$$

et la filtration $(\mathcal{F}_t, t\ge 0)$ comme la famille des tribus $\mathcal{F}_t$ engendrée par

$$\{\mathbb{1}_{\{T_A\le u\}},\mathbb{1}_{\{T_B\le u\}},u\le t,\delta_A,\delta_B,Z_A,Z_B\}.$$

Le calcul des compensateurs $\mathcal{F}_t$-prévisibles des processus de comptage N_A et N_B permet de définir un critère de log-vraisemblance partielle

$$l_n(\beta) = \sum_{h=A,B}\sum_{i=1}^{n}\int_0^\tau \ln\left(\frac{e^{\beta_h^T W_{h,i}(s)}}{S_{n,h}(\beta,s)}\right) dN_{h,i}(s)$$

avec les notations

$$\begin{aligned} \beta_A &= (\rho_A,\theta_A), & \beta_B &= (\rho_B,\theta_B),\\ W_{A,i}(t) &= \begin{pmatrix}\mathbb{1}_{\{T_{B,i}<t\}}\\ Z_{A,i}\end{pmatrix}, & W_{B,i}(t) &= \begin{pmatrix}\mathbb{1}_{\{T_{A,i}<t\}}\\ Z_{B,i}\end{pmatrix}, \end{aligned}$$

$$\begin{aligned} Y_{A,i}(t) &= \mathbb{1}_{\{T_{A,i}\geq t\}}(\delta_{B,i}\,\mathbb{1}_{\{T_{B,i}<t\}} + \mathbb{1}_{\{T_{B,i}\geq t\}}), \\ Y_{B,i}(t) &= \mathbb{1}_{\{T_{B,i}\geq t\}}(\delta_{A,i}\,\mathbb{1}_{\{T_{A,i}<t\}} + \mathbb{1}_{\{T_{A,i}\geq t\}}), \\ S_{n,h}(\beta,s) &= n^{-1}\sum_{i=1}^{n} e^{\beta_h^T W_{h,i}(s)} Y_{h,i}(s), \text{ pour } h = A,B. \end{aligned}$$

Nous définissons alors un estimateur du maximum de log-vraisemblance partielle

$$\hat{\beta}_n = \underset{\beta}{\operatorname{argmax}}\, l_n(\beta),$$

et nous vérifions les hypothèses du théorème d'Andersen et Gill [6] pour obtenir la consistance et la normalité asymptotique de notre estimateur

$$n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0,(\Sigma_\tau)^{-1}\right),$$

où Σ_τ est la matrice définie par

$$\begin{aligned} \Sigma_\tau &= \sum_{h=A,B}\int_0^\tau V_h(\beta_0,s)S_h(\beta_0,s)\alpha_{h,0}(s)ds, \\ S_h(\beta,s) &= \mathbb{E}\left(S_{n,h}(s)\right), \\ E_h &= DS_h/S_h \\ V_h &= D^2S_h/S_h - (E_h)^{\otimes 2}, \end{aligned}$$

où $x^{\otimes 2} = xx^T$ et DF et D^2F sont les dérivées première et seconde de F. Nous pouvons de plus estimer cette matrice par $-\frac{1}{n}D^2l_n(\hat{\beta}_n)$, qui tend en probabilité vers Σ_τ, quand n tend vers l'infini.

Nous illustrons nos résultats par des simulations : celles-ci montrent en particulier que les estimations sont en fait déjà bonnes pour des petites tailles d'échantillon ($n = 30$ ou 50).

Plan

Le premier chapitre présente un travail en collaboration avec Gwénaëlle Castellan sur l'estimation par sélection de modèle de la fonction de régression dans le modèle de Cox.

Le second chapitre est consacré à la présentation d'un modèle de chocs en analyse de survie bivariée censurée.

Dans l'annexe A, nous présentons les premiers résultats d'une étude de simulations, réalisée par Marie-Laure Martin en collaboration avec Gwénaëlle Castellan et moi-même, qui tente d'appliquer la méthode de sélection de modèle proposée dans le chapitre 1.

L'annexe B donne une preuve des théorèmes de consistance et de normalité asymptotique du chapitre 2.

L'annexe C présente les programmes Matlab utilisés pour les simulations du chapitre 2.

Chapitre 1

Estimation of the Cox regression function via model selection [1]

Contents

1. Ce chapitre présente un travail en collaboration avec G. Castellan.

Abstract

The Cox model is considered when the regression function of the covariates is not necessarily linear. To estimate this regression function, we devise a nonparametric estimation by model selection. A model is defined as a L_∞ ball of some finite-dimensional linear space of functions. In each model, the regression function is estimated by maximizing the Cox partial log-likelihood. A collection of estimators is thus obtained. Our aim is to select the best one, in some sense, in this collection, using a data-driven criterion.

We define a penalized maximum partial log-likelihood estimator, minimizing over all models the sum of the Cox criterion at the estimator and of some penalty term depending among others on the model dimension. We give a risk bound in the sense of the Kullback-Leibler Information for our estimator, to be compared to smallest risk bound of the estimators in the collection. We calculate the rate of convergence of our estimator in some particular case (regular histogram estimation).

Keywords and phrases: Cox model, Kullback-Leibler Information, Model selection, Nonparametric estimation, Penalization.

Résumé

Dans ce chapitre, nous considérons le modèle de Cox non paramétrique, défini quand la fonction de régression n'est pas nécessairement linéaire, et nous cherchons à estimer cette fonction de régression par sélection de modèle. Un modèle est défini comme une boule L_∞ d'un espace linéaire de fonctions de dimension finie. Dans chaque modèle, la fonction de régression est estimée en maximisant la log-vraisemblance partielle de Cox. On obtient ainsi une collection d'estimateurs. Le but est de choisir le meilleur estimateur de cette collection, en un certain sens, à partir d'un critère issu des données uniquement.

Nous définissons un estimateur du maximum de log-vraisemblance partielle pénalisé en minimisant sur tous les modèles la somme du critère de Cox en l'estimateur et d'un terme de pénalité dépendant entre autres de la dimension du modèle. Nous donnons une borne de risque au sens de l'Information de Kullback-Leibler de notre estimateur, qui est comparable à la plus petite des bornes de risques des estimateurs de la collection. Nous calculons la vitesse de convergence de notre estimateur dans un cas particulier (estimation par histogrammes réguliers).

Mots clés : Estimation non paramétrique, Information de Kullback-Leibler, Modèle de Cox, Pénalisation, Sélection de modèle.

This chapter presents a joint work with G. Castellan.

1.1 Introduction

In medical studies, the Cox model is often used, while studying a duration (for instance, between the disease diagnosis and the death of the individual) with respect to covariates. In such a case, the conditional hazard rate of the duration X given the d-dimensional covariates vector W is modelled by $\alpha_W(t) = e^{\theta_0^T W}\alpha_0(t)$, where θ_0 is some d-dimensional vector to be estimated, and α_0 is some baseline hazard rate, that plays the role of a nuisance parameter. Sometimes, the entire observation of the duration of interest is not possible, in which case the data are said to be censored. If U is the random variable representing the censoring duration (between the beginning of the study till accidental death), we only observe (W, T, δ), where $T = \min(X, U)$ and $\delta = 1\!\!1_{\{X \leq U\}}$.

In [25], Cox proposes a semi-parametric maximum partial likelihood estimator for the parameter θ_0 from a n-sample of independent and identically distributed random variables $(W_i, T_i, \delta_i)_{1 \leq i \leq n}$ with same distribution as (W, T, δ). This estimator has been widely studied in the literature, and in particular, it is now well-known that it is consistent and asymptotically normal with rate of convergence $n^{-1/2}$. An estimation of the asymptotic covariance matrix is also established (see Andersen *et al.*[5]). It is also well-known that $2n(l_n(\hat{\theta}) - l_n(\theta_0))$ has asymptotically a Khi-square distribution with d degrees of freedom, where l_n denotes the Cox log partial likelihood. These results established by Andersen and Gill in [6] are proved introducing counting processes and using results from the martingale theory.

In [53], Senoussi proposes an estimator for the number of coordinates of the parameter θ, which are different from zero. This problem is the same as choosing the "right" coordinates of the covariates vector W, that is to say, to select the covariates that are really significant for the duration of interest. To do so, he uses Akaike's likelihood method to construct a penalized Cox partial log-likelihood estimator. He proves the consistency of this estimator and runs simulations that show that penalty functions such as

$$\frac{D_m}{\sqrt{n}}, \quad \frac{D_m \log(n)}{n} \quad \text{and} \quad \frac{D_m \log(\log(n))}{n}$$

perform well, whereas a penalty function of the form D_m is too heavy and systematically chooses the model with smaller dimension. His results are asymptotic in the sense that the dimension d of the covariate W is fixed while the sample size n tends to infinity, and he does not propose any risk bound.

In some applications, it may be interesting to introduce non necessarily linear functions of the covariates. In [33], Hastie and Tibshirani apply the so-called local likelihood estimation method to generalized additive model, that is to say, they assume a model of the form

$$\alpha_W = e^{s_0 + \sum_{j=1}^{d} s_j(W_j)}\alpha_0,$$

where the s_j are smooth functions. O'Sullivan [49] proposes a model where the conditional hazard rate given the possibly time dependent covariates is of the form

$$\alpha_{W(t)} = e^{f(W(t))}\alpha_0(t),$$

where the regression function f is to be estimated. In the sequel, we shall refer to this model as the nonparametric model. He proposes a penalized maximum partial likelihood estimator for the regression function f, applying methods developped by Wahba [61]. The penalized partial likelihood functional is the sum of two terms, the first being the opposite of the Cox log partial likelihood and the second being proportional to some bounded linear operator on some class of possible regression functions. Typically, this linear operator is often taken as the standard Laplacian operator (see Cox [24]). He obtains upper bounds on the rates of convergence which correspond to the optimal rates of convergence in the framework of nonparametric regression and density estimation in Hölder spaces (see Stone [54]).

Our aim in this chapter consists in applying one method of nonparametric estimation by model selection to the estimation of the regression function f. This method is explained in Barron *et al.*[8] in the framework of density estimation and regression function estimation, with various criteria like penalized maximum likelihood, projection or least square estimation. The ideas of this method are the following.

We choose a list of models (for instance, linear subspaces) which are supposed to have good properties for the approximation of some classes of smoothness. We estimate the regression function on each model to obtain a collection of estimators. Then, adding some penalty term roughly proportional to the number of parameters to be estimated divided by the number of observations, we select one model among the collection. The penalty must be chosen such that the selected model realizes the best trade-off between the approximation error and the estimation error. If this is true, then the risk bound of this estimator should be of the order of the smallest risk over all estimators in the collection.

An example of application of this method has been studied by Castellan [19, 21] in the framework of histogram density estimation. She proposes a penalized maximum likelihood estimator and the authorized forms of the penalty function depends on the complexity of the considered collection of models. Here, each model is a partition. In the case of regular histograms, i.e. when all partitions have intervals with the same size, then there is only one partition in the collection with a given number of intervals, and the penalty function should be chosen as

$$c_1\frac{D_m}{n} + c_2\frac{D_m^{3/4}}{n},$$

with $c_1 > 1/2$ and D_m the number of intervals of the partition m. In the case of irregular histograms, a model is a partition based on a given grid with N_n intervals. Then, there are $\binom{N_n}{D}$ partitions with D intervals built on this grid and the penalty function should

be taken as

$$c\frac{D_m}{n}\log\left(\frac{c'N_n}{n}\right).$$

Note that the first case include the famous Akaike's Information Criterion (see Akaike [4]), while taking $c_1 = 1$ and $c_2 = 0$. The risk bounds are expressed in terms of Kullback-Leibler information and of Hellinger distance. Such penalty functions allow to obtain precise and explicit risk bounds. Furthermore, her results can be adapted for exponential of piecewise polynomials density estimation. A simulation study in the context of histogram estimation performed by Birgé and Rozenholc [15] confirm these theoretical results, and they propose adequate choices for the constants.

In this chapter, we apply the same method for the estimation of the regression function, assuming a nonparametric Cox model with time-constant covariates,

$$\alpha_W(t) = e^{f(W)}\alpha_0(t),$$

and assuming that the regression function f verifies

$$\mathbb{E}\left(f(W)\right) = 0. \tag{1.1.1.1}$$

This restriction, that was already used by O'Sullivan [49], is due to the fact that the Cox model is not identifiable: the model characterized by $(f + c, e^{-c}\alpha_0)$ is the same as the one characterized by (f, α_0), for any real number c. Observing a n-sample of independent and identically distributed random variables $(W_i, T_i, \delta_i)_{1\leq i\leq n}$ with same distribution as (W, T, δ), one way to estimate f is to maximize the Cox log partial likelihood (see Cox [25]) denoted by $-\gamma_n$ on some L_∞-ball of some finite-dimensional linear subspace. We consider a collection $\{\mathcal{S}_m, m \in \mathcal{M}_n\}$ of such subspaces, called models. The estimation on each model leads to a collection of estimators $\{\hat{f}_m, m \in \mathcal{M}_n\}$, defined by

$$\hat{f}_m = \operatorname*{argmin}_{g\in\mathcal{S}_m}\gamma_n(g).$$

Then, we propose some penalty function pen_n on the set of all possible models, and we define an estimator $\tilde{f} = \hat{f}_{\hat{m}}$ where $\hat{m}$ is defined by

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n}\left\{\gamma_n(\hat{f}_m) + \mathrm{pen}_n(m)\right\}.$$

To measure the quality of this estimator, we introduce some loss function $K(f, g)$ between the true function f and any function g such that $\mathbb{E}\left(g(W)\right) = 0$, which appears to play the same role as the Kullback-Leibler information in the density estimation framework. We would like to prove that our estimator has a risk bound of the order of the smallest risk bounds over all estimators $\hat{f}_m$, $m \in \mathcal{M}_n$, but since we do not reach this target, we content ourselves to give a significant risk bound for our estimator.

We give illustrations in two particular cases, namely histogram estimation of the re-

gression function f and covariates selection when the covariate W is some d-dimensional vector, that is to say, selection of the covariates which are actually involved in the estimation of the regression function.

In the next section, we introduce the contrast function, the models and the loss function. In section 3, we define our estimator and we present our main result. Section 4 is devoted to the proof which relies on an exponential inequality which stands in subsection 4.2. Technical points can be found in section 5.

1.2 Statistical framework

1.2.1 Notations and Assumptions

Let us recall our notations and introduce some new ones. The covariate is denoted by W, varies in some probability space E and has probability measure μ on this set. The variable of interest is denoted by X, which conditional hazard rate given the covariate is $e^{f(W)}\alpha_0$, where f belongs to the space $\mathcal{G}$ defined by

$$\mathcal{G} = \{g : E \to \mathbb{R},\ \mathbb{E}\,(g(W)) = 0\},$$

and α_0 is some baseline hazard rate. The censoring duration is denoted by U and we observe T defined by $T = \min(X, U)$ in some deterministic time interval $[0, \tau]$, and $\delta = \mathbb{1}_{\{X \leq U\}}$. We consider a sample of independent and identically distributed variables $(W_i, T_i, \delta_i)_{1 \leq i \leq n}$ with same distribution as (W, T, δ).

Like in the classical Cox model, we assume the following hypotheses:

Assumption 1.1.

- *The random variables X and U are conditionally independent given the covariate W.*
- *The baseline hazard rate α_0 is such that $\int_0^\tau \alpha_0(s)ds$ is finite.*
- *The conditional distribution of random variable U given W verifies that $\mathbb{P}\,(U \geq \tau \mid W)$ is positive.*

All these assumptions assure in particular that $\mathbb{P}\,(T \geq \tau)$ is positive.

1.2.2 The contrast

Let γ_n be the empirical contrast defined for any function g in $\mathcal{G}$ by:

$$\gamma_n(g) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{e^{g(W_i)}}{S_n(g,s)}\right) dN_i(s), \tag{1.2.1.1}$$

where τ is the final time of observation, N_i is the censored counting process relative to T_i defined by $N_i(s) = \delta_i\, \mathbb{1}_{\{T_i \leq s\}},\ s \geq 0$, and $S_n(g,s) = \frac{1}{n}\sum_{i=1}^{n} e^{g(W_i)}\, \mathbb{1}_{\{T_i \geq s\}},\ s \geq 0$. Note

that γ_n is the opposite of the Cox log partial likelihood (see Cox [25]), adapted to non necessarily linear functions.

The next lemma establishes the fact that γ_n is a contrast.

Lemma 1.2.1. *For any deterministic function g of $\mathcal{G}$, $\mathbb{E}(\gamma_n(g) - \gamma_n(f))$ is non-negative and $\mathbb{E}(\gamma_n(g) - \gamma_n(f)) = 0$ implies $g = f$ almost everywhere.*

A proof of this lemma is given in section 1.5.1.

In the sequel, we denote by $\mathcal{E}$ the operator which coincides with the expectation on the deterministic functions, but we shall extend this operator to random functions. We keep the notation $\mathbb{E}$ for the expectation. So, for any deterministic function h, it is true that $\mathcal{E}[h(W,T,\delta)] = \mathbb{E}(h(W,T,\delta))$, whereas it is not the case for random functions.

Let us now define the following empirical process, for any function h from $E \times [0,\tau]$ to $\mathbb{R}$:

$$\mathbb{P}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau h(W_i,s)dN_i(s) = \frac{1}{n}\sum_{i=1}^{n} h(W_i,T_i)N_i(\tau).$$

We introduce now the operator $\mathbb{P}$ defined for any deterministic function h from $E\times[0,\tau]$ to $\mathbb{R}$ by

$$\mathbb{P}(h) = \mathcal{E}\left[\int_0^\tau h(W,s)e^{f(W)}\,\mathbb{1}_{\{T\geq s\}}\,\alpha_0(s)ds\right].$$

Denoting by $\mathcal{F}_t, 0 \leq t \leq \tau$ the σ-algebra generated by $(W_i, N_i(s), 0 \leq s \leq t)_{1\leq i\leq n}$, and since any deterministic function h is predictable with respect to the filtration $(\mathcal{F}_t, 0 \leq t \leq \tau)$, we can apply the Claim 1.5.4.1

$$\mathbb{P}(h) = \mathbb{E}\left(\int_0^\tau h(W,s)e^{f(W)}\,\mathbb{1}_{\{T\geq s\}}\,\alpha_0(s)ds\right) = \mathbb{E}\left(\int_0^\tau h(W,s)dN(s)\right) = \mathbb{E}\left(\mathbb{P}_n(h)\right).$$

So, we remark that this operator coincides with the expectation of the previous empirical process on the deterministic functions.

1.2.3 The loss function

In order to evaluate the estimation error, we construct some loss function K between the function f and any function g in $\mathcal{G}$. Like previously, we define it as an operator on the deterministic functions, and we extend it to random functions.

$$K(f,g) = \mathcal{E}\left[\int_0^\tau \left(\log\frac{A(f)}{A(g)}\right)(s,W)A(f)(s,W)S(f,s)\,\mathbb{1}_{\{T\geq s\}}\,\alpha_0(s)ds\right], \qquad (1.2.1.2)$$

where

$$S(g,s) = \mathcal{E}\left[e^{g(W)}\,\mathbb{1}_{\{T\geq s\}}\right],\ 0 \leq s \leq \tau,$$

and

$$A(g)(s,w) = \frac{e^{g(w)}}{S(g,s)},\ 0 \leq s \leq \tau,\ w \in E.$$

Using the Claim 1.5.4.1 and since the function $\log \frac{A(f)}{A(g)}$ is $\mathcal{F}_t$-predictable for any deterministic function g, the operator K can also be written as

$$K(f,g) = \mathcal{E}\left[\int_0^\tau \left(\log \frac{e^{f(W)}}{S(f,s)} - \log \frac{e^{g(W)}}{S(g,s)}\right) dN(s)\right].$$

The loss function K is similar to the classical Kullback-Leibler information. It is non-negative and equals zero if and only if $f = g$, μ almost surely (see Senoussi [53]). The quantities $A(g)$ are similar to the densities in the nonparametric density estimation framework and they verify for every g in $\mathcal{G}$,

$$\mathcal{E}\left[\int_0^\tau A(g)(s,W)S(f,s)\, 1\!\!1_{\{T\geq s\}}\, \alpha_0(s)ds\right] = \mathbb{E}\left(N(\tau)\right),$$

that is analogous to $\int s d\mu = 1$ in the density case.

1.2.4 The models

Linear subspaces

Suppose that the distribution μ of W is known. In order to define what we call a model, we begin with considering a finite collection of linear subspaces $\{\mathcal{G}_\Lambda, \Lambda \in \Lambda_n\}$ of $\mathcal{G}$, Λ_n being a finite collection of sets of indexes. The dimension of any linear subspace $\mathcal{G}_\Lambda$ is denoted by D and equals the cardinality of Λ. Let us consider the following assumption:

Assumption 1.2. *There exists some nonnegative real numbers a and $\bar{b}$ such that, for any linear subspace $\mathcal{G}_\Lambda$, there exists some basis $\{\varphi_\lambda, \lambda \in \Lambda\}$ belonging to $L_\infty(E,\mu) \cap L_2(E,\mu)$ and verifying that for any $\beta \in \mathbb{R}^\Lambda$,*

$$\|\sum_{\lambda\in\Lambda} \beta_\lambda \varphi_\lambda\|_\infty \leq \bar{b} D^a |\beta|_\infty.$$

We have in particular in mind two examples of such models:

- *Covariates selection:* Suppose that $E = \mathbb{R}^d$, that the coordinates W_λ are centered. We define φ_λ for any $\lambda \in \{1,\ldots,N_n\}$ as the coordinate applications: $\varphi_\lambda(W) = W_\lambda$. A linear subspace $\mathcal{G}_\Lambda$ is generated by the functions $\{\varphi_\lambda, \lambda \in \Lambda\}$, where Λ is some subset of $\{1,\ldots,N_n\}$ and we denote by D the cardinality of Λ. In this case, we are in the classical linear Cox regression framework, where f is estimated by some function of the form $< \theta, W >$. We are here interested in choosing the appropriate coordinates of W.

  Furthermore, we have

$$|\sum_{\lambda\in\Lambda} \beta_\lambda \varphi_\lambda(W)| \leq |\beta|_\infty |\sum_{\lambda\in\Lambda} W_\lambda|.$$

As soon as the covariates W verifies that $|W|_\infty$ is smaller than some constant $\bar{w}$ μ-almost surely, Assumption 1.2 is satisfied with $a = 1$ and $\bar{b} = \bar{w}$.

- *Histogram estimation:* Suppose now that $E = [0, 1]$. Given some partition Λ on $[0, 1]$, the linear subspace $\mathcal{G}_\Lambda$ is defined as the linear subspace of piecewise constant functions g on Λ such that $\mathbb{E}(g(W)) = 0$. For instance, if Λ is a dyadic partition, we can choose the Haar basis on E as the basis functions φ_λ.

  In that case, Assumption 1.2 is satisfied with $a = 1/2$ and $\bar{b} = 1$.

Definition of the models

For technical reasons, we need bounds on the functions we manipulate. Therefore, we do not consider the linear subspaces described before, but some L_∞-balls of these linear subspaces.

Definition 1.1. *A model m is defined as a pair (Λ, B) where Λ is some set of indexes and B is an integer. Given a model $m = (\Lambda, B)$, we denote by $\mathcal{S}_m$ the following set:*

$$\mathcal{S}_m = \left\{ g = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda, |\beta|_\infty \leq \frac{B}{\bar{b} D^a} \right\},$$

where D is the cardinality of Λ.

Note that any function g of $\mathcal{S}_m$ has its L_∞-norm bounded by B.
Writing the contrast of some function $g = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda$ in $\mathcal{S}_m$, we get

$$\Gamma_n(\beta) = \gamma_n\Big(\sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda\Big) = -\frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \log\left(\frac{e^{\sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda(W_i)}}{S_n(\sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda, s)} \right) dN_i(s),$$

with $S_n(\sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda, s) = \frac{1}{n} \sum_{i=1}^{n} e^{\sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda(W_i)} \mathbb{1}_{\{T_i \geq s\}}$, that is clearly continuous with respect to β. Therefore, there always exists some minimizer of Γ_n on the closed, bounded L_∞-ball with radius $B/\bar{b}D^a$. Doing so, we define a minimum contrast estimator.

Definition 1.2. *The maximum log partial likelihood estimator $\hat{f}_m$ of the regression function f on the model m is the unique function in $\mathcal{S}_m$ which minimizes $\gamma_n(g)$ over all functions g in $\mathcal{S}_m$:*

$$\hat{f}_m = \operatorname*{argmin}_{g \in \mathcal{S}_m} \gamma_n(g).$$

By definition, the estimator $\hat{f}_m$ is both an element of $\mathcal{G}_\Lambda$ and bounded by B.

1.3 Model selection

Calculating an estimator $\hat{f}_m$ for each model m, we obtain a collection of estimators $(\hat{f}_m, m \in \mathcal{M}_n)$, where $\mathcal{M}_n = \{m = (\Lambda, B), \Lambda \in \Lambda_n, B \geq 2\}$. Ideally, what we would

like to do, is to select the "best" estimator in the collection, that is to say the ideal estimator which minimizes the risk over the collection of estimators. Unfortunately, this estimator depends on the unknown function f and we are even not able to calculate theoretically the risk of the Cox estimator on one model. Therefore, we can try to calculate at least an upper bound for this risk. Usually, it is possible to get such a bound which decomposes in two terms, one of it being a bias term and representing the approximation error of the model, while the second is a variance term representing the estimation error within the model. So, our aim is now to try to select a model from the data only, which realizes the best trade-off between these two errors, and the risk of the estimator over this model should be bounded by the smallest risk bound over the collection of estimators.

1.3.1 The main theorem

Definition 1.3. *Given some penalty function* $\mathrm{pen}_n(m)$ *on* $\mathcal{M}_n$*, the penalized maximum log partial likelihood estimator is defined as:*

$$\tilde{f} = \hat{f}_{\hat{m}} \quad \textit{where} \quad \hat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \right\}. \tag{1.3.1.1}$$

The problem now consists in choosing some penalty function pen_n such that the risk of the penalized estimator $\tilde{f}$ is of the required order.

Theorem 1.3.1. *Recall Definitions 1.1, 1.2 and 1.3, Assumptions 1.1 and 1.2. Suppose that the regression function f is bounded by some known constant M. Suppose that the dimension of each model is not larger than $n/\log n$. Choose some ε in $[0,1]$ and some numbers $L_\Lambda \geq 1, \Lambda \in \Lambda_n$ such that there exists some positive number Σ verifying*

$$\sum_{\Lambda \in \Lambda_n} e^{-L_\Lambda |\Lambda|} \leq \Sigma < +\infty, \tag{1.3.1.2}$$

where $|\Lambda|$ denotes the cardinality of Λ.

There exists some absolute constant c such that if we choose some penalty function pen_n *such that*

$$\mathrm{pen}_n(m) = c' \left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^n \mathrm{II}_{\{T_i \geq \tau\}}} \right)^2 \frac{B \log(B) L_\Lambda D \log(n)}{n},$$

with $c' \geq c$ for any model $m = (\Lambda, B)$ with D the cardinality of Λ, then there exist constants C depending on c' and M, C' depending on c', C'' depending on c' and ε such that the

penalized log partial likelihood estimator $\tilde{f} = \hat{f}_{\hat{m}}$ *verifies:*

$$\mathbb{E}\left(K(f,\tilde{f})\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C \inf_{m\in\mathcal{M}_n} \inf_{\substack{f_m^* \in \mathcal{S}_m,\\ \|f_m^*\|_\infty \leq M}} \left(K(f,f_m^*) + \mathbb{E}\left(\mathrm{pen}_n(m)\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right)\right) + \frac{C'\Sigma}{n}, \quad (1.3.1.3)$$

where

$$\mathbb{P}\left(\Omega_n^C[\varepsilon]\right) \leq 2e^{-2n\varepsilon^2\,\mathbb{P}(T\geq\tau)}. \quad (1.3.1.4)$$

Furthermore, we have the control

$$\mathbb{E}\left(\mathrm{pen}_n(m)\,\mathbb{1}_{\{\Omega_n\}}[\varepsilon]\right) \leq C''\left(\frac{e^{2(B+M)}}{\mathbb{P}(T\geq\tau)}\right)^2 \frac{B\log(B)L_\Lambda D\log(n)}{n}.$$

A proof of this theorem is presented in the next section. Let us now comment this theorem.

Remark 1.1. One consequence of the proof of the theorem is that a bound of the risk of one estimator on one model m, which is a new result up to our knowledge, is given by

$$\mathbb{E}\left(K(f,\hat{f}_m)\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C \inf_{\substack{f_m^* \in \mathcal{S}_m,\\ \|f_m^*\|_\infty \leq M}} \left(K(f,f_m^*) + C''\left(\frac{e^{2(B+M)}}{\mathbb{P}(T\geq\tau)}\right)^2 \frac{BD\log(n)}{n}\right) + \frac{C'\Sigma}{n}.$$

Here, we can see the decomposition in two terms, the bias term $K(f,f_m^*)$ representing an approximation error, whereas the second term of the bound represents an estimation error. But, comparing with the density estimation framework (see Castellan [19, 20, 21]), this bound is probably too rough and we suspect in particular that the $\log n$ term is not necessary in reality and that it only appears because of technical reasons. Furthermore, we do not have any lower bound.

Remark 1.2. Choice of the weights $\{L_\Lambda, \Lambda \in \Lambda_n\}$**:**

The choice of these weights depends in fact on the complexity of the collection Λ_n and must verify the constraint 1.3.1.2. Therefore, we consider two different hypotheses for the collection of models.

Case H_1: polynomial number of models per dimension

Assumption 1.3. *There exist some integer* r *and some fixed constant* Υ *such that the number of sets of indexes* Λ *with cardinality* D *is bounded by* ΥD^r.

In this case, we are allowed to choose constant weights $L_\Lambda = L, \Lambda \in \Lambda_n$, for some $L \geq 1$, so that our aim is reached. Indeed, we can write

$$\sum_{\Lambda} e^{-L_\Lambda |\Lambda|} \leq \Upsilon \sum_{D=1}^{+\infty} D^r e^{-LD} = \Sigma < +\infty.$$

Case H_2: exponential number of models per dimension

Assumption 1.4. *The collection of sets of indexes Λ_n is composed of all subsets Λ of the set $\{1, 2, \ldots, N_n\}$, where N_n verifies $N_n \leq \frac{n}{\log n}$.*

In order to assure the constraint 1.3.1.2, we shall choose $L_\Lambda = \log n$, $\Lambda \in \Lambda_n$. Indeed, we can write

$$\sum_{\Lambda} e^{-L_\Lambda |\Lambda|} = \sum_{D=1}^{N_n} \binom{N_n}{D} e^{-\log(n) D} = \left(1 + \frac{1}{n}\right)^{N_n} - 1 \leq e - 1 = \Sigma.$$

Doing so, we loose another $\log n$ term in the risk bound, but this one seems to be necessary, since it already appeared in the density estimation framework (see Castellan [19, 20, 21]).

Remark 1.3. Given inequality 1.3.1.4, we would like to take $\varepsilon = 1$, but unfortunately, the constant C'', that depends on ε, tends to infinity when ε tends to 1. At the contrary, minimizing C'' in ε would lead to the choice $\varepsilon = 0$, but that gives a very bad bound for $\mathbb{P}\left(\Omega_n^C[\varepsilon]\right)$.

Remark 1.4. The main disadvantage in this theorem is the presence of the constant M which is supposed to be known, whereas this is not the case in reality. One way to solve this problem would be to exhibit some estimator of this constant with good properties (for instance, choosing one model with high dimension, estimating f on this model and taking the sup norm of this estimator). In practice, it seems more relevant to estimate the regression function f on non-bounded linear spaces, to choose some penalty function proportional to D/n and to try to estimate the proportionality constant (see Appendix A for an example of application).

Remark 1.5. The penalty function is not always well defined since the term $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}$ may equal zero, even if this only happens with very small probability (see inequality 1.4.1.7). Nevertheless, if this happens, it is always possible in practice to choose some τ' smaller than τ such that $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau'\}}$ is positive.

1.3.2 Applications

In this section, we apply our main theorem to the examples cited in section 1.2.4.

Covariate selection

In this application, $E = \mathbb{R}^{N_n}$, $W = (W_\lambda)_{1\leq\lambda\leq N_n}$, and we suppose that the number of covariates N_n is smaller than $n/\log n$, and that $\mathbb{E}(W_\lambda) = 0$, for $1 \leq \lambda \leq N_n$. Suppose furthermore than $|W|_\infty \leq \bar{w}$ μ-almost surely, so that

$$|\sum_{\lambda\in\Lambda} \beta_\lambda W_\lambda| \leq \bar{w} D |\beta|_\infty,$$

for any subset Λ of $\{1,\dots,N_n\}$ with cardinality D. We have now to distinguish between two cases.

Ordered covariate selection:

Suppose that there is some natural order on the coordinates $(W_\lambda)_{1\leq\lambda\leq N_n}$ of the covariate W, that is to say, that we suspect that the first coordinate W_1 should be more significant for the phenomenon we are studying than the coordinate W_2, and so on. In this case, we consider the collection of models $\mathcal{M}_n = \{(\Lambda, B),\ \Lambda \in \Lambda_n,\ B \geq 2\}$ where $\Lambda_n = \{\Lambda,\ |\Lambda| \leq N_n\}$ and Λ has the form $\{1,\dots,D\}$. The purpose is now to select the number D of significant covariates. Thus, we define

$$\begin{aligned} \mathcal{S}_m &= \left\{ g = \sum_{\lambda\in\Lambda} \beta_\lambda W_\lambda,\ |\beta|_\infty \leq \frac{B}{\bar{w} D} \right\} \\ \hat{f}_m &= \operatorname*{argmin}_{g\in\mathcal{S}_m} \gamma_n(g). \end{aligned}$$

Remark now that we are in the case of Assumption 1.3 with $\Upsilon = 1$ and $r = 0$: for a given B, there is only one model $m = (\Lambda, B)$ with dimension D. We are allowed to take constant weights $L_\Lambda = L \geq 1$. Let us now define the penalty function by

$$\mathrm{pen}_n(m) = c' \left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}} \right)^2 \frac{B\log(B) D \log(n)}{n},$$

with $c' \geq c$. Our estimator is then defined by $\tilde{f} = \hat{f}_{\hat{m}}$, where

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \right\}.$$

Complete covariate selection:

Suppose now that there is no natural order between the N_n coordinates of W anymore. The collection Λ_n is now composed of all subset Λ corresponding to the choice of D coordinates among the N_n possible coordinates, that is to say, Λ can be any subset of the set $\{1,\dots,N_n\}$. There are $\binom{N_n}{D}$ such possible choices. An estimator on one model is still defined by

$$\hat{f}_m = \operatorname*{argmin}_{g\in\mathcal{S}_m} \gamma_n(g),$$

where

$$\mathcal{S}_m = \left\{ g = \sum_{\lambda \in \Lambda} \beta_\lambda W_\lambda,\ |\beta|_\infty \leq \frac{B}{\bar{w} D} \right\}.$$

Now, since Assumption 1.4 is fulfilled, we need to take weights equal to $\log(n)$ (see Remark 1.2). The penalty function is defined by

$$\mathrm{pen}_n(m) = c' \left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}} \right)^2 \frac{B \log(B) D \log^2(n)}{n},$$

with $c' \geq c$, and our estimator is defined by

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\mathrm{argmin}} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \right\},$$

and $\tilde{f} = \hat{f}_{\hat{m}}$.

Histogram estimation

We suppose now that $E = [0,1]$, and that W is uniformly distributed on $[0,1]$. We propose here three different choices for the histograms.

Dyadic histograms

Let us first give an example with dyadic partitions: the natural basis is then the Haar basis: let us denote by $I_{j,k}$ for $j \geq 1$ and $0 \leq k \leq 2^j - 1$ the interval

$$I_{j,k} = \left[\frac{k}{2^j}, \frac{k+1}{2^j} \right[,$$

and by $\varphi_{j,k}$ for $j \geq 1$ and $0 \leq k \leq 2^{j-1} - 1$ the functions

$$\varphi_{j,k}(x) = 2^{\frac{j-1}{2}} (\mathbb{1}_{\{I_{j,2k}\}} - \mathbb{1}_{\{I_{j,2k+1}\}}).$$

We define now Λ^j as the following set of indexes

$$\Lambda^j = \{(j,k),\ 0 \leq k \leq 2^{j-1} - 1\},$$

and the collection Λ_n as

$$\Lambda_n = \left\{ \Lambda^j,\ 2^j \leq N_n \right\},$$

where we suppose that $N_n \leq \frac{n}{\log n}$. Let us consider the following linear spaces for $\Lambda^j \in \Lambda_n$,

$$\mathcal{G}_{\Lambda^j} = \left\{ \sum_{\lambda \in \Lambda^j} \beta_\lambda \varphi_\lambda,\ \beta \in \mathbb{R}^{\Lambda^j} \right\}.$$

The set $\mathcal{G}_{\Lambda^j}$ is the linear space of the piecewise constant functions based on the partition

$(I_{j,k})_{0\leq k\leq 2^j-1}$ and has dimension $D = 2^{j-1}$. Since $\mathbb{E}(\varphi_{j,k}(W)) = 0$ for any $(j,k) \in \Lambda^j$, we have $\mathbb{E}(g(W)) = 0$ for any $g \in \mathcal{G}_{\Lambda^j}$. By definition,

$$\sum_{\lambda\in\Lambda^j} \beta_\lambda \varphi_\lambda = \sum_{k=0}^{2^{j-1}-1} \beta_{j,k}\varphi_{j,k} = \sum_{k=0}^{2^{j-1}-1} \beta_{j,k} 2^{\frac{j-1}{2}} (\mathbb{1}_{\{I_{j,2k}\}} - \mathbb{1}_{\{I_{j,2k+1}\}}),$$

so that

$$\| \sum_{\lambda\in\Lambda^j} \beta_\lambda W_\lambda \|_\infty \leq 2^{\frac{j-1}{2}} |\beta|_\infty = D^{1/2} |\beta|_\infty,$$

Assumption 1.2 is thus fulfilled with $\bar{b} = 1$ and $a = 1/2$. The sets $\mathcal{S}_m$, for $m = (\Lambda^j, B)$ are now defined by

$$\mathcal{S}_m = \left\{ g \in \mathcal{G}_{\Lambda^j},\ |\beta|_\infty \leq \frac{B}{2^{\frac{j-1}{2}}} \right\},$$

and the estimators on each model by

$$\hat{f}_m = \operatorname*{argmin}_{g\in\mathcal{S}_m} \gamma_n(g).$$

Assumption 1.3 is fulfilled with $\Upsilon = 1$ and $r = 0$, since there is only one model $m = (\Lambda^j, B)$ with dimension 2^{j-1} for a given B. Thus, we are allowed to take constant weights, $L_\Lambda = L \geq 1$, and a penalty function of the form

$$\mathrm{pen}_n(m) = c' \left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}} \right)^2 \frac{B\log(B) D \log(n)}{n},$$

with $c' \geq c$. Our estimator is then defined by $\tilde{f} = \hat{f}_{\hat{m}}$, with

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \right\}.$$

Regular histograms

We consider now general regular histograms, that is to say, piecewise constant functions built on partitions where all pieces have the same size. Let us denote by $I_{D,k}$ for $D \geq 1$ and $0 \leq k \leq D$ the interval

$$I_{D,k} = \left] \frac{k}{D+1}, \frac{k+1}{D+1} \right[,$$

and by $\varphi_{D,k}$ for $D \geq 1$ and $0 \leq k \leq D$ the functions

$$\varphi_{D,k}(x) = \mathbb{1}_{\{I_{D,k}\}}(x).$$

Let us denote by Λ^D the set of indexes

$$\Lambda^D = \{(D,k), 0 \leq k \leq D\},$$

and by Λ_n the collection

$$\Lambda_n = \left\{\Lambda^D,\ |\Lambda^D| - 1 = D \leq N_n\right\},$$

where $N_n \leq \frac{n}{\log n}$, where $|\Lambda|$ denotes the cardinality of Λ. Remark that

$$\|\sum_{\lambda\in\Lambda} \beta_\lambda \varphi_\lambda\|_\infty = \|\sum_{k=0}^{D} \beta_{D,k}\, \mathbb{1}_{\{I_{D,k}\}}\|_\infty = |\beta_D|_\infty,$$

so that Assumption 1.2 is fulfilled with $\bar{b} = 1$ and $a = 0$. Let us now define by $\mathcal{S}_m$ the following set for $m = (\Lambda, B)$

$$\mathcal{S}_m = \left\{ g = \sum_{\lambda\in\Lambda} \beta_\lambda \varphi_\lambda = \sum_{k=0}^{D} \beta_{D,k}\, \mathbb{1}_{\{I_{D,k}\}},\ |\beta_D|_\infty \leq B,\ \sum_{k=0}^{D} \beta_{D,k} = 0 \right\},$$

the last constraint $\sum_{k=0}^{D} \beta_{D,k} = 0$ being necessary to impose $\mathbb{E}\left(g(W)\right) = 0$. The linear subspace $\mathcal{S}_m$ has dimension D. The estimator on the model m is defined by

$$\hat{f}_m = \operatorname*{argmin}_{g\in\mathcal{S}_m} \gamma_n(g).$$

Once again, Assumption 1.3 is fulfilled with $\Upsilon = 1$ and $r = 0$, since there is only one regular partition of $[0,1]$ with D pieces. Therefore, we can take constant weights $L_\Lambda = L \geq 1$ and the penalty function pen_n can be defined by

$$\mathrm{pen}_n(m) = c' \left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}}} \right)^2 \frac{B\log(B) D \log(n)}{n},$$

with $c' \geq c$. Our estimator is still defined by $\tilde{f} = \hat{f}_{\hat{m}}$, with

$$\hat{m} = \operatorname*{argmin}_{m\in\mathcal{M}_n} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \right\}.$$

Irregular histograms

In this last example, we consider irregular histograms, that is to say, piecewise constant functions built on partitions where pieces may have different sizes. Suppose that $N_n \leq n/\log n$. For any sequence $\bar{k} = (k_0, k_1, \ldots, k_{D+1})$ of integers such that $k_0 = 0 < k_1 < \cdots < k_{D+1} = N_n$, we define the intervals $I_{\bar{k},k_i}$ for $0 \leq i \leq D$ by

$$I_{\bar{k},k_i} = \left]\frac{k_i}{N_n+1}, \frac{k_{i+1}}{N_n+1}\right[,$$

and the functions $\varphi_{\bar{k},k_i}$ for $0 \leq i \leq D$ by

$$\varphi_{\bar{k},k_i} = \mathbb{1}_{\{I_{\bar{k},k_i}\}}.$$

The sets of indexes $\Lambda^{\bar{k}}$ can now be written as

$$\Lambda^{\bar{k}} = \{(\bar{k}, k_i),\ 0 \leq i \leq D-1\},$$

and the collection Λ_n as

$$\Lambda_n = \{\Lambda^{\bar{k}},\ |\Lambda| - 1 = |\bar{k}| - 2 \leq N_n\},$$

where $|\Lambda|$ denotes the cardinality of Λ and $|\bar{k}|$ the length of the sequence $\bar{k}$. We have

$$\| \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda \|_\infty = \| \sum_{i=0}^{D} \beta_{\bar{k},k_i} \mathbb{1}_{\{I_{\bar{k},k_i}\}} \|_\infty = |\beta_{\bar{k}}|_\infty :$$

Assumption 1.2 is fulfilled with $\bar{b} = 1$ and $a = 0$. The sets $\mathcal{S}_m$ can therefore be defined as

$$\mathcal{S}_m = \left\{ g = \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda = \sum_{i=0}^{D} \beta_{\bar{k},k_i} \mathbb{1}_{\{I_{\bar{k},k_i}\}},\ |\beta_{\bar{k}}|_\infty \leq B,\ \sum_{i=0}^{D} \beta_{\bar{k},k_i}(k_{i+1} - k_i) = 0 \right\},$$

the constraint $\sum_{i=0}^{D} \beta_{\bar{k},k_i}(k_{i+1} - k_i) = 0$ assuring that $\mathbb{E}\,(g(W)) = 0$ for any function g in $\mathcal{S}_m$. The linear subspace $\mathcal{S}_m$ has dimension D. We estimate the function f in the model m in minimizing γ_n over the set $\mathcal{S}_m$

$$\hat{f}_m = \underset{g \in \mathcal{S}_m}{\operatorname{argmin}}\, \gamma_n(g).$$

Here, Assumption 1.4 is fulfilled, so that we need to take weights L_Λ equals to $\log n$. The penalty function is therefore defined by

$$\mathrm{pen}_n(m) = c' \left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}}} \right)^2 \frac{B \log(B) D \log^2(n)}{n},$$

with some $c' \geq c$, and our estimator is defined by

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \right\},$$

and $\tilde{f} = \hat{f}_{\hat{m}}$.

1.3.3 Rate of convergence for the penalized maximum partial likelihood estimator

We consider here the framework of histogram estimation of the regression function f, like presented in section 1.2.4.

Proposition 1.3.1. *Suppose that we are in a regular case, that is to say in the framework of Assumption 1.3 with constant weights ($L_\Lambda = L$). Suppose that the regression function*

f is Hölderian of order α, i.e. there exist some positive number H and some real number α in $(0,1]$ such that, for any w_1, w_2 in $E = [0,1]$,

$$|f(w_1) - f(w_2)| \leq H|w_1 - w_2|^\alpha,$$

and that $\|f\|_\infty \leq M$.

Then, there exists some constant C''' depending on M and α such that

$$\mathbb{E}\left(K(f,\tilde{f})\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C'''(M,\alpha)H^{\frac{2}{2\alpha+1}}\left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + \frac{C'\Sigma}{n}.$$

Note that we necessarily have $M \leq H$.

Proof. Let us first recall Lemma 1.5.4:

$$K(f,g) \leq 2\,\mathbb{E}\,(N(\tau))\,\|g-f\|_\infty^2 \leq 2\|g-f\|_\infty^2.$$

Now, defining for any linear subspace Λ the function $\bar{f}_\Lambda$ by

$$\bar{f}_\Lambda = \operatorname*{argmin}_{g\in\mathcal{G}_\Lambda} \|g-f\|_\infty,$$

we have that $\|\bar{f}_\Lambda\|_\infty \leq M$, and for any model $m = (\Lambda, B)$ with $B \geq M$, $\bar{f}_\Lambda$ is in the set $\mathcal{S}_m$. Therefore, from the risk bound 1.3.1.3, we deduce

$$\mathbb{E}\left(K(f,\tilde{f})\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C \inf_{\substack{m=(\Lambda,B),\\ B\geq M}} \left(K(f,\bar{f}_\Lambda) + C''(B)\frac{L_\Lambda D\log(n)}{n}\right) + \frac{C'\Sigma}{n},$$

denoting by $C''(B)$ the constant $C''\left(\frac{e^{2(B+M)}}{\mathbb{P}(T\geq\tau)}\right)^2 B\log(B)$.

In that case, we can affirm that

$$K(f,\bar{f}_\Lambda) \leq 2\|\bar{f}_\Lambda - f\|_\infty^2 \leq 2\frac{H^2}{(2D)^{2\alpha}}.$$

For a given set Λ, it is clear that the infimum over all models $m = (\Lambda, B)$ such that $B \geq M$ is reached for $B = M$, so that

$$\mathbb{E}\left(K(f,\tilde{f})\,\mathbb{1}_{\{\Omega_n[\varepsilon]\}}\right) \leq C \inf_{D\geq 1}\left(2\frac{H^2}{(2D)^{2\alpha}} + C''(M)\frac{LD\log(n)}{n}\right) + \frac{C'\Sigma}{n},$$

the infimum being reached for D proportional to $(H^2 n/\log n)^{1/(2\alpha+1)}$. Replacing D by this expression in the previous inequality, we get the result. □

1.4 Proof of the main theorem

Let us first introduce some new notations needed in the proof. We shall call γ'_n and R_n the following empirical processes

$$\gamma'_n(g) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{e^{g(W_i)}}{S(g,s)}\right) dN_i(s), \tag{1.4.1.1}$$

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{S_n(g,s)}{S(g,s)}\right) dN_i(s). \tag{1.4.1.2}$$

Remark that we only replaced S_n by the operator S in the definition of γ_n to define γ'_n, and that for any $g \in \mathcal{G}$,

$$\gamma_n(g) = \gamma'_n(g) + R_n(g). \tag{1.4.1.3}$$

Note also that, with the previous notations, $\gamma'_n = \mathbb{P}_n(-\log A(g))$, and thus, $K(f,g) = \mathcal{E}\left[\gamma'_n(g) - \gamma'_n(f)\right]$.

1.4.1 A fundamental inequality

By definition, $K(f,\hat{f}_{\hat{m}}) = \mathcal{E}\left[\gamma'_n(\hat{f}_{\hat{m}}) - \gamma'_n(f)\right]$, where the expectation must be understood as an operator, that is to say, it only applies to the hazard contained in the empirical process γ'_n, not to the one contained in the $\hat{f}_{\hat{m}}$.

Let $m = (\Lambda, B)$ be some model that we fix for the sequel. Let f^*_m be some function in $\mathcal{S}_m$ such that $\|f^*_m\|_\infty \leq M$. We can write

$$K(f,\hat{f}_{\hat{m}}) = \mathcal{E}\left[\gamma'_n(\hat{f}_{\hat{m}}) - \gamma'_n(f^*_m)\right] + \mathcal{E}\left[\gamma'_n(f^*_m) - \gamma'_n(f)\right].$$

In the second term, we recognize $K(f,f^*_m)$. Introducing the empirical processes γ'_n, we can write

$$K(f,\hat{f}_{\hat{m}}) = K(f,f^*_m) + \gamma'_n(f^*_m) - \mathcal{E}\left[\gamma'_n(f^*_m)\right] - \gamma'_n(\hat{f}_{\hat{m}}) + \mathcal{E}\left[\gamma'_n(\hat{f}_{\hat{m}})\right] + \gamma'_n(\hat{f}_{\hat{m}}) - \gamma'_n(f^*_m).$$

Using the previous notations and the decomposition 1.4.1.3 of γ_n, we can write

$$K(f,\hat{f}_{\hat{m}}) = K(f,f^*_m) + (\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{\hat{m}})}{A(f^*_m)}\right) + \gamma_n(\hat{f}_{\hat{m}}) - \gamma_n(f^*_m) + R_n(f^*_m) - R_n(\hat{f}_{\hat{m}}).$$

Using Definitions 1.2 and 1.3,

$$\gamma_n(\hat{f}_{\hat{m}}) + \mathrm{pen}_n(\hat{m}) \leq \gamma_n(\hat{f}_m) + \mathrm{pen}_n(m) \leq \gamma_n(f^*_m) + \mathrm{pen}_n(m).$$

From this inequality, we get

$$K(f,\hat{f}_{\hat{m}}) = K(f,f^*_m) + (\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{\hat{m}})}{A(f^*_m)}\right) + \mathrm{pen}_n(m) - \mathrm{pen}_n(\hat{m}) + R_n(f^*_m) - R_n(\hat{f}_{\hat{m}}).$$

The proof consists now in the control of the terms $(\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{\hat{m}})}{A(f^*_m)}\right)$ and $R_n(f^*_m) - R_n(\hat{f}_{\hat{m}})$. In order to control these terms, we establish in the next section some exponential inequality.

1.4.2 A useful exponential inequality

Let us begin to give a theorem due to Birgé and Massart, that can be found in [12] (see [13] for another version of this theorem and a proof). We cite it here in an independent and identically distributed framework.

Theorem 1.4.1. *Let $\mathcal{F}$ be some class of real valued and measurable functions on $\mathcal{X}$. Let P_n be the empirical process $P_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and P be its expectation $P(f) = \mathbb{E}(f(X))$. Assume that there exist some positive numbers σ and b such that for any $f \in \mathcal{F}$ and all integers $k \geq 2$*

$$\mathbb{E}\left(|f(X)|^k\right) \leq \frac{k!}{2}\sigma^2 b^{k-2}. \tag{1.4.1.4}$$

Assume furthermore that for any positive number η, there exists some finite set $\mathcal{B}_\eta$ of brackets covering $\mathcal{F}$ such that, for any brackets $[g_1,g_2] \in \mathcal{B}_\eta$ and any integer $k \geq 2$,

$$\mathbb{E}\left(|g_2 - g_1|^k(X)\right) \leq \frac{k!}{2}\eta^2 b^{k-2}. \tag{1.4.1.5}$$

Let $e^{H(\eta)}$ denote the minimal cardinality of such a covering. There exists some absolute constant κ such that, for any measurable set A with $P(A) > 0$, and denoting by $\mathbb{E}^A$ the conditional expectation given A, we have

$$\mathbb{E}^A\left(\sup_{f\in\mathcal{F}}(P_n - P)(f)\right) \leq E + 7\sigma\sqrt{\frac{2}{n}\log\left(\frac{1}{P(A)}\right)} + \frac{2b}{n}\log\left(\frac{1}{P(A)}\right),$$

where

$$E = \frac{\kappa}{\sqrt{n}}\int_0^\sigma H^{1/2}(u)du + 2\frac{b+\sigma}{n}H(\sigma),$$

and κ is some absolute constant.

Note that the hypotheses of this theorem are satisfied with 1.4.1.4 and 1.4.1.5 replaced by

$$\begin{array}{rcrcl} \|f\|_\infty & \leq & 3b, & \mathbb{E}\left(|f(X)|^2\right) & \leq & \sigma^2 \\ \|g_2 - g_1\|_\infty & \leq & 3b, & \mathbb{E}\left(|g_2 - g_1|^2\right) & \leq & \eta^2 \end{array}$$

In the following, we use this theorem in order to bound $\sup_{f\in\mathcal{F}_m}(\mathbb{P}_n-\mathbb{P})(f)$ for various classes of functions $\mathcal{F}_m$ depending on the model $m=(\Lambda,B)$ and this, uniformly in m.

Proposition 1.4.1. *Let $\mathcal{F}_m$ be some class of functions*

$$\mathcal{F}_m=\{F_{g,s}(w,u,d)/g\in\mathcal{S}_m, 0\le s\le\tau\}.$$

Let P_n be some empirical process $P_n(F_{g,s})=\frac{1}{n}\sum_{i=1}^n F_{g,s}(W_i,T_i,\delta_i)$ and P be the operator expectation $P(F_{g,s})=\mathbb{E}\left(F_{g,s}(W,T,\delta)\right)$. Assume that there exists some positive number b depending on B such that for any $g\in\mathcal{S}_m$ and $0\le s\le\tau$, $\|F_{g,s}\|_\infty\le b$. Assume furthermore that for any positive number η, there exists some finite set $\mathcal{T}_\eta$ of brackets covering $\mathcal{F}_m$ such that, for any brackets $[G_l,G_u]\in\mathcal{T}_\eta$ and any integer $k\ge 2$,

$$\mathbb{E}\left(|G_u-G_l|^k\right)\le\frac{k!}{2}\eta^2b^{k-2}.$$

Assume that the log *of the minimal cardinality of such a covering $H(\eta)$, called the entropy number, is of the form $H(\eta)\le(D+c)\left(\log(b'/\eta)\right)_+$, for some constants c and b' and where $(x)_+$ denotes the positive part of x. Let $x_m, m\in\mathcal{M}_n$ be positive real numbers and ϕ_m be the function defined by*

$$\phi_m(\sigma,x)=\kappa\sigma\sqrt{\frac{D+c}{n}}\left(\sqrt{\log\frac{b'}{\min(b',\sigma)}}+c_1\right)+2\frac{b+\sigma}{n}(D+c)\left(\log\frac{b'}{\sigma}\right)_+ +7\sigma\sqrt{\frac{2x}{n}}+\frac{2bx}{n},$$

where κ and c_1 are some absolute constants. Then, for any model m and any positive number x_m, we have

$$\mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\ge\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)\le e^{-x_m}$$

$$\mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{|P_n-P|(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\ge\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)\le 2e^{-x_m}$$

Proof. The proof consists of three steps. In the first one, we simply apply Birgé and Massart's Theorem 1.4.1 to some subset of $\mathcal{F}_m$, in the second, we go from a supremum over functions which variance is bounded to a supremum of renormalized empirical processes, in the third step, we construct our exponential inequality.

First step

Let us begin with applying Birgé and Massart's theorem 1.4.1 to the subset of all

functions F in $\mathcal{F}_m$ such that $\mathbb{E}\left(F^2\right) \leq \sigma^2$. Calculating the integral of the entropy, we find

$$\begin{aligned}\int_0^\sigma H^{1/2}(u)du &\leq \int_0^{\min(b',\sigma)} \sqrt{(D+c)\log\frac{b'}{u}}du \\ &= \int_1^{+\infty} \sqrt{(D+c)\log\frac{b'v}{\min(b',\sigma)}}\frac{\min(b',\sigma)}{v^2}dv \\ &\leq \sqrt{D+c}\min(b',\sigma)\left(\sqrt{\log\frac{b'}{\min(b',\sigma)}} + \int_1^{+\infty}\frac{\sqrt{\log v}}{v^2}dv\right).\end{aligned}$$

Denoting by c_1 the universal constant $c_1 = \int_1^\infty \frac{\sqrt{\log v}}{v^2}dv$, we obtain that, for any positive σ,

$$\mathbb{E}^A\left(\sup_{F_{g,s}/P(F_{g,s}^2)\leq\sigma^2}(P_n - P)(F_{g,s})\right) \leq \phi_m\left(\sigma, \log\frac{1}{P(A)}\right),$$

where

$$\phi_m(\sigma,x) = \kappa\sigma\sqrt{\frac{(D+c)}{n}}\left(\sqrt{\log\frac{b'}{\min(b',\sigma)}} + c_1\right) + 2\frac{b+\sigma}{n}(D+c)\left(\log\frac{b'}{\sigma}\right)_+ + 7\sigma\sqrt{\frac{2x}{n}} + \frac{2bx}{n}.$$

Second step

Let σ_0 be some positive real number and for any integer j, $\sigma_j = 2^j\sigma_0$. We want now to bound

$$\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n - P)(F_{g,s})}{\sigma_0^2 + P(F_{g,s}^2)}\right).$$

The set $\mathcal{F}_m$ can be decomposed in $\mathcal{F}_m = \mathcal{B}_{\sigma_0} \cup \left(\cup_{j=0}^{+\infty}\mathcal{C}_j\right)$, where

$$\mathcal{B}_\sigma = \{F_{g,s}/P(F_{g,s}^2) \leq \sigma^2\} \text{ and } \mathcal{C}_j = \{F_{g,s}/\sigma_j^2 \leq P(F_{g,s}^2) \leq \sigma_{j+1}^2\}.$$

Trying to bound the supremum over $\mathcal{B}_{\sigma_0}$, we obtain

$$\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{B}_{\sigma_0}}\frac{(P_n - P)(F_{g,s})}{\sigma_0^2 + P(F_{g,s}^2)}\right) \leq \frac{1}{\sigma_0^2}\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{B}_{\sigma_0}}(P_n - P)(F_{g,s})\right) \leq \frac{\phi_m(\sigma_0, \log\frac{1}{P(A)})}{\sigma_0^2}.$$

To do the same over the set $\mathcal{C}_j$, we remark that on this set, the denominator is bounded lower by $P(F_{g,s}^2)$, and therefore by σ_j^2. Then, the supremum over the set $\mathcal{C}_j$ is smaller than the supremum over the ball $\mathcal{B}_{\sigma_{j+1}}$.

$$\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{C}_j}\frac{(P_n - P)(F_{g,s})}{\sigma_0^2 + P(F_{g,s}^2)}\right) \leq \frac{1}{\sigma_j^2}\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{B}_{\sigma_{j+1}}}(P_n - P)(F_{g,s})\right) \leq \frac{\phi_m(\sigma_{j+1}, \log\frac{1}{P(A)})}{\sigma_j^2}.$$

Since the supremum over all $\mathcal{F}_m$ is attained at least on one of these sets and since these bounds are positive, we can bound this supremum by the sum of the corresponding bounds.

$$\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)}{\sigma_0^2+P(F_{g,s}^2)}(F_{g,s})\right)\leq\frac{\phi_m(\sigma_0,\log\frac{1}{P(A)})}{\sigma_0^2}+\sum_{j=0}^{+\infty}\frac{\phi_m(\sigma_{j+1},\log\frac{1}{P(A)})}{\sigma_j^2}.$$

Note now that the function $\phi_m(\sigma)/\sigma$ is non-increasing over $\mathbb{R}^+$ so that, for any integer j,

$$\frac{\phi_m(\sigma_{j+1},\log\frac{1}{P(A)})}{\sigma_{j+1}}=\frac{\phi_m(2^{j+1}\sigma_0,\log\frac{1}{P(A)})}{2^{j+1}\sigma_0}\leq\frac{\phi_m(\sigma_0,\log\frac{1}{P(A)})}{\sigma_0},$$

and therefore,

$$\phi_m(\sigma_{j+1},\log\frac{1}{P(A)})\leq 2^{j+1}\phi_m(\sigma_0,\log\frac{1}{P(A)}).$$

Thus, we can bound the sum by

$$\sum_{j=0}^{+\infty}\frac{\phi_m(\sigma_{j+1},\log\frac{1}{P(A)})}{\sigma_j^2}\leq\sum_{j=0}^{+\infty}\frac{2^{j+1}\phi_m(\sigma_0,\log\frac{1}{P(A)})}{(2^j\sigma_0)^2}=\frac{\phi_m(\sigma_0,\log\frac{1}{P(A)})}{\sigma_0^2}\sum_{j=0}^{+\infty}\frac{2}{2^j},$$

and we obtain

$$\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\right)\leq\frac{5\phi_m(\sigma_0,\log\frac{1}{P(A)})}{\sigma_0^2}.$$

Note now that the class $-\mathcal{F}_m$ verifies the same hypotheses as $\mathcal{F}_m$, so that

$$\mathbb{E}^A\left(\sup_{F_{g,s}\in-\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\right)=\mathbb{E}^A\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P-P_n)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\right)\leq\frac{5\phi_m(\sigma_0,\log\frac{1}{P(A)})}{\sigma_0^2}.$$

Third step

From this inequality, we can build the following exponential inequality, using lemma 1.5.5. Thus, we can affirm that, for any $m\in\mathcal{M}_n$ and any positive x_m,

$$\begin{aligned}\mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\geq\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)&\leq e^{-x_m},\\ \mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P-P_n)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\geq\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)&\leq e^{-x_m}.\end{aligned}$$

Now, note that $|P_n-P|=\max((P_n-P),(P-P_n))$, so that

$$\sup_{F_{g,s}\in\mathcal{F}_m}\frac{|P_n-P|(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\leq\max\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)},\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P-P_n)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\right).$$

Finally, we can write

$$\mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{|P_n-P|(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\geq\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)$$
$$\leq\mathbb{P}\left(\max\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)},\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P-P_n)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\right)\geq\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)$$
$$\leq\mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P_n-P)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\geq\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)$$
$$+\mathbb{P}\left(\sup_{F_{g,s}\in\mathcal{F}_m}\frac{(P-P_n)(F_{g,s})}{\sigma_0^2+P(F_{g,s}^2)}\geq\frac{5\phi_m(\sigma_0,x_m)}{\sigma_0^2}\right)\leq 2e^{-x_m}$$

□

We turn out now to the proof of the main theorem.

1.4.3 Control of the term $(\mathbb{P}_n-\mathbb{P})\left(\log\frac{A(\hat{f}_{\hat{m}})}{A(f_m^*)}\right)$

The control of this term consists in fact in the control of $(\mathbb{P}_n-\mathbb{P})\left(\log\frac{A(\hat{f}_{m'})}{A(f_m^*)}\right)$, uniformly in $m'=(\Lambda',B')$. We denote by D' the dimension of Λ'. To realize this control, we introduce the class of functions $\mathcal{F}_{1,m'}$

$$\mathcal{F}_{1,m'}=\left\{F_g(s,w)=\log\frac{A(g)}{A(f_m^*)}(s,w),g\in\mathcal{S}_{m'}\right\},$$

and we write

$$(\mathbb{P}_n-\mathbb{P})\left(\log\frac{A(\hat{f}_{m'})}{A(f_m^*)}\right)=\frac{(\mathbb{P}_n-\mathbb{P})(F_{\hat{f}_{m'}})}{\sigma_1^2+\mathbb{P}(F_{\hat{f}_{m'}}^2)}\left(\sigma_1^2+\mathbb{P}(F_{\hat{f}_{m'}}^2)\right)$$
$$\leq\sup_{F_g\in\mathcal{F}_{1,m'}}\frac{(\mathbb{P}_n-\mathbb{P})(F_g)}{\sigma_1^2+\mathbb{P}(F_g^2)}\left(\sigma_1^2+\mathbb{P}(F_{\hat{f}_{m'}}^2)\right),$$

since $\hat{f}_{m'}$ is in $\mathcal{S}_{m'}$.

Remark 1.6. For any function $F_g\in\mathcal{F}_{1,m'}$ with $g=\sum_{\lambda\in\Lambda'}\beta_\lambda\varphi_\lambda$, $\|F_g\|_\infty\leq 2\|g-f_m^*\|_\infty$. Indeed, we can write

$$F_g=\log\frac{A(g)}{A(f_m^*)}=g-f_m^*-\log\frac{\mathcal{E}\left[e^{g-f_m^*+f_m^*}\mathbb{1}_{\{T\geq s\}}\right]}{\mathcal{E}\left[e^{f_m^*}\mathbb{1}_{\{T\geq s\}}\right]}.$$

Bounding $e^{g-f_m^*}$ up by $e^{\|g-f_m^*\|_\infty}$ and down by $e^{-\|g-f_m^*\|_\infty}$, we find

$$e^{-\|g-f_m^*\|_\infty}\leq\frac{\mathcal{E}\left[e^{g-f_m^*+f_m^*}\mathbb{1}_{\{T\geq s\}}\right]}{\mathcal{E}\left[e^{f_m^*}\mathbb{1}_{\{T\geq s\}}\right]}\leq e^{\|g-f_m^*\|_\infty},$$

and the result follows. This bound will be used repeatedly in the sequel.

Now, for any F_{g^1} and F_{g^2} functions of $\mathcal{F}_{1,m'}$, with $g^1 = \sum_{\lambda\in\Lambda'}\beta^1_\lambda\varphi_\lambda$ and $g^2 = \sum_{\lambda\in\Lambda'}\beta^2_\lambda\varphi_\lambda$, we can prove in the same manner that

$$\begin{aligned}\|F_g\|_\infty &\leq 2\|g - f^*_m\|_\infty,\\ \|F_{g^1} - F_{g^2}\|_\infty &\leq 2\bar{b}D'^a|\beta^1_\lambda - \beta^2_\lambda|_\infty,\\ \mathbb{P}\left(F_{g^1} - F_{g^2}\right) &\leq 4\bar{b}^2D'^{2a}|\beta^1_\lambda - \beta^2_\lambda|^2_\infty.\end{aligned}$$

In order to obtain a L_∞ covering of the class $\mathcal{F}_{1,m'}$, we need to cover the L_∞-ball with radius $B'/\bar{b}(D')^a$ with L_∞-balls with radius $\eta/2\bar{b}(D')^a$. The L_2-L_∞-entropy number of the class $\mathcal{F}_{1,m'}$ is $H_1(\eta) = D'(\log\frac{2B'}{\eta})_+$. Thus, we can apply Proposition 1.4.1 with the operators $P_n = \mathbb{P}_n$, $P = \mathbb{P}$, and the constants $b = 2(B'+M)$, $c = 0$, $b' = 2B'$ and $\phi = \phi_{1,m'}$ where

$$\begin{aligned}\phi_{1,m'}(\sigma,x) = \kappa\sigma\sqrt{\frac{D'}{n}}\left(\sqrt{\log\frac{2B'}{\min(2B',\sigma)}} + c_1\right)\\ + 2\frac{2(B'+M)+\sigma}{n}D'\left(\log\frac{2B'}{\sigma}\right)_+ + 7\sigma\sqrt{\frac{2x}{n}} + \frac{4(B'+M)x}{n}.\end{aligned}$$

Let $x_{m'}, m' \in \mathcal{M}_n$ be positive real numbers such that $\sum_{m'\in\mathcal{M}_n} e^{-x_{m'}}$ is finite and let Ω_1 be $\Omega_1 = \cup_{m'}\Omega_{1,m'}(x_{m'})$, where $\Omega_{1,m'}(x_{m'})$ denotes the following event

$$\Omega_{1,m'}(x_{m'}) = \left\{\sup_{\mathcal{F}_{1,m'}}\frac{(\mathbb{P}_n - \mathbb{P})(F_g)}{\sigma_1^2 + \mathbb{P}(F_g^2)} \geq \frac{5\phi_{1,m'}(\sigma_1,x_{m'})}{\sigma_1^2}\right\}.$$

Then $P(\Omega_1) \leq \sum_{m'\in\mathcal{M}_n} e^{-x_{m'}}$. On Ω_1^C, we can affirm that, for all m',

$$(\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right) \leq \frac{5\phi_{1,m'}(\sigma_1,x_{m'})}{\sigma_1^2}\left(\sigma_1^2 + \mathbb{P}\left(\log^2\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right)\right). \qquad (1.4.1.6)$$

The quantity $\mathbb{P}\left(\log^2\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right)$ plays the role of a variance term and can be link to the Kullback-Leibler information in the following way.

$$\mathbb{P}\left(\log^2\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right) \leq 2\left(\mathbb{P}\left(\log^2\frac{A(\hat{f}_{m'})}{A(f)}\right) + \mathbb{P}\left(\log^2\frac{A(f)}{A(f^*_m)}\right)\right).$$

Now, note that $\|\hat{f}_{m'} - f\|_\infty \leq B' + M$ and $\|f - f^*_m\|_\infty \leq 2M$. Using lemma 1.5.1, we get

$$\mathbb{P}\left(\log^2\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right) \leq 4e^{2(B'+M)}K(f,\hat{f}_{m'}) + 4e^{4M}K(f,f^*_m),$$

and factorizing $4e^{2(B'+M)}$ in 1.4.1.6,

$$(\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right) \leq \frac{20e^{2(B'+M)}\phi_{1,m'}(\sigma_1, x_{m'})}{\sigma_1^2} \left(\frac{e^{-2(B'+M)}}{4}\sigma_1^2 + K(f, \hat{f}_{m'}) + e^{2(M-B')}K(f, f^*_m)\right).$$

Now, we want to choose $x_{m'}$ and σ_1 in such a way that $\frac{20e^{2(B'+M)}\phi_{1,m'}(\sigma_1,x_{m'})}{\sigma_1^2}$ is smaller than some universal constant $\theta_1 < 1$, that will be choosen later. To do so, we set $x_{m'} = \xi + L_{\Lambda'}D' + 2\log(B')$ and

$$\sigma_1^2 = C_1^2 e^{2(B'+M)}(B'+M)\left(\frac{\xi}{n} + \frac{\log(B')L_{\Lambda'}D'\log(n)}{n}\right),$$

for some positive constant C_1. So, it is possible to bound the quantity $20e^{2(B'+M)}\frac{\phi_{1,m'}(\sigma_1,x_{m'})}{\sigma_1^2}$ by $\frac{K_1}{C_1}$ where K_1 is some universal constant. Just choose $C_1 = \frac{K_1}{\theta_1}$ to conclude.

Thus, on the set Ω_1, we have for any model m'

$$(\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{m'})}{A(f^*_m)}\right) \leq \frac{K_1^2}{4\theta_1}(B'+M)\log(B')\frac{L_{\Lambda'}D'\log(n)}{n} + \theta_1 K(f, \hat{f}_{m'}) + \theta_1 e^{2M}K(f, f^*_m) + \frac{K_1^2(B'+M)\xi}{4\theta_1 n}.$$

Using the fact that $2(B'+M)\xi \leq (B'+M)^2 + \xi^2$ and applying the previous inequality for $m' = \hat{m}$, we get

$$(\mathbb{P}_n - \mathbb{P})\left(\log\frac{A(\hat{f}_{\hat{m}})}{A(f^*_m)}\right) \leq \frac{K_1^2}{4\theta_1}(\hat{B}+M)\log\hat{B}\left(\frac{L_{\hat{m}}\hat{D}\log(n)}{n} + \frac{\hat{B}+M}{2n}\right) + \theta_1 K(f, \hat{f}_{\hat{m}}) + \theta_1 e^{2M}K(f, f^*_m) + \frac{K_1^2\xi^2}{8\theta_1 n}.$$

where $\hat{m} = (\hat{\Lambda}, \hat{B})$ and $\hat{D}$ is the cardinality of $\hat{\Lambda}$.

1.4.4 Control of the term $R_n(f^*_m) - R_n(\hat{f}_{\hat{m}})$

Like previously, we will control the term $R_n(f^*_m) - R_n(\hat{f}_{m'})$, uniformly in $m' = (\Lambda', B')$. Rewriting the term $R_n(f^*_m) - R_n(\hat{f}_{m'})$, we have

$$R_n(f^*_m) - R_n(\hat{f}_{m'}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\log\frac{S_n(f^*_m, s)}{S(f^*_m, s)} - \log\frac{S_n(\hat{f}_{m'}, s)}{S(\hat{f}_{m'}, s)}\right) dN_i(s).$$

We begin with linearizing the log, then we control the difference

$$\frac{S_n(f_m^*, s)}{S(f_m^*, s)} - \frac{S_n(\hat{f}_{m'}, s)}{S(\hat{f}_{m'}, s)}$$

uniformly in $m' \in \mathcal{M}_n$ and in $0 \leq s \leq \tau$, applying Proposition 1.4.1 and finally, we conclude applying this Proposition once again to another class of functions.

Linearisation

Let ε be some real number such that $0 < \varepsilon < 1$ and let define the event $\Omega[\varepsilon]$ by

$$\Omega[\varepsilon] = \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}} - \mathbb{P}(T \geq \tau) \right| \leq \varepsilon \, \mathbb{P}(T \geq \tau) \right\}.$$

A classical result on the binomial law tells us that

$$\mathbb{P}(\Omega[\varepsilon]^C) \leq 2e^{-2n\varepsilon^2 \mathbb{P}(T \geq \tau)^2}. \tag{1.4.1.7}$$

On the set $\Omega[\varepsilon]$, $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}} \geq (1-\varepsilon)\,\mathbb{P}(T \geq \tau) > 0$. Using that for any positive real numbers x and y,

$$\log x - \log y \leq \frac{(x-y)_+}{y} \leq \frac{|x-y|}{y}$$

and that, on the set $\Omega[\varepsilon]$, for any $0 \leq s \leq \tau$,

$$\frac{S_n}{S}(\hat{f}_{m'}, s) \geq e^{-2B'} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}} > 0.$$

Finally, on $\Omega[\varepsilon]$, we can write that

$$\begin{aligned}(R_n(f_m^*) - R_n(\hat{f}_{m'}))\, &\mathbb{1}_{\{\Omega[\varepsilon]\}} \\ &\leq \frac{e^{2B'}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}}} \mathbb{1}_{\{\Omega[\varepsilon]\}} \frac{1}{n} \sum_{i=1}^{n} \int_0^\tau \left| \frac{S_n}{S}(f_m^*, s) - \frac{S_n}{S}(\hat{f}_{m'}, s) \right| dN_i(s).\end{aligned}$$

Uniform control of the term $\left|\frac{S_n}{S}(\hat{f}_{m'}, s) - \frac{S_n}{S}(f_m^*, s)\right|$

As previously, we will control

$$\left| \frac{S_n}{S}(\hat{f}_{m'}, s) - \frac{S_n}{S}(f_m^*, s) \right| = \left| \frac{1}{n} \sum_{i=1}^{n} A(\hat{f}_{m'})(W_i, s)\, \mathbb{1}_{\{T_i \geq s\}} - A(f_m^*)(W_i, s)\, \mathbb{1}_{\{T_i \geq s\}} \right|,$$

uniformly in m', but we also need a uniform control in $s \in [0, \tau]$, in order to integrate this difference with respect to the counting processes N_i. Let us define the empirical process

Q_n for any function h from $\mathcal{G} \times [0,\tau]$ by

$$Q_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(W_i, T_i),$$

and let Q denote its expectation operator $Q(h) = \mathcal{E}[h(W,T)]$. Let us consider the class of functions

$$\mathcal{F}_{2,m'} = \left\{F_{g,r}(w,v) = A(g)(w,r)\,\mathbb{1}_{\{v\geq r\}} - A(f_m^*)(w,r)\,\mathbb{1}_{\{v\geq r\}}, g \in \mathcal{S}_{m'}, 0 \leq r \leq \tau\right\}$$

Since $\mathbb{E}\left(A(g)(W,r)\,\mathbb{1}_{\{T\geq r\}}\right) = 1$ for any deterministic function g in $\mathcal{G}$ and any $0 \leq r \leq \tau$, we can write, using these notations, that for any model m' and any $0 \leq s \leq \tau$,

$$\left|\frac{S_n}{S}(\hat{f}_{m'}, s) - \frac{S_n}{S}(f_m^*, s)\right| = |Q_n - Q|(F_{\hat{f}_{m'},s}) = \frac{|(Q_n - Q)(F_{\hat{f}_{m'},r})|}{\sigma_2^2 + Q(F^2_{\hat{f}_{m'},r})}\left(\sigma_2^2 + Q(F^2_{\hat{f}_{m'},s})\right)$$
$$\leq \sup_{F_{g,r}\in\mathcal{F}_{2,m'}} \frac{|(Q_n - Q)(F_{g,r})|}{\sigma_2^2 + Q(F^2_{g,r})}\left(\sigma_2^2 + Q(F^2_{\hat{f}_{m'},s})\right).$$

Let us now control the term

$$\sup_{F_{g,r}\in\mathcal{F}_{2,m'}} \frac{|(Q_n - Q)(F_{g,r})|}{\sigma_2^2 + Q(F^2_{g,r})},$$

applying Proposition 1.4.1.

Bounding the functions

For any function $F_{g,r}$ in $\mathcal{F}_{2,m'}$, we have

$$|F_{g,r}(w,v)| = \left|A(g)(w,r)\,\mathbb{1}_{\{v\geq r\}} - A(f_m^*)(w,r)\,\mathbb{1}_{\{v\geq r\}}\right|$$
$$\leq \frac{e^{2B'}}{\mathbb{P}(T\geq\tau)} + \frac{e^{2M}}{\mathbb{P}(T\geq\tau)} \leq \frac{2e^{2(B'+M)}}{\mathbb{P}(T\geq t)} = b_2,$$

so that $\|F_{g,s}\|_\infty \leq b_2$.

Construction of the brackets

Given some function $F_{g,r}$, some functions g_1 and g_2 in $\mathcal{S}_{m'}$ such that $g_1 \leq g \leq g_2$ and some positive real number r_1 and r_2 such that $r_1 \geq r \geq r_2$ (and therefore $\mathbb{1}_{\{v\geq r_1\}} \leq \mathbb{1}_{\{v\geq r\}} \leq \mathbb{1}_{\{v\geq r_2\}}$), it is possible to bound $F_{g,r}$ lower and upper by

$$G_l(w,v) = \frac{e^{g_1(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_2,r_2)} - \frac{e^{f_m^*(w)}\,\mathbb{1}_{\{v\geq r_2\}}}{S(f_m^*,r_1)} \leq F_{g,r}(w,v)$$
$$\leq \frac{e^{g_2(w)}\,\mathbb{1}_{\{v\geq r_2\}}}{S(g_1,r_1)} - \frac{e^{f_m^*(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(f_m^*,r_2)} = G_u(w,v).$$

In order to apply Proposition 1.4.1, we have now to bound $\mathcal{E}\left[|G_u - G_l|^2\right]$. The difference $G_u - G_l$ can be decomposed as $G_u - G_l = A_1 + A_2 + A_3 + A_4 + A_5 + A_6$ where

$$A_1 = \frac{e^{g_2(w)}\,\mathbb{1}_{\{v\geq r_2\}}}{S(g_1,r_1)} - \frac{e^{g_2(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_1,r_1)} \qquad A_2 = \frac{e^{g_2(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_1,r_1)} - \frac{e^{g_1(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_1,r_1)}$$

$$A_3 = \frac{e^{g_1(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_1,r_1)} - \frac{e^{g_1(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_2,r_1)} \qquad A_4 = \frac{e^{g_1(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_2,r_1)} - \frac{e^{g_1(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(g_2,r_2)}$$

$$A_5 = \frac{e^{f_m^*(w)}\,\mathbb{1}_{\{v\geq r_2\}}}{S(f_m^*,r_1)} - \frac{e^{f_m^*(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(f_m^*,r_1)} \qquad A_6 = \frac{e^{f_m^*(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(f_m^*,r_1)} - \frac{e^{f_m^*(w)}\,\mathbb{1}_{\{v\geq r_1\}}}{S(f_m^*,r_2)}$$

In A_1 and A_5, the difference between the two indicators can be rewritten as $\mathbb{1}_{\{v\geq r_2\}} - \mathbb{1}_{\{v\geq r_1\}} = \mathbb{1}_{\{r_2\leq v<r_1\}}$, whereas in all other terms, $\mathbb{1}_{\{v\geq r_1\}}$ can be factorized. Therefore,

$$\mathcal{E}\left[|G_u - G_l|^2\right] = \mathcal{E}\left[|A_1 + A_5|^2\right] + \mathcal{E}\left[|A_2 + A_3 + A_4 + A_6|^2\right].$$

Now, using the fact that

$$\Big|\sum_{i=1}^{I} A_i\Big|^2 \leq (I \sup_i |A_i|)^2 = I^2 \sup_i |A_i|^2 \leq I^2 \sum_{i=1}^{I} |A_i|^2,$$

we have that

$$\mathcal{E}\left[|G_u - G_l|^2\right] \leq 4\,\mathcal{E}\left[|A_1|^2 + |A_5|^2\right] + 16\,\mathcal{E}\left[|A_2|^2 + |A_3|^2 + |A_4|^2 + |A_6|^2\right],$$

and we only have to bound each of the terms $\mathcal{E}\left[|A_j|^2\right]$.

$$\mathcal{E}\left[|A_1|^2\right] = \mathcal{E}\left[\left|\tfrac{e^{g_2}}{S(g_1,r_1)}\,\mathbb{1}_{\{r_2\leq T<r_1\}}\right|^2\right] \leq \left(\tfrac{e^{2B'}}{\mathbb{P}(T\geq t)}\right)^2 \mathbb{P}\left(r_2 \leq T < r_1\right)$$

$$\mathcal{E}\left[|A_2|^2\right] = \mathcal{E}\left[\left|\tfrac{(e^{g_2}-e^{g_1})\,\mathbb{1}_{\{T\geq r_1\}}}{S(g_1,r_1)}\right]\right|^2 \leq \left(\tfrac{e^{2B'}}{\mathbb{P}(T\geq \tau)}\right)^2 \|g_2 - g_1\|_\infty^2$$

$$\mathcal{E}\left[|A_3|^2\right] = \mathcal{E}\left[\left|e^{g_1}\,\mathbb{1}_{\{T\geq r_1\}}\left(\tfrac{1}{S(g_1,r_1)} - \tfrac{1}{S(g_2,r_1)}\right)\right]\right|^2 \leq \left(\tfrac{e^{2B'}}{\mathbb{P}(T\geq \tau)}\right)^4 \|g_2 - g_1\|_\infty^2$$

$$\mathcal{E}\left[|A_4|^2\right] = \mathcal{E}\left[\left|e^{g_1}\,\mathbb{1}_{\{T\geq r_1\}}\left(\tfrac{1}{S(g_2,r_1)} - \tfrac{1}{S(g_2,r_2)}\right)\right]\right|^2 \leq \left(\tfrac{e^{2B'}}{\mathbb{P}(T\geq \tau)}\right)^4 \mathbb{P}\left(r_2 \leq T < r_1\right)$$

$$\mathcal{E}\left[|A_5|^2\right] = \mathcal{E}\left[\left|\tfrac{e^{f_m^*}}{S(f_m^*,r_1)}\,\mathbb{1}_{\{r_2\leq T<r_1\}}\right|^2\right] \leq \left(\tfrac{e^{2M}}{\mathbb{P}(T\geq \tau)}\right)^2 \mathbb{P}\left(r_2 \leq T < r_1\right)$$

$$\mathcal{E}\left[|A_6|^2\right] = \mathcal{E}\left[\left|e^{f_m^*}\,\mathbb{1}_{\{T\geq r_1\}}\left(\tfrac{1}{S(f_m^*,r_1)} - \tfrac{1}{S(f_m^*,r_2)}\right)\right]\right|^2 \leq \left(\tfrac{e^{2M}}{\mathbb{P}(T\geq \tau)}\right)^4 \mathbb{P}\left(r_2 \leq T < r_1\right)$$

Finally, adding all these terms, we can bound $\mathcal{E}\left[|G_u - G_l|^2\right]$ by

$$\mathcal{E}\left[|G_u - G_l|^2\right] \leq 32\left(\frac{e^{2B'}}{\mathbb{P}(T \geq \tau)}\right)^4 \|g_2 - g_1\|_\infty^2 + 40\left(\frac{e^{2(B'+M)}}{\mathbb{P}(T \geq t)}\right)^4 \mathbb{P}(r_2 \leq T < r_1)$$
$$\leq 40\left(\frac{e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)}\right)^4 \left(\|g_2 - g_1\|_\infty^2 + \mathbb{P}(r_2 \leq T < r_1)\right).$$

Calculation of the entropy number

In order to obtain a bracketing covering of the set $\mathcal{F}_{2,m'}$, we need to cover, on one hand, the L_∞-ball with radius $B'/\bar{b}(D')^a$ with L_∞-balls with radius $\eta/2b_2'\bar{b}(D')^a$, where $b_2' = \sqrt{40}\left(\frac{e^{2(B'+M)}}{\mathbb{P}(T\geq\tau)}\right)^2$, and on the other hand, we need to cover the set of indicators of the form $\mathbb{1}_{\{[r,+\infty[\}}$ with $L_2(\mathbb{P})$-balls with radius $\eta/2b_2'$. The entropy number with bracketing of such a set can be found in Van der Vaart and Wellner [59]. The bracketing entropy number H_2 corresponding to such a covering is given by

$$H_2(\eta) \leq \left(D' \log \frac{\frac{B'}{\bar{b}(D')^a}}{\frac{\eta}{2b_2'\bar{b}(D')^a}} + \log \frac{2}{\frac{\eta^2}{4{b_2'}^2}}\right)_+ \leq (D'+2)\left(\log \frac{4B'b_2'}{\eta}\right)_+.$$

Application of Proposition 1.4.1

We can now apply Proposition 1.4.1 with $P_n = Q_n$, $P = Q$, $b = b_2 = \frac{2e^{2(B'+M)}}{\mathbb{P}(T\geq\tau)}$, $c = 2$, $b' = 4B'b_2' = 8\sqrt{10}B'\left(\frac{e^{2(B'+M)}}{\mathbb{P}(T\geq\tau)}\right)^2$ and $\phi = \phi_{2,m'}$,

$$\phi_{2,m'}(\sigma, x) = \kappa\sigma\sqrt{\frac{D'+2}{n}}\left(\sqrt{\log \frac{4B'b_2'}{\min(4B'b_2', \sigma)}} + c_1\right)$$
$$+ 2\frac{b_2 + \sigma}{n}(D'+2)\left(\log \frac{4B'b_2'}{\sigma}\right)_+ + 7\sigma\sqrt{\frac{2x}{n}} + \frac{2b_2 x}{n}.$$

Let Ω_2 be $\Omega_2 = \cup_{m'}\Omega_{2,m'}(x_{m'})$, where $\Omega_{2,m'}(x_{m'})$ denotes the following event

$$\Omega_{2,m'}(x_{m'}) = \left\{\sup_{F_{g,r}\in\mathcal{F}_{2,m'}} \frac{|(Q_n - Q)(F_{g,r})|}{\sigma_2^2 + Q(F_{g,r}^2)} \geq \frac{5\phi_{2,m'}(\sigma_2, x_{m'})}{\sigma_2^2}\right\}.$$

Then, $\mathbb{P}(\Omega_2) \leq 2\sum_{m'\in\mathcal{M}_n} e^{-x_{m'}}$. On Ω_2^C, we can affirm that, for any m' in $\mathcal{M}_n$ and any $0 \leq s \leq \tau$,

$$\left|\frac{S_n}{S}(f_m^*, s) - \frac{S_n}{S}(\hat{f}_{m'}, s)\right| \leq \frac{5\phi_{2,m'}(\sigma_2, x_{m'})}{\sigma_2^2}\left(\sigma_2^2 + Q(F^2_{\hat{f}_{m'},s})\right),$$

and in particular, this is true for $m' = \hat{m} = (\hat{\Lambda}, \hat{B})$.

Thus, on the same set Ω_2^C, we have

$$(R_n(f_m^*) - R_n(\hat{f}_{m'}))\, \mathbb{1}_{\{\Omega[\varepsilon]\}} \leq \frac{e^{2B'}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}} \mathbb{1}_{\{\Omega[\varepsilon]\}} \left(\frac{5\phi_{2,m'}(\sigma_2, x_{m'})}{\sigma_2^2} \left(\sigma_2^2 + \frac{1}{n}\sum_{i=1}^n \int_0^\tau Q(F^2_{\hat{f}_{m'},s}) dN_i(s) \right) \right).$$

It remains to control, uniformly in m', the terms $\frac{1}{n}\sum_{i=1}^n \int_0^\tau Q(F^2_{\hat{f}_{m'},s}) dN_i(s)$, that we can write

$$(\mathbb{P}_n - \mathbb{P})(Q(F^2_{\hat{f}_{m'},s})) + \mathbb{P}(Q(F^2_{\hat{f}_{m'},s})).$$

Control of the term $(\mathbb{P}_n - \mathbb{P})(Q(F^2_{\hat{f}_{m'},s}))$

We want to apply once again Proposition 1.4.1.

Let us consider the class of functions

$$\mathcal{F}_{3,m'} = \left\{ G_g(s) = \mathcal{E}\left[(A(g)(W,s) - A(f_m^*)(W,s))^2 \mathbb{1}_{\{T \geq s\}} \right], g \in \mathcal{S}_{m'} \right\}.$$

For any functions G_{g^1} and G_{g^2} of $\mathcal{F}_{3,m'}$, we have

$$\|G_{g^1}\|_\infty \leq \left(\frac{2e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)} \right)^2, \tag{1.4.1.8}$$

using the same techniques as in Remark 6.

$$\begin{aligned} |G_{g^2} - G_{g^1}| &= \left| \mathcal{E}\left[\left(A(g^2) - A(f_m^*)\right)^2 \mathbb{1}_{\{T \geq s\}} - \left(A(g^1) - A(f_m^*)\right)^2 \mathbb{1}_{\{T \geq s\}} \right] \right| \\ &\leq \mathcal{E}\left[\left| \left(A(g^1) + A(g^2) - 2A(f_m^*)\right) \left(A(g^2)\mathbb{1}_{\{T \geq s\}} - A(g^1)\mathbb{1}_{\{T \geq s\}}\right) \right| \right] \\ &\leq \frac{4e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)} \mathcal{E}\left[\left| A(g^2)\mathbb{1}_{\{T \geq s\}} - A(g^1)\mathbb{1}_{\{T \geq s\}} \right| \right] \end{aligned}$$

so that

$$\|G_{g^2} - G_{g^1}\|_\infty \leq 8 \left(\frac{e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)} \right)^2 \|g^2 - g^1\|_\infty.$$

Setting $b_3 = 8\left(\frac{e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)}\right)^2$, we can deduce the following inequality from the previous one,

$$\mathbb{P}((G_{g^2} - G_{g^1})^2) \leq b_3^2 \|g^2 - g^1\|_\infty^2.$$

The entropy number of the set $\mathcal{F}_{3,m'}$ is then given by $H_3(\eta) \leq D'\left(\log \frac{B'b_3}{\eta}\right)_+$. Thus, we can apply Proposition 1.4.1 with $P_n = \mathbb{P}_n$, $P = \mathbb{P}$, $b = b_3$, $c = 0$, $b' = B'b_3$ and $\phi = \phi_{3,m'}$

given by

$$\phi_{3,m'}(\sigma,x)=\kappa\sigma\sqrt{\frac{D'}{n}}\left(\sqrt{\log\frac{B'b_3}{\min(B'b_3,\sigma)}}+c_1\right)$$
$$+2\frac{b_3+\sigma}{n}D'\left(\log\frac{B'b_3}{\sigma}\right)_+ +7\sigma\sqrt{\frac{2x}{n}}+\frac{2b_3x}{n}.$$

Let Ω_3 be $\Omega_3=\cup_{m'}\Omega_{3,m'}(x_{m'})$, where $\Omega_{3,m'}(x_{m'})$ denotes the following event

$$\Omega_{3,m'}(x_{m'})=\left\{\sup_{\mathcal{F}_{3,m'}}\frac{(\mathbb{P}_n-\mathbb{P})(G_g)}{\sigma_3^2+\mathbb{P}(G_g^2)}\geq\frac{5\phi_{3,m'}(\sigma_3,x_{m'})}{\sigma_3^2}\right\}.$$

Then $P(\Omega_3)\leq\sum_{m'\in\mathcal{M}_n}e^{-x_{m'}}$. On Ω_3^C, we can affirm that, for all m',

$$(\mathbb{P}_n-\mathbb{P})\left(G_{\hat{f}_{m'}}\right)\leq\frac{5\phi_{3,m'}(\sigma_3,x_{m'})}{\sigma_3^2}\left(\sigma_3^2+\mathbb{P}\left(G_{\hat{f}_{m'}}^2\right)\right).$$

Conclusion

On $\Omega_2^C\cap\Omega_3^C$, we can write

$$\left(R_n(f_m^*)-R_n(\hat{f}_{m'})\right)\mathbb{1}_{\{\Omega[\varepsilon]\}}\leq\frac{e^{2B'}}{\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{\{T_i\geq\tau\}}}\mathbb{1}_{\{\Omega[\varepsilon]\}}\frac{5\phi_{2,m'}(\sigma_2,x_{m'})}{\sigma_2^2}$$
$$\left(\sigma_2^2+\frac{5\phi_{3,m'}(\sigma_3,x_{m'})}{\sigma_3^2}\left(\sigma_3^2+\mathbb{P}\left(G_{\hat{f}_{m'}}^2\right)\right)+\mathbb{P}(Q(F_{\hat{f}_{m'},s}^2))\right).$$

We have now to retrieve the Kullback-Leibler informations from the variances terms. For any deterministic function g in $\mathcal{S}_{m'}$, we have

$$\begin{aligned}G_g&=\mathcal{E}\left[(A(g)(W,s)-A(f_m^*)(W,s))^2\,\mathbb{1}_{\{T\geq s\}}\right]\\&=\mathcal{E}\left[\left(\sqrt{A(g)(W,s)}+\sqrt{A(f_m^*)(W,s)}\right)^2\right.\\&\qquad\left.\left(\sqrt{A(g)(W,s)}-\sqrt{A(f_m^*)(W,s)}\right)^2\mathbb{1}_{\{T\geq s\}}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}G_g&\leq\frac{2e^{2(B'+M)}}{\mathbb{P}(T\geq\tau)}\mathcal{E}\left[\left(\sqrt{A(g)(W,s)}-\sqrt{A(f_m^*)(W,s)}\right)^2\mathbb{1}_{\{T\geq s\}}\right]\\&\leq\frac{2e^{2(B'+M)}}{\mathbb{P}(T\geq\tau)}\left(\mathcal{E}\left[2\left(\sqrt{A(g)(W,s)}-\sqrt{A(f)(W,s)}\right)^2\mathbb{1}_{\{T\geq s\}}\right]\right.\\&\qquad\left.+2\mathcal{E}\left[\left(\sqrt{A(f_m^*)(W,s)}-\sqrt{A(f)(W,s)}\right)^2\mathbb{1}_{\{T\geq s\}}\right]\right)\end{aligned}$$

Thus, using 1.4.1.8 and bounding one G_g by its L_∞-norm,

$$\mathbb{P}\left(G_g^2\right) \leq \left(\frac{2e^{2(B'+M)}}{\mathbb{P}\left(T \geq \tau\right)}\right)^3 \mathcal{E}\left[\int_0^\tau \left(2\,\mathcal{E}\left[\left(\sqrt{A(g)(W,s)} - \sqrt{A(f)(W,s)}\right)^2 \mathbb{1}_{\{T\geq s\}}\right]\right.$$
$$\left. +2\,\mathcal{E}\left[\left(\sqrt{A(f_m^*)(W,s)} - \sqrt{A(f)(W,s)}\right)^2 \mathbb{1}_{\{T\geq s\}}\right]\right) A(f)S(f,s)\,\mathbb{1}_{\{T\geq s\}}\,\alpha_0(s)ds\Bigg].$$

Since $\mathcal{E}\left[\left(\sqrt{A(g)(W,s)} - \sqrt{A(f)(W,s)}\right)^2 \mathbb{1}_{\{T\geq s\}}\right]$ is deterministic, the first expectation only applies to $A(f)\,\mathbb{1}_{\{T\geq s\}}$, which expectation equals 1, so that

$$\mathbb{P}\left(G_g^2\right) \leq \left(\frac{2e^{2(B'+M)}}{\mathbb{P}\left(T \geq \tau\right)}\right)^3$$
$$\left(2\,\mathcal{E}\left[\int_0^\tau \left(\sqrt{A(g)(W,s)} - \sqrt{A(f)(W,s)}\right)^2 \mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right]\right.$$
$$\left. +2\,\mathcal{E}\left[\int_0^\tau \left(\sqrt{A(f_m^*)(W,s)} - \sqrt{A(f)(W,s)}\right)^2 \mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right]\right).$$

We recognize the Hellinger distance that we can bound by one half of the Kullback-Leibler information (see Lemma 1.5.2)

$$\mathbb{P}\left(G_g^2\right) \leq \left(\frac{2e^{2(B'+M)}}{\mathbb{P}\left(T \geq \tau\right)}\right)^3 \left(K(f,g) + K(f,f_m^*)\right).$$

Since $\hat{f}_{m'}$ is also in $\mathcal{S}_{m'}$, we can also write

$$\mathbb{P}\left(G_{\hat{f}_{m'}}^2\right) \leq \left(\frac{2e^{2(B'+M)}}{\mathbb{P}\left(T \geq \tau\right)}\right)^3 \left(K(f,\hat{f}_{m'}) + K(f,f_m^*)\right).$$

In the same manner, it is easy to prove that

$$\mathbb{P}\left(Q(F_{\hat{f}_{m'},s})\right) \leq \frac{2e^{2(B'+M)}}{\mathbb{P}\left(T \geq \tau\right)} \left(K(f,\hat{f}_{m'}) + K(f,f_m^*)\right),$$

so that we can write that, on $\Omega_2^C \cap \Omega_3^C$,

$$\left(R_n(f_m^*) - R_n(\hat{f}_{m'})\right) \mathbb{1}_{\{\Omega[\varepsilon]\}} \leq \frac{e^{2B'}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}} \mathbb{1}_{\{\Omega[\varepsilon]\}} \frac{5\phi_{2,m'}(\sigma_2, x_{m'})}{\sigma_2^2}$$
$$\left(\sigma_2^2 + \frac{5\phi_{3,m'}(\sigma_3, x_{m'})}{\sigma_3^2}\left(\sigma_3^2 + \left(\frac{2e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)}\right)^3 \left(K(f, \hat{f}_{m'}) + K(f, f_m^*)\right)\right)\right.$$
$$\left. + \frac{2e^{2(B'+M)}}{\mathbb{P}(T \geq \tau)}\left(K(f, \hat{f}_{m'}) + K(f, f_m^*)\right)\right).$$

Note that, on the set $\Omega[\varepsilon]$,

$$\frac{1}{\mathbb{P}(T \geq \tau)} \leq \frac{1+\varepsilon}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}.$$

Factorizing $\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}$, we get

$$(R_n(f_m^*) - R_n(\hat{f}_{m'}))\, \mathbb{1}_{\{\Omega[\varepsilon]\}} \leq 2(1+\varepsilon)\left(\frac{e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right)^2 \mathbb{1}_{\{\Omega[\varepsilon]\}} \frac{5\phi_{2,m'}(\sigma_2, x_{m'})}{\sigma_2^2}$$
$$\left(\left(\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right)^{-1} \sigma_2^2 + \frac{5\phi_{3,m'}(\sigma_3, x_{m'})}{\sigma_3^2}\left(\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right)^{-1}\right.$$
$$\left.\left(\sigma_3^2 + \left(\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right)^3 \left(K(f, \hat{f}_{m'}) + K(f, f_m^*)\right)\right) + K(f, \hat{f}_{m'}) + K(f, f_m^*)\right).$$

Now, we choose σ_2 in such a way that

$$10(1+\varepsilon)\left(\frac{e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right)^2 \frac{\phi_{2,m'}(\sigma_2, x_{m'})}{\sigma_2^2} = \theta_2 < 1,$$

where θ_2 is some universal constant smaller than 1, to be choosen later. Recall that $x_{m'} = \xi + L_{\Lambda'}D' + 2\log(B')$. Set

$$\sigma_2^2 = C_2^2(1+\varepsilon)^2\left(\frac{e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right)^3$$
$$\left(\frac{\xi}{n} + (1+\varepsilon)\left(\frac{e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i \geq \tau\}}}\right) B' \log(B') \frac{L_{\Lambda'}D' \log n}{n}\right).$$

So, it is possible to bound the quantity $10(1+\varepsilon)^2\left(\frac{e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{\phi_{2,m'}(\sigma_2,x_{m'})}{\sigma_2^2}$ by $\frac{K_2}{C_2}$ where K_2 is some universal constant. Just choose $C_2=\frac{K_2}{\theta_2}$, and we have

$$
\begin{aligned}
(R_n(f_m^*)-R_n(\hat{f}_{m'}))\,\mathbb{1}_{\{\Omega[\varepsilon]\}} &\leq \theta_2\,\mathbb{1}_{\{\Omega[\varepsilon]\}} \\
&\left(\frac{K_2^2}{2\theta_2^2}\left(\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{B'L_{\Lambda'}D'\log(n)}{n}+\frac{K_2^2}{2\theta_2^2}\left(\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)\frac{\xi}{n}\right. \\
&\qquad +\frac{5\phi_{3,m'}(\sigma_3,x_{m'})}{\sigma_3^2}\left(\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \\
&\left.\left(\left(\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^{-3}\sigma_3^2+K(f,\hat{f}_{m'})+K(f,f_m^*)\right)+K(f,\hat{f}_{m'})+K(f,f_m^*)\right).
\end{aligned}
$$

Once again, set

$$
\sigma_3^2=C_3^2\left(\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^4\left(\frac{\xi}{n}+\frac{B'\log(B')L_{\Lambda'}D'\log(n)}{n}\right).
$$

Like previously, it is possible to bound the quantity $\frac{5\phi_{3,m'}(\sigma_3,x_{m'})}{\sigma_3^2}\left(\frac{2(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq t\}}}\right)^2$ by $\theta_3<1$ to be choosen later, bounding it by $\frac{K_3}{C_3}$, where K_3 is some universal constant, and choosing $C_3=\frac{K_3}{\theta_3}$. We obtain

$$
\begin{aligned}
(R_n(f_m^*)-R_n(\hat{f}_{m'}))\,\mathbb{1}_{\{\Omega[\varepsilon]\}} &\leq \theta_2\,\mathbb{1}_{\{\Omega[\varepsilon]\}} \\
&\left(\frac{K_2^2}{2\theta_2^2}\left(\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{B'\log(B')L_{\Lambda'}D'\log(n)}{n}+\frac{K_2^2}{2\theta_2^2}\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq t\}}}\frac{\xi}{n}\right. \\
&+\theta_3\left(\frac{K_3^2}{8\theta_3^2}\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\frac{B'\log(B')L_{\Lambda'}D'\log(n)}{n}+\frac{K_3^2}{8\theta_3^2}\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\frac{\xi}{n}\right. \\
&\qquad \left.\left.+K(f,\hat{f}_{m'})+K(f,f_m^*)\right)+K(f,\hat{f}_{m'})+K(f,f_m^*)\right).
\end{aligned}
$$

Using that

$$
2\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\xi\leq\left(\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2+\xi^2,
$$

and gathering terms, we can affirm that, on the set $\Omega_2^C \cap \Omega_3^C$,

$$(R_n(f_m^*) - R_n(\hat{f}_{m'}))\,\mathbb{1}_{\{\Omega[\varepsilon]\}} \leq \mathbb{1}_{\{\Omega[\varepsilon]\}}\,\theta_2(1+\theta_3)K(f,\hat{f}_{m'}) + \theta_2(1+\theta_3)K(f,f_m^*)$$
$$+\left(\frac{K_2^2}{4\theta_2}+\frac{\theta_2 K_3^2}{16\theta_3}\right)\left(\frac{(1+\varepsilon)e^{2(B'+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{B'\log(B')L_{\Lambda'}D'\log(n)+1}{n} + \left(\frac{K_2^2}{4\theta_2}+\frac{\theta_2 K_3^2}{16\theta_3}\right)\frac{\xi^2}{n}$$

1.4.5 Conclusion of the proof

Let $\Omega = \cup_{k=1}^3 \Omega_k$. Recall that $\mathbb{P}(\Omega) \leq 4\sum_{m'} e^{-x_{m'}} \leq \kappa' \Sigma e^{-\xi}$, where κ' is an absolute constant. We can write that on the set Ω^C,

$$K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\{\Omega[\varepsilon]\}} \leq (1+\theta_1 e^{2M}+\theta_2(1+\theta_3))K(f,f_m^*)$$
$$+(\mathrm{pen}_n(m)-\mathrm{pen}_n(\hat{m}))\,\mathbb{1}_{\{\Omega[\varepsilon]\}} + (\theta_1+\theta_2(1+\theta_3))K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\{\Omega[\varepsilon]\}}$$
$$+\left(\frac{K_1^2}{4\theta_1}+\frac{K_2^2}{4\theta_2}+\frac{\theta_2 K_3^2}{16\theta_3}\right)\left(\frac{e^{2(\hat{B}+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{(\hat{B}+M)\log\hat{B}L_{\hat{m}}\hat{D}\log(n)+1}{n}$$
$$+\left(\frac{K_1^2}{8\theta_1}+\frac{K_2^2}{4\theta_2}+\frac{\theta_2 K_3^2}{16\theta_3}\right)\frac{\xi^2}{n}$$

Let $\Delta = \{(\theta_1,\theta_2,\theta_3)\in\mathbb{R}^3, \theta_1>0, \theta_2>0, \theta_3>0, \theta_1+\theta_2(1+\theta_3)<1\}$, and let consider the function Z from Δ to $\mathbb{R}$ such that for any $(\theta_1,\theta_2,\theta_3)\in\Delta$,

$$Z(\theta_1,\theta_2,\theta_3) = \frac{1}{1-(\theta_1+\theta_2(1+\theta_3))}\left(\frac{K_1^2}{4\theta_1}+\frac{K_2^2}{4\theta_2}+\frac{\theta_2 K_3^2}{16\theta_3}\right).$$

The function Z is clearly continuous and bounded down by $\frac{K_1^2}{4}+\frac{K_2^2}{4}>0$. Let define by c the infimum of this function on Δ. Since $Z(\theta_1,\theta_2,\theta_3)$ tends to infinity when $(\theta_1,\theta_2,\theta_3)$ tends to $(1,0,1)$, staying in Δ, we can affirm that, for any $c'>c$, there exists $(\theta_1,\theta_2,\theta_3)$ such that $Z(\theta_1,\theta_2,\theta_3)=c'$. Recall that the penalty function is defined by

$$\mathrm{pen}_n(m) = c'\left(\frac{e^{2(B+M)}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i\geq\tau\}}}\right)^2 \frac{(B+M)\log(B)L_m D\log(n)+1}{n}.$$

Choose now θ_1, θ_2 and θ_3 such that $Z(\theta_1,\theta_2,\theta_3)=c'$. Writing θ for $\theta_1+\theta_2(1+\theta_3)$, we obtain

$$K(f,\hat{f}_{\hat{m}})\,\mathbb{1}_{\{\Omega[\varepsilon]\}} \leq \frac{1+\theta_1 e^{2M}+\theta_2(1+\theta_3)}{1-\theta}K(f,f_m^*) + \frac{1}{1-\theta}\mathrm{pen}_n(m)\,\mathbb{1}_{\{\Omega[\varepsilon]\}}$$
$$+\frac{\frac{K_1^2}{8\theta_1}+\frac{K_2^2}{4\theta_2}+\frac{\theta_2 K_3^2}{16\theta_3}}{1-\theta}\frac{\xi^2}{n}.$$

Applying lemma 1.5.6, we obtain the risk bound announced in the theorem. Furthermore, since on the set $\Omega[\varepsilon]$

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq \tau\}}} \leq \frac{1}{(1-\varepsilon)\,\mathbb{P}(T \geq \tau)},$$

we also get the bound

$$\mathbb{E}\left(\mathrm{pen}_n(m)\,\mathbb{1}_{\{\Omega[\varepsilon]\}}\right) \leq \frac{1}{1-\theta}\left(\frac{K_1^2}{4\theta_1} + \frac{K_2^2}{4\theta_2} + \frac{\theta_2 K_3^2}{16\theta_3}\right)\left(\frac{(1+\varepsilon)e^{2(B+M)}}{(1-\varepsilon)\,\mathbb{P}(T \geq \tau)}\right)^2 \frac{(B+M)\log(B)L_m D\log(n)+1}{n}.$$

1.5 Technical lemmas

1.5.1 Proof of Lemma 1.2.1

Let us introduce some notations. For any deterministic function g of $\mathcal{G}$, let $A_n(g)$ denote the following expression:

$$A_n(g)(s,w) = \frac{e^{g(w)}}{S_n(g,s)}.$$

With this notation, we can write:

$$\mathbb{E}\left(\gamma_n(g) - \gamma_n(f)\right) = \mathbb{E}\left(-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{A_n(g)}{A_n(f)}(s,W_i)\right) dN_i(s)\right).$$

Using the Claim 1.5.4.1, this also equals

$$\mathbb{E}\left(-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{A_n(g)}{A_n(f)}(s,W_i)\right) e^{f(W_i)}\,\mathbb{1}_{\{T_i \geq s\}}\,\alpha_0(s)ds\right),$$

that we can rewrite

$$\mathbb{E}\left(-\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left(\frac{A_n(g)}{A_n(f)}(s,W_i)\right) A_n(f)(s,W_i)S_n(f,s)\,\mathbb{1}_{\{T_i \geq s\}}\,\alpha_0(s)ds\right). \quad (1.5.1.1)$$

Let us now consider the following expression:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\frac{A_n(g)}{A_n(f)}(s,W_i) - 1\right) A_n(f)(s,W_i)S_n(f,s)\,\mathbb{1}_{\{T_i \geq s\}}\,\alpha_0(s)ds\right), \quad (1.5.1.2)$$

that also equals

$$\mathbb{E}\left(\int_0^\tau \left(\frac{1}{n}\sum_{i=1}^{n} A_n(g)(s,W_i)\,\mathbb{1}_{\{T_i \geq s\}} - \frac{1}{n}\sum_{i=1}^{n} A_n(f)(s,W_i)\,\mathbb{1}_{\{T_i \geq s\}}\right) S_n(f,s)\alpha_0(s)ds\right).$$

Note now that, for any function g,

$$\frac{1}{n}\sum_{i=1}^{n} A_n(g)(s,W_i)\,\mathbb{1}_{\{T_i\geq s\}} = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{g(W_i)}}{S_n(g,s)}\,\mathbb{1}_{\{T_i\geq s\}} = 1.$$

Therefore, expression 1.5.1.2 equals 0 and we can add it to expression 1.5.1.1. We obtain then,

$$\mathbb{E}\left(\gamma_n(g)-\gamma_n(f)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Phi\left(\log\frac{A_n(g)}{A_n(f)}(s,W_i)\right)e^{f(W_i)}\,\mathbb{1}_{\{T_i\geq s\}}\,\alpha_0(s)ds\right),$$

where Φ denotes the function $\Phi(x) = e^x - 1 - x$. Since this function is nonnegative, $\mathbb{E}\left(\gamma_n(g)-\gamma_n(f)\right)$ also is nonnegative.

Furthermore, if $\mathbb{E}\left(\gamma_n(g)-\gamma_n(f)\right)=0$, then for any $1\leq i\leq n$,

$$\int_0^\tau \Phi\left(\log\frac{A_n(g)}{A_n(f)}(s,W_i)\right)e^{f(W_i)}\,\mathbb{1}_{\{T_i\geq s\}}\,\alpha_0(s)ds = 0$$

almost surely, for any $1\leq i\leq n$ and $0\leq s\leq\tau$,

$$\Phi\left(\log\frac{A_n(g)}{A_n(f)}(s,W_i)\right)=0$$

almost surely. Since $\Phi(x)=0$ implies $x=0$, that implies that for any $1\leq i\leq n$ and $0\leq s\leq\tau$, $A_n(g)(s,W_i)=A_n(f)(s,W_i)$ almost surely. This can be rewritten

$$e^{g(W_i)-f(W_i)} = n^{-1/2}\sum_{j=1}^{n} e^{g(W_j)-f(W_j)}A_n(f)(s,W_j)Y_j(s)$$

almost surely. That means that $e^{g(W_i)-f(W_i)}$ is a constant almost surely, that is to say $g=f+cste$ almost surely. The condition $\mathbb{E}\left(g(W)\right)=0$ implies $g=f$ almost surely.

1.5.2 Properties of the Kullback-Leibler information

The next lemma, that provides a link between the variance term and the Kullback-Leibler information, can be found in the framework of density estimation in Barron and Sheu [9] and in Castellan [21].

Lemma 1.5.1. *For any functions f and g in $\mathcal{G}$ and any deterministic function c from $[0,\tau]$ to $\mathbb{R}$, we have*

$$\frac{e^{-2\|g-f\|_\infty}}{2}\,\mathbb{P}\left(\log^2\frac{A(g)}{A(f)}\right)\leq K(f,g)\leq\frac{e^{\|\log\frac{A(g)}{A(f)}+c(s)\|_\infty}}{2}\,\mathbb{P}\left(\log^2\left(\frac{A(g)}{A(f)}+c(s)\right)\right).$$

Proof. By definition,

$$K(f,g) = \mathcal{E}\left[\int_0^\tau \log\frac{A(f)}{A(g)} A(f)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right].$$

Using the Claim 1.5.4.2, we find

$$\mathcal{E}\left[\int_0^\tau \left(\frac{A(g)}{A(f)} - 1\right) A(f)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right] = 0,$$

and thus, we can add this expression to the previous one. Denoting by Φ and the function defined by $\Phi(x) = e^x - 1 - x$, for any $x \in \mathbb{R}$, we have

$$K(f,g) = \mathcal{E}\left[\int_0^\tau \Phi\left(\log\frac{A(g)}{A(f)}\right) A(f)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right].$$

Now, using a Taylor expansion, it is clear that for any $x \in \mathbb{R}$,

$$\Phi(x) \geq \frac{e^{-|x|}}{2} x^2.$$

Recalling that

$$\log\frac{A(g)}{A(f)} = g - f - \log\frac{S(g,s)}{S(f,s)},$$

so that $\|\log\frac{A(g)}{A(f)}\|_\infty \leq 2\|g-f\|_\infty$, we can write

$$K(f,g) \geq \mathcal{E}\left[\int_0^\tau \frac{e^{-2\|g-f\|_\infty}}{2}\left(\log\frac{A(g)}{A(f)}\right)^2 A(f)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right],$$

which gives the first inequality.

Now, we can also write

$$K(f,g) = \mathcal{E}\left[\int_0^\tau \Phi(\log\frac{A(g)}{A(f)} + c(s))A(f)S(f,s)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right]$$
$$+ \mathcal{E}\left[\int_0^\tau \left((1-e^{c(s)})\frac{A(g)}{A(f)} + c(s)\right) A(f)S(f,s)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right].$$

Using Claim 1.5.4.2, the second term also equals

$$\mathcal{E}\left[\int_0^\tau \left(1 - e^{c(s)} + c(s)\right) A(f)S(f,s)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right],$$

that is nonpositive. Using a Taylor expansion once again, we have that for any $x \in \mathbb{R}$,

$$\Phi(x) \leq \frac{e^{|x|}}{2} x^2,$$

from which the second inequality follows. □

In the same way that we introduced some quantity comparable to a Kullback-Leibler information, we can define some Hellinger-like distance, which have the same link with the Kullback-Leibler information as in the density framework. That is what we prove here.

Lemma 1.5.2. *Let us define the Hellinger-like distance by*

$$H^2(f,g) = \frac{1}{2}\mathcal{E}\left[\int_0^\tau \left(\sqrt{A(g)(s,W)} - \sqrt{A(f)(s,W)}\right)^2 S(f,s)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right],$$

for any function g in $\mathcal{G}$. Then, $K(f,g) \geq 2H^2(f,g)$.

Proof. For any function g in $\mathcal{G}$,

$$\begin{aligned}
&K(f,g) - 2H^2(f,g)\\
&= \mathcal{E}\left[\int_0^\tau \left(\log\frac{A(f)}{A(g)} - \left(\sqrt{\frac{A(g)}{A(f)}} - 1\right)^2\right) A(f)S(f,s)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right]\\
&= \mathcal{E}\left[\int_0^\tau \left(-2\log\sqrt{\frac{A(g)}{A(f)}} - \frac{A(g)}{A(f)} + 2\sqrt{\frac{A(g)}{A(f)}} - 1\right) A(f)S(f,s)\, \mathbb{1}_{\{T\geq s\}}\, \alpha_0(s)ds\right]
\end{aligned}$$

From Claim 1.5.4.2, we have that

$$\begin{aligned}
&\mathcal{E}\left[\int_0^\tau \frac{A(g)}{A(f)} A(f)S(f,s)\, \mathbb{1}_{\{T\geq\tau\}}\, \alpha_0(s)ds\right]\\
&\qquad = \mathcal{E}\left[\int_0^\tau A(f)S(f,s)\, \mathbb{1}_{\{T\geq\tau\}}\, \alpha_0(s)ds\right] = \mathbb{E}\left(N(\tau)\right),
\end{aligned}$$

so that

$$K(f,g) - 2H^2(f,g) = \mathcal{E}\left[\int_0^\tau 2\Phi\left(\log\sqrt{\frac{A(g)}{A(f)}}\right) A(f)S(f,s)\, \mathbb{1}_{\{T\geq\tau\}}\, \alpha_0(s)ds\right].$$

Since the function Φ is nonnegative on $\mathbb{R}$, we get the result. □

We give now a link between the Kullback-Leibler information and the L_2 distance between two functions, via the L_∞ distance.

Lemma 1.5.3. *For any functions f and g in $\mathcal{G}$,*

$$K(f,g) \leq \frac{e^{\|g-f\|_\infty}}{2}\mathbb{E}\left((g-f)^2(W)N(\tau)\right).$$

Proof. Apply the second inequality of Lemma 1.5.1 with

$$c(s) = \log\frac{S(g,s)}{S(f,s)}.$$

We get

$$\log \frac{A(g)}{A(f)} + c(s) = g - f,$$

so that

$$K(f,g) \leq \frac{e^{\|g-f\|_\infty}}{2} \mathcal{E}\left[\int_0^\tau (g-f)^2(W) A(f) S(f,s)\, \mathbb{1}_{\{T \geq s\}}\, \alpha_0(s) ds\right],$$

which is exactly the result. □

The next lemma will be useful while bounding the distance between the true function f and any function in some model.

Lemma 1.5.4. *For any functions f and g in $\mathcal{E}$,*

$$K(f,g) \leq 2\,\mathbb{E}\left(N(\tau)\right) \|f-g\|_\infty^2.$$

Proof. Recall the definition of $K(f,g)$

$$K(f,g) = \mathcal{E}\left[\int_0^\tau \log \frac{A(f)}{A(g)} A(f) S(f,s)\, \mathbb{1}_{\{T \geq s\}}\, \alpha_0(s) ds\right].$$

Using Remark 6, we can bound $\log \frac{A(f)}{A(g)}$ by $2\|f-g\|_\infty$, so that

$$K(f,g) \leq 2\|f-g\|_\infty \mathcal{E}\left[\int_0^\tau A(f) S(f,s)\, \mathbb{1}_{\{T \geq s\}}\, \alpha_0(s) ds\right] = 2\|f-g\|_\infty \mathbb{E}\left(N(\tau)\right).$$

In particular, when $\|f-g\|_\infty \geq 1$, we have

$$K(f,g) \leq 2\|f-g\|_\infty^2 \mathbb{E}\left(N(\tau)\right). \tag{1.5.1.3}$$

On the other hand, using Lemma 1.5.3, we have

$$K(f,g) \leq \frac{e^{\|g-f\|_\infty}}{2} \mathbb{E}\left((g-f)^2(W) N(\tau)\right) \leq \frac{e^{\|g-f\|_\infty}}{2} \|g-f\|_\infty^2 \mathbb{E}\left(N(\tau)\right).$$

In particular, when $\|f-g\|_\infty \leq 1$, we have

$$K(f,g) \leq \frac{e}{2}\|g-f\|_\infty^2 \mathbb{E}\left(N(\tau)\right) \leq 2\|g-f\|_\infty^2 \mathbb{E}\left(N(\tau)\right). \tag{1.5.1.4}$$

Gathering inequalities 1.5.1.3 and 1.5.1.4, we get the result. □

1.5.3 Other lemmas

Lemma 1.5.5. *Let Z be some real valued integrable variable and ψ some increasing function on $\mathbb{R}^+$ such that, for any measurable set A with $P(A) > 0$, $\mathbb{E}^A(Z) \leq \psi(\log(\frac{1}{P(A)}))$. Then, for any positive x, $\mathbb{P}(Z \geq \psi(x)) \leq e^{-x}$.*

Proof. Using Markov's inequality,

$$\mathbb{P}\left(Z \geq \psi(x)\right) \leq \frac{\mathbb{E}\left(Z \, \mathbb{1}_{\{Z \geq \psi(x)\}}\right)}{\psi(x)}.$$

Applying the hypothesis with $A = \{Z \geq \psi(x)\}$, we have that

$$\mathbb{E}\left(Z \, \mathbb{1}_{\{Z \geq \psi(x)\}}\right) = \mathbb{E}^A(Z)\mathbb{P}(A),$$

so that $\psi(x) \leq \mathbb{E}^A(Z) \leq \psi(\log \frac{1}{\mathbb{P}(A)})$. Since the function ψ is increasing, we get the conclusion. □

Lemma 1.5.6. *Let X, Y be two positive random variables. Assume that there exist constants κ_1 and κ_2 such that for any positive ξ, $\mathbb{P}\left(X \geq Y + \kappa_1 \xi^2\right) \leq \kappa_2 e^{-\xi}$, then $\mathbb{E}(X) \leq \mathbb{E}(Y) + 2\kappa_1\kappa_2$.*

Proof. By definition,

$$\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}\left(X \geq t\right) dt = \int_\Omega \int_0^{+\infty} \mathbb{1}_{\{X \geq t\}} \, dt d\mathbb{P}.$$

Decomposing the indicator, we get

$$\mathbb{E}(X) \leq \int_\Omega \int_0^{+\infty} \mathbb{1}_{\{Y \geq t\}} \, dt d\mathbb{P} + \int_\Omega \int_0^{+\infty} \mathbb{1}_{\{Y < t \leq X\}} \, dt d\mathbb{P}.$$

Setting $t = Y + \kappa_1 \xi^2$ and integrating, we found

$$\mathbb{E}(X) \leq \mathbb{E}(Y) + \int_\Omega \int_0^{+\infty} \mathbb{1}_{\{X \geq Y + \kappa_1 \xi^2\}} \, 2\kappa_1 \xi d\xi d\mathbb{P} \leq \mathbb{E}(Y) + \int_0^{+\infty} 2\kappa_1\kappa_2 \xi e^{-\xi} d\xi,$$

that enables us to conclude. □

1.5.4 Claims

In this section, we present Claims that are used all along the chapter.

Claim 1.5.4.1. *For any predictable function H with respect to the filtration generated by the observation,*

$$\mathbb{E}\left(\int_0^\tau H(s) dN s\right) = \mathbb{E}\left(\int_0^\tau H(s) e^{f(W)} \, \mathbb{1}_{\{T \geq s\}} \, \alpha_0(s) ds\right).$$

Proof. The counting process N has predictable compensator

$$\tilde{N}(t) = \int_0^\tau e^{f(W)} \, \mathbb{1}_{\{T \geq s\}} \, \alpha_0(s) ds,$$

with respect to the filtration generated by the observations. That means that the process $M = N - \tilde{N}$ is a martingale with respect to the same filtration. Since the function H is predictable with respect to the same filtration, we have that the process $\int_0^\tau H(s)dM(s)$ also is a martingale. In particular, its expectation is 0, that enables us to conclude. □

Claim 1.5.4.2. *For any deterministic function h from $[0,\tau]$ to $\mathbb{R}$ and any function $g \in \mathcal{G}_\Lambda$,*

$$\mathbb{E}\left(\int_0^\tau h(s)A(g)\,\mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right) = \mathbb{E}\left(\int_0^\tau h(s)A(f)\,\mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right).$$

Proof. Since the only random terms are $A(g)\,\mathbb{1}_{\{T\geq s\}}$ and interchanging the expectation and the integral, we have that

$$\mathbb{E}\left(\int_0^\tau h(s)A(g)\,\mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right) = \int_0^\tau h(s)\,\mathbb{E}\left(A(g)\,\mathbb{1}_{\{T\geq s\}}\right)S(f,s)\alpha_0(s)ds.$$

Now,

$$\mathbb{E}\left(A(g)(s,W)\,\mathbb{1}_{\{T\geq s\}}\right) = \mathbb{E}\left(\frac{e^{g(W)}\,\mathbb{1}_{\{T\geq s\}}}{S(g,s)}\right) = 1.$$

Recalling the definition of $S(f,s)$,

$$\mathbb{E}\left(\int_0^\tau h(s)A(g)\,\mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right) = \int_0^\tau h(s)\,\mathbb{E}\left(e^{f(W)}\,\mathbb{1}_{\{T\geq s\}}\right)\alpha_0(s)ds.$$

Interchanging the integral and the expectation once again and introducing $S(f,s)$, we finally have

$$\mathbb{E}\left(\int_0^\tau h(s)A(g)\,\mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right) = \mathbb{E}\left(\int_0^\tau h(s)A(f)\,\mathbb{1}_{\{T\geq s\}}\,S(f,s)\alpha_0(s)ds\right).$$

□

Chapitre 2

A semiparametric shock model in censoring bivariate survival analysis

Contents

Abstract

In this chapter, we propose a semiparametric shock model in order to model situations in demography where the biographies of a pair of individuals cannot be considered as independent. For that purpose, we construct two dependent counting processes representing these biographies in such a way that, whenever either one of both counting processes jumps, the hazard rate of the other one is instantaneously multiplied by a constant. These constants are called shock parameters. Moreover, these counting processes may be censored.

In such a context, assuming a Cox model, our aim is to estimate the shock parameters and the Cox regression parameters, from a sample of independent and identically distributed, possibly censored pairs. Maximum log partial likelihood estimators for the shock constants and for the Cox regression parameters are proposed. Consistency and asymptotic normality of these estimators are established. We illustrate our method with simulations.

Keywords and phrases: Bivariate censored data, Bivariate survival analysis, Cox model.

Résumé

Dans ce chapitre, nous proposons un modèle de chocs semi-paramétrique pour modéliser des situations courantes en démographie, quand les biographies d'un couple d'individus ne peuvent pas être considérées comme indépendantes. Pour cela, nous construisons deux processus de comptage dépendants représentant ces biographies de telle façon que, quand l'un des processus saute, le processus de comptage de l'autre est instantanément multiplié par une constante. Ces constantes sont appelées constantes de chocs. De plus, ces processus de comptage peuvent être supposés censurés.

Dans ce contexte, supposant de plus un modèle de Cox, notre but est d'estimer les paramètres de chocs ainsi que les paramètres de régression de Cox, à partir d'un échantillon de couples éventuellement censurés. Nous proposons des estimateurs du maximum de log-vraisemblance partielle pour les paramètres de chocs et pour les paramètres de régression de Cox. Nous établissons la consistance et la normalité asymptotique de nos estimateurs. Nous illustrons notre méthode par des simulations.

Mots clés: Analyse de survie bivariée, Données censurées bivariées, Modèle de Cox.

2.1 Introduction

In demography studies, the Cox model (see Cox [25]) is often used in order to study the individual biographies with respect to covariates, depending or not of the time (see Kalbfleisch and Prentice [38]). This model cannot be used anymore when introducing covariates which may be influenced by the individual in study himself. Consider for instance the case of a man looking for a job: we cannot take as a covariate the process indicating whether his wife has a job or not, since both processes are strongly linked to each other. In such cases, we are led to consider multivariate models, where the involved failure times are supposed not to be independent.

One famous class of multivariate survival models is composed of the so-called frailty models (see for instance Vaupel *et al.*[60], Clayton and Cuzick [23], Hougaard [34], Oakes [48], Nielsen *et al.*[47]). In such models, the individual hazard rates are supposed to be proportional to a frailty variable that is common to all variables of a group and that is not observed. Usually, this frailty variable is supposed to have a Gamma distribution, which parameters are to be estimated.

Copula models are also used as bivariate survival models. They consist in assuming that the joint distribution function can be written as a function of the two marginals distribution functions. One particular copula model is Clayton's shock model (see Clayton [22], Cox and Oakes [26]). He considers a father and his son who are supposed to be subject to some event at ages T and S. The dependency between both random variables is supposed to be defined by the following relationship between the conditional hazard rates of the father λ_f and of the son λ_s:

$$\frac{\lambda_s(s \mid T = t_0)}{\lambda_s(s \mid T > t_0)} = \frac{\lambda_f(t \mid S = s_0)}{\lambda_f(t \mid S > s_0)} = \theta, \quad \text{for any positive} \quad s, t, s_0, t_0.$$

In such a model, the joint distribution of (S,T) is necessarily of the form:

$$f(s,t) = \frac{\theta a(s) b(t)}{\left[1 + (\theta - 1)(\int_0^s a + \int_0^t b)\right]^{2+\frac{1}{\theta-1}}}$$

for some nonnegative functions a and b.

In this paper, we propose a bivariate survival model which corresponds to the following situation: in a pair of individuals denoted by A and B, each one is likely to be subject to an event at times $\tilde{X}_A$ and $\tilde{X}_B$. The model is constructed in such a way that whenever either one of both events happens, the hazard rate of the other one is instantaneously multiplied by a constant. That is the reason why it is called a shock model. A lot of situations in demography studies can be modelled in this way (see Lelièvre *et al.*[42]). Let us begin with defining this model.

Let (X_A, X_B) be a pair of independent positive random variables with hazard rate α_A and α_B.

If $X_A < X_B$, we define $\tilde{X}_A = X_A$ and $\tilde{X}_B = X_A + X'_B$ where X'_B is some positive

random variable, independent of X_B, with hazard rate $e^{\rho_{B,0}}\alpha_B(X_A+s)$, where $\rho_{A,0}$ is some real number (see Fig. 2.1).

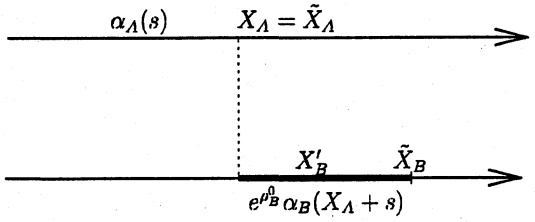

Figure 2.1: Case $X_A < X_B$: $\tilde{X}_B = X_A + X'_B$

If $X_B < X_A$, we define $\tilde{X}_B = X_B$ and $\tilde{X}_A = X_B + X'_A$ where X'_A is some positive random variable, independent of X_A, with hazard rate $e^{\rho_{A,0}}\alpha_A(X_B+s)$, where $\rho_{B,0}$ is some real number real (see Fig. 2.2).

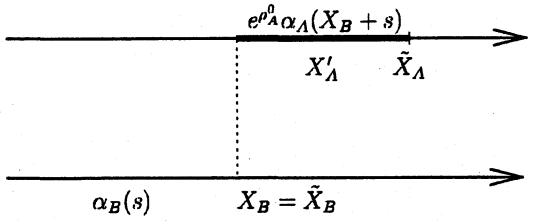

Figure 2.2: Case $X_B < X_A$: $\tilde{X}_A = X_B + X'_A$

Random variables $\tilde{X}_A$ and $\tilde{X}_B$ can also be written as

$$\begin{aligned} \tilde{X}_A &= \inf(X_A, X_B) + X'_A \, \mathbb{1}_{\{X_B<X_A\}} \\ \tilde{X}_B &= \inf(X_B, X_A) + X'_B \, \mathbb{1}_{\{X_A<X_B\}} . \end{aligned} \tag{2.1.2.1}$$

Their joint distribution is given by:

$$\begin{aligned} f(s,t) &= \frac{e^{\rho^0_B}\alpha_A(s)\alpha_B(t)}{\exp\left[\int_0^s \alpha_A + \int_0^t e^{\rho^0_B \mathbb{1}_{\{u\geq s\}}}\alpha_B(u)du\right]}, \ \forall s < t, \\ f(s,t) &= \frac{e^{\rho^0_A}\alpha_A(s)\alpha_B(t)}{\exp\left[\int_0^s e^{\rho^0_A \mathbb{1}_{\{u\geq t\}}}\alpha_A(u)du + \int_0^t \alpha_B\right]}, \ \forall t < s. \end{aligned} \tag{2.1.2.2}$$

One can see from the construction of the random variables and with help of Fig. 2.1 and 2.2 that random variable $\tilde{X}_A$ has hazard rate α_A so long it is smaller than $\tilde{X}_B$ and hazard rate $e^{\rho_{A,0}}\alpha_A$ whenever it is greater than $\tilde{X}_B$. In the same way, random variable $\tilde{X}_B$ has hazard rate α_B so long it is smaller than $\tilde{X}_A$ and hazard rate $e^{\rho_{B,0}}\alpha_B$ whenever it is greater than $\tilde{X}_A$. In the following, reals $\rho_{A,0}$ and $\rho_{B,0}$ will be called the shock parameters of the model.

Another way to illustrate this situation is shown in Fig. 2.3. From a state 0 where none of both events has occurred, we can go with hazard rate α_A to a state A where only A's event has occurred or with hazard rate α_B to a state B where only B's event has occurred. From these two states, we can go to a state AB where both events have occurred, and we go there with hazard rate $e^{\rho_{A,0}}\alpha_A$ if we come from state B and with hazard rate $e^{\rho_{B,0}}\alpha_B$ if we come from state A. Such cases happen in medical studies, when having one disease modifies the hazard rate for another disease (see Aalen *et al.*[3]).

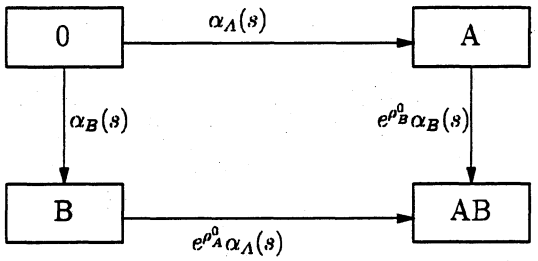

Figure 2.3: A four states-model: state 0: no event has occurred, state A: only A's event has occurred, state B: only B's event has occurred, state AB: both event have occurred.

In such a context, it may be interesting to include covariates. Let (Z_A, Z_B) be a D-dimensional vector of such covariates. These vectors may be nonindependent. In particular, they can have common coordinates or be equal. We can assume a Cox model and replace the hazard rates α_A and α_B by conditional hazard rates given (Z_A, Z_B) $\alpha_{Z_A} = e^{\theta_{A,0}{}^T Z_A}\alpha_{A,0}$ and $\alpha_{Z_B} = e^{\theta_{B,0}{}^T Z_B}\alpha_{B,0}$, where $\theta_{A,0}$ and $\theta_{B,0}$ are D-dimensional vectors called the Cox regression parameters.

Suppose now that the random variables $\tilde{X}_A$ and $\tilde{X}_B$ may be censored: we observe the random variables in some deterministic time interval $[0,\tau]$ and there exist two right-censoring times U_A and U_B verifying the following assumption:

Assumption 2.1. *$(\tilde{X}_A, \tilde{X}_B)$ and (U_A, U_B) are conditionally independent given (Z_A, Z_B). The distribution of (U_A, U_B) is such that:*

$$\mathbb{P}\left(U_A \geq \tau,\ U_B \geq \tau \mid (Z_A, Z_B)\right) > 0.$$

We allow, for instance, the case $U_A = U_B$. Instead of observing $(\tilde{X}_A, \tilde{X}_B)$, we actually only observe $(T_A, T_B, \delta_A, \delta_B, Z_A, Z_B)$, where

$$\begin{array}{rclcrcl} T_A &=& \inf(\tilde{X}_A, U_A), & & T_B &=& \inf(\tilde{X}_B, U_B), \\ \delta_A &=& \mathbb{1}_{\{\tilde{X}_A \leq U_A\}} & & \delta_B &=& \mathbb{1}_{\{\tilde{X}_B \leq U_B\}}. \end{array} \tag{2.1.2.3}$$

Note that the baseline hazard functions $\alpha_{A,0}$ and $\alpha_{B,0}$ are supposed to be unknown.

Our model can be considered as a particular bivariate Cox model, and our aim is now to estimate $\beta_0 = (\rho_{A,0}, \theta_{A,0}, \rho_{B,0}, \theta_{B,0})^T$ (where x^T denotes the transpose of the vector x), from a sample of independent and identically distributed variables

$(T_{A,i}, T_{B,i}, \delta_{A,i}, \delta_{B,i}, Z_{A,i}, Z_{B,i})_{1\leq i\leq n}$, in the presence of the nuisance parameters $\alpha_{A,0}$ and $\alpha_{B,0}$, belonging to some functional space.

Let us introduce some notations needed to define our estimators. Let $N_{h,i}, h = A, B, 1 \leq i \leq n$ denote the following counting processes:

$$\begin{array}{rcl} N_{A,i}(t) & = & \delta_{A,i}\, \mathbb{1}_{\{T_{A,i}\leq t\}}(\delta_{B,i}\, \mathbb{1}_{\{T_{B,i}<T_{A,i}\}} + \mathbb{1}_{\{T_{A,i}\leq T_{B,i}\}}), \\ N_{B,i}(t) & = & \delta_{B,i}\, \mathbb{1}_{\{T_{B,i}\leq t\}}(\delta_{A,i}\, \mathbb{1}_{\{T_{A,i}<T_{B,i}\}} + \mathbb{1}_{\{T_{B,i}\leq T_{A,i}\}}), \text{ for any } t > 0, \end{array} \tag{2.1.2.4}$$

let $W_{h,i}(t), t > 0, h = A, B, 1 \leq i \leq n$ be the following covariates vectors:

$$W_{A,i}(t) = \begin{pmatrix} \mathbb{1}_{\{T_{B,i}<t\}} \\ Z_{A,i} \end{pmatrix}, W_{B,i}(t) = \begin{pmatrix} \mathbb{1}_{\{T_{A,i}<t\}} \\ Z_{B,i} \end{pmatrix}, \tag{2.1.2.5}$$

and let $Y_{h,i}(t), t > 0, h = A, B, 1 \leq i \leq n$ denote the following risk indicators:

$$\begin{array}{rcl} Y_{A,i}(t) & = & \mathbb{1}_{\{T_{A,i}\geq t\}}(\delta_{B,i}\, \mathbb{1}_{\{T_{B,i}<t\}} + \mathbb{1}_{\{T_{B,i}\geq t\}}), \\ Y_{B,i}(t) & = & \mathbb{1}_{\{T_{B,i}\geq t\}}(\delta_{A,i}\, \mathbb{1}_{\{T_{A,i}<t\}} + \mathbb{1}_{\{T_{A,i}\geq t\}}). \end{array} \tag{2.1.2.6}$$

In the sequel, for any vector $\beta = (\rho_A, \theta_A, \rho_B, \theta_B)^T$, we denote β_A and β_B the vectors $\beta_A = (\rho_A, \theta_A)^T$ and $\beta_B = (\rho_B, \theta_B)^T$. We define the log partial likelihood of our model as

$$l_n(\beta) = \sum_{h=A,B} \sum_{i=1}^{n} \int_0^\tau \log\left(\frac{e^{\beta_h^T W_{h,i}(s)}}{S_{n,h}(\beta, s)}\right) dN_{h,i}(s) \tag{2.1.2.7}$$

where $S_{n,h}, h = A, B$ denotes for any s such that $0 \leq s \leq \tau$:

$$S_{n,h}(\beta, s) = n^{-1} \sum_{i=1}^{n} e^{\beta_h^T W_{h,i}(s)} Y_{h,i}(s) \tag{2.1.2.8}$$

Our estimator $\hat{\beta}_n$ is then defined by

$$\hat{\beta}_n = (\hat{\rho}_{A,n}, \hat{\theta}_{A,n}, \hat{\rho}_{B,n}, \hat{\theta}_{B,n}) = \underset{\beta}{\operatorname{argmax}}\, l_n(\rho_A, \theta_A, \rho_B, \theta_B). \tag{2.1.2.9}$$

We prove consistency of the estimator $\hat{\beta}_n$, we show that it is asymptotically normal with rate of convergence $n^{-1/2}$ and that $-n^{-1}D^2 l_n(\hat{\beta}_n)$ is a consistent estimator of the inverse of the asymptotic covariance matrix (where $D^2 l_n(\beta)$ denotes the second derivative of function l_n at point β).

We derive the asymptotic properties of our estimators by applying Theorems VII.2.1 and VII.2.2 in Andersen *et al.*[5]. A complete presentation of the proofs is given by Andersen *et al.*[5] (see also Appendix B).

Next section is devoted to the asymptotic properties of our estimators. In section 3, we illustrate our purpose with some simulation study. Proofs are to be found in section 4.

2.2 Large sample properties

2.2.1 Construction of the likelihood

We follow Andersen and Gill's approach that is based on the martingale theory. We begin with defining the counting processes that represent the model best, we calculate their compensators with respect to the σ-algebra generated by the observations, and finally, we apply Jacod's formula (see Jacod [35, 36], Andersen *et al.*[5]) in order to obtain the likelihood of the model.

The basic counting processes

Before introducing the censoring, let us consider the basic counting processes $\tilde{N}_h$ for $h = A, B$, that is to say, the counting processes that would be considered, if there had been no censoring. These counting processes are defined by: $\tilde{N}_h(t) = \mathbb{1}_{\{\tilde{X}_h \leq t\}}$, for $t > 0$. Let also define the functions $\tilde{\alpha}_A$ and $\tilde{\alpha}_B$ by

$$\begin{aligned} \tilde{\alpha}_A(t) &= e^{\rho_{A,0}\,\mathbb{1}_{\{\tilde{X}_B < \tilde{X}_A, \tilde{X}_B < t\}} + \theta_{A,0}^T Z_A} \alpha_{A,0}(t), \\ \tilde{\alpha}_B(t) &= e^{\rho_{B,0}\,\mathbb{1}_{\{\tilde{X}_A < \tilde{X}_B, \tilde{X}_A < t\}} + \theta_{B,0}^T Z_B} \alpha_{B,0}(t), \text{ for } t > 0. \end{aligned} \tag{2.2.2.1}$$

Let define the risk indicator $\tilde{Y}_h$ for $h = A, B$ by $\tilde{Y}_h(t) = \mathbb{1}_{\{\tilde{X}_h \geq t\}}$, and the filtration $(\mathcal{N}_t, t \geq 0)$ as the family of the σ-algebras $\mathcal{N}_t$ generated by $\{\tilde{N}_A(u), \tilde{N}_B(u), u \leq t, (Z_A, Z_B)\}$.

Assumption 2.2. $\int_0^\tau \alpha_{h,0}(s)ds < \infty$, $h = A, B$

Assumption 2.3. *There exists a neighbourhood* $\Theta_A \times \Theta_B$ *of* $(\theta_{A,0}, \theta_{B,0})$ *such that,* $\mathbb{E}\left(\sup_{\theta_h \in \Theta_h} |Z_h|^2 \exp(\theta_h^T Z_h)\right) < \infty$, $h = A, B$,

where $|Z_h|$ denotes the Euclidean norm of the vector Z_h.

Proposition 2.2.1. *Under Assumptions 2.2 and 2.3, the counting processes* $\tilde{N}_h$ *for* $h = A, B$ *have predictable compensators* $\tilde{\Lambda}_h$, *where* $\tilde{\Lambda}_h(t) = \int_0^\tau \tilde{\alpha}_h(s)\tilde{Y}_h(s)ds$ *for any positive* t, *with respect to the filtration* $(\mathcal{N}_t, t \geq 0)$.

Let denote by $\tilde{M}_h$ the difference $\tilde{N}_h - \tilde{\Lambda}_h$. We just need to prove that the processes $\tilde{M}_h$ are $\mathcal{N}_t$-martingales, which is done in 2.4.1.

The censored counting processes

In reality, the previous counting processes are not observed because of the presence of the censoring. We now have to define new counting processes, that take account of the censoring.

For that purpose, let us consider individual A. We expect some counting process of the form $\delta_A \mathbb{1}_{\{T_A \geq t\}}$, like in the classical Cox model. So, let us consider the case when $\delta_A = 1$. Whenever A's event occurs before B's one, that is to say whenever T_A is smaller than T_B,

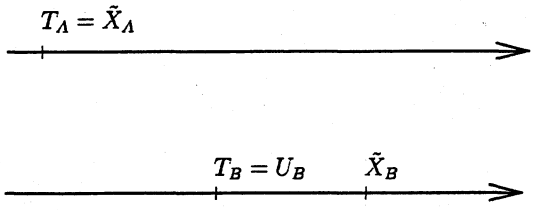

Figure 2.4: Case $T_A < T_B$, $\delta_A = 1$, $\delta_B = 0$.

$\tilde{X}_A$ is necessarily smaller than $\tilde{X}_B$, so that A cannot have been influenced by B (see Fig. 2.4).

This situation is illustrated by the counting process N_1 defined by:

$$N_1(t) = \delta_A \, \mathbb{1}_{\{T_A \leq T_B, T_A \leq t\}} .$$

On the contrary, whenever A's event occurs after B's one, if B is individually censored ($\delta_B = 0$), A also have to be considered as censored because in this case, we do not observe the time $\tilde{X}_B$ when A's hazard rate is modified by the constant $e^{\rho_{A,0}}$ (see Fig. 2.5).

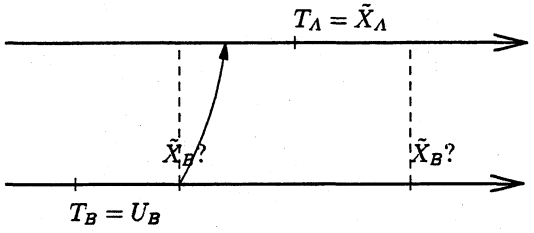

Figure 2.5: Case $T_B < T_A$, $\delta_A = 1$, $\delta_B = 0$.

That leads to the counting process N_2:

$$N_2(t) = \delta_A \delta_B \, \mathbb{1}_{\{T_B < T_A, T_A \leq t\}} .$$

We can now define the counting process N_A corresponding to individual A by $N_A = N_1 + N_2$ which implies formula (2.1.2.4). In the same manner, we define

$$N_3(t) = \delta_B \, \mathbb{1}_{\{T_B \leq T_A, T_B \leq t\}},$$
$$N_4(t) = \delta_A \delta_B \, \mathbb{1}_{\{T_A < T_B, T_B \leq t\}},$$

and $N_B = N_3 + N_4$.

In order to calculate easily the compensator of these counting processes with respect to the filtration generated by the observations, we can try to interpret N_A and N_B as integrals of predictable processes with respect to the previous basic counting processes $\tilde{N}_A$ and $\tilde{N}_B$.

For this purpose, let us recall the definition (2.1.2.4) of N_A:

$$N_A(t) = \delta_A \, \mathbb{1}_{\{T_A \le t\}} (\delta_B \, \mathbb{1}_{\{T_B < T_A\}} + \mathbb{1}_{\{T_A \le T_B\}}).$$

Recalling the definitions (2.1.2.3), we can write

$$\begin{aligned} N_A(t) &= \mathbb{1}_{\{\tilde{X}_A \le U_A, \tilde{X}_A \le t\}} (\mathbb{1}_{\{\tilde{X}_B \le U_B, \tilde{X}_B < \tilde{X}_A\}} + \mathbb{1}_{\{\tilde{X}_A \le \tilde{X}_B, \tilde{X}_A \le U_B\}}) \\ &= \int_0^t \mathbb{1}_{\{u \le U_A\}} (\mathbb{1}_{\{\tilde{X}_B \le U_B, \tilde{X}_B < u\}} + \mathbb{1}_{\{u \le \tilde{X}_B, u \le U_B\}}) d\tilde{N}_A(u), \end{aligned}$$

since $T_h = \tilde{X}_h$ as soon as $\delta_h = 1$. In the same way, we have

$$N_B(t) = \int_0^t \mathbb{1}_{\{u \le U_B\}} (\mathbb{1}_{\{\tilde{X}_A \le U_A, \tilde{X}_A < u\}} + \mathbb{1}_{\{u \le \tilde{X}_A, u \le U_A\}}) d\tilde{N}_B(u).$$

Define now the censoring processes C_A and C_B by

$$\begin{aligned} C_A(u) &= \mathbb{1}_{\{u \le U_A\}} (\mathbb{1}_{\{\tilde{X}_B \le U_B, \tilde{X}_B < u\}} + \mathbb{1}_{\{u \le \tilde{X}_B, u \le U_B\}}), \\ C_B(u) &= \mathbb{1}_{\{u \le U_B\}} (\mathbb{1}_{\{\tilde{X}_A \le U_A, \tilde{X}_A < u\}} + \mathbb{1}_{\{u \le \tilde{X}_A, u \le U_A\}}). \end{aligned} \tag{2.2.2.2}$$

Integrating the censoring processes C_h with respect to the compensators $\tilde{\Lambda}_h$, we define the processes $\Lambda_h(t) = \int_0^t C_h(u) d\tilde{\Lambda}_h(u)$. Recalling the definitions of the $\tilde{\Lambda}_h$ given in Proposition 2.2.1, we obtain:

$$\begin{aligned} \Lambda_A(t) &= \int_0^t \mathbb{1}_{\{u \le U_A\}} (\mathbb{1}_{\{\tilde{X}_B \le U_B, \tilde{X}_B < u\}} + \mathbb{1}_{\{u \le \tilde{X}_B, u \le U_B\}}) \\ &\qquad e^{\rho_{A,0} \mathbb{1}_{\{X_B < X_A, X_B < u\}} + \theta_{A,0}^T Z_A} \alpha_{A,0}(u) \, \mathbb{1}_{\{\tilde{X}_A \ge u\}} \, du \\ &= \int_0^t (e^{\rho_{A,0} + \theta_{A,0}^T Z_A} \delta_B \, \mathbb{1}_{\{T_B < u\}} + e^{\theta_{A,0}^T Z_A} \, \mathbb{1}_{\{T_B \ge u\}}) \, \mathbb{1}_{\{T_A \ge u\}} \, \alpha_{A,0}(u) du, \\ \Lambda_B(t) &= \int_0^t (e^{\rho_{B,0} + \theta_{B,0}^T Z_B} \delta_A \, \mathbb{1}_{\{T_A < u\}} + e^{\theta_{B,0}^T Z_B} \, \mathbb{1}_{\{T_A \ge u\}}) \, \mathbb{1}_{\{T_B \ge u\}} \, \alpha_{B,0}(u) du. \end{aligned}$$

After gathering terms, we can write the processes Λ_h as

$$\Lambda_h(t) = \int_0^t e^{\beta_{h,0}^T W_h(u)} Y_h(u) \alpha_{h,0}(u) du, \tag{2.2.2.3}$$

with notations (2.1.2.5) and (2.1.2.6).

Finally, define the filtration $\mathcal{F}_t$ as the filtration generated by one observation, namely by $\{\mathbb{1}_{\{T_A \le u\}}, \mathbb{1}_{\{T_B \le u\}}, u \le t, \delta_A, \delta_B, Z_A, Z_B\}$.

Proposition 2.2.2. *Under Assumption 2.2, the counting processes N_h for $h = A, B$ have predictable compensators Λ_h with respect to the filtration $(\mathcal{F}_t, t \ge 0)$.*

Proof. Define first the filtrations $\mathcal{U}_t$ as the filtration generated by $\{\mathbb{1}_{\{U_A \ge u\}}, \mathbb{1}_{\{U_B \ge u\}}, u \le t\}$ and $\mathcal{G}_t$ as the filtration generated by both $\mathcal{N}_t$ and $\mathcal{U}_t$. The censoring processes C_h are

cag-lad and clearly $\mathcal{G}_t$-adapted, so that they are $\mathcal{G}_t$-predictable.

Since the processes $\tilde{N}_t$ and $\tilde{\Lambda}_t$ are independent with respect to the filtration $\mathcal{U}_t$ and since they are $\mathcal{G}_t$-adapted, the $\mathcal{N}_t$-martingales $\tilde{M}_h$ are also $\mathcal{G}_t$-martingales.

Since the censoring processes C_h are bounded and $\mathcal{G}_t$-predictable, we can affirm that the processes $M_h(t) = \int_0^t C_h(u)d\tilde{M}_h(u)$ are also $\mathcal{G}_t$-martingales. Each σ-algebra $\mathcal{F}_t$ is contained in the σ-algebra $\mathcal{G}_t$, and it is clear from formulas (2.1.2.4) to (2.1.2.6) together with (2.2.2.3) that these martingales are also $\mathcal{F}_t$-measurable. Thus, they also are $\mathcal{F}_t$-martingales. □

The likelihood

We have now to write the likelihood of the model thanks to Jacod's formula (see Jacod [35, 36], Andersen *et al.*[5]). Since the counting processes $N_{h,i}, h = A, B, 1 \leq i \leq n$ cannot jump simultaneously, the likelihood is proportional to:

$$
\begin{aligned}
L_n(\beta) &= \prod_{h=A,B}\prod_{i=1}^{n}\prod_{s\leq t}(\alpha_{h,i}(s)Y_{h,i}(s))^{\Delta N_{h,i}(s)}\exp\left(-\int_0^\tau \alpha_{h,i}(u)Y_{h,i}(u)du\right)\\
&=\prod_{i=1}^{n}\prod_{s\leq t}\left(e^{\beta_A^T W_{A,i}}Y_{A,i}(s)\right)^{\Delta N_{A,i}(s)}\left(e^{\beta_B^T W_{B,i}}Y_{B,i}(s)\right)^{\Delta N_{B,i}(s)}\\
&\quad(\alpha_{A,0}(s))^{\Delta N_{A,i}(s)}(\alpha_{B,0}(s))^{\Delta N_{B,i}(s)}\\
&\quad\exp\left(-\int_0^\tau nS_{n,A}(\beta,u)\alpha_{A,0}(u)du\right)\exp\left(-\int_0^\tau nS_{n,B}(\beta,u)\alpha_{B,0}(u)du\right),
\end{aligned}
$$

where $S_{n,A}$ and $S_{n,B}$ have been defined by (2.1.2.8).

Maximizing in $\alpha_{A,0}$ and $\alpha_{B,0}$ for a fixed $\beta = (\rho_A, \theta_A, \rho_B, \theta_B)^T$, we finally find the log partial likelihood defined in (2.1.2.7).

Remark 2.1. Note that in the case where the two individuals are subject to the same kind of event (think of the case of two married people searching for a job), we may assume that $\alpha_{A,0} = \alpha_{B,0}$ and the maximizing of the likelihood leads to another form of the log partial likelihood. In that case, just replace the sums $S_{n,h}$ by $S_{n,A} + S_{n,B}$.

2.2.2 The theorems

At that point, we are exactly in the framework described by Andersen and Gill [6] and by Andersen *et al.*(Chapter VII of [5]). Therefore, we can apply Theorems VII.2.1 and VII.2.2 in order to establish our results. Recalling the definition (2.1.2.9) of our estimator, we begin with announcing the consistency of our estimator. Consider the following assumption:

Assumption 2.4. *Random variables $Z_{A,i}$ and $Z_{B,i}$ are not concentrated in a hyperplane.*

These assumption is classical. Assumption 2.4 ensures the asymptotic existence of the estimator.

Theorem 2.2.1. *Under Assumptions 2.1 to 2.4, the maximal log partial likelihood $\hat{\beta}_n$ tends to the true parameter β_0, when n tends to infinity.*

In the next theorem, the asymptotic behaviour of the estimator is established and an estimator of the inverse of the asymptotic covariance matrix is proposed.

Let us introduce some notations. Let us denote by DF and D^2F the vector of first derivatives and the matrix of second derivatives of the function F with respect to β. Let $S_h, h = A, B$ be defined by $S_h(\beta, s) = \mathbb{E}(S_{n,h}(s))$, let $E_h, h = A, B$ be the vectors $E_h = DS_h/S_h$ and let $V_h, h = A, B$ be the matrices $V_h = D^2S_h/S_h - (E_h)^{\otimes 2}$, where $x^{\otimes 2}$ denotes the vector product xx^T. Finally, let Σ_τ be the matrix:

$$\sum_{h=A,B} \int_0^\tau V_h(\beta_0, s) S_h(\beta_0, s) \alpha_{h,0}(s) ds. \tag{2.2.2.4}$$

Theorem 2.2.2. *Under Assumptions (2.1-2.4), the variable $n^{1/2}(\hat{\beta}_n - \beta_0)$ converges in distribution to a Gaussian variable $\mathcal{N}\left(0, (\Sigma_\tau)^{-1}\right)$, and the matrix $n^{-1}D^2l_n(\hat{\beta}_n)$ tends in probability to Σ_τ.*

The proofs of these theorems were first established in Andersen and Gill [6], they can also be found in Andersen *et al.*[5] (see also Appendix B). In the same paper, the authors propose some simple conditions to be verified in the special case where the observations are independent and identically distributed (Theorem 4.1 in [6]). Since we are in such a case, we only verify the assumptions of this theorem in section 2.4.2.

2.3 A simulation study

We carry out a simulation study in order to illustrate the theoretical properties of our estimator $\hat{\beta}_n$ and to investigate its finite sample properties (see the routines in Appendix C). As the behaviour of the Cox regression parameters estimators are now well-known, we focus our attention on the shock parameters. Thus, we run our simulation experiments without any covariates.

For different sample sizes ($n = 30, 50, 100$), we simulate a sample of independent and identically distributed variables $(T_{A,i}, T_{B,i}, \delta_{A,i}, \delta_{B,i}, Z_{A,i}, Z_{B,i})_{1 \leq i \leq n}$ of size n, following the method indicated in the introduction: we begin with simulating a n-sample of independent and identically distributed variables $(X_{A,i}, X_{B,i})_{1 \leq i \leq n}$ such that each pair has the law of two independent exponential random variables with parameter 1. That corresponds to constant hazard rates $\alpha_A = \alpha_B = 1$. Then, we simulate a n-sample of independent and identically distributed variables $(X'_{A,i}, X'_{B,i})_{1 \leq i \leq n}$ such that each pair has the law of two independent exponential random variables with parameters $e^{\rho_{A,0}}$ and $e^{\rho_{B,0}}$. Here, both $\rho_{A,0}$ and $\rho_{B,0}$ equal 1. Random variables $(\tilde{X}_{A,i}, \tilde{X}_{B,i})_{1 \leq i \leq n}$ result from formula (2.1.2.1). Right-censoring random variables $(U_{A,i}, U_{B,i})_{1 \leq i \leq n}$ are also generated as independent exponential random variables with parameter μ taking successively the values 0.0001, 0.1,

0.4 and 1. The approximate rate of censoring in each case is given in Tab. 2.1. The duration of study t is taken to be 10.

For each experiment, we calculate the estimators following remark 1 (remind that $\alpha_A = \alpha_B$), we calculate the estimator of the asymptotic covariance matrix and a 95%-confidence interval. We run $L = 500$ replications of such an experiment.

Table 2.1: Approximate rate of censoring with respect to the parameter of the censoring random variable μ

μ	0.0001	0.1	0.4	1
Approximate rate of censoring	0	0.06	0.21	0.42

For the different values of the sample size n and of the censoring parameter μ, we give the mean of the estimates over all replications, the approximate bias, the mean of asymptotic standard deviation, the coverage probabilities of the 95%-confidence interval calculated with the previous estimates. The results are given in Tab. 2.2.

Note that in all experiments, the bias is small. Not surprisingly, the asymptotic standard deviation decreases with the sample size for a given censoring parameter, and it increases with the censoring parameter for a given sample size. The coverage probabilities are closed to the one expected (0.025%, 0.950%, 0.025%).

To examine the performance of the estimators, we draw the empirical distribution functions of the conveniently renormalized estimates and we compare them to the distribution function of a standard Gaussian random variable. This is shown in Fig. 2.6 for a sample size equal to 50 and a censoring parameter equal to 0.1. We note that even with small sample sizes the distribution of the estimators are very closed to the asymptotic distribution.

2.4 Proofs

2.4.1 Proof of Proposition 2.2.1

The processes $\tilde{\Lambda}_h$ are nondecreasing and $\mathcal{N}_t$-adapted, thus they are $\mathcal{N}_t$-predictable. It suffices to prove that the processes $\tilde{M}_h$ defined just after Proposition 2.2.1 are $\mathcal{N}_t$-martingales, that is to say, for $\tilde{M}_A$, that for every $s < t$, $\mathbb{E}\left(\tilde{M}_A(t) - \tilde{M}_B(s) \mid \mathcal{N}_s\right) = 0$. In fact, we have to prove the five following points: for every $u \leq s < t$, $v \leq s < t$,

i) $\mathbb{E}\left(|\tilde{M}_A(t)|\right) < \infty$

ii) $\mathbb{E}\left(\left(\tilde{M}_A(t) - \tilde{M}_A(s)\right) \mathbb{1}_{\{\tilde{X}_A \leq u\}}\right) = 0$

iii) $\mathbb{E}\left(\left(\tilde{M}_A(t) - \tilde{M}_A(s)\right) \mathbb{1}_{\{\tilde{X}_B \leq v\}}\right) = 0$

iv) $\mathbb{E}\left(\left(\tilde{M}_A(t) - \tilde{M}_A(s)\right) \mathbb{1}_{\{\tilde{X}_A \leq u\}} \mathbb{1}_{\{\tilde{X}_B \leq v\}}\right) = 0$

Table 2.2: Small sample performance of the estimator for the shock parameters for various values of the censoring parameter μ and of the sample size n.

μ	n	$\hat{\rho}_A$ $\hat{\rho}_B$	$\hat{\rho}_A - \rho_{A,0}$ $\hat{\rho}_B - \rho_{B,0}$	$\hat{\sigma}_A$ $\hat{\sigma}_B$	< IC	∈ IC	> IC
0.0001	30	0.9862	-0.0138	0.3553	0.024	0.952	0.024
		1.0019	0.0019	0.3550	0.026	0.950	0.024
	50	1.0135	0.0135	0.2684	0.028	0.948	0.024
		0.9885	-0.0115	0.2688	0.028	0.0936	0.036
	100	1.0012	0.0012	0.1864	0.030	0.950	0.020
		0.9949	-0.0051	0.1868	0.032	0.956	0.012
0.1	30	0.9651	-0.0349	0.3790	0.024	0.950	0.026
		0.9879	-0.0121	0.3809	0.028	0.950	0.022
	50	0.9875	-0.0125	0.2871	0.040	0.932	0.028
		1.0094	0.0094	0.2876	0.026	0.950	0.024
	100	0.9857	-0.0143	0.1990	0.032	0.0940	0.028
		0.9925	-0.0075	0.1992	0.022	0.944	0.034
0.4	30	1.0155	0.0155	0.4583	0.040	0.938	0.022
		0.9785	-0.0215	0.4588	0.030	0.944	0.026
	50	0.9985	-0.0015	0.3411	0.026	0.944	0.030
		1.0094	0.0094	0.3405	0.038	0.944	0.018
	100	0.9974	-0.0026	0.2355	0.042	0.930	0.028
		0.9961	-0.0039	0.2360	0.030	0.930	0.040
1	30	0.9669	-0.0331	1.5212	0.040	0.934	0.026
		0.9826	-0.0174	1.1174	0.044	0.940	0.016
	50	1.0191	0.0191	0.4526	0.022	0.958	0.020
		1.0349	0.0349	0.4531	0.034	0.954	0.012
	100	1.0017	0.0017	0.3019	0.032	0.948	0.020
		0.9993	-0.0007	0.3068	0.034	0.948	0.018

v) $\mathbb{E}\left(\tilde{M}_A(t) - \tilde{M}_A(s)\right) = 0$

First of all, note the equality of the following events appearing in the definition (2.2.2.1) of the functions $\tilde{\alpha}_H, h = A, B$:

$$\begin{aligned} \{X_B < X_A, X_B < t\} &= \{\tilde{X}_B < \tilde{X}_A, \tilde{X}_B < t\}, \\ \{X_A < X_B, X_A < t\} &= \{\tilde{X}_A < \tilde{X}_B, \tilde{X}_A < t\}. \end{aligned}$$

The direct implication is clear. For the reverse implication, just recall the definition (2.1.2.1). If $\tilde{X}_B < \tilde{X}_A$, then $\tilde{X}_B \, \mathbb{1}_{\{X_A<X_B\}} < \tilde{X}_A \, \mathbb{1}_{\{X_B<X_A\}}$, and since only one of both indicators can equal 1, necessarily $X_B < X_A$, and the conclusion follows.

Point i) is fulfilled with Assumption 2.2.

Points ii) and iv) result from the equalities $(\tilde{M}_A(t) - \tilde{M}_A(s)) \, \mathbb{1}_{\{\tilde{X}_A \le u\}} = 0$ and $(\tilde{M}_A(t) - \tilde{M}_A(s)) \, \mathbb{1}_{\{\tilde{X}_A \le u\}} \, \mathbb{1}_{\{\tilde{X}_B \le v\}} = 0$.

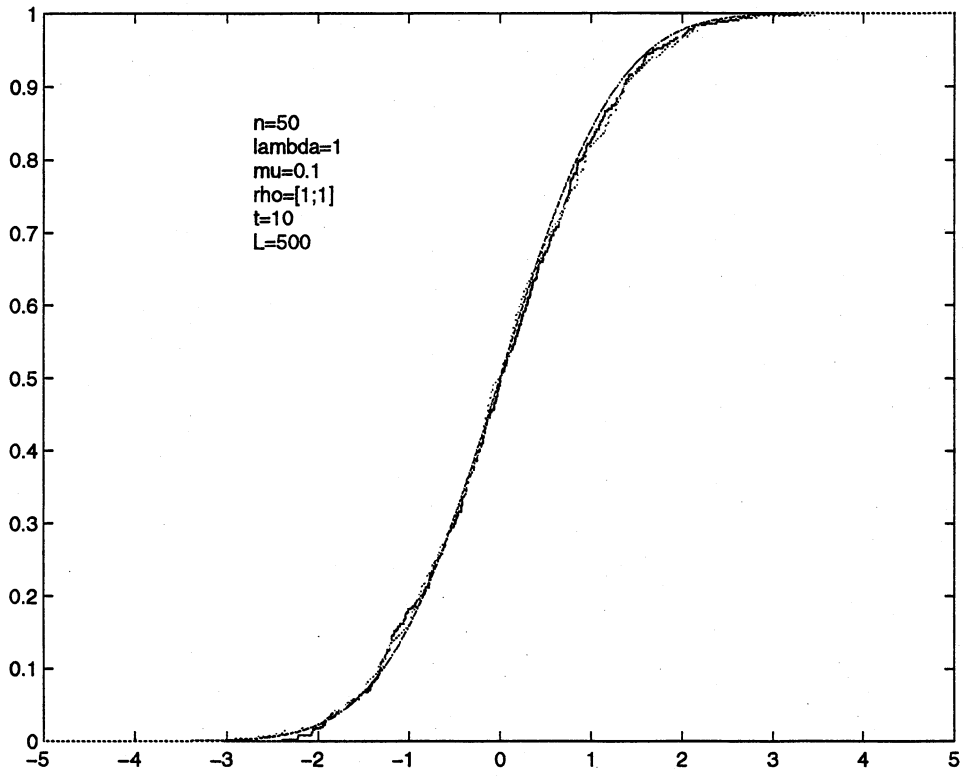

Figure 2.6: Empirical distribution functions of the estimators in comparison with the standard Gaussian distribution function

To prove point iii), by definition of $\tilde{X}_A$ and $\tilde{X}_B$, we have:

$$\left(\tilde{M}_A(t) - \tilde{M}_A(s)\right) \mathbb{1}_{\{\tilde{X}_B \leq v\}} = \mathbb{1}_{\{s < X_B + X'_A \leq t, X_B \leq v\}} - \int_s^t e^{\rho_{A,0}} \alpha_A(x) \, \mathbb{1}_{\{X_B + X_{A'} \geq x, X_B \leq v\}} \, dx.$$

The expectation of the first term is $\mathbb{P}\left(s < X_B + X'_A \leq t, X_B \leq v\right)$ that we calculate conditioning on X_B and (Z_A, Z_B):

$$\mathbb{P}\left(s < X_B + X'_A \leq t, X_B \leq v\right) = \mathbb{E}\left(\int_s^t e^{\rho_{A,0}} \alpha_A(x) \exp\left(-\int_{X_B}^x e^{\rho_{A,0}} \alpha_A\right) dx \, \mathbb{1}_{\{X_B \leq v\}}\right).$$

For the expectation of the second term, we also condition on X_B and (Z_A, Z_B), and we find:

$$\mathbb{E}\left(\int_s^t e^{\rho_{A,0}}\alpha_A(x)\,\mathbb{1}_{\{X_B+X_{A'}\geq x, X_B\leq v\}}\,dx\right) = \mathbb{E}\left(\int_s^t e^{\rho_{A,0}}\alpha_A(x)\exp\left(-\int_{X_B}^x e^{\rho_{A,0}}\alpha_A\right)dx\,\mathbb{1}_{\{X_B\leq v\}}\right),$$

so that $\mathbb{E}\left(\left(\tilde{M}_A(t)-\tilde{M}_A(s)\right)\mathbb{1}_{\{\tilde{X}_B\leq v\}}\right)=0$.

We handle in the same way for point v).

2.4.2 Verification of Conditions VII.2.1 and VII.2.2 of Andersen *et al.*[5]

Since the observations $(T_{A,i}, T_{B,i}, \delta_{A,i}, \delta_{B,i}, Z_{A,i}, Z_{B,i})_{1\leq i\leq n}$ are independent and identically distributed, we are typically in the framework given by Theorem 4.1 of [6] (see also Example VII.2.7 of [5]). We only have to verify the following points:

i) $\int_0^\tau \alpha_{h,0}(s)ds$, $h=A,B$ are finite

ii) There exists a neighbourhood $\mathcal{B}$ of β_0 such that for any $h=A,B$, $\mathbb{E}\left(\sup_{\beta\in\mathcal{B},0\leq s\leq t} Y_h(s)|W_h(s)|^2\exp(\beta_h^T W_h(s))\right)$ is finite

iii) For any $h=A,B$, $\mathbb{P}\left(Y_h(s)=1,\text{ for any } 0\leq s\leq\tau\right)$ is positive

iv) The asymptotic covariance matrix Σ_τ is positive definite

Point i) is exactly Assumption 2.2.

For point ii), note that

$$Y_h(s)|W_h(s)|^2\exp(\beta_h^T W_h(s)) \leq (1+|Z_h|^2)e^{\rho_h+\theta_h^T Z_h},$$

and define $\mathcal{B}$ as $\mathbb{R}^2\times\Theta_A\times\Theta_B$, where Θ has been defined in Assumption 2.3.

For point iii), note that in our definition (2.1.2.6), the functions $Y_{h,i}$ are nonincreasing with respect to s, so that the event in the probability is the same as $Y_h(\tau)=1$.

$$\begin{aligned}\mathbb{P}(Y_A(\tau)=1) &= \mathbb{P}\left(\tilde{X}_B\leq U_B, \tilde{X}_B<\tau, \tilde{X}_A\geq\tau, U_A\geq\tau\right)\\ &\quad+\mathbb{P}\left(\tilde{X}_A\geq\tau, U_A\geq\tau, \tilde{X}_B\geq\tau, U_B\geq\tau\right)\\ &\geq \mathbb{E}\left(\mathbb{P}\left(\tilde{X}_A\geq\tau, U_A\geq\tau, \tilde{X}_B\geq\tau, U_B\geq\tau\mid(Z_A,Z_B)\right)\right).\end{aligned}$$

Thanks to Assumption 2.1, the conditional probability also equals

$$\mathbb{E}\left(\mathbb{P}(U_A\geq\tau, U_B\geq\tau\mid(Z_A,Z_B))\,\mathbb{P}\left(\tilde{X}_A\geq\tau,\tilde{X}_B\geq\tau\mid(Z_A,Z_B)\right)\right).$$

The first conditional probability is positive from Assumption 2.1. For the second, we can write that

$$\mathbb{P}\left(\tilde{X}_A \geq \tau, \tilde{X}_B \geq \tau \mid (Z_A, Z_B)\right) = \mathbb{P}\left(\inf(\tilde{X}_A, \tilde{X}_B) \geq \tau \mid (Z_A, Z_B)\right).$$

Since $\inf(\tilde{X}_A, \tilde{X}_B) = \inf(X_A, X_B)$, this probability also equals

$$\mathbb{P}\left(X_A \geq \tau, X_B \geq \tau \mid (Z_A, Z_B)\right) = \exp(-\sum_{h=A,B} \int_0^\tau \alpha_{h,0}(sds)),$$

which is positive from Assumption 2.2. The same arguments hold for Y_B.

The last point is the object of the following lemma:

Lemma 2.4.1. *Suppose that Assumption 2.4 is verified, then Σ_τ is positive definite.*

Proof. Recall the expression of Σ_τ (2.2.2.4):

$$\Sigma_\tau = \sum_{h=A,B} \int_0^\tau (D^2 S_h / S_h - (DS_h/S_h)^{\otimes 2})(\beta_0, s) S_h(\beta_0, s) \alpha_{h,0}(s) ds,$$

of S_h, $h = A, B$ $S_h(\beta, s) = \mathbb{E}\left(e^{\beta_h^T W_h(s)} Y_h(s)\right)$ and of their derivatives

$$DS_A(\beta, s) = \begin{pmatrix} \mathbb{E}\left(e^{\beta_A^T W_A(s)} Y_A(s) W_A(s)\right) \\ 0 \end{pmatrix}, \tag{2.4.2.1}$$

$$DS_B(\beta, s) = \begin{pmatrix} 0 \\ \mathbb{E}\left(e^{\beta_B^T W_B(s)} Y_B(s) W_B(s)\right) \end{pmatrix}, \tag{2.4.2.2}$$

$$D^2 S_A(\beta, s) = \begin{pmatrix} \mathbb{E}\left(e^{\beta_A^T W_A(s)} Y_A(s) W_A(s)^{\otimes 2}\right) & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.4.2.3}$$

$$D^2 S_B(\beta, s) = \begin{pmatrix} 0 & 0 \\ 0 & \mathbb{E}\left(e^{\beta_B^T W_B(s)} Y_B(s) W_B(s)^{\otimes 2}\right) \end{pmatrix}. \tag{2.4.2.4}$$

Since all our hypotheses are formulated in term of the conditional distribution of the random variables given the covariates, we need to express the matrix Σ_τ as the expectation in some sense of the conditional expectation given the covariates of some random matrix.

For this purpose, we introduce some positive random variable R, with distribution

$$d\mu = \tau^{-1} \mathbb{1}_{\{[0,\tau]\}} ds,$$

where ds denotes the Lebesgue measure, and H be some random variable in $\{A, B\}$ with

conditional distribution given R,

$$\begin{aligned}\mathbb{P}(H=A \mid R) &= \frac{S_A\alpha_{0,A}}{S_A\alpha_{0,A}+S_B\alpha_{0,B}}(\beta_0,R),\\ \mathbb{P}(H=B \mid R) &= \frac{S_B\alpha_{0,B}}{S_A\alpha_{0,A}+S_B\alpha_{0,B}}(\beta_0,R).\end{aligned}$$

Denoting by $d\Pi$ the law of the pair (R,H), we can write

$$\Sigma_\tau = \tau\mathbb{E}_\Pi\left((S_A\alpha_{0,A}+S_B\alpha_{0,B})(\beta_0,R)M_H(\beta_0,R)\right),$$

where

$$M_A=\begin{pmatrix}\frac{D^2S_A}{S_A}-\left(\frac{DS_A}{S_A}\right)^{\otimes 2} & 0\\ 0 & 0\end{pmatrix},\quad M_B=\begin{pmatrix}0 & 0\\ 0 & \frac{D^2S_B}{S_B}-\left(\frac{DS_B}{S_B}\right)^{\otimes 2}\end{pmatrix}.$$

Now, if F denotes the law of the pair (Z_A,Z_B), we remark that $S_h(\beta_0,s) = \mathbb{E}_F\left(\mathbb{E}\left(e^{\beta_{h,0}^T W_h(s)}Y_h(s) \mid (Z_A,Z_B)\right)\right)$. Denoting by $dP_{s,h}$ the distribution

$$dP_{s,h}=\frac{\mathbb{E}\left(e^{\beta_{h,0}^T W_h(s)}Y_h(s) \mid (Z_A,Z_B)\right)}{S_h(\beta_0,s)}dF,$$

we have

$$\begin{aligned}\Sigma_\tau = \tau\mathbb{E}_\Pi\big[&(S_A+S_B)(\beta_0,R)\\ &\mathbb{E}_{P_{s,h}}\left[\left(\mathcal{Z}_h-\mathbb{E}_{P_{s,h}}\left(\mathcal{Z}_h \mid R=s,H=h\right)\right)^{\otimes 2} \mid R=s,H=h\right]\big],\end{aligned}$$

where

$$\begin{aligned}\mathcal{Z}_A &= \begin{pmatrix}\frac{\mathbb{E}\left(e^{\beta_{A,0}^T W_A(s)}Y_A(s)\,\mathbb{1}_{\{T_B<s\}}|(Z_A,Z_B)\right)}{\mathbb{E}\left(e^{\beta_{A,0}^T W_A(s)}Y_A(s)|(Z_A,Z_B)\right)}\\ Z_A\\ 0\end{pmatrix},\\ \mathcal{Z}_B &= \begin{pmatrix}0\\ \frac{\mathbb{E}\left(e^{\beta_{B,0}^T W_B(s)}Y_B(s)\,\mathbb{1}_{\{T_A<s\}}|(Z_A,Z_B)\right)}{\mathbb{E}\left(e^{\beta_{B,0}^T W_B(s)}Y_B(s)|(Z_A,Z_B)\right)}\\ Z_B\end{pmatrix}.\end{aligned}$$

Let $x=(x_A,x_B)$ be a $(2D+2)$-dimensional vector. Let us calculate $x^T\Sigma_\tau x$:

$$\begin{aligned}x^T\Sigma_\tau x = \tau\mathbb{E}_\Pi\big(&(S_A\alpha_{0,A}+S_B\alpha_{0,B})(\beta_0,R)\\ &\mathbb{E}_{P_{s,h}}\left[x_h^T\left(\mathcal{Z}_h-\mathbb{E}_{P_{s,h}}\left(\mathcal{Z}_h \mid R=s,H=h\right)\right)^{\otimes 2}x_h \mid R=s,H=h\right]\big).\end{aligned}$$

Now,

$$\mathbb{E}\left(e^{\beta_{A,0}^T W_A(s)} Y_A(s) \mid (Z_A, Z_B)\right) =$$
$$e^{\theta_{A,0}^T Z_A}\, \mathbb{E}\left(e^{\rho_{A,0}\, 1\!\!1_{\{T_B<s\}}}\, 1\!\!1_{\{T_A\geq s\}}(\delta_B\, 1\!\!1_{\{T_B<s\}} + 1\!\!1_{\{T_B\geq s\}}) \mid (Z_A, Z_B)\right)$$
$$\geq e^{\theta_{A,0}^T Z_A}\, \mathbb{P}\left(T_A \geq s, T_B \geq s \mid (Z_A, Z_B)\right),$$

the last probability being positive (see point iii) just before). Therefore, the measures $dP_{s,h}$ and dF are equivalent. Thus, the equation $x^T \Sigma_\tau x = 0$ is equivalent to

$$x_h^T \left(\mathcal{Z}_h - \mathbb{E}_{P_{s,h}}\left(\mathcal{Z}_h \mid \tau = s, H = h\right)\right)^{\otimes 2} x_h = 0,$$

F-almost surely, Π-almost surely, that is to say,

$$\left(\mathcal{Z}_h - \mathbb{E}_{P_{s,h}}\left(\mathcal{Z}_h \mid \tau = s, H = h\right)\right)^T x_h = 0,$$

F-almost surely, Π-almost surely. That implies that $\mathcal{Z}_h^T x_h$ degenerates. Since

$$\mathbb{E}\left(e^{\beta_{A,0}^T W_A(s)} Y_A(s)\, 1\!\!1_{\{T_B<s\}} \mid (Z_A, Z_B)\right) =$$
$$e^{\rho_{A,0}+\theta_{A,0}^T Z_A}\, \mathbb{P}\left(\tilde{X}_A \geq s, U_A \geq s, \tilde{X}_B < s, \tilde{X}_B \leq U_B \mid (Z_A, Z_B)\right),$$

which is positive, and since the distribution of (Z_A, Z_B) does not degenerate, the last assertion implies that $x_h = 0$ for $h = A, B$, which concludes the proof. □

Annexe A

Application de la méthode de sélection de modèle au modèle de Cox[1]

1. Ce chapitre présente un travail en collaboration avec M.-L. Martin et G. Castellan.

Dans le chapitre 1, nous proposons une méthode d'estimation de la fonction de régression dans le modèle de Cox par sélection de modèle. Notre théorème 1.3.1 donne une fonction de pénalité pour laquelle nous sommes capables de prouver la borne de risque 1.3.1.3. Cette fonction de pénalité est assez compliquée et est définie pour des espaces peu naturels: l'introduction des bornes B, par exemple, semble peu naturelle et on aimerait pouvoir estimer la fonction de régression f sur les espaces linéaires tout entiers et pas seulement sur des boules L_∞ notées $\mathcal{S}_m$ dans le chapitre 1 (voir remarque 1.4).

Une étude de simulations a donc été commencée par M.-L. Martin, en collaboration avec G. Castellan et moi-même, pour tenter d'appliquer concrètement la méthode de sélection de modèle à l'estimation de la fonction de régression dans le modèle de Cox. Il s'agit de voir si une pénalité proportionnelle au rapport de la dimension de l'espace sur la taille de l'échantillon permet de sélectionner un "bon" modèle, le cas échéant, quelle doit être la constante de proportionalité et si cette constante dépend notamment de la censure, comme le suggère notre théorème 1.3.1.

Dans cette annexe, nous présentons les tout premiers résultats de notre étude de simulation, qui est encore en cours.

A.1 Cadre et notations

Nous considérons ici des histogrammes réguliers, c'est-à-dire, des fonctions constantes par morceaux, dont les morceaux ont tous la même taille. La contrainte d'identifiabilité 1.1.1.1 sera ici remplacée par la suivante:

La fonction de régression f est supposée nulle sur un intervalle $[0,\varepsilon]$ (dans toute la suite, $\varepsilon = 0{,}02$).

Nous notons donc cette fois $I_{D,k}$ pour $D \geq 1$ et $0 \leq k \leq D$ l'intervalle défini par

$$\begin{aligned} I_{D,0} &= [0,\varepsilon] \\ I_{D,k} &= \left]\varepsilon + \frac{k-1}{D}(1-\varepsilon), \varepsilon + \frac{k}{D}(1-\varepsilon)\right[, \text{ si } 1 \leq k \leq D \end{aligned}$$

et par $\varphi_{D,k}$ pour $D \geq 1$ et $0 \leq k \leq D$ les fonctions $\varphi_{D,k}(x) = \mathbb{1}_{\{I_{D,k}\}}(x)$. Nous notons Λ^D l'ensemble d'indices $\Lambda^D = \{(D,k), 0 \leq k \leq D\}$, et Λ_n la collection $\Lambda_n = \{\Lambda^D, |\Lambda^D| - 1 = D \leq N_n\}$, où $N_n \leq \frac{n}{\log n}$, et $|\Lambda^D|$ est le cardinal de Λ^D. En suivant la remarque 1.4 du chapitre 1, nous considérons comme modèles les espaces linéaires engendrés par les fonctions $\varphi_{D,k}$. Les espaces sur lesquels on minimise, notés $\mathcal{S}_{\Lambda^D}$, sont définis par

$$\mathcal{S}_{\Lambda^D} = \left\{ g = \sum_{\lambda \in \Lambda^D} \beta_\lambda \varphi_\lambda = \sum_{k=0}^{D} \beta_{D,k} \mathbb{1}_{\{I_{D,k}\}}, \beta_0 = 0 \right\},$$

la contrainte $\beta_0 = 0$ étant la contrainte d'identifiabilité. La dimension de l'espace linéaire

S_{Λ^D} est bien D. L'estimateur sur le modèle Λ^D est défini par

$$\hat{f}_{\Lambda^D} = \underset{g \in S_{\Lambda^D}}{\operatorname{argmin}} \gamma_n(g),$$

où le constraste γ_n a été défini par 1.2.1.1. Dans la suite, nous définirons la fonction de pénalité par

$$\operatorname{pen}_n(\Lambda^D) = \mu \frac{D}{n},$$

pour différentes valeurs possibles de μ. Notre estimateur est défini par $\tilde{f} = \hat{f}_{\hat{\Lambda}}$, avec

$$\hat{\Lambda} = \underset{\Lambda^D \in \Lambda_n}{\operatorname{argmin}} \left\{ \gamma_n(\hat{f}_{\Lambda^D}) + \operatorname{pen}_n(\Lambda^D) \right\}.$$

A.2 Simulations

Cette étude de simulations (qui est encore en cours) a été réalisée dans les conditions suivantes :

- La covariable W suit la loi uniforme sur l'intervalle [0,1].
- Le taux de risque de base α_0 est constant et égal à 1.
- La variable aléatoire de censure U suit la loi exponentielle de paramètre $\xi_0 \in \{0.01, 0.03, 0.5, 0.75, 1, 1.5\}$.
- Les tailles d'échantillon n peuvent prendre les valeurs 250, 500, 1000.
- Le temps de fin d'étude τ est 10.

Nous considérons trois fonctions de régression possibles notées f_1, f_2 et f_3 dont les graphes sont visibles sur la figure A.1.

Remarquons que les fonctions f_1 et f_2 définissent toutes les deux des histogrammes réguliers à 4 morceaux sur l'intervalle $[\varepsilon,1]$, elles appartiennent donc à l'espace S_{Λ^4}. En revanche, la fonction f_3, si elle est bien constante par morceaux, n'appartient à aucun modèle de la collection Λ_n, puisque ses morceaux sur $[\varepsilon,1]$ ne sont pas réguliers.

A.3 Premiers résultats

A.3.1 Modèles sélectionnés

Pour savoir si la méthode de sélection de modèle fonctionne avec une pénalité proportionelle à D/n, nous pouvons commencer par choisir une fonction de régression dans l'un des modèles de la collection et voir si la méthode permet de retrouver ce modèle. C'est l'objet de ce paragraphe.

Nous cherchons donc à estimer les fonctions f_1, f_2 et f_3 par des histogrammes réguliers sur l'intervalle $[\varepsilon,1]$, pour différents paramètres de censure ξ_0. Pour une taille d'échantillon $n = 250$, une constante de pénalité $\mu = 2$ et pour $L = 100$ simulations, nous comptons

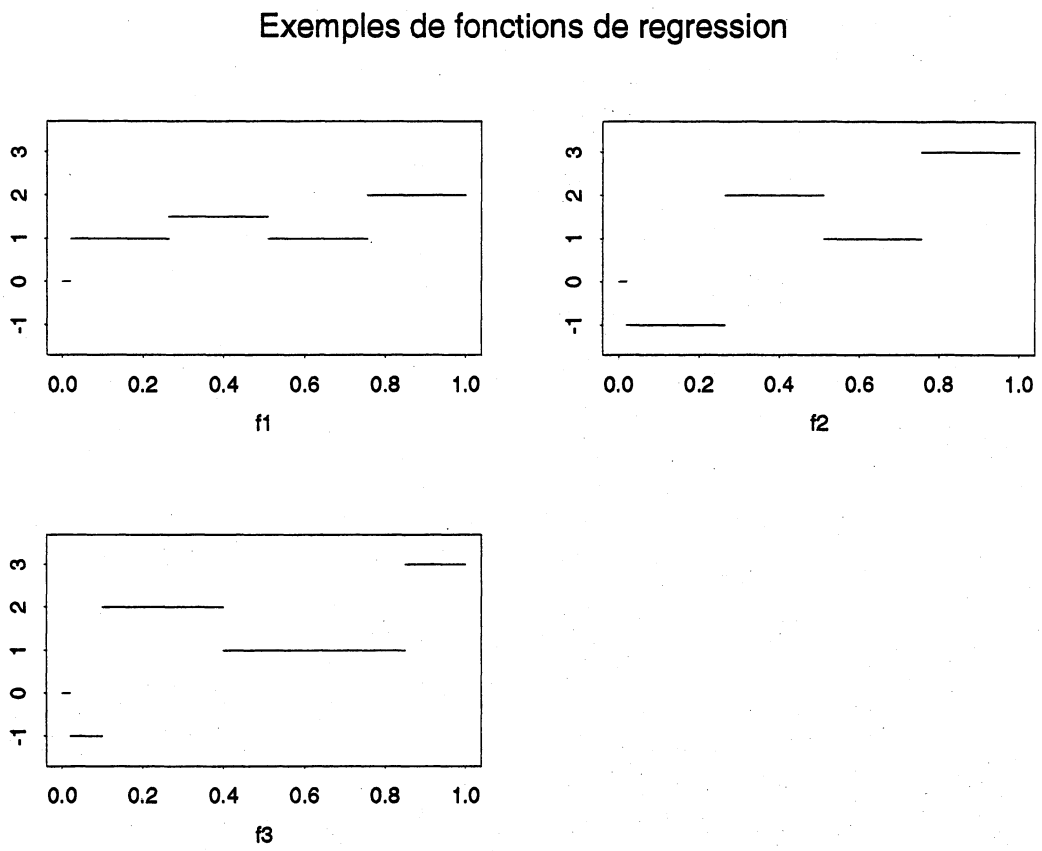

FIG. A.1 – *Graphes des fonctions de régression f_1, f_2 et f_3.*

le nombre de modèles sélectionnés $\hat{\Lambda}$, de dimension D. Les résultats sont donnés dans le tableau A.1.

f	ξ_0	3	4	5	6	7	8	9	10	11	12	13	14
f_1	1	4	**93**	1	0	0	2	0	0	0	0	0	0
f_2	0.5	0	**100**	0	0	0	0	0	0	0	0	0	0
f_3	1	0	0	0	0	0	0	2	1	5	**34**	**57**	1

TAB. A.1 – *Nombre de modèles de dimension D sélectionnés sur 100 simulations*

On remarque que, pour les fonctions f_1 et f_2, une pénalité en $2D/n$ permet de choisir très souvent le vrai modèle (Λ^4) : 93 cas sur 100 pour f_1 avec un paramètre de censure $\xi_0 = 1$ (environ 20 % de censure) et 100 % pour f_2 avec un paramètre de censure $\xi_0 = 1$ (environ 10 % de censure). La fonction f_3 n'appartient à aucun modèle de la collection : il faut donc plus de morceaux pour l'estimer correctement (12 ou 13).

Cependant, regarder quel modèle a été sélectionné permet de tester la méthode quand

la vraie fonction de régression appartient à l'un des modèles, mais est difficilement interprétable quand la fonction de régression n'appartient à aucun modèle de la collection. Dans ce cas, la grandeur significative est le risque de l'estimateur sélectionné, puisque la méthode de sélection de modèle prétend choisir un modèle dont le risque est de l'ordre du plus petit risque des estimateurs de la collection. Il s'agit donc de comparer le risque de l'estimateur sélectionné au plus petit risque des estimateurs de la collection.

A.3.2 Rapport de risques

Dans ce paragraphe, nous cherchons à évaluer le rapport du risque de l'estimateur sélectionné sur le plus petit risque des estimateurs de la collection :

$$\frac{\mathbb{E}\left(K(f,\tilde{f})\right)}{\inf_{\Lambda\in\Lambda_n}\mathbb{E}\left(K(f,\hat{f}_\Lambda)\right)}. \tag{A.3.A.1}$$

Pour cela, nous estimons la fonction f_1 pour un paramètre de censure $\xi_0 = 0.5$, pour des tailles d'échantillon $n = 250, 500, 1000$ et pour différentes constantes de pénalité μ variant de 1 à 4, par pas de 0.2. Pour chaque valeur de μ, nous estimons le rapport A.3.A.1. Ici, la fonction $K(f,g)$ est calculée par une approximation de Riemann pour toute fonction de régression f et toute fonction g appartenant à l'un des modèles, alors que les espérances, qui portent ici sur l'estimateur $\hat{f}_{\hat{\Lambda}}$ au numérateur et sur les estimateurs $\hat{f}_\Lambda$ au dénominateur, sont estimées par simulation : le numérateur est estimé par la moyenne de 200 simulations et le dénominateur par la moyenne de 1500 simulations. Les résultats sont visibles sur la figure A.2.

On remarque que ce rapport des risques varie avec la constante de pénalité μ : il semble augmenter pour des petites et des grandes valeurs de μ. Une constante de pénalité μ comprise entre 2.4 et 2.6 semble être optimale, quelle que soit la taille de l'échantillon n. Cependant, cette valeur n'est vraie que pour la fonction f_1 et le taux de censure $\xi_0 = 0.5$. Une étude de simulation complète demanderait de calculer le rapport des risques A.3.A.1 pour un grand nombre de fonctions de régression f et pour un grand nombre de paramètres de censure ξ_0. Notre théorème 1.3.1 suggère que la constante de pénalité "idéale" dépend du paramètre de la censure ξ_0. Or, celui-ci n'étant pas connue, il faudrait trouver un moyen de l'estimer et de voir comment il intervient dans la pénalité.

A.4 Comment trouver la "bonne" fonction de pénalité ?

A.4.1 C_p de Mallow et heuristique

Nous présentons ici une heuristique proposée par Birgé et Massart, qui consiste à tenter de trouver la bonne constante de pénalité à partir des données elles-même. Cette heuristique, bâtie à partir du cadre des processus linéaires gaussiens (voir Birgé et Massart [11, 14]) est basée sur celle du C_p de Mallow, que nous rappelons ici.

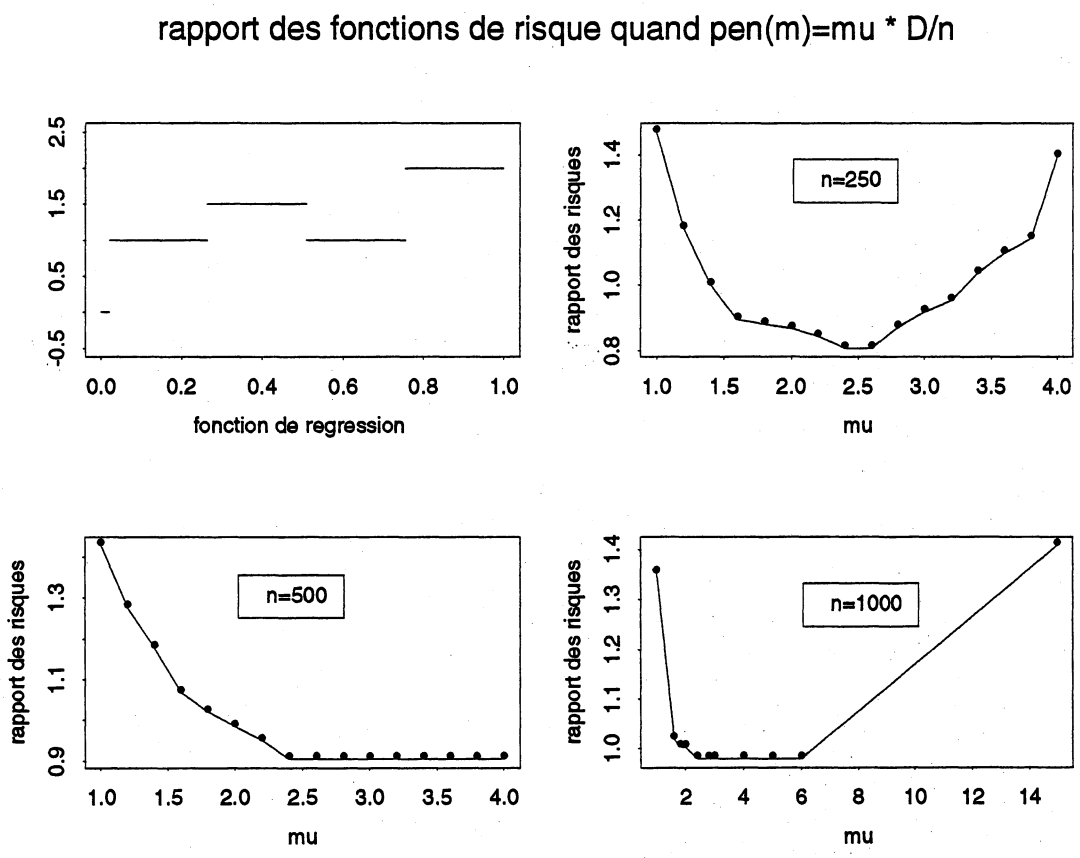

FIG. A.2 – *Rapport des risques pour l'estimation de* f_1

On se place dans le cadre d'un processus linéaire gaussien Y défini sur un sous-espace linéaire $\mathbb{S}$ d'un espace de Hilbert $\mathbb{H}$:

$$Y(t) = \langle s,t \rangle + \varepsilon Z(t), t \in \mathbb{S},$$

où $s \in \mathbb{H}$ est la moyenne du processus Y, ε^2 sa variance et Z un processus linéaire isonormal indexé par $\mathbb{S}$ (voir Dudley [29] pour les définitions classiques). Ce cadre contient notamment ceux de la régression linéaire gaussienne et du bruit blanc gaussien. On cherche à estimer s à partir des observations Y, la quantité ε étant connue (typiquement, $\varepsilon = 1/\sqrt{n}$ où n est la quantité d'observations disponibles).

On considère maintenant une collection $\{S_m\}_{m \in \mathcal{M}}$ de sous-espaces linéaires de dimension finie de $\mathbb{H}$, et on suppose que ces espaces linéaires S_m sont engendrés par des fonctions $\{\varphi_j, 1 \leq j \leq m\}$ où le système $\{\varphi_j, 1 \leq j \leq m\}$ est supposé orthonormal. On définit le

contraste γ pour tout $t \in \mathbb{S}$ par

$$\gamma(t) = \|t\|^2 - 2Y(t),$$

où $\|.\|$ désigne la norme L_2. L'estimateur sur le modèle S_m de dimension D_m est défini par

$$\hat{s}_m = \underset{t \in S_m}{\operatorname{argmin}} \gamma(t).$$

Dans ce cas, on peut toujours décomposer le risque quadratique de l'estimateur $\hat{s}_m$ en deux termes

$$\mathbb{E}\left(\|s - \hat{s}_m\|^2\right) = \|s - s_m\|^2 + \theta \frac{D_m}{n},$$

où s_m est le projeté orthogonal de s sur le modèle S_m et θ est un paramètre de variance inconnu. Le modèle idéal m_0 serait celui qui minimise le risque quadratique, mais celui-ci est indisponible, car il dépend de s inconnu. Par l'inégalité de Pythagore, on peut encore écrire:

$$\mathbb{E}\left(\|s - \hat{s}_m\|^2\right) = \|s\|^2 - \|s_m\|^2 + \theta \frac{D_m}{n},$$

il suffit donc de minimiser $-\|s_m\|^2 + \theta \frac{D_m}{n}$. L'heuristique du C_p de Mallow consiste à remplacer le terme $\|s_m\|^2$ par son estimateur non biaisé $\|\hat{s}_m\|^2 - \theta \frac{D_m}{n}$. Le bon critère semble donc être $-\|\hat{s}_m\|^2 + 2\theta \frac{D_m}{n}$. Mais θ est ici toujours inconnu, il faut donc l'estimer.

Supposons maintenant que s appartienne à un modèle S_D de dimension raisonnable. L'idée de Birgé et Massart consiste à dire que la quantité $\|\hat{s}_m\|^2$ est une fonction affine de D_m/n de pente θ pour les modèles $m \geq D$. Il suffirait alors pour estimer θ, de tracer la droite de la quantité $\|\hat{s}_m\|^2$ en fonction du ratio D_m/n et d'estimer la pente de cette droite sur les grandes dimensions. Si $\hat{p}$ est un estimateur de cette pente, la "bonne" pénalité devrait être

$$\text{pen}_n(m) = 2\hat{p} \frac{D_m}{n}.$$

A.4.2 Application au modèle de Cox

Revenons maintenant au modèle de Cox et tentons de lui appliquer cette heuristique pour définir la "bonne" pénalité.

Pour pouvoir appliquer l'heuristique de Birgé et Massart, il faut dans un premier temps vérifier que la quantité $\gamma_n(\hat{f}_{\Lambda D})$ est bien une fonction affine de D/n. Dans cette nouvelle étude de simulations, la fonction de régression est la suivante:

$$\begin{aligned} f_4(x) &= 0, \text{ si } 0 \leq x \leq \varepsilon \\ &= 1, \text{ si } \varepsilon < x \leq 0.51 \\ &= 2, \text{ si } 0.51 < x \leq 1 \end{aligned}$$

Avec cette fonction de régression, nous donnons dans le tableau A.2 les taux de censure approximatifs, selon le paramètre de la censure.

ξ_0	0.01	0.03	0.5	0.75	1	1.5
Taux de censure approximatif	0	0.03	0.1	0.15	0.2	0.25

TAB. A.2 – *Taux de censure approximatif pour la fonction de régression f_4 et diverses valeurs du paramètre de censure*

La figure A.3 montre la quantité $-\gamma_n(\hat{f}_{\Lambda D})$ en fonction de D/n pour des valeurs de D variant de 1 à 60 et pour deux valeurs du paramètre de censure $\xi_0 = 0.01$ et $\xi_0 = 1$.

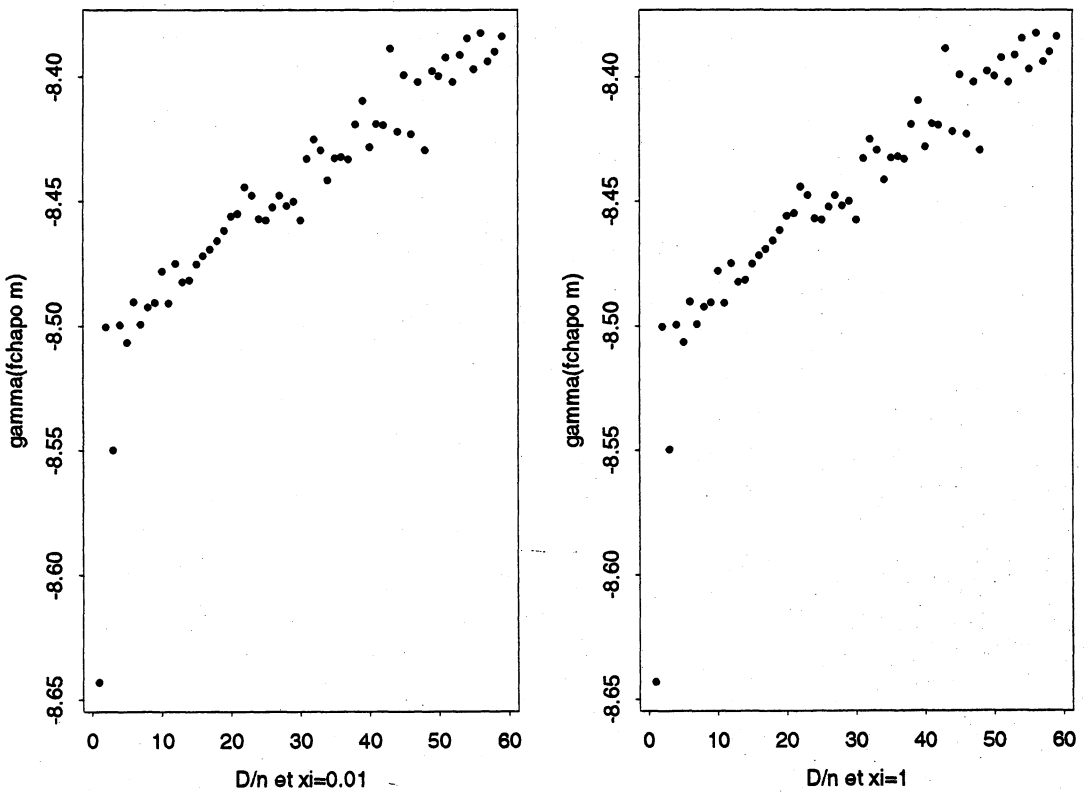

FIG. A.3 – $-\gamma_n(\hat{f}_{\Lambda D})$ *en fonction de* D/n *pour* $\xi_0 = 0.01$ *et* $\xi_0 = 1$

Le fait que $-\gamma_n(\hat{f}_{\Lambda D})$ soit affine en D/n est ici clairement vérifié, quelle que soit la valeur de la censure $\xi_0 = 0.01$ (pratiquement pas de censure) et $\xi_0 = 1$ (environ 20% de censure).

La figure A.4 permet de voir les variations de la pente $\hat{p}$ en fonction de la censure. Pour cinq valeurs de la censure, on y voit la valeur de $\hat{p}$, ainsi que des intervalles de confiance. On remarque que la pente semble varier avec le paramètre de la censure ξ_0, ce qui semble confirmer le fait que la constante de pénalité doive dépendre de la censure.

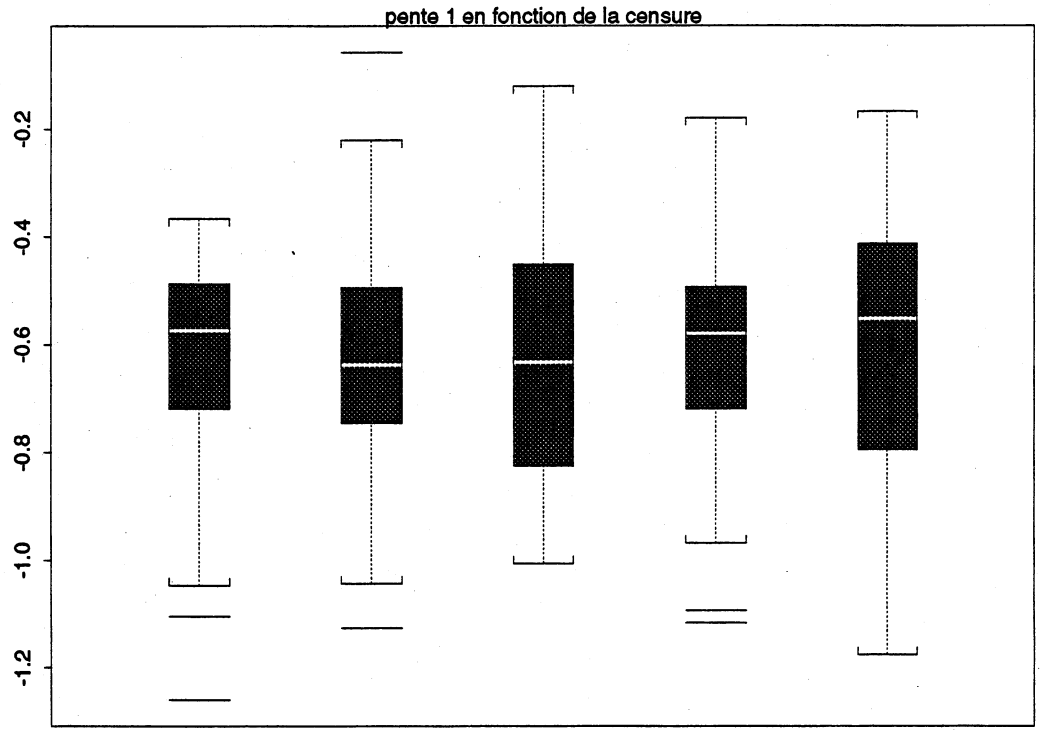

FIG. A.4 – *Valeur de $\hat{p}$ en fonction des valeurs du paramètre de la censure* $\xi_0 \in \{0.03,0.5,0.75,1,1.5\}$

Le critère $-\gamma_n(\hat{f}_{\Lambda^D})$ étant linéaire en D/n, on peut tenter d'appliquer l'heuristique de Birgé et Massart. Pour cela, pour chacune des valeurs possibles de la censure, nous estimons la pente de la droite $-\gamma_n(\hat{f}_{\Lambda^D})$ en fonction de D/n par une simple méthode de régression linéaire. Cette régression se fait sur des dimensions allant de 10 à 20, qui sont déjà de grandes dimensions, puisque la fonction f_4 n'a que 2 morceaux sur l'intervalle $[\varepsilon,1]$. Puis nous sélectionnons notre estimateur dans la collection avec une pénalité égale à

$$\text{pen}_n(\Lambda^D) = -2\hat{p}\frac{D}{n},$$

où $\hat{p}$ est la pente estimée par régression linéaire. Cette étude est faite avec une taille d'échantillon $n = 500$ et $L = 50$ simulations. Sur les 50 simulations, nous comptons le nombre de fois où le modèle sélectionné est le bon, c'est-à-dire le modèle de dimension 2. Les résultats sont donnés dans le tableau A.3.

ξ_0	0.01	0.03	0.5	0.75	1	1.5
Nb de $\hat{\Lambda}$ de dimension 2	43	40	38	38	38	38

TAB. A.3 – *Nombre de modèles sélectionnés* $\hat{\Lambda}$ *de dimension 2*

La méthode semble fonctionner puisqu'elle permet de sélectionner le bon modèle dans 76 à 86 % des cas. On remarque de plus que les résultats semblent meilleurs quand la censure est faible: le taux de réussite dans le choix du modèle diminue avec le paramètre de la censure.

Là aussi, l'étude de simulations mettant en oeuvre l'heuristique de Birgé et Massart n'a pour l'instant été faite que pour une seule fonction de régression f_4, qui est de plus très simple. Une étude de simulations complète demanderait de tester la méthode sur un grand nombre de fonctions de régression, et toujours avec plusieurs valeurs de la censure. Il pourrait être alors intéressant de comparer la constante de pénalité donnée par l'heuristique de Birgé et Massart avec la constante qui minimise le rapport des risques A.3.A.1.

Annexe B

Preuves des théorèmes du chapitre 2

Dans cette annexe, nous donnons une preuve des théorèmes du chapitre 2, à savoir la consistance et la normalité asymptotique de l'estimateur défini en 2.1.2.9. Cette preuve s'inspire à la fois du cours de DEA de Bretagnolle [18] et de l'article original d'Andersen et Gill [5]. Les notations sont les mêmes que celles du chapitre 2.

B.1 Preuve du théorème 2.2.1

Proposition B.1.1. *l_n est une fonction concave de β.*

Preuve. Commençons par dériver les fonctions $S_{n,h}$, $h = A,B$ par rapport à $\beta = (\beta_A,\beta_B) = (\rho_A,\theta_A,\rho_B,\theta_B)$. Les dérivées premières et secondes sont, pour tout $0 \leq s \leq \tau$:

$$\begin{aligned}
DS_{n,A}(\beta,s) &= \begin{pmatrix} n^{-1}\sum_{i=1}^n e^{\beta_A^T W_{A,i}(s)} Y_{A,i}(s) W_{A,i}(s) \\ 0 \end{pmatrix}, \\
DS_{n,B}(\beta,s) &= \begin{pmatrix} 0 \\ n^{-1}\sum_{i=1}^n e^{\beta_B^T W_{B,i}(s)} Y_{B,i}(s) W_{B,i}(s) \end{pmatrix}, \\
D^2S_{n,A}(\beta,s) &= \begin{pmatrix} n^{-1}\sum_{i=1}^n e^{\beta_A^T W_{A,i}(s)} Y_{A,i}(s) W_{A,i}^{\otimes 2}(s) & 0 \\ 0 & 0 \end{pmatrix}, \\
D^2S_{n,B}(\beta,s) &= \begin{pmatrix} 0 & 0 \\ 0 & n^{-1}\sum_{i=1}^n e^{\beta_B^T W_{B,i}(s)} Y_{B,i}(s) W_{B,i}^{\otimes 2}(s) \end{pmatrix}.
\end{aligned}$$

Définissons $\bar{N}_{n,h}$ pour $h = A,B$ par $\bar{N}_{n,h} = \sum_{i=1}^n N_{h,i}$ et $\mathcal{W}_{h,i}$, pour $h = A,B$, et $1 \leq i \leq n$ les vecteurs de dimension $2D+2$ définis pour $0 \leq s \leq \tau$ par

$$\mathcal{W}_{A,i}(s) = \begin{pmatrix} W_{A,i}(s) \\ 0 \end{pmatrix}, \mathcal{W}_{B,i}(s) = \begin{pmatrix} 0 \\ W_{B,i}(s) \end{pmatrix}.$$

En dérivant deux fois la fonction l_n, on obtient les dérivées première et seconde suivantes :

$$\begin{aligned}
Dl_n(\beta) &= \sum_{h=A,B} \left(\sum_{i=1}^n \int_0^\tau \mathcal{W}_{h,i}(s) dN_{h,i}(s) - \int_0^\tau \frac{DS_{n,h}(\beta,s)}{S_{n,h}(\beta,s)} d\bar{N}_{n,h}(s) \right) \\
-D^2l_n(\beta) &= \sum_{h=A,B} \int_0^\tau \left(\frac{D^2S_{n,h}(\beta,s)}{S_{n,h}(\beta,s)} - \left(\frac{DS_{n,h}(\beta,s)}{S_{n,h}(\beta,s)} \right)^{\otimes 2} \right) d\bar{N}_{n,h}(s)
\end{aligned}$$

On remarque que les matrices $\left(\frac{D^2S_{n,h}}{S_{n,h}} - (\frac{DS_{n,h}}{S_{n,h}})^{\otimes 2} \right)$, $h = A,B$ ont une structure de variance, elles sont donc semi-définies positives. En les intégrant par rapport à $\bar{N}_{n,h}$, on obtient la matrice $-D^2l_n(\beta)$ qui est aussi semi-définie positive, donc l_n est une fonction concave de β. □

Calculons maintenant la limite de la fonction $n^{-1}l_n$.

Proposition B.1.2. *Soit l la fonction définie par:*

$$l(\beta) = \sum_{h=A,B} \mathbb{E}\left(\int_0^\tau \beta_h^T W_h(s) e^{\beta_{h,0}{}^T W_h(s)} Y_h(s)\alpha_{h,0}(s)ds\right) - \int_0^\tau \ln S_h(\beta,s) S_h(\beta_0,s)\alpha_{h,0}(s)ds.$$

Alors, $n^{-1}l_n(\beta)$ tend presque sûrement vers $l(\beta)$ quand n tend vers l'infini.

Preuve. Avec les notations précédentes, on peut écrire:

$$n^{-1}l_n(\beta) = \sum_{h=A,B}\left(n^{-1}\sum_{i=1}^n \int_0^\tau \beta_h^T W_{h,i}(s)dN_{h,i}(s) - \int_0^\tau \ln\left(\frac{S_{n,h}(\beta,s)}{S_h(\beta,s)}\right) n^{-1}d\bar{N}_{n,h}(s)\right.$$
$$\left. - \int_0^\tau \ln\left(S_h(\beta,s)\right) n^{-1}d\bar{N}_{n,h}(s)\right).$$

D'après la loi des grands nombres, le premier terme de cette somme tend vers

$$\sum_{h=A,B} \mathbb{E}\left(\int_0^\tau \beta_h^T W_h(s)dN_h(s)\right),$$

quand n tend vers l'infini. D'après la définition 2.2.2.3 des compensateurs prévisibles Λ_h des processus N_h, cette limite est aussi égale à

$$\sum_{h=A,B} \mathbb{E}\left(\int_0^\tau \beta_h^T W_h(s) e^{\beta_{h,0}{}^T W_h(s)} Y_h(s)\alpha_{h,0}(s)ds\right).$$

Pour le second terme, rappelons que pour $h = A,B$, S_h est strictement positive. On peut donc utiliser le lemme B.3.1, qui implique

$$\forall h = A,B,\ \sup_{s\le\tau}\left|\ln\frac{S_{n,h}(\beta,s)}{S_h(\beta,s)}\right| \xrightarrow[n\to\infty]{\text{p.s.}} 0.$$

Comme pour $h = A,B$ et tout $s \le \tau$, $0 \le n^{-1}\bar{N}_{n,h}(s) \le 1$, on a aussi que, pour $h = A,B$, $\int_0^\tau \ln\frac{S_{n,h}(\beta,s)}{S_h(\beta,s)} n^{-1}d\bar{N}_{n,h}(s)$ tend presque sûrement vers 0 quand n tend vers l'infini. Pour le troisième terme, la loi des grands nombres nous permet d'affirmer que $\int_0^\tau \ln S_h(\beta,s) n^{-1}d\bar{N}_{n,h}(s)$ tend presque sûrement vers son espérance

$$\mathbb{E}\left(\int_0^\tau \ln S_h(\beta,s)dN_h(s)\right).$$

Les fonctions S_h sont des fonctions déterministes, continues par rapport au temps s, si bien que les processus $\int_0^t \ln S_h(\beta,s)(dN_h(s) - d\Lambda_h(s))$ sont aussi des $\mathcal{F}_t$-martingales (voir Andersen *et al.*[5]). En particulier, leurs espérances sont nulles et la limite précédente est égale à $\mathbb{E}\left(\int_0^\tau \ln S_h(\beta,s)d\Lambda_h(s)\right)$. On conclut en utilisant l'expression des compensateurs Λ_h, et en remarquant que les fonctions $\ln S_h$ sont déterministes. □

Pour terminer la preuve du théorème, remarquons que la fonction l est concave comme limite de fonctions concaves. En calculant ses dérivées première et seconde par rapport à β, et en utilisant le fait que pour toute fonction prévisible F du temps,

$$\mathbb{E}\left(\int_0^\tau F(s)dN_{h,i}(s)\right) = \mathbb{E}\left(\int_0^\tau F(s)d\Lambda_{h,i}(s)\right),$$

on obtient :

$$Dl(\beta) = \sum_{h=A,B}\int_0^\tau DS_h(\beta_0,s)\alpha_{h,0}(s)ds - \int_0^\tau \frac{DS_h(\beta,s)}{S_h(\beta,s)}S_h(\beta_0,s)\alpha_{h,0}(s)ds,$$

$$-D^2l(\beta) = \sum_{h=A,B}\int_0^\tau \left(\frac{D^2S_h}{S_h}(\beta,s) - \left(\frac{DS_h}{S_h}(\beta,s)\right)^{\otimes 2}\right)S_h(\beta_0,s)\alpha_{h,0}(s)ds,$$

avec les notations 2.4.2.1. Notons que $Dl(\beta_0) = 0$, si bien que la fonction l est maximale au point β_0. Grâce à l'hypothèse 2.4, nous prouvons dans le lemme 2.4.1 que les matrices $\left(\frac{D^2S_h}{S_h} - \left(\frac{DS_h}{S_h}\right)^{\otimes 2}\right)(\beta_0,s)$, pour $h = A,B$ sont définies positives. C'est pourquoi la matrice $-D^2l(\beta_0)$ est aussi définie positive, et l est strictement concave en β_0, où elle est maximale. Pour conclure, on utilise le corollaire II.2 (Appendix II) d'Andersen et Gill (voir [6]), qui stipule que toute suite de variables aléatoires X_n qui maximisent une suite de fonctions concaves F_n tendant point par point en probabilité vers une fonction F, tend en probabilité vers le point x qui maximise la fonction limite F.

B.2 Preuve du théorème 2.2.2

La preuve de ce théorème est basé sur un développement de Taylor au premier ordre du vecteur dérivé de la fonction l_n. On écrit :

$$Dl_n(\hat{\beta}_n) = 0 = Dl_n(\beta^0) + D^2l_n(\beta_n^*)(\hat{\beta}_n - \beta^0),$$

où β_n^* appartient au segment $[\beta^0,\hat{\beta}_n]$, et on en déduit que :

$$\sqrt{n}(\hat{\beta}_n - \beta^0) = \left(-\frac{D^2l_n(\beta_n^*)}{n}\right)^{-1}\frac{Dl_n(\beta^0)}{\sqrt{n}}.$$

Il s'agit ensuite de prouver la normalité asymptotique du vecteur des scores $\xi_n(\tau) = \frac{Dl_n(\beta^0)}{\sqrt{n}}$ et la convergence en probabilité de la matrice $\frac{D^2l_n(\beta_n^*)}{n}$.

Proposition B.2.1. *Le vecteur des scores $\xi_n(\tau)$ converge en loi vers une variable aléatoire gaussienne $\mathcal{N}\left(0, -D^2l(\beta_0)\right)$.*

Preuve. Considérons le processus

$$\xi_n(t) = \sum_{h=A,B} n^{-1/2} \sum_{i=1}^{n} \int_0^t \left(\mathcal{W}_{h,i}(s) - \frac{DS_{n,h}}{S_{n,h}}(\beta_0, s) \right) dN_{h,i}(s),$$

et calculons son compensateur prévisible:

$$\sum_{h=A,B} n^{-1/2} \sum_{i=1}^{n} \int_0^t \left(\mathcal{W}_{h,i}(s) - \frac{DS_{n,h}}{S_{n,h}}(\beta_0, s) \right) e^{\beta_{h,0}^T W_{h,i}(s)} Y_{h,i}(s) \alpha_{h,0}(s) ds = 0.$$

Ce processus est donc une $\mathcal{F}_t$-martingale. Nous allons montrer sa convergence en utilisant le théorème de la limite centrale pour des martingales dû à Rebolledo (voir [52, 5]). Pour cela, montrons d'abord la convergence de son processus de variation prévisible $< \xi_n(t) >$:

$$< \xi_n(t) >= \sum_{h=A,B} n^{-1} \sum_{i=1}^{n} \int_0^t \left(\mathcal{W}_{h,i}(s) - \frac{DS_{n,h}}{S_{n,h}}(\beta_0, s) \right)^{\otimes 2} e^{\beta_0^T W_{h,i}(s)} Y_{h,i}(s) \alpha_{h,0}(s) ds.$$

En développant le carré, on peut écrire:

$$< \xi_n(t) >= \sum_{h=A,B} \int_0^t \left(\frac{D^2 S_{n,h}}{S_{n,h}} - \left(\frac{DS_{n,h}}{S_{n,h}} \right)^{\otimes 2} \right) S_{n,h} \alpha_{h,0}(s) ds.$$

En utilisant la convergence uniforme des fonctions $S_{n,h}$, $DS_{n,h}$ et $D^2 S_{n,h}$, pour $h = A,B$, on montre que le processus de variation prévisible $< \xi_n(t) >$ tend en probabilité vers

$$\sum_{h=A,B} \int_0^t \left(\frac{D^2 S_h}{S_h} - \left(\frac{DS_h}{S_h} \right)^{\otimes 2} \right) S_h \alpha_{h,0}(s) ds.$$

Remarquons que pour $t = \tau$, cette matrice limite est $-D^2 l(\beta_0)$.

Vérifions maintenant les conditions de Lindeberg du théorème de Rebolledo: il suffit de prouver, pour chacune des coordonnées (j), $1 \leq j \leq 2D + 2$, que:

$$< \xi_{\varepsilon,n}^{(j)}(t) >= \sum_{h=A,B} n^{-1} \sum_{i=1}^{n} \int_0^t \left(\mathcal{W}_{h,i}(s)^{(j)} - \frac{\frac{\partial S_{n,h}}{\partial \beta^{(j)}}}{S_{n,h}}(\beta_0, s) \right)^2 \mathbb{1}_{\left\{ n^{-1/2} \left| \mathcal{W}_{h,i}(s)^{(j)} - \frac{\frac{\partial S_{n,h}}{\partial \beta^{(j)}}}{S_{n,h}} \right| > \varepsilon \right\}}$$
$$e^{\beta_{h,0}^T W_{h,i}(s)} Y_{h,i}(s) \alpha_{h,0}(s) ds$$

tend vers 0 quand n tend vers l'infini, pour tout $\varepsilon > 0$. En utilisant l'inégalité suivante, pour tous réels a,b:

$$(a-b)^2 \mathbb{1}_{\{|a-b|>\varepsilon\}} \leq 4a^2 \mathbb{1}_{\{|a|>\varepsilon/2\}} + 4b^2 \mathbb{1}_{\{|b|>\varepsilon/2\}},$$

on obtient :

$$< \xi_{\varepsilon,n}^{(j)}(t) > \leq$$
$$4 \sum_{h=A,B} n^{-1} \sum_{i=1}^{n} \int_0^t \left(\mathcal{W}_{h,i}(s)^{(j)} \right)^2 \mathbb{1}_{\{n^{-1/2}|\mathcal{W}_{h,i}(s)^{(j)}|>\varepsilon/2\}} e^{\beta_{h,0}^T W_{h,i}(s)} Y_{h,i}(s) \alpha_{h,0}(s) ds$$
$$+ \int_0^t \left(\frac{\frac{\partial S_{n,h}}{\partial \beta^{(j)}}}{S_{n,h}}(\beta_0, s) \right)^2 \mathbb{1}_{\{n^{-1/2}|\frac{\frac{\partial S_{n,h}}{\partial \beta^{(j)}}}{S_{n,h}}|>\varepsilon/2\}} S_{n,h}(\beta_0, s) \alpha_{h,0}(s) ds.$$

En utilisant une inégalité de type Chebychev, il découle que :

$$< \xi_{\varepsilon,n}^{(j)}(t) > \leq 4 \sum_{h=A,B} n^{-1} \sum_{i=1}^{n} \int_0^t \left(\mathcal{W}_{h,i}(s)^{(j)} \right)^3 2n^{-1/2} \varepsilon^{-1} e^{\beta_{h,0}^T W_{h,i}(s)} Y_{h,i}(s) \alpha_{h,0}(s) ds$$
$$+ \int_0^t \left(\frac{\frac{\partial S_{n,h}}{\partial \beta^{(j)}}}{S_{n,h}} \right)^3 2n^{-1/2} \varepsilon^{-1} S_{n,h}(\beta_0, s) \alpha_{h,0}(s) ds,$$

et donc,

$$< \xi_{\varepsilon,n}^{(j)}(t) > \leq 8n^{-1/2} \varepsilon^{-1} \sup_{h,i,j,s} \left(|\mathcal{W}_{h,i}(s)^{(j)}| \right) \sum_{h=A,B} \int_0^t \frac{\partial^2 S_{n,h}}{\partial \beta^{(j)2}} \alpha_{h,0}(s) ds$$
$$+ 8n^{-1/2} \varepsilon^{-1} \sum_{h=A,B} \int_0^t \left(\frac{\frac{\partial S_{n,h}}{\partial \beta^{(j)}}}{S_{n,h}} \right)^3 S_{n,h}(\beta_0, s) \alpha_{h,0}(s) ds.$$

Chaque somme $S_{n,h}$ et ses dérivées tendent uniformément vers leurs espérances (voir lemme B.3.1), donc le second terme de la borne précédente tend vers 0 quand n tend vers l'infini. Pour le premier terme, remarquons que

$$\sup_{i,h,j,s} \left(|\mathcal{W}_{h,i}(s)^{(j)}| \right) = \sup_{h,i} \max(1, |Z_{h,i}|_\infty).$$

Dans le lemme B.3.2, nous prouvons que

$$\sup_{h,i} (|Z_{h,i}|_\infty) = O_{p.s.}(\ln n),$$

ce qui permet d'affirmer que le premier terme tend aussi vers 0 quand n tend vers l'infini. Finalement, le processus $< \xi_{\varepsilon,n}^{(j)}(t) >$ tend presque sûrement vers 0 quand n tend vers l'infini. Les conditions du théorème de la limite centrale pour des martingales dû à Rebolledo sont donc remplies, et donc $\xi_n(\tau) = n^{-1/2} Dl_n(\beta_0)$ tend en loi vers une gaussienne $\mathcal{N}\left(0, -D^2 l(\beta_0)\right)$. □

Il reste maintenant à prouver la proposition suivante :

Proposition B.2.2. *Pour toute suite β_n^* qui tend en probabilité vers β_0 quand n tend vers l'infini, la matrice $n^{-1}D^2l_n(\beta_n^*)$ tend vers la matrice $D^2l(\beta_0)$ quand n tend vers l'infini.*

Preuve. La preuve se passe en deux étapes :

La première étape est la convergence de $n^{-1}D^2l_n(\beta_0)$ vers $D^2l(\beta_0)$, ce qui est immédiat en utilisant la convergence uniforme des fonctions $S_{n,h}$ et de ses dérivées vers S_h et ses dérivées (cf B.3.1). C'est pourquoi $\left\|\frac{D^2l_n(\beta_0)}{n} - D^2l(\beta_0)\right\|_\infty$ tend en probabilité vers 0 quand n tend vers l'infini.

La seconde étape est la preuve du lemme suivant :

Lemme B.2.1.

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\sup_{|\beta-\beta_0|_\infty\leq\varepsilon} n^{-1}\left\|D^2l_n(\beta) - D^2l_n(\beta_0)\right\|_\infty = 0.$$

Preuve. Notons $S'(\varepsilon)$, $DS'(\varepsilon)$, $D^2S'(\varepsilon)$ les expressions suivantes :

$$\begin{aligned}
S'(\varepsilon) &= \max_h \sup_{|\beta_h-\beta_{h,0}|_\infty<\varepsilon} \sup_s |e^{\beta_h^T W_h(s)} - e^{\beta_{h,0}{}^T W_h(s)}| \\
DS'(\varepsilon) &= \max_h \sup_{|\beta_h-\beta_{h,0}|_\infty<\varepsilon} \sup_s |(e^{\beta_h^T W_h(s)} - e^{\beta_{h,0}{}^T W_h(s)})\mathcal{W}_h(s)|_\infty \\
D^2S'(\varepsilon) &= \max_h \sup_{|\beta_h-\beta_{h,0}|_\infty<\varepsilon} \sup_s \|(e^{\beta_h^T W_h(s)} - e^{\beta_{h,0}{}^T W_h(s)})\mathcal{W}_h(s)^{\otimes 2}\|_\infty
\end{aligned}$$

Il est clair que $\mathbb{E}(S'(\varepsilon))$, $\mathbb{E}(DS'(\varepsilon))$ and $\mathbb{E}\left(D^2S'(\varepsilon)\right)$ sont bornés par $\varphi(\varepsilon) = \mathbb{E}(\phi(\varepsilon))$, où :

$$\phi(\varepsilon) = \max_h \sup_{|\beta_h-\beta_{h,0}|_\infty<\varepsilon} \sup_s |e^{\beta_h^T W_h(s)} - e^{\beta_{h,0}{}^T W_h(s)}|(1 + |W_h(s)|_\infty + |W_h(s)|_\infty^2).$$

Prouvons par convergence dominée que $\lim_{\varepsilon\to 0}\varphi(\varepsilon) = 0$.

Il est clair que $\lim_{\varepsilon\to 0}\phi(\varepsilon) = 0$. Soit ε_0 un réel strictement positif. Il existe un réel strictement positif C_{ε_0} tel que pour tout $z > 0$,

$$1 + z + z^2 \leq C_{\varepsilon_0}e^{\varepsilon_0 z}. \tag{B.2.B.1}$$

Soit ε un réel tel que $0 < \varepsilon < \varepsilon_0$. Pour tout β tel que $|\beta - \beta_0|_\infty < \varepsilon$, en utilisant un développement de Taylor, on peut majorer $|e^{\beta_h^T W_h(s)} - e^{\beta_{h,0}{}^T W_h(s)}|$ par $\varepsilon|W_h(s)|_\infty e^{\varepsilon|W_h(s)|_\infty}$. En utilisant l'inégalité B.2.B.1, nous obtenons :

$$\begin{aligned}
\phi(\varepsilon) &\leq \max_h \varepsilon|W_h(s)|_\infty e^{\varepsilon|W_h(s)|_\infty} e^{\beta_{h,0}{}^T W_h(s)} C_{\varepsilon_0} e^{\varepsilon_0|W_h(s)|_\infty} \\
&\leq \sum_{h=A,B} \varepsilon_0|W_h(s)|_\infty e^{\varepsilon_0|W_h(s)|_\infty} e^{\beta_{h,0}{}^T W_h(s)} C_{\varepsilon_0} e^{\varepsilon_0|W_h(s)|_\infty}
\end{aligned}$$

En prenant l'espérance et en appliquant une inégalité de Hölder, on obtient que, pour tous

P et Q tels que $\frac{1}{P} + \frac{1}{Q} = 1$:

$$\mathbb{E}\left(\sum_{h=A,B} \varepsilon_0 |W_h(s)|_\infty e^{\varepsilon_0 |W_h(s)|_\infty} e^{\beta_{h,0}{}^T W_h(s)} C_{\varepsilon_0} e^{\varepsilon_0 |W_h(s)|_\infty}\right)$$
$$\leq \sum_{h=A,B} \mathbb{E}\left(\varepsilon_0 C_{\varepsilon_0} |W_h(s)|_\infty e^{2Q\varepsilon_0 |W_h(s)|_\infty}\right)^{\frac{1}{Q}} \mathbb{E}\left(e^{P\beta_{h,0}^T W_h(s)}\right)^{\frac{1}{P}}$$

Comme l'ensemble $\Theta_A \times \Theta_B$ défini en 2.3 est un ensemble ouvert, il est toujours possible de choisir P tel que pour $h = A,B$ et $0 \leq s \leq \tau$, $\mathbb{E}\left(e^{P\beta_{h,0}^T W_h(s)}\right) < \infty$, puis Q tel que $\frac{1}{P} + \frac{1}{Q} = 1$. Ensuite, en choisissant ε_0 tel que $\mathbb{E}\left(\varepsilon_0 |W_h(s)|_\infty e^{2Q\varepsilon_0 |W_h(s)|_\infty}\right) < \infty$ (voir B.3.3), cette espérance est finie. Par convergence dominée, $\varphi(\varepsilon) = \mathbb{E}(\phi(\varepsilon))$ tend vers 0 quand n tend vers l'infini.

Soit $M_n(\beta,s)$ défini par :

$$M_n(\beta,s) = \max_{h=A,B} \max(|S_{n,h}(\beta,s) - S_{n,h}(\beta_0,s)|,$$
$$|DS_{n,h}(\beta,s) - DS_{n,h}(\beta_0,s)|_\infty, \|D^2 S_{n,h}(\beta,s) - D^2 S_{n,h}(\beta_0,s)\|_\infty).$$

Par définition de $S'(\varepsilon)$, $DS'(\varepsilon)$, $D^2S'(\varepsilon)$, il est facile d'obtenir que

$$\sup_{|\beta-\beta_0|_\infty < \varepsilon} \sup_{s \leq \tau} M_n(\beta,s) \leq 2n^{-1} \sum_{i=1}^{n} S_i'(\varepsilon) + DS_i'(\varepsilon) + D^2 S_i'(\varepsilon).$$

Par la loi des grands nombres, le terme majorant tend vers $2\,\mathbb{E}\left(S_i'(\varepsilon) + DS_i'(\varepsilon) + D^2 S_i'(\varepsilon)\right)$ qui est lui-même borné par $6\varphi(\varepsilon)$. Finalement,

$$\limsup_{n\to\infty} \sup_{|\beta-\beta_0|_\infty < \varepsilon} \sup_{s \leq \tau} M_n(\beta,s) \leq 7\varphi(\varepsilon). \qquad \text{(B.2.B.2)}$$

Soit $c = \min_{h=A,B} \inf_{s \leq \tau} S_h(\beta_0,s)$. Les hypothèses 2.1 et 2.2 permettent d'affirmer que c est strictement positif. Soit Ω_n l'événement $\{\forall s \leq \tau,\ h = A,B\ S_{n,h}(\beta_0,s) > c/2\}$. En utilisant la convergence uniforme des $S_{n,h}$ vers les S_h, il est clair que $\mathbb{P}(\Omega_n)$ tend vers 1 quand n tend vers l'infini. Sur l'événement Ω_n, il est maintenant possible, pour tout $s \leq \tau$, de borner $\left\|D^2 S_{n,h}/S_{n,h}(\beta,s) - D^2 S_{n,h}/S_{n,h}(\beta_0,s)\right\|_\infty$ par $2/c \left\|D^2 S_{n,h}(\beta,s) - D^2 S_{n,h}(\beta_0,s))\right\|_\infty$, et donc par $2M_n(\beta,s)/c$. En suivant la même idée, nous obtenons

$$\left\|\left(\frac{DS_{n,h}}{S_{n,h}}\right)^{\otimes 2}(\beta,s) - \left(\frac{DS_{n,h}}{S_{n,h}}\right)^{\otimes 2}(\beta_0,s)\right\|_\infty \leq$$
$$4/c^2 \left\|(DS_{n,h})^{\otimes 2}(\beta,s) - (DS_{n,h})^{\otimes 2}(\beta_0,s)\right\|_\infty$$

En factorisant le membre de droite et en utilisant la définition de M_n, nous avons aussi,

$$\left\| \left(\frac{DS_{n,h}}{S_{n,h}}\right)^{\otimes 2}(\beta,s) - \left(\frac{DS_{n,h}}{S_{n,h}}\right)^{\otimes 2}(\beta_0,s) \right\|_\infty \leq$$
$$4/c^2\,|DS_{n,h}(\beta,s) - DS_{n,h}(\beta_0,s)|_\infty\,(|DS_{n,h}(\beta,s)|_\infty + |DS_{n,h}(\beta_0,s)|_\infty)$$
$$\leq 4M_n(\beta,s)/c^2(|DS_{n,h}(\beta,s)|_\infty + |DS_{n,h}(\beta_0,s)|_\infty).$$

Comme $DS_{n,h}(\beta_0,s)$ tend presque sûrement vers $DS_h(\beta_0,s)$, $|DS_{n,h}(\beta_0,s)|_\infty \leq 3/2|DS_h(\beta_0,s)|_\infty$ pour n assez grand. Pour la même raison, $|DS_{n,h}(\beta,s)| \leq 3/2|DS_h(\beta,s)|_\infty$. Pour $h = A,B$, pour tout $0 \leq s \leq \tau$ et β tel que $|\beta - \beta_0|_\infty < \varepsilon$, nous avons:

$$\sup_{s\leq\tau}|DS_h(\beta,s)|_\infty \leq \sum_{h=A,B}\mathbb{E}\left(e^{\beta_h^T W_h(s)}|W_h(s)|_\infty\right).$$

En décomposant $e^{\beta_h^T W_h(s)}$ en $e^{(\beta_h-\beta_{h,0})^T W_h(s)}e^{\beta_{h,0}{}^T W_h(s)}$ et en majorant $(\beta_h - \beta_{h,0})^T W_h(s)$ par $\varepsilon|W_h(s)|_\infty$, nous pouvons conclure que

$$\sup_{\beta:|\beta-\beta_0|_\infty<\varepsilon}\sup_{s\leq\tau}|DS_h(\beta,s)|_\infty < \infty.$$

Finalement, il existe une constante C telle que, pour tout $0 \leq s \leq t$ et β tel que $|\beta-\beta_0|_\infty < \varepsilon$,

$$\left\| \left(\frac{D^2S_{n,h}}{S_{n,h}} - \left(\frac{DS_{n,h}}{S_{n,h}}\right)^{\otimes 2}\right)(\beta,s) - \left(\frac{D^2S_{n,h}}{S_{n,h}} - (\frac{DS_{n,h}}{S_{n,h}})^{\otimes 2}\right)(\beta_0,s) \right\|_\infty \leq CM_n(\beta,s).$$

En intégrant par rapport à $\bar{N}_{n,h}$,

$$\|D^2l_n(\beta) - D^2l_n(\beta_0)\|/n \leq C\int_0^\tau M_n(\beta,s)d\bar{N}(s)/n \leq C\sup_{s\leq t}M_n(\beta,s)\bar{N}(\tau)/n,$$

avec $\bar{N} = \bar{N}_{A,n} + \bar{N}_{B,n}$. Notons que $\bar{N}(\tau)/n \leq 2$, si bien qu'en utilisant B.2.B.2

$$\limsup_{n\to\infty}\sup_{|\beta-\beta_0|_\infty<\varepsilon}\sup_{s\leq\tau}\|(D^2l_n(\beta,s) - D^2l_n(\beta_0,s))/n\|_\infty \leq 14C\varphi(\varepsilon),$$

et $\varphi(\varepsilon)$ tend vers 0 quand n tend vers l'infini. □ □

B.3 Lemmes techniques

Nous montrons ici des résultats très classiques et bien connus, utilisés dans les preuves précédentes.

Lemme B.3.1.

$$\max\Big(\max_{h=A,B}\sup_{s\leq\tau}|S_{n,h}(\beta,s)-S_h(\beta,s)|\,,\ \max_{h=A,B}\sup_{s\leq\tau}|DS_{n,h}(\beta,s)-DS_h(\beta,s)|_\infty\,,\ \max_{h=A,B}\sup_{s\leq\tau}\left\|D^2S_{n,h}(\beta,s)-D^2S_h(\beta,s)\right\|_\infty\Big)$$

tend presque sûrement vers 0 quand n tend vers l'infini.

Preuve. Prouvons par exemple que $\sup_{s\leq\tau}|S_{n,A}(\beta,s)-S_A(\beta,s)|$ tend vers 0 quand n tend vers l'infini. Rappelons la définition 2.1.2.8 de $S_{n,A}$:

$$S_{n,A}(\beta,s)=n^{-1}\sum_{i=1}^{n}e^{\beta_A^TW_{A,i}(s)}Y_{A,i}(s),$$

et des indicatrices de risque 2.1.2.6. En notant V_i^k, $1\leq k\leq 3$ les expressions suivantes:

$$\begin{aligned}V_i^1 &= e^{\rho_A+\theta_A^TZ_{A,i}}\delta_{B,i}\,\mathbb{1}_{\{T_{B,i}<T_{A,i}\}}+e^{\theta_A^TZ_{A,i}}\,\mathbb{1}_{\{T_{A,i}<T_{B,i}\}}\\ V_i^2 &= e^{\theta_A^TZ_{A,i}}\,\mathbb{1}_{\{T_{B,i}<T_{A,i}\}}\\ V_i^3 &= e^{\rho_A+\theta_A^TZ_{A,i}}\delta_{B,i}\,\mathbb{1}_{\{T_{B,i}<T_{A,i}\}},\end{aligned}$$

on peut décomposer $S_{n,A}$ en:

$$S_{n,A}(\beta,s)=n^{-1}\sum_{i=1}^{n}V_i^1\,\mathbb{1}_{\{T_{A,i}\geq s\}}+n^{-1}\sum_{i=1}^{n}V_i^2\,\mathbb{1}_{\{T_{B,i}\geq s\}}-n^{-1}\sum_{i=1}^{n}V_i^3\,\mathbb{1}_{\{T_{B,i}\geq s\}}.$$

Chacun des V_i^k est strictement positif, et leurs espérances sont finies. Il suffit de prouver que chacune de ces sommes tend vers son espérance uniformément sur $[0,\tau]$. Montrons-le par exemple pour la première:

Définissons pour tout $s\leq\tau$, $F_{1,n}(s)=n^{-1}\sum_{i=1}^{n}V_i^1\,\mathbb{1}_{\{T_{A,i}\geq s\}}$. Pour tout $s\leq\tau$, $F_{1,n}(s)$ tend presque sûrement vers $F_1(s)=\mathbb{E}\left(V^1\,\mathbb{1}_{\{T_A\geq s\}}\right)$. Soit N un entier et $t_{k,N}$ des réels tels que $t_{k,N}=\sup\{s,F_1(s)>\mathbb{E}\left(V^1\right)\frac{k}{N}\}$. On a donc, pour tout s tel que $t_{k+1,N}\leq s\leq t_{k,N}$:

$$F_1(t_{k,N})=F_1(t_{k+1,N})-n^{-1}\leq F_1(s)\leq F_1(t_{k+1,N})=F_1(t_{k,N})+n^{-1}.$$

Comme $F_{1,n}$ est une fonction décroissante de s, nous avons aussi:

$$F_{1,n}(t_{k,N})\leq F_{1,n}(s)\leq F_{1,n}(t_{k+1,N}).$$

En utilisant les deux inégalités, on trouve que:

$$F_{1,n}(t_{k,N})-F_1(t_{k,N})-n^{-1}\leq F_{1,n}(s)-F_1(s)\leq F_{1,n}(t_{k+1,N})-F_1(t_{k+1,N})+n^{-1},$$

si bien que,

$$\sup_{s\leq\tau}|F_{1,n}(s)-F_1(s)|\leq\max_{0\leq k\leq N}|F_{1,n}(t_{k,N})-F_1(t_{k,N})|+n^{-1}.$$

Soit ε un réel strictement positif. Choisissons d'abord N tel que $1/N<\varepsilon/2$, puis n assez grand pour que $\max_{0\leq k\leq N}|F_{1,n}(t_{k,N})-F_1(t_{k,N})|<\varepsilon/2$, ce qui permet de conclure. Nous procédons de la même manière avec les autres sommes. □

Lemme B.3.2. *La suite* $(\ln n)^{-1}\sup_{1\leq i\leq n}|Z_i|_\infty$ *est tendue.*

Preuve. Soit λ un réel strictement positif tel que $\mathbb{E}\left(e^{\lambda|Z_i|_\infty}\right)<\infty$ et M un réel strictement positif. Commençons par majorer $\mathbb{P}\left(\sup_{1\leq i\leq n}|Z_i|_\infty>M\ln n\right)$. Il est clair que cette probabilité est plus petite que $\mathbb{P}\left(\exists i|Z_i|_\infty>M\ln n\right)$, qui peut être bornée par la somme des probabilités $\sum_{i=1}^n\mathbb{P}\left(|Z_i|_\infty>M\ln n\right)$. En utilisant une inégalité de type Chebychev, cette somme est plus petite que $\sum_{i=1}^n e^{-\lambda M\ln n}\,\mathbb{E}\left(e^{\lambda|Z_i|_\infty}\right)$ et finalement,

$$\mathbb{P}\left(\sup_{1\leq i\leq n}|Z_i|_\infty>M\ln n\right)\leq\frac{\mathbb{E}\left(e^{\lambda|Z|_\infty}\right)}{n^{\lambda M-1}}.$$

Pour M assez grand, $\lambda M-1$ est strictement positif, et

$$\limsup_{M\to\infty}\lim_{n\to\infty}\mathbb{P}\left(\sup_{1\leq i\leq n}\frac{|Z_i|_\infty}{\ln n}>M\right)=0,$$

ce qui termine la preuve. □

Lemme B.3.3. *Supposons qu'il existe* $\varepsilon_0>0$ *tel que pour tout* θ *tel que* $|\theta|_\infty<\varepsilon_0$, $\mathbb{E}\left(|Z|_\infty^2e^{\theta^TZ}\right)<\infty$. *Alors, il existe* $\eta_0>0$ *tel que pour tout* $0<\eta<\eta_0$, $\mathbb{E}\left(|Z|_\infty^2e^{\eta|Z|_\infty}\right)<\infty$.

Preuve. Notons $Z^{(j)}$ les coordonnées de Z, nous avons

$$|Z|_\infty=\max_{1\leq j\leq D}Z^{(j)}\leq\sum_{j=1}^D|Z^{(j)}|.$$

En utilisant une inégalité de Hölder, nous avons

$$\mathbb{E}\left(|Z|_\infty^2e^{\eta\sum_{j=1}^D|Z^{(j)}|}\right)=\mathbb{E}\left(\prod_{j=1}^D|Z|_\infty^{2/D}e^{\eta|Z^{(j)}|}\right)\leq\left(\prod_{j=1}^D\mathbb{E}\left(|Z|_\infty^2e^{D\eta|Z^{(j)}|}\right)\right)^{\frac{1}{D}}.$$

Chaque facteur du produit peut être borné par $\max_{1\leq j\leq D}\mathbb{E}\left(|Z|_\infty^2e^{D\eta|Z^{(j)}|}\right)$, si bien qu'on peut affirmer que

$$\mathbb{E}\left(|Z|_\infty^2e^{\eta|Z|_\infty}\right)\leq\max_{1\leq j\leq D}\mathbb{E}\left(|Z|_\infty^2e^{D\eta Z^{(j)}}\right)+\mathbb{E}\left(|Z|_\infty^2e^{-D\eta Z^{(j)}}\right).$$

Soit θ tel que $|\theta|_\infty < \varepsilon_0$. Alors, pour toute coordonnées $\theta^{(j)}$ de θ, $|\theta^{(j)}| = |-\theta^{(j)}| < \varepsilon_0$, donc on a aussi $\mathbb{E}\left(|Z|_\infty^2 e^{\theta^{(j)} Z^{(j)}}\right) < \infty$ et $\mathbb{E}\left(|Z|_\infty^2 e^{-\theta^{(j)}, Z^{(j)}}\right) < \infty$. Finalement, $\mathbb{E}\left(|Z|_\infty^2 e^{\eta |Z|_\infty}\right)$ est finie dès que η est plus petit que $\frac{\varepsilon_0}{D}$. □

Annexe C

Programmes Matlab pour les simulations du chapitre 2

Cette annexe présente les fichiers Matlab utilisés pour les simulations du chapitre 2. Les noms de fichiers apparaissent en gras (**fichier.m**), alors que les commandes Matlab apparaissent en type machine à écrire (`y=function(x)`).

C.1 Fichiers de simulation des variables aléatoires

Les simulations ont été réalisées dans les conditions suivantes :

- La taille de l'échantillon est n.
- Les variables X_A et X_B sont indépendantes et suivent des lois exponentielles de même paramètre λ.
- Les variables de censure U_A et U_B sont indépendantes et suivent des lois exponentielles de même paramètre μ.
- Les paramètres de chocs sont $\rho = [\rho_A^0, \rho_B^0]$.
- La durée d'étude est t.
- Le nombre de simulations est L.

Tous ces paramètres peuvent être réglés par le fichier **Parametres.m**. Celui-ci demande à l'utilisateur les valeurs citées ci-dessus dans ce même ordre, et les stocke dans un fichier **parametres.mat** que l'on peut charger par la commande `load parametres`. Les noms des paramètres sont `n, lambda, mu, rho, t, L`.

Pour une simulation, les variables aléatoires sont engendrées par le fichier **Simulation.m**. C'est une fonction `[T,Delta]=Simulation(n,lambda,mu,rho)`, où `n` est la taille de la population, `lambda` est le paramètre des variables $X_{h,i}$, `mu` est le paramètre des variables de censure et `rho` est le vecteur $[\rho_A^0, \rho_B^0]$. Ce fichier donne en sortie un tableau `T` à deux lignes et n colonnes contenant les variables $[T_{A,i}, T_{B,i}]$ et un tableau `Delta` à deux lignes et n colonnes contenant les variables $[\delta_{A,i}, \delta_{B,i}]$. Il affiche à l'écran

- l'espérance des $X_{h,i}$
- leur moyenne empirique
- leur écart-type théorique
- leur écart-type empirique
- des intervalles de confiance à 95%
- la même chose pour les $X'_{h,i}$
- la même chose pour les $\tilde{X}_{h,i}$
- la même chose pour les $U_{h,i}$
- la moyenne et l'écart-type empiriques des $\delta_{h,i}$
- la moyenne et l'écart-type empiriques des $T_{h.i}$.

Le fichier réellement utilisé dans les autres programmes est **Simulation2.m**. Il fait la même chose que **Simulation.m** sans rien afficher. Il se déclare par `[T,Delta]= Simulation2(n,lambda,mu,rh`

C.2 Fichiers nécessaires à la construction de la vraisemblance

Le fichier calculant la vraisemblance, sa dérivée et sa dérivée seconde à partir des données est **Vraisemb.m**. Il fait appel aux fichiers suivants : **Compte.m**, **Sn.m**, **DSn.m**, **Variance.m** et indirectement **Matrice.m**.

Le fichier **Compte.m** est une fonction `[Nprime, Nseconde]=Compte(T, Delta,t)` qui prend en argument le tableau `T` des dates $[T_{A,i},T_{B,i}]$, le tableau `Delta` des indicatrices de censure $[\delta_{A,i},\delta_{B,i}]$ et une date quelconque `t`. Il donne en retour deux tableaux `Nprime` et `Nseconde` de même taille que `T`. `Nprime` contient dans sa première ligne les valeurs $N_{2,i}(t)$ et sur sa deuxième ligne les valeurs $N_{4,i}(t)$ (mêmes notations que la partie 2.2.1). `Nseconde` contient dans sa première ligne les valeurs $N_{1,i}(t)$ et sur sa deuxième ligne les valeurs $N_{2,i}(t)$.

Remarque C.1. On a choisi le même paramètre λ pour les deux dates X_A et X_B. On fera donc l'estimation dans le cas où l'on sait que les deux événements étudiés ont même risque de base (voir Remarque 1).

Le fichier **Sn.m** est une fonction `s=Sn(T,Delta,t,rho)` qui prend en argument le tableau `T` des dates $[T_{A,i},T_{B,i}]$, le tableau `Delta` des indicatrices de censure $[\delta_{A,i},\delta_{B,i}]$, une date quelconque `t` et un vecteur `rho` de dimension 2 $[\rho_A,\rho_B]$. Il donne en sortie la somme S_n intervenant dans la vraisemblance.

Le fichier **DSn.m** est une fonction `D=DSn(T,Delta,t,rho)` qui prend en argument le tableau `T` des dates $[T_{A,i},T_{B,i}]$, le tableau `Delta` des indicatrices de censure $[\delta_{A,i},\delta_{B,i}]$, une date quelconque `t` et un vecteur `rho` de dimension 2 : $[\rho_A,\rho_B]$. Il donne en sortie le vecteur `D` contenant les dérivées de la somme S_n par rapport à ρ_A et à ρ_B.

Le fichier **Matrice.m** est une fonction `M=Matrice(T,Delta,t,rho)` qui prend en argument les tableaux `T` et `Delta` de taille $2 \times n$, une date quelconque `t` et le vecteur de dimension 2 `rho`. Il donne en sortie une matrice `M` qui est égale à $\left(\frac{D^2S_n}{S_n} - (\frac{DS_n}{S_n})^2\right)$`(rho,t)`. Il fait appel aux fichiers **Sn.m** et **DSn.m**.

Le fichier **Variance.m** est une fonction `v=Variance(T,Delta,t,rho)` qui prend en argument les tableaux `T` et `Delta` de taille $2 \times n$, la date finale d'étude `t` et le vecteur de dimension 2 `rho`. Il donne en sortie la matrice `v` égale à $-D^2l_n$`(rho)`. Il fait appel aux fichiers **Compte.m** et **Matrice.m**.

Le fichier **Vraisemb.m** est une fonction `[v,grad,hess]=Vraisemb(rho,T,Delta,t)` qui prend en argument les tableaux `T` et `Delta` de taille $2 \times n$, la date finale d'étude `t` et le vecteur de dimension 2 `rho`. Il donne en sortie la valeur `v` de la vraisemblance au point `rho`, le vecteur `grad`, gradient de la vraisemblance en ce point et la matrice hessienne `hess` des dérivées secondes de la vraisemblance au même point. Le gradient et la matrice hessienne sont nécessaires à Matlab pour maximiser la vraisemblance.

C.3 Fichiers d'estimation des paramètres

Le fichier **Estim.m** est une fonction `Estim(T,Delta,t,rho)` qui prend en argument les tableaux `T` et `Delta` de taille $2 \times n$, la date finale d'étude `t` et le vecteur de dimension 2 `rho`. Il maximise la vraisemblance à partir de la valeur donnée par `rho`. Il affiche les différents résultats suivant les options de maximisation choisies ainsi que le temps de calcul. Il fait appel au fichier **Vraisemb.m**.

Le fichier **Estim2.m** est une fonction `[res,crit,I]=Estim2(T,Delta,t,rho)` qui prend en argument les tableaux `T` et `Delta` de taille $2 \times n$, la date finale d'étude `t` et le vecteur de dimension 2 `rho`. Il calcule en sortie un vecteur d'estimateurs `res` qui contient les estimateurs du maximum de vraisemblance $(\hat{\rho}_A,\hat{\rho}_B)$, un critère de maximisation qui vaut 1 si l'algorithme de maximisation a convergé et 0 sinon, et une matrice `I` qui est la matrice $-D^2 l_n(\hat{\rho}_A,\hat{\rho}_B)$ qui estime la matrice asymptotique de variance-covariance.

Le fichier **Prog-princ.m** est le fichier de commandes principal. Il commence par charger les paramètres stockés dans le fichier **parametres.mat**. Il affiche les valeurs de ces paramètres à l'écran. Il réalise `L` simulations en appelant le fichier **Simulation2.m** et pour chaque simulation, il calcule un vecteur d'estimateurs en appelant le fichier **Estim2.m**. Il stocke dans un fichier **resultats.mat** les grandeurs `n`, `L`, `Rho` qui est un tableau de taille $2 \times L$ contenant les estimateurs des `L` simulations, `Critere` qui est un tableau de taille $2 \times L$ contenant les critères d'optimisation donnés par **Estim2.m** et `Sigma` qui est un tableau de taille $3 \times L$ contenant les écarts-types asymptotiques de $\hat{\rho}_A$ et $\hat{\rho}_B$ et la covariance de $(\hat{\rho}_A,\hat{\rho}_B)$.

C.4 Fichiers d'affichage graphique

Le fichier **Repart.m** est une fonction `res=Repart(t,Rho,rho,Sigma)` qui prend en argument une date quelconque `t`, un tableau `Rho` de taille $2 \times L$ contenant les estimateurs $\hat{\rho}_A$ et $\hat{\rho}_B$, le vecteur `rho` de dimension 2 contenant les vraies valeurs des parametres ρ_A^0 et ρ_B^0 et le tableau `Sigma` de taille $3 \times L$ contenant les écarts-types asymptotiques de $\hat{\rho}_A$ et $\hat{\rho}_B$ et la covariance de $(\hat{\rho}_A,\hat{\rho}_B)$. Il calcule le vecteur `res` les valeurs des fonctions de répartition empirique au point `t` des variables $\sqrt{\sigma_{h,h}}(\hat{\rho}_h - \rho_h^0)$.

Le fichier **Gaussienne.m** est une fonction `Gaussienne(Rho,rho,Sigma)` qui prend en argument un tableau `Rho` de taille $2 \times L$ contenant les estimateurs $\hat{\rho}_A$ et $\hat{\rho}_B$, le vecteur `rho` de dimension 2 contenant les vraies valeurs des parametres ρ_A^0 et ρ_B^0 et le tableau `Sigma` de taille $3 \times L$ contenant les écarts-types asymptotiques de ($\hat{\rho}_A$ et $\hat{\rho}_B$) et la covariance de $(\hat{\rho}_A,\hat{\rho}_B)$. Il donne en sortie les graphes de la fonction de répartition d'une Gaussienne centrée réduite et des fonctions de répartition empirique des estimateurs ρ_A^0 et ρ_B^0. Il fait appel au fichier **Repart.m** et nécessite la boîte à outils **stixbox**.

C.5 Programmes

```
% Parametres.m
```

```
%
% FONCTION : Ce programme permet d'enregistrer les parametres necessaires
% a la simulation des variables aleatoires dans le fichier
% parametres.mat. Les donnees sont demandees a l'ecran.
%
% APPEL DE FONCTIONS : aucune
%
% ENTREES : tout est demande a l'ecran
%
% SORTIE : fichier parametres.mat
%
%----------------
% PROGRAMME
%----------------

n=input('Taille des echantillons:');
lambda=input('Risque de base:');
mu=input('Risque de la censure:');
rho=input('Parametres de dependance ([rho1;rho2]):');
t=input('Date de fin d''etude:');
L=input('Nombre des simulations:');

save parametres n lambda mu  rho t L;
```

```
% Simulation.m
% function [T,Delta]=Simulation(n,lambda,mu,rho);
%
% FONCTION : Ce programme permet de simuler les variables
% aleatoires T_{A,i}, T_{B,i}, delta_{A,i} et delta_{B,i} et de les
% stocker dans des tableaux T et Delta.
%
% APPEL DE FONCTIONS : aucune
%
% ENTREES : n : taille de la population
%           lambda : parametre des lois exponentielles de depart
%           mu : parametre des lois exponentielles de la censure
%           rho : parametre de chocs a estimer
%
% SORTIE : T : tableau des dates T_{A,i}, T_{B,i}
%          delta : tableau des indicatrices de censure delta_{A,i} et
%          delta_{B,i}
%----------------
```

```
% PROGRAMME
%----------------

function [T,Delta]=Simulation(n,lambda,mu,rho);

% Simulation des Xi
% -----------------

U=rand(2,n); % Variables uniformes independantes sur [0 1]
A=-log(U)/lambda; % Variables exp(lambda) independantes

% Verification
disp('Esperance des Xi:');
disp(1/lambda);
MA=mean(A,2); % Moyenne empirique des Xi
  disp('Moyenne empirique des Xi:');
  disp(MA);
  disp('Ecart-type theorique des Xi:');
  disp(1/lambda);
  VA=std(A,0,2);
  disp('Ecart-type empirique des Xi:');
  disp(VA);
  IA=[MA-2*VA/sqrt(n) MA+2*VA/sqrt(n)];
  disp('Intervalle a 95 pour cent:');
  disp(IA);
disp('----------------------------------------');

B=A<flipud(A); % Tableau des Xi1<Xi2 et Xi2<Xi1

% Simulation des X'i
% ------------------

V=rand(2,n); % Variables uniformes independantes sur [0 1]
R=diag(exp(rho))*lambda;
C=-inv(R)*log(V); % Variables exp(e^rho1*lambda) et
                  % exp(e^rho2*lambda) independantes

% Verification
CC=diag(inv(R));
disp('Esperance des X''i:');
disp(CC);
MC=mean(C,2); % Moyenne empirique des X'i
```

```
  disp('Moyenne empirique des X''i:');
  disp(MC);
  disp('Ecart-type theorique des X''i:');
  disp(diag(inv(R)));
  VC=std(C,0,2);
  disp('Ecart-type empirique des X''i:');
  disp(VC);
  IC=[MC-2*VC/sqrt(n) MC+2*VC/sqrt(n)];
  disp('Intervalle a 95 pour cent:');
  disp(IC);
  disp('_____________________________________');

I=ones(2,n); % Tableau de 1
X=A.*B+(flipud(A)+C).*(I-B);
% Tableau des  Xi1*1{Xi1<Xi2}+(Xi2+X'i2)*1{Xi1>=Xi2}
% Xi2*1{Xi2<Xi1}+(Xi1+X'i1)*1{Xi2>=Xi1}

% Verification
CX=[(1/lambda+exp(-rho(1)))/2;(1/lambda+exp(-rho(2)))/2];
disp('Esperance des Xi-tilde:');
disp(CX);
MX=mean(X,2); % Moyenne empirique des X~i
  disp('Moyenne empirique des X-tilde:');
  disp(MX);
  ECX=[sqrt(1/lambda^2+3*exp(-2*rho(1)))/2 ;
     sqrt(1/lambda^2+3*exp(-2*rho(2)))/2];
disp('Ecart-type theorique des Xi-tilde:');
  disp(ECX);
  VX=std(X,0,2);
  disp('Ecart-type empirique des Xi-tilde:');
  disp(VC);
  IX=[MX-2*VX/sqrt(n) MX+2*VX/sqrt(n)];
  disp('Intervalle a 95 pour cent:');
  disp(IX);
  disp('_____________________________________');

% Simulation de la censure

W=rand(2,n); % Variables uniformes independantes sur [0 1]
E=-log(W)/mu; % Variables exp(mu) independantes

  % Verification
```

```
  disp('Esperance des Ci:');
disp(1/mu);
ME=mean(E,2); % Moyenne empirique des Ci
  disp('Moyenne empirique des Ci:');
  disp(ME);
  disp('Ecart-type theorique des Ci:');
  disp(1/mu);
  VE=std(E,0,2);
  disp('Ecart-type empirique des Ci:');
  disp(VE);
  IE=[ME-2*VE/sqrt(n) ME+2*VE/sqrt(n)];
  disp('Intervalle a 95 pour cent:');
  disp(IE);
  disp('_____________________________________________');

% Definition des variables Ti et delta-i

Delta=(X<=E); % Tableau des 1{Xi1<=Ui1}
 disp('Moyenne des Delta-i:');
 disp(mean(Delta,2));
 disp('Ecart-type empirique des Delta-i:');
 disp(std(Delta,0,2));
 disp('_____________________________________________');
T=(X-E).*Delta+E; % Tableau des inf(X~i,Ui)
 disp('Moyenne des Ti:');
 disp(mean(T,2));
 disp('Ecart-type empirique des T-i:');
 disp(std(T,0,2));
```

```
% Simulation2.m
% function [T,Delta]=Simulation2(n,lambda,mu,rho);
%
% FONCTION : Ce programme permet de simuler les variables
% aleatoires T_{A,i}, T_{B,i}, delta_{A,i} et delta_{B,i} et de les
% stocker dans des tableaux T et Delta.
%
% APPEL DE FONCTIONS : aucune
%
% ENTREES : n : taille de la population
%           lambda : parametre des lois exponentielles de depart
%           mu : parametre des lois exponentielles de la censure
%           rho : parametre de chocs a estimer
```

```
%
% SORTIE : T : tableau des dates T_{A,i}, T_{B,i}
%          delta : tableau des indicatrices de censure delta_{A,i} et
%          delta_{B,i}
%----------------
% PROGRAMME
%----------------
function [T,Delta]=Simulation2(n,lambda,mu,rho);

% Simulation des Xi

U=rand(2,n); % Variables uniformes independantes sur [0 1]
A=-log(U)/lambda; % Variables exp(lambda) independantes

B=A<flipud(A); % Tableau des Xi1<Xi2 et Xi2<Xi1

% Simulation des Xi-tilde

V=rand(2,n); % Variables uniformes independantes sur [0 1]
R=diag(exp(rho))*lambda;
C=-inv(R)*log(V); % Variables exp(e^rho1*lambda) et
                        % exp(e^rho2*lambda) independantes

I=ones(2,n); % Tableau de 1
X=A.*B+(flipud(A)+C).*(I-B);
% Tableau des  Xi1*1{Xi1<Xi2}+(Xi2+X'i2)*1{Xi1>=Xi2}
% Xi2*1{Xi2<Xi1}+(Xi1+X'i1)*1{Xi2>=Xi1}

% Simulation de la censure

W=rand(2,n); % Variables uniformes independantes sur [0 1]
E=-log(W)/mu; % Variables exp(mu) independantes

% Definition des variables Ti et delta-i

Delta=X<=E; % Tableau des 1{Xi1<=Ui1}

T=(X-E).*Delta+E; % Tableau des inf(X~i,Ui)
```

```
% Compte.m
% function [Nprime, Nseconde]=Compte(T, Delta,t);
%
```

```
% FONCTION : Ce programme permet de calculer les processus de comptages
% N_1, N_2, N_3, et N_4 a la date t a partir des donnees contenues dans
% les tableaux T et Delta.
%
% APPEL DE FONCTION : aucune
%
% ENTREES : T : tableau des dates T_{A,i}, T_{B,i}
%           delta : tableau des indicatrices de censure delta_{A,i} et
%           delta_{B,i}
%           t : date a laquelle on calcule les processus de comptage
%
% SORTIE : Nprime : tableau d'indicatrices contenant les N_{2,i} sur la
% premiere ligne et les  N_{4,i} sur la deuxieme ligne
%          Nseconde : tableau d'indicatrices contenant les N_{1,i} sur la
% premiere ligne et les  N_{3,i} sur la deuxieme ligne
%----------------
% PROGRAMME
%----------------
function [Nprime, Nseconde]=Compte(T, Delta,t);

[m,n]=size(T); % taille du tableau T

A=Delta.*flipud(Delta); % tableau des di1*di2
B=T<=flipud(T); % tableau des 1{Ti1<=Ti2} et 1{Ti2<=Ti1}
I=ones(m,n); % tableau de 1
C=T<=(t*I); % tableau des 1{Ti<t}
Nprime=A.*(I-B).*C; % N'1(t)=di1*di2*1{Ti1>Ti2}*1{Ti1<=t}
% N'2(t)=di2*di1*1{Ti2>Ti1}*1{Ti2<t}
Nseconde=Delta.*B.*C; % N''1(t)=di1*1{Ti1<=Ti2}*1{Ti1<=t}
% N''2(t)=di2*1{Ti2<=Ti1}*1{Ti2<=t}
```

```
% Sn.m
% function s=Sn(T,Delta,t,rho);
%
% FONCTION : Cette fonction permet de calculer la somme S_n(rho,t) a
% partir des donnees T et Delta et en fonction de la date t et des
% parametres rho.
%
% APPEL DE FONCTIONS : vide
%
% ENTREE : T : tableau des dates T_{A,i} et T_{B,i}
%          Delta : tableau des indicatrices Delta_{A,i} et Dleta_{B,i}
```

```
%          t : date finale
%          rho : parametres de chocs
%
% SORTIE : s : valeur de S_n(rho,t)
%----------------
% PROGRAMME
%----------------
function s=Sn(T,Delta,t,rho);

[m,n]=size(T); % Taille du tableau T

I=ones(size(Delta)); % Matrice de 1
A=T>=(t*I); % Matrice des indicatrices Ti>=t
R=diag(exp(rho)); % Matrice diagonale (exp(rho2),exp(rho1))
B=R*flipud((Delta.*(I-A)))+flipud(A); % Matrice des
% exp(rho)*Delta-i*1{Ti<t}+1{Ti>=t}
C=B.*A; % Matrice des Bi1*1{Ti2>=t} et Bi2*1{Ti1>=t}
s=sum(sum(C))/n; % Somme de tous les termes de C divisee par n
```

```
% DSn.m
% function D=DSn(T,Delta,t,rho);
%
% FONCTION : Cette fonction permet de calculer le vecteur-derive de la
% somme S_n(rho,t) par rapport a rho_A et rho_B a partir des donnees T
% et Delta, en fonction de la date t et des parametres rho.
%
% APPEL DE FONCTIONS : vide
%
% ENTREE : T : tableau des dates T_{A,i} et T_{B,i}
%          Delta : tableau des indicatrices Delta_{A,i} et Dleta_{B,i}
%          t : date finale
%          rho : parametres de chocs
%
% SORTIE : D : vecteur des derivees de S_n(rho,t) par rapport a rho_A
% et rho_B.
%
%----------------
% PROGRAMME
%----------------
function D=DSn(T,Delta,t,rho);

[m,n]=size(T); % Taille du tableau T
```

```
I=ones(size(Delta)); % Matrice de 1
A=T>=(t*I); % Matrice des indicatrices Ti>=t
R=diag(exp(rho)); % Matrice diagonale (exp(rho2),exp(rho1))
B=R*flipud((Delta.*(I-A))).     % Matrice des
% exp(rho1)*Delta-i2*1{Ti2<t}*1{Ti1>=t}
D=sum(B,2)/n;
```

```
% Variance.m
% function v=Variance(T,Delta,t,rho);
%
% FONCTION : Cette fonction permet de calculer la matrice
% -D^2ln(rho_A,rho_B) a partir des donnees T et Delta, avec la date
% finale t et au point rho.
%
% APPEL DE FONCTION : Compte.m, Matrice.m
%
% ENTREE : T : tableau des dates T_{A,i} et T_{B,i}
%          Delta : tableau des indicatrices Delta_{A,i} et Delta_{B,i}
%          t : date finale
%          rho : parametres de chocs
%
% SORTIE : v : matrice des derivees secondes de la vraisemblance
%----------------
% PROGRAMME
%----------------
function v=Variance(T,Delta,t,rho);

[m,n]=size(T); % taille de T

[Nprime,Nseconde]=Compte(T,Delta,t); % Tableau des N' et N''
B=Nprime+Nseconde; % Tableau des N'+N''

% Calcul de la somme des D2Sn/Sn-(DSn/Sn)2 aux points Ti1 et Ti2
% divisee par n

M=[0 0;0 0];
for j=1:n
M=M+Matrice(T,Delta,T(1,j),rho)*B(1,j);
M=M+Matrice(T,Delta,T(2,j),rho)*B(2,j);
end;
v=M/n;
```

```
% Matrice.m
% function M=Matrice(T,Delta,t,rho);
%
% FONCTION : Cette fonction permet de calculer la matrice
% D^2Sn/Sn-(DSn/Sn)^2 au point rho et a la date t a partir des donnees.
%
% APPEL DE FONCTIONS : Sn.m, DSn.m
%
% ENTREE : T : tableau des dates T_{A,i} et T_{B,i}
%          Delta : tableau des indicatrices Delta_{A,i} et Dleta_{B,i}
%          t : date finale
%          rho : parametres de chocs
%
% SORTIE : M : matrice D^2Sn/Sn-(DSn/Sn)^2 au point rho et a la date t
%----------------
% PROGRAMME
%----------------
function M=Matrice(T,Delta,t,rho);

[m,n]=size(T); % taille de T

D=DSn(T,Delta,t,rho);
s=Sn(T,Delta,t,rho);

% Si Sn=0 cette matrice est egale a zero car
% alors les processus contre lesquels on l'integre sont nuls
if s==0
M=[0 0;0 0];
else
M=diag(D)/s-(D/s)*(D'/s);
end;
```

```
% Vraisemb.m
% function [v,grad,hess]=Vraisemb(rho,T,Delta,t);
%
% FONCTION : Cette fonction permet de calculer l'opposee de la valeur de
% la fonction de vraisemblance, sa derivee premiere et sa derivee seconde
% au point rho a la date finale t a partir des donnees de T et Delta.
%
% APPEL DE FONCTIONS : Compte.m, Sn.m, DSn.m, Variance.m,
%
```

```
% ENTREE : rho : valeur du point ou l'on calcule la vraisemblance.
%          T : tableau des dates
%          Delta : tableau des indicatrices
%          t : date finale d'etude
%
% SORTIE : v : valeur de la vraisemblance  au point rho
%          grad : vecteur des derivees premieres au point rho
%          hess : matrice des derivees secondes au point rho
%----------------
% PROGRAMME
%----------------
function [v,grad,hess]=Vraisemb(rho,T,Delta,t);

[m,n]=size(T); % taille de T

% Calcul des Sn(Ti1), Sn(Ti2), DSn(Ti1) et DSn(Ti2)

for j=1:n % tableau des Sn(Ti1) et Sn(Ti2)
S(1,j)=Sn(T,Delta,T(1,j),rho);
S(2,j)=Sn(T,Delta,T(2,j),rho);
V1=DSn(T,Delta,T(1,j),rho);
V2=DSn(T,Delta,T(2,j),rho);
DS1(1,j)=V1(1);
DS1(2,j)=V1(2);
DS2(1,j)=V2(1);
DS2(2,j)=V2(2);
end;

[Nprime,Nseconde]=Compte(T,Delta,t); % Calcul des N'(t) et N''(t)

R=diag(rho); % Matrice diagonale de rho1 et rho2

A=R*Nprime;   % Matrice des rho1*N'1(t) et
% rho2*N''2(t)

B=Nprime+Nseconde; % Matrice des N'+N''

O=zeros(size(S));
C=A-log(S+(S==0)).*B; % Matrice des rho1*N'1-log(Sn(T1))*N1 et des
% rho2*N'2-log(Sn(T2))*N2

% Si Sn(Ti1)=0 alors N'(Ti1) et N''(Ti1) sont nuls aussi, donc avec la
```

```
% convention 0*log(0)=0, il n'y a pas de probleme. On peut donc remplacer
% tous les coeffs nuls de la matrice par des 1. D'ou le (S==0).

v=-sum(sum(C))/n; % Opposee de la somme des termes de C divisee par n

% Calcul du gradient

E1=DS1./([1 0;1 0]*(S+(S==0))).*([1 0;1 0]*B);
E2=DS2./([0 1;0 1]*(S+(S==0))).*([0 1;0 1]*B);
G=Nprime-E1-E2; % Matrice des N'-DSn/Sn*(N'+N'')
grad=-sum(G,2)/n; % Opposee de la somme ligne
% par ligne de G

% Calcul du hessien

hess=Variance(T,Delta,t,rho);
```

```
% Estim.m
% function Estim(T,Delta,t,rho);
%
% FONCTION : Cette fonction permet de calculer l'estimateur du maximum de
% vraisemblance a partir des donnees T et Delta, de la date finale t et
% du parametres rho de depart. Le programme compare les estimateurs
% suivant differentes options de maximisation.
%
% APPEL DE FONCTIONS : Vraisemb.m
%
% ENTREE : T : tableau des dates T_{A,i} et T_{B,i}
%          Delta : tableau des indicatrices Delta_{A,i} et Dleta_{B,i}
%          t : date finale
%          rho : vraies valeurs des parametres de chocs, servent a
%          initialiser le programme de maximisation.
%
% SORTIE : aucune
%----------------
% PROGRAMME
%----------------
function Estim(T,Delta,t,rho);

[m,n]=size(T); % Taille du tableau T

options=optimset([],'GradObj','off','Hessian','off','Diagnostics','on');
```

```
% Options d'optimisation
disp('Options choisies: GradObj=off, Hessian=off, Diagnostics=on');

debut=clock;          % Heure de debut de calcul
res1=fminunc('Vraisemb',rho,options,T,Delta,t);
% Maximisation de la vraisemblance
% a partir de la vraie valeur du parametre
disp(res1);

disp('Temps de calcul:');
disp(fix(etime(clock,debut)));

disp('_____________________________________________________________');

options=optimset([],'GradObj','on','Hessian','off','Diagnostics','on');
% Options d'optimisation
disp('Options choisies: GradObj=on, Hessian=off, Diagnostics=on');
debut=clock;          % Heure de debut de calcul
res2=fminunc('Vraisemb',rho,options,T,Delta,t);
disp(res2);

disp('Temps de calcul:');
disp(fix(etime(clock,debut)));

disp('_____________________________________________________________');

options=optimset([],'GradObj','on','Hessian','on','Diagnostics','on');
% Options d'optimisation
disp('Options choisies: GradObj=on, Hessian=on, Diagnostics=on');
debut=clock;          % Heure de debut de calcul
res3=fminunc('Vraisemb',rho,options,T,Delta,t);
disp(res3);

disp('Temps de calcul:');
disp(fix(etime(clock,debut)));
```

```
% Estim2.m
% function [res,crit,I]=Estim2(T,Delta,t,rho);
%
% FONCTION : Cette fonction permet de calculer l'estimateur du max de
% vraisemblance, donne un critere de convergence et un estimateur de la
```

```
% matrice asymptotique.
%
% APPEL DE FONCTIONS : Vraisemb.m, Variance.m,
%
% ENTREE : T : tableau des dates
%          Delta : tableau des indicatrices
%          t : date finale d'etude
%          rho : vraies valeurs des parametres rho_A et rho_B servant a
%          initialiser la maximisation.
%
% SORTIE : res: vecteur des estimateurs rho_A_chap et rho_B_chap
%          crit : critere d'optimisation de la convergence (=1 si OK)
%          I : matrice de variance-covariance asymptotique -D^2ln(rho_chap)
%----------------
% PROGRAMME
%----------------
function [res,crit,I]=Estim2(T,Delta,t,rho);

[m,n]=size(T);                  % Taille du tableau T

options=optimset([],'GradObj','on','Hessian','on','Display','off');
                                % Options d''optimisation

[res,eval,crit]=fminunc('Vraisemb',rho,options,T,Delta,t);
                                % Maximisation de la vraisemblance a
                                % partir de la vraie valeur du parametre

I=Variance(T,Delta,t,res);      % Matrice de la variance limite de
% sqrt(n)(rho_chapeau-rho_0)
```

```
% Programme principal: Prog_princ
%
% FONCTION : Fichier pour plusieurs simulations gardant en memoire les
% estimateurs des rho, le critere de convergence et la matrice de
% variance covariance de la loi limite de
% racine(n)(rho_chap-rho_0). Calcule la moyenne de ces estimateurs et la
% somme des criteres convergents.
%
% APPEL DE FONCTION : Simulation2.m, Estim2.m
%
% ENTREE : parametres.mat contenant les variables n, lambda, mu, rho, t ,
% L (nombre de simulations).
```

```
%
% SORTIE : resultats.mat gardant en memoire les variables n, L, Rho,
% Critere, Sigma. Rho contient les L vecteurs-estimateurs, Critere les L
% criteres de convergence, Sigma les L vecteurs contenant l'ecart-type de
% A, l'ecart-type de B et la covariance de A et B.
%----------------
% PROGRAMME
%----------------
clear;

load parametres;
disp('n=');
disp(n);
disp('lambda=');
disp(lambda);
disp('mu=');
disp(mu);
disp('rho=');
disp(rho);
disp('t=');
disp(t);
disp('L=');
disp(L);

debut=fix(clock); % Heure de debut de calcul
disp('Date et heure de debut:');
disp(debut);

for l=1:L     % pour chaque simulation
  [T,Delta]=Simulation2(n,lambda,mu,rho); % Simulation des variables
  [res,crit,I]=Estim2(T,Delta,t,rho);     % Estimation de rho1 rho2
  % critere de convergence
  % variance et covariance
  Rho(1,l)=res(1);    % Stockage de rho1 dans le tableau Rho1
  Rho(2,l)=res(2);    % Stockage de rho2 dans le tableau Rho2
  Critere(l)=crit;    % Stockage du critere
  M=inv(I);
  Sigma(1,l)=sqrt(M(1,1)/n); % Stockage du terme de ecart-type pour rho1
  Sigma(2,l)=sqrt(M(2,2)/n); % Stockage du terme de ecart-type pour rho2
  Sigma(3,l)=M(1,2)/n;    % Stockage du terme de covariance
end;
```

```
save resultats n L Rho Critere Sigma;  % Sauvegarde des donnees dans
      % resultats

r=sum(Rho,2)/L;        % Moyennes des rho1 et rho2
disp('Moyenne des rho_chapeau:');
disp(r);

c=sum(Critere);        % Nombre de criteres convergents
disp('Nombre de criteres convergents:');
disp(c);

b=(r-rho)./rho;
disp('Biais relatif estime:');
disp(b);

v=std(Rho,0,2);
disp('Ecart-type empirique:');
disp(v);

S=[1 0 0;0 1 0]*Sigma;
s=sum(S,2)/L;
disp('Ecart-type asymptotique estime:');
disp(s);

disp('Biais relatif des ecarts-type:');
disp((v-s)./v);

O=ones(size(Rho));
D=diag(rho);
Inf=(D*O<Rho-S*1.96);
IC=(D*O>=Rho-S*1.96).*(D*O<=Rho+S*1.96);
Sup=(D*O>Rho+S*1.96);

disp('Nombre de cas en dessous de l''IC:');
disp(sum(Inf,2));
disp('Nombre de cas dans l''IC:');
disp(sum(IC,2));
disp('Nombre de cas au dessus de l''IC:');
disp(sum(Sup,2));

% Gaussienne(Rho,rho);
```

```
fin=fix(clock); % Heure de fin de calcul
disp('Date et heure de fin:');
disp(fin);
```

```
% Repart.m
% function res=Repart(t,Rho,rho,Sigma);
%
% FONCTION : Ce programme permet de calculer la fonction de repartition
% des rho_A_chap et rho_B_chap au point t.
%
% APPEL DE FONCTIONS : aucune
%
% ENTREES : t : point de la fonction de repartition
%           Rho : tableau des estimateurs
%           rho : vraies valeurs des parametres de chocs
%           Sigma : matrice des variance-covariance
%
% SORTIE : res : valeur de la fonction de repartition au point t
%----------------
% PROGRAMME
%----------------
function res=Repart(t,Rho,rho,Sigma);

[m,L]=size(Rho); % donne la taille de Rho

I=ones(m,L); % matrice de 1
R=diag(rho);
A=(Rho-R*I)./([1 0 0;0 1 0]*Sigma);

B=A<=(t.*I); % matrice des indicatrices
% (rho_chapeau-rho_0)/se(rho_chapeau)
res=sum(B,2)/L; % fonctions de repartitions empiriques
```

```
% Gaussienne.m
% function Gaussienne(Rho,rho,Sigma);
%
% FONCTION : Ce programme permet de tracer la fonction de repartition
% d'une gaussienne N(0,1) et les deux fonctions de repartition empiriques
% des estimateurs rho_A_chap et rho_B_chap sur le meme graphique.
%
% APPEL DE FONCTIONS : Repart.m, stixbox
%
```

```
% ENTREES : Rho : tableau des estimateurs
%           rho : vraies valeurs des parametres de chocs a estimer
%           Sigma : tableau des variance-covariance
%
% SORTIE : un graphique. Ce programme ne peut pas etre lancer en nohup.
%----------------
% PROGRAMME
%----------------
function Gaussienne(Rho,rho,Sigma);

addpath  /home/letue/stixbox;

[m,L]=size(Rho); % donne la taille de Rho

I=ones(m,L); % matrice de 1

borne_inf=-5;
borne_sup=5;
X=(0:1000)*(borne_sup-borne_inf)/1000+borne_inf;
for i=1:1001
Y=Repart(X(i),Rho,rho,Sigma);
Y1(i)=Y(1);
Y2(i)=Y(2);
Z(i)=pnorm(X(i));
end;
[E1,H1]=stairs(X,Y1);
[E2,H2]=stairs(X,Y2);
plot(E1,H1,'b',E2,H2,'g',X,Z,'r');
```

Bibliographie

[1] AALEN, O. Nonparametric inference in connection with multiple decrement models. *Scand. J. Statist. 3*, 1 (1976), 15–27.

[2] AALEN, O. Nonparametric inference for a family of counting processes. *Ann. Statist. 6*, 4 (1978), 701–726.

[3] AALEN, O. O., BORGAN, Ø., KEIDING, N., AND THORMANN, J. Interaction between life history events. Nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scand. J. Statist. 7*, 4 (1980), 161–171.

[4] AKAIKE, H. Information theory and an extension of the maximum likelihood principle. 267–281.

[5] ANDERSEN, P. K., BORGAN, Ø., GILL, R. D., AND KEIDING, N. *Statistical models based on counting processes.* Springer-Verlag, New York, 1993.

[6] ANDERSEN, P. K., AND GILL, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Statist. 10*, 4 (1982), 1100–1120.

[7] BAILEY, K. R. Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model. *Ann. Statist. 12*, 2 (1984), 730–736.

[8] BARRON, A. R., BIRGÉ, L., AND MASSART, P. Risk bounds for model selection via penalization. *Probab. Theory Related Fields 113*, 3 (1999), 301–413.

[9] BARRON, A. R., AND SHEU, C.-H. Approximation of density functions by sequences of exponential families. *Ann. Statist. 19*, 3 (1991), 1347–1369.

[10] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y., AND WELLNER, J. A. *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins University Press, Baltimore, MD, 1993.

[11] BIRGÉ, L., AND MASSART, P. A generalized C_p criterion for gaussian model selection. Tech. rep., En préparation.

[12] BIRGÉ, L., AND MASSART, P. *Model selection from a non-asymptotic point of view.* En préparation.

[13] BIRGÉ, L., AND MASSART, P. Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields 97*, 1-2 (1993), 113–150.

[14] BIRGÉ, L., AND MASSART, P. Gaussian model selection. Preprint 2000.05, Université de Paris-Sud, 2000.

[15] BIRGÉ, L., AND ROZENHOLC, Y. How many bins should be put in a regular histogram. Technical report, Université Paris VI, 1999.

[16] BRESLOW, N. E. Covariance analysis of censored survival data. *Biometrics 30* (1974), 89–99.

[17] BRESLOW, N. E., AND CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist. 2* (1974), 437–453. Collection of articles dedicated to Jerzy Neyman on his 80th birthday.

[18] BRETAGNOLLE, J. Processus ponctuels. Cours de DEA, Université de Paris Sud, 1997–98.

[19] CASTELLAN, G. Modified Akaike's criterion for histogram density estimation. Preprint 99.61, Université de Paris-Sud, 1999.

[20] CASTELLAN, G. Density estimation via exponential model selection. Preprint 2000.25, Université de Paris-Sud, 2000.

[21] CASTELLAN, G. Sélection d'histogrammes ou de modèles exponentiels de polynômes par morceaux à l'aide d'un critère de type Akaike. Thèse 6039, Université de Paris-Sud, 2000.

[22] CLAYTON, D. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 1 (1978), 141–151.

[23] CLAYTON, D., AND CUZICK, J. Multivariate generalizations of the proportional hazards model. *J. Roy. Statist. Soc. Ser. A 148*, 2 (1985), 82–117.

[24] COX, D. D. Multivariate smoothing spline functions. *SIAM J. Numer. Anal. 21*, 4 (1984), 789–813.

[25] COX, D. R. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B 34* (1972), 187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

[26] COX, D. R., AND OAKES, D. *Analysis of survival data.* Chapman & Hall, London, 1984.

[27] DABROWSKA, D. M. Kaplan-Meier estimate on the plane. *Ann. Statist. 16*, 4 (1988), 1475–1489.

[28] DABROWSKA, D. M. Kaplan-Meier estimate on the plane: weak convergence, LIL, and the bootstrap. *J. Multivariate Anal. 29*, 2 (1989), 308–325.

[29] DUDLEY, R. M. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Functional Analysis 1* (1967), 290–330.

[30] GILL, R. D., VAN DER LAAN, M. J., AND WELLNER, J. A. Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. H. Poincaré Probab. Statist. 31*, 3 (1995), 545–597.

[31] GROSS, A. J., AND CLARK, V. A. *Survival distributions: Reliability applications in the biomedical sciences.* Wiley Series in Probability and Mathematical Statistics., New York, 1975.

[32] GROSS, S. T., AND HUBER, C. Matched pair experiments: Cox and maximum likelihood estimation. *Scand. J. Statist. 14*, 1 (1987), 27–41.

[33] HASTIE, T. J., AND TIBSHIRANI, R. J. *Generalized additive models.* Chapman and Hall Ltd., London, 1990.

[34] HOUGAARD, P. Modelling multivariate survival. *Scand. J. Statist. 14*, 4 (1987), 291–304.

[35] JACOD, J. On the stochastic intensity of a random point process over the half line. Technical report 15, Department of Statistics, Princeton University, 1973.

[36] JACOD, J. Multivariate point processes: predictable projection, Radon-Nikodým derivatives, representation of martingales. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 31* (1974/75), 235–253.

[37] JOHANSEN, S. An extension of Cox's regression model. *Internat. Statist. Rev. 51*, 2 (1983), 165–174.

[38] KALBFLEISCH, J. D., AND PRENTICE, R. L. *The statistical analysis of failure time data.* John Wiley and Sons, New York-Chichester-Brisbane, 1980. Wiley Series in Probability and Mathematical Statistics.

[39] KAPLAN, E. L., AND MEIER, P. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc. 53* (1958), 457–481.

[40] KEIDING, N. Statistical inference in the Lexis diagram. *Philos. Trans. Roy. Soc. London Ser. A 332*, 1627 (1990), 487–509.

[41] LAWLESS, J. F. *Statistical models and methods for lifetime data.* John Wiley & Sons Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[42] LELIÈVRE, E., BONVALET, C., AND BRY, X. Analyse biographique des groupes, les avancées d'une recherche en cours. *Population 4* (1997), 803–830.

[43] MCKEAGUE, I. W., AND UTIKAL, K. J. Inference for a nonlinear counting process regression model. *Ann. Statist. 18*, 3 (1990), 1172–1187.

[44] NÆS, T. The asymptotic distribution of the estimator for the regression parameter in Cox's regression model. *Scand. J. Statist. 9*, 2 (1982), 107–115.

[45] NELSON, W. Hazard plotting for incomplete failure data. *J. Qual. Technol. 1* (1969), 27–52.

[46] NELSON, W. Theory and applications of hazard plotting for censored failure data. *Technometrics 14* (1972), 945–965.

[47] NIELSEN, G. G., GILL, R. D., ANDERSEN, P. K., AND SØRENSEN, T. I. A. A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist. 19*, 1 (1992), 25–43.

[48] OAKES, D. Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc. 84*, 406 (1989), 487–493.

[49] O'SULLIVAN, F. Nonparametric estimation in the Cox model. *Ann. Statist. 21*, 1 (1993), 124–145.

[50] PONS, O. A test of independence between two censored survival times. *Scand. J. Statist. 13*, 3 (1986), 173–185.

[51] PONS, O., AND DE TURCKHEIM, È. Tests of independence for bivariate censored data based on the empirical joint hazard function. *Scand. J. Statist. 18*, 1 (1991), 21–37.

[52] REBOLLEDO, R. La méthode des martingales appliquée à l'étude de la convergence en loi de processus. *Bull. Soc. Math. France Mém.*, 62 (1979), v+125 pp. (1980).

[53] SENOUSSI, R. Problème d'identification dans le modèle de Cox. *Ann. Inst. H. Poincaré Probab. Statist. 26*, 1 (1990), 45–64.

[54] STONE, C. J. Optimal global rates of convergence for nonparametric regression. *Ann. Statist. 10*, 4 (1982), 1040–1053.

[55] TSIATIS, A. A. A large sample study of Cox's regression model. *Ann. Statist. 9*, 1 (1981), 93–108.

[56] VAN DER LAAN, M. J. Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist. 24*, 2 (1996), 596–627.

[57] VAN DER VAART, A. W. *Asymptotic statistics.* Cambridge University Press, Cambridge, 1998.

[58] VAN DER VAART, A. W. Semiparametric statistics. Cours, École d'Été de Probabilités de Saint-Flour XXIX, 1999.

[59] VAN DER VAART, A. W., AND WELLNER, J. A. *Weak convergence and empirical processes.* Springer-Verlag, New York, 1996. With applications to statistics.

[60] VAUPEL, J. W., MANTON, K. G., AND STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16 (1979), 439–454.

[61] WAHBA, G. *Spline models for observational data.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.