UNIVERSITÉ TOULOUSE III - PAUL SABATIER

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE TOULOUSE
délivré par l'Université Toulouse III - Paul Sabatier

en Mathématiques Appliquées

présentée par

Amélie DETAIS

intitulée

Maximum de vraisemblance et moindres carrés pénalisés dans des modèles de durées de vie censurées.

Directeurs de thèse :
Jean-François Dupuy et Jean-Claude Fort

Soutenue le 19 novembre 2008 devant le jury composé de Messieurs :

Laurent Bordes	Université de Pau et des Pays de l'Adour	Rapporteur
Daniel Commenges	Université Bordeaux II	Examinateur
Jean-François Dupuy	Université Toulouse III	Directeur
Jean-Claude Fort	Université Toulouse III	Directeur
Jean-Michel Loubes	Université Toulouse III	Examinateur
Carlos Gabriel Matrán Bea	Universidad de Valladolid	Rapporteur

Institut de Mathématiques de Toulouse

Unité mixte de recherche C.N.R.S. - U.M.R. 5219

Université Paul Sabatier Toulouse 3 - Bât 1R3, 31062 TOULOUSE cedex 9, France

Remerciements

Je tiens en premier lieu à remercier Jean-François Dupuy de m'avoir offert la possibilité de commencer cette thèse et qui m'a donc initiée aux joies de la recherche. Il m'a proposé des sujets d'étude très intéressants et a été d'une grande aide pour m'expliquer la façon de les aborder. Je le remercie de tous ses conseils lors de la direction de ce travail de recherche.

Je souhaite également remercier Jean-Claude Fort de m'avoir accompagnée au cours de ces trois années et d'avoir réfléchi avec moi sur certains obstacles mathématiques. Son expérience m'a été tout à fait profitable.

Je tiens bien entendu à exprimer ma sincère reconnaissance à Laurent Bordes et Carlos Gabriel Matrán Bea qui ont accepté, malgré toutes leurs occupations, d'être les rapporteurs de ce travail de thèse. Je leur sais gré de l'intérêt qu'ils ont porté à mon travail. Je souhaite les remercier pour leur relecture de ce mémoire ainsi que pour leurs remarques et commentaires qui me donnent notamment de nouvelles perspectives de recherche. Je suis également reconnaissante à Daniel Commenges de me faire l'honneur de présider le jury.

Je souhaite remercier Jean-Michel Loubes pour sa présence dans le jury, d'autant qu'il a assisté à mes débuts dans la recherche en encadrant mon stage de master 2. Je lui sais gré de m'avoir fait profité de son expérience lors de diverses discussions.

J'exprime aussi mes remerciements plus généralement à toute l'équipe du Laboratoire de Statistique et Probabilités de Toulouse que j'ai eu l'opportunité de côtoyer au cours de ces trois années de thèse. Je pense notamment à Jérémie Bigot qui m'a éveillée à un certain Van der Vaart, à Fabrice Gamboa qui m'a orientée à l'issue de mon master recherche et à Serge Cohen avec lequel il a été très agréable de partager certains enseignements. Je remercie aussi Jean-Marc Azaïs pour avoir accepté de partager son guide des Pyrénées avec moi ! Je n'oublie pas Marie-Laure Ausset, Françoise Michel et Agnès Requis qui ont été d'un grand secours pour venir à bout de certaines questions administratives.

Ces quatre années dans la ville rose m'ont permis de rencontrer de nombreux Toulousains, d'origine ou d'adoption. Je me souviendrai des doctorants que j'ai côtoyés, capables de comprendre les affres traversées durant la thèse : Agnès, Myriam, Christophe, Ignacio, Renaud, qui m'ont fait partager

leur expérience, et Mary-Ana, Jean-Paul, Matthieu, Maxime F. et Michel. Je remercie aussi Maxime D., notamment pour son aide efficace sur tous les problèmes de l'informatique en réseau, et Laurent, pour sa gentillesse et son aide pour les dernières formalités de thèse. Je pense également à Nolwen et Marion, pour les diverses activités partagées (LaTeX, nature, sport, rencontres avec les animaux !). L'écoute de Florent et nos diverses discussions (plongée, jeux, relations humaines...) m'ont souvent permis de relativiser et j'en garderai de très bons souvenirs. La double compréhension de Lionel et Séverine et leur amitié m'ont également beaucoup aidée durant ces trois ans. Je remercie aussi Aurélien et Julie pour leurs oreilles attentives et tous les bons moments partagés, que ce soit autour d'un jeu, d'un morceau de canard ou en altitude. Je n'oublie pas mes amis d'Orsay, Antoine, Stéphanie et Mickaël dont je me réjouis de partager plus souvent la bonne humeur alors que j'ai retraversé la Loire, et les amis d'ailleurs, Virginie, Julien.

Je souhaite enfin exprimer toute ma reconnaissance à ma mère et ma soeur pour leur écoute et leur soutien durant ces années d'éloignement. Je remercie ma mère de m'avoir permis de réaliser mes études jusqu'au bout et d'être venue me voir exposer ce charabia "sur la vraisemblance" à Toulouse. À Antoine, un énorme MERCI de m'avoir si patiemment accompagnée à travers ces trois années dans la ville rose et de partager sa vie avec moi. Je lui dédie ce mémoire de thèse.

Résumé

L'analyse de durées de vie censurées est utilisée dans des domaines d'application variés et différentes possibilités ont été proposées pour la modélisation de telles données. Nous nous intéressons dans cette thèse à deux types de modélisation différents, le modèle de Cox stratifié avec indicateurs de strates aléatoirement manquants et le modèle de régression linéaire censuré à droite. Nous proposons des méthodes d'estimation des paramètres et établissons les propriétés asymptotiques des estimateurs obtenus dans chacun de ces modèles.

Dans un premier temps, nous considérons une généralisation du modèle de Cox qui permet à différents groupes de la population, appelés strates, de posséder des fonctions d'intensité de base différentes tandis que la valeur du paramètre de régression est commune. Dans ce modèle à intensité proportionnelle stratifié, nous nous intéressons à l'estimation des paramètres lorsque l'indicateur de strate est manquant pour certains individus de la population. Des estimateurs du maximum de vraisemblance non paramétrique pour les paramètres du modèle sont proposés et nous montrons leurs consistance et normalité asymptotique. L'efficacité asymptotique du paramètre de régression est établie et des estimateurs consistants de sa variance asymptotique sont également obtenus. Pour l'évaluation des estimateurs du modèle, nous proposons l'utilisation de l'algorithme Espérance-Maximisation et le développons dans ce cas particulier.

Dans un second temps, nous nous intéressons au modèle de régression linéaire lorsque la donnée réponse est censurée aléatoirement à droite. Nous introduisons un nouvel estimateur du paramètre de régression minimisant un critère des moindres carrés pénalisé et pondéré par des poids de Kaplan-Meier. Des résultats de consistance et normalité asymptotique sont obtenus et une étude de simulations est effectuée pour illustrer les propriétés de cet estimateur de type LASSO. La méthode bootstrap est utilisée pour l'estimation de la variance asymptotique.

Mots clés

Durées de vie censurées, régression linéaire, modèle à intensité proportionnelle stratifié, estimateurs LASSO, bridge, estimateur du maximum de vraisemblance non paramétrique, poids de Kaplan-Meier, estimation bootstrap, consistance, normalité asymptotique, données manquantes, estimation de la variance.

Abstract

Life data analysis is used in various application fields. Different methods have been proposed for modelling such data. In this thesis, we are interested in two distinct modelisation types, the stratified Cox model with randomly missing strata indicators and the right-censored linear regression model. We propose methods for estimating the parameters and establish the asymptotic properties of the obtained estimators in each of these models.

First, we consider a generalization of the Cox model, allowing different groups, named strata, of the population to have distinct baseline intensity functions, whereas the regression parameter is shared by all the strata. In this stratified proportional intensity model, we are interested in the parameters estimation when the strata indicator is missing for some of the population individuals. Nonparametric maximum likelihood estimators are proposed for the model parameters and their consistency and asymptotic normality are established. We show the asymptotic efficiency of the regression parameter and obtain consistent estimators of its variance. The Expectation-Maximization algorithm is proposed and developed for the evaluation of the estimators of the model parameters.

Second, we are interested in the regression linear model when the response data is randomly right-censored. We introduce a new estimator of the regression parameter, which minimizes a Kaplan-Meier-weighted penalized least squares criterion. Results of consistency and asymptotic normality are obtained and a simulation study is conducted in order to investigate the small sample properties of this LASSO-type estimator. The bootstrap method is used for the estimation of the asymptotic variance.

Key words

Right-censored survival data, linear regression, stratified proportional intensity model, LASSO estimator, bridge estimator, nonparametric maximum likelihood estimator, Kaplan-Meier weights, bootstrap estimation, consistency, asymptotic normality, missing data, variance estimation.

Table des matières

Introduction

L'analyse de durée de vie est utilisée dans de nombreux domaines, comme la médecine, la fiabilité industrielle, l'économie ou encore la psychologie, et l'étude de données issues de ces secteurs se développe depuis plusieurs décennies. Il existe de nombreuses manières de modéliser des données de survie et ce travail s'intéresse à deux modélisations différentes dans le cas de données censurées, le **modèle de Cox** défini par sa fonction de risque

$$\lambda(t) = \lambda_0(t)e^{\beta' X},$$

où λ_0 est une fonction de risque de base non paramétrée, et le **modèle de régression linéaire** défini par une égalité liant la variable réponse Y aux covariables X

$$Y = \beta' X + \varepsilon,$$

ε étant une variable aléatoire d'erreur.

Dans le premier chapitre, nous introduisons et motivons la notion de durée de survie et présentons les outils de modélisation utilisés dans les études d'analyse de durée de vie.

Après avoir défini le cadre de l'étude, dans la première partie de cet exposé, nous nous intéressons au modèle de Cox censuré à droite. Nous commençons, dans le chapitre 2, par présenter ce modèle et sa généralisation au modèle de Cox stratifié, ainsi que les résultats existants pour l'estimation des paramètres du modèle. Nous expliquons pourquoi ces résultats ne peuvent être utilisés lorsque les strates sont partiellement manquantes. Dans le chapitre 3, nous proposons des estimateurs non paramétriques par maximum de vraisemblance pour le modèle de Cox stratifié censuré avec strates aléatoirement manquantes. Nous présentons les résultats asymptotiques obtenus pour ces estimateurs, et proposons une méthode d'estimation de la variance asymptotique du paramètre β. Passant aux aspects numériques, le chapitre 4

décrit l'algorithme Espérance-Maximisation et l'applique à l'estimation des paramètres proposés dans le chapitre 3.

Dans la deuxième partie, nous nous intéressons au modèle de régression linéaire censuré à droite. Le chapitre 5 permet d'introduire le modèle de durée de vie à temps accéléré issu du modèle de régression linéaire à travers une transformation logarithmique de la donnée réponse. Il introduit les résultats existants sur les estimateurs de type Kaplan-Meier et rappelle la théorie des moindres carrés contraints par une pénalité bridge dans le cas du modèle de régression linéaire. Dans le chapitre 6, nous proposons un estimateur de Kaplan-Meier pénalisé de type LASSO dans le modèle de régression linéaire censuré à droite. Nous étudions ses propriétés asymptotiques et présentons les résultats de simulations visant à illustrer les bonnes propriétés d'un tel estimateur à taille d'échantillon fixée. Enfin, le chapitre 7 s'intéresse à la méthode bootstrap permettant d'estimer la variance asymptotique de l'estimateur proposé dans le chapitre 6. Il décrit les mécanismes de cette méthode.

Chapitre 1

Modèles de durée de vie et processus

Sommaire

L'objectif de ce chapitre est de motiver et d'introduire la notion de durée de vie censurée ainsi que l'approche statistique au travers de laquelle nous allons étudier ce type de données. Dans ce premier chapitre, nous présentons les outils de modélisation utilisés en analyse de durées de vie et nous introduisons le problème de la censure qui affecte ce type de données.

1.1 Motivations

En analyse de survie, on s'intéresse à un groupe d'individus associés à un événement d'intérêt, souvent appelé *échec* ou *mort*, survenant après une durée appelée *durée de vie* ou *donnée de survie*. Cet événement d'intérêt intervient au plus une seule fois pour chaque individu. Des exemples classiques

sont la panne de composants électroniques en fiabilité industrielle, la fin d'une grève ou d'une période de chômage en économie, la résolution d'une tâche spécifique en expérimentation psychologique ou, en médecine, la rechute ou la mort d'un patient, ce qui a donné son nom au domaine. Dans la plupart des cas, l'événement d'intérêt symbolise la transition d'un état à un autre.

Pour déterminer précisément la survenue de l'échec, il est nécessaire de définir sans ambiguïté l'origine des temps et le terme d'échec, ainsi que choisir une échelle de temps. L'origine des temps n'est pas nécessairement la même pour tous les individus et doit être définie précisément pour chacun d'entre eux. Dans la plupart des cas, le temps 0 est choisi comme étant le moment d'une transition. Par exemple, lors de la mesure d'un âge, l'origine des temps est la naissance. Pour un essai de traitement médical, l'origine naturelle des temps est le début du traitement, mais en ce qui concerne l'évolution d'une maladie, la date de contamination n'étant en général pas connue, on peut considérer le moment du diagnostic comme origine. Cela peut paraître une alternative convenable même si cela entraînera des approximations par la suite. En ce qui concerne l'échelle des temps, le cas le plus fréquent est une mesure horaire, mais d'autres possibilités existent, comme le kilométrage d'un véhicule ou le temps cumulé d'utilisation d'un système.

Dans beaucoup de domaines d'application, on dispose, en plus de l'observation de durées de vie, d'informations supplémentaires suspectées d'influer sur les durées étudiées. Ces informations supplémentaires, appelées *covariables* ou *variables explicatives*, peuvent être différentes pour chaque individu. Cela peut être une caractéristique de l'individu (groupe sanguin, sexe, domaine professionnel, âge...) ou une observation dépendant de l'étude (posologie d'un traitement médical, type de greffe, durée d'hospitalisation...). Deux objectifs majeurs de l'analyse de survie sont l'évaluation de l'influence des covariables et la prédiction d'une durée de survie.

La suite de ce chapitre permet d'introduire les notations et outils classiques de modélisation utilisés en analyse de survie.

1.2 Modèles de durée de vie

L'analyse statistique des durées de vie étudie les lois d'instants d'occurence d'événements, à partir d'observations de durées et éventuellement de

variables explicatives, faites de manière discrète ou continue dans le temps.

Ainsi, nous désignons par T une variable aléatoire positive définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et représentant une durée jusqu'à un événement d'intérêt, l'origine des temps étant prédéfinie. Dans le domaine médical, cet événement peut être la mort, la guérison, la rechute d'un individu ; dans le domaine économique, la perte d'un emploi ; en fiabilité, l'instant de première panne. Par la suite, la durée T sera appelée *durée de vie*. Nous notons F_T sa fonction de répartition. La loi de T peut également être caractérisée par d'autres fonctions facilement interprétables en considérant T en terme de durée de vie.

Définition 1.1. *On appelle* fonction de survie S_T *la probabilité que la durée de vie T soit supérieure à un temps t :*

$$\forall t \in \mathbb{R}, S_T(t) = \mathbb{P}(T > t) = 1 - F_T(t).$$

Notons que si la loi de T admet une densité f_T par rapport à la mesure de Lebesgue,

$$\forall t \in \mathbb{R}, S_T(t) = \int_t^{+\infty} f_T(t)\mathrm{d}t \quad \text{et} \quad f_T(t) = -S'(t) \quad p.p.$$

Définition 1.2. *On appelle* fonction de risque instantané λ_T *la fonction définie pour t dans $\mathbb{R}^+$ par*

$$\lambda_T(t) = \begin{cases} \lim_{h\to 0} \frac{1}{h}\mathbb{P}(t < T \leq t + h | T > t) & \textit{si } t \textit{ est tel que } \mathbb{P}(T > t) > 0 \\ +\infty & \textit{sinon} \end{cases}$$

La fonction de risque peut avoir des formes très différentes mais est nécessairement positive sur $\mathbb{R}$.
Supposons maintenant que T soit une variable continue, on observe alors que

$$\forall t \in \mathbb{R}^+, \lambda_T(t) = \frac{f_T(t)}{S_T(t)} = -\frac{\partial}{\partial t} \ln(S_T(t)),$$

en posant $c/0 = +\infty$ pour tout $c > 0$. La définition de λ_T montre que pour h assez petit, $h\lambda_T(t)$ s'interprète comme la probabilité de survenue de l'événement d'intérêt dans l'intervalle $]t, t + h]$ sachant que cet événement ne s'est pas encore produit à l'instant t. Cette fonction traduit donc l'évolution dans le temps du risque de survenue de l'événement d'intérêt.

Définition 1.3. *On appelle* fonction de risque cumulé Λ_T *la fonction définie pour* t *dans* $\mathbb{R}^+$ *par*

$$\Lambda_T(t) = \int_0^t \lambda_T(s)\mathrm{d}s = -\ln(S(t)),$$

qui vaut $+\infty$ *quand* $S(t) = 0$.

Des définitions précédentes on déduit que pour tout $t \in \mathbb{R}^+$, on a la relation

$$f_T(t) = \lambda_T(t)\exp(-\Lambda_T(t)).$$

En conclusion, les cinq fonctions précédentes permettent de caractériser la loi de T et les unes sont déductibles des autres. Cependant, c'est l'interprétation de la fonction de risque instantané qui permettra le plus souvent de guider le choix d'un modèle pour des données de durée de vie. Nous illustrons les définitions précédentes avec des exemples de distributions de durée de vie parmi les plus utilisés.

Exemple 1.1. Dans une étude de survie, si nous pouvons supposer qu'il n'y a pas d'effet d'usure ou de vieillissement, la probabilité de survenue de l'événement d'intérêt, sachant qu'il n'est pas encore survenu, ne varie pas au cours du temps. Nous pouvons donc caractériser la loi de la durée T par une fonction de risque instantané constante :

$$\forall t \in \mathbb{R}^+, \lambda(t) = \lambda,$$

où λ est une constante strictement positive.
On obtient ainsi les fonctions définies ci-dessus pour $t \in \mathbb{R}^+$:

$$f_T(t) = \lambda e^{-\lambda t}, F_T(t) = 1 - e^{-\lambda t}, S_T(t) = e^{-\lambda t} \text{ et } \Lambda_T(t) = \lambda t.$$

La variable T suit donc une loi exponentielle de paramètre λ. Cette loi est connue pour son "absence de mémoire" car le temps d'attente jusqu'à l'occurence de l'événement d'intérêt ne dépend pas du passé de l'individu :

$$\mathbb{P}(T \geq x + t | T \geq x) = \mathbb{P}(T \geq t).$$

Si historiquement la loi exponentielle a été très étudiée en raison de ses nombreuses propriétés mathématiques, sa fonction de risque constante apparaît trop restrictive dans la majorité des applications médicales ou industrielles.

Exemple 1.2. La loi exponentielle fut généralisée par Weibull en 1939 à la loi du même nom en introduisant un nouveau paramètre, de manière à ce que la fonction de risque soit la suivante :

$$\forall t \in \mathbb{R}^+, \lambda(t) = \lambda\alpha t^{\alpha-1},$$

où λ et α sont deux constantes strictement positives.
Le paramètre λ est appelé paramètre d'échelle et α paramètre de forme. En effet, λ donne l'amplitude de la fonction de risque, et la position de α par rapport à 1 définit la monotonie de la fonction de risque : si $\alpha = 1$, on retrouve la fonction de risque constant et donc la loi exponentielle ; si $\alpha > 1$ (respectivement $\alpha < 1$), λ_T est croissante (respectivement décroissante) dans le temps et il y a donc phénomène d'usure, vieillissement (respectivement rajeunissement).
Grâce à l'expression de la fonction de risque, on obtient les expressions suivantes pour $t \in \mathbb{R}^+$:

$$f_T(t) = \lambda\alpha t^{\alpha-1}e^{-\lambda t^\alpha}, F_T(t) = 1 - e^{-\lambda t^\alpha}, S_T(t) = e^{-\lambda t^\alpha} \text{ et } \Lambda_T(t) = \lambda t^\alpha.$$

La loi de Weibull est très largement utilisée dans les domaines industriel (fiabilité) et biomédical (analyse de durée de vie). En effet, cette loi est apparue comme étant le choix de modèle le plus approprié dans la description de données portant sur la durée de vie de composants manufacturés ou l'apparition d'une tumeur chez l'animal. Son succès est également dû au fait que cette loi a un spectre assez large, couvrant à la fois le cas d'une fonction de risque croissante et celui d'une fonction de risque décroissante.

Prise en compte de covariables

Nous nous sommes intéressés jusqu'ici à la modélisation de données de survie d'une population homogène. Cependant, dans la plupart des domaines d'application, on constate que les individus ont des caractéristiques différentes observables qui peuvent influer sur la donnée de survie qui nous intéresse. Ces caractéristiques sont modélisables par des covariables qui donnent une information supplémentaire sur chaque individu et sont soit fixes dans le temps (sexe, catégorie socio-professionnelle, appartenance à une population à risque...) ou au contraire dépendantes du temps (mesure d'une quantité biologique...), ce qui est le cas lorsqu'on souhaite évaluer l'influence d'un phénomène individuel sur la durée précédant la survenue d'un événement. On se restreindra par la suite à des covariables, aussi appelées variables explicatives, fixes dans le temps.

Nous considérons ainsi une durée de vie aléatoire T et un vecteur de p variables explicatives réelles $X = (X_1, ..., X_p)'$ associée à la durée de survie T.
Il existe deux approches différentes pour modéliser l'influence des covariables sur la durée de survie T. La première est analogue à l'approche de la régression linéaire classique. Dans celle-ci, on modélise le logarithme népérien de la durée de survie T de manière à transformer la durée, positive, en une variable prenant l'ensemble de ses valeurs dans $\mathbb{R}$, et on suppose un modèle linéaire pour $\ln(T)$:

$$\ln(T) = m + \beta' X + \sigma\varepsilon,$$

où m est une constante réelle, β est un vecteur de coefficients de régression, $\sigma > 0$ et ε représente une erreur aléatoire. Cette modélisation a la propriété d'introduire un phénomène d'accélération ou décélération du temps selon un facteur $e^{-\beta' X}$ dans la fonction de survie de T : si $S_{T|0}$ désigne la fonction de survie de T quand les covariables sont fixées égales à 0 et $S_{T|X=x}$ celle de T quand les covariables sont fixées égales à x, alors

$$S_{T|X=x}(t) = S_{T|0}(te^{-\beta' x}).$$

Si cette approche permet une extension du modèle linéaire classique aux données de survie, son utilisation est restreinte par le choix de la distribution de l'erreur ε.

Ainsi, l'approche classique de modélisation de l'effet des covariables sur une donnée de survie est de modéliser la fonction de risque conditionnelle aux covariables comme une fonction de celles-ci. Deux classes de modèles généraux sont utilisés pour cette modélisation : la famille des modèles à risque multiplicatif et celle des modèles à risque additif. Les modèles à risque additif ne seront pas abordés ici, le lecteur pourra se référer à Buckley (1984), Lin et al. (1998), Martinussen & Scheike (2002), Klein & Moeschberger (1997).

Les modèles à risque multiplicatif sont définis à partir d'une fonction de risque conditionnelle au vecteur de covariables X s'écrivant comme le produit d'une fonction de risque dite *de base* λ_0 par une fonction positive des covariables $c(\beta' x)$:

$$\lambda_{T|X=x}(t) = \lambda_0(t)c(\beta' x).$$

La fonction de risque de base s'interprète comme le risque instantané d'occurence de l'événement d'intérêt pour un individu associé à des covariables telles que $X = 0$. On choisit généralement comme fonction de lien $c = \exp$.

EXEMPLE 1.3. Le modèle multiplicatif associé au modèle de Weibull est défini par la fonction de risque conditionnelle

$$\lambda_{T|X}(t) = \lambda\alpha t^{\alpha-1} e^{\beta' X},$$

où $\beta, X \in \mathbb{R}^p$. $\lambda\alpha t^{\alpha-1}$ correspond à la fonction de risque de base de l'individu. Le vecteur de paramètres à estimer est donc (α, λ, β).
Les fonctions conditionnelles de densité, répartition, survie et risque cumulé sont donc données respectivement, pour tout $t > 0$, par :

$$f_{T|X}(t) = \alpha\lambda t^{\alpha-1} e^{\beta' X} e^{-\lambda t^\alpha \exp(\beta' X)} \quad , \quad F_{T|X}(t) = 1 - e^{-\lambda t^\alpha \exp(\beta' X)},$$
$$S_{T|X}(t) = e^{-\lambda t^\alpha \exp(\beta' X)} \quad , \quad \Lambda_{T|X}(t) = \lambda t^\alpha e^{\beta' X}.$$

1.3 Censure aléatoire

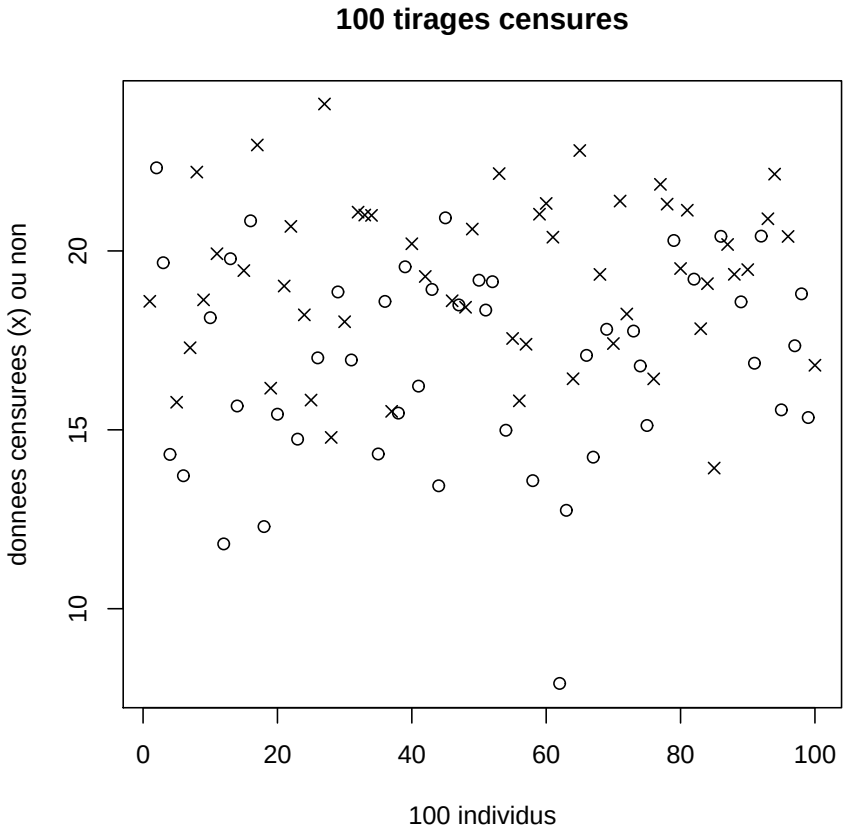
L'analyse des durées de vie pose des problèmes particuliers dûs au fait que les observations des durées de vie sont le plus souvent censurées. Par exemple, les données de survie de certains individus sont susceptibles de n'avoir pas été observées à la fin de l'étude, et donc la seule information dont nous disposons pour celles-ci est une borne inférieure. Cela correspond à un mécanisme de censure à droite de type I, type de censure que nous allons traiter dans la suite de ce travail. Le lecteur pourra se référer à Klein & Moeschberger (1997) pour des exposés sur la censure de type II, ou la censure de type I à gauche ou par intervalles.

Nous notons donc T^0 une durée de vie aléatoire. Introduisons une variable aléatoire C indépendante de T^0 à valeurs dans $\mathbb{R}^+ \cup \{+\infty\}$ dite *variable aléatoire de censure*. Dans le modèle de censure à droite, la durée de vie n'est observée que si elle est inférieure à la variable de censure ; sinon, on observe la valeur de la variable de censure. De plus, le caractère de la variable observée est connu, c'est-à-dire que l'on sait si la variable observée est la variable d'intérêt (durée de vie), ou la variable de censure. En résumé, dans le modèle de censure à droite, on observe

$$T = \min(T^0, C) \quad \text{et} \quad \Delta = \mathbb{1}_{T^0 \leq C}.$$

Le cas où C est une variable aléatoire dégénérée constante correspond au cas où l'observation des durées de vie s'arrête à la fin d'une durée fixée (l'étude a une durée limitée et fixe).

Voici par exemple sur la figure suivante un jeu de données de 100 individus auquel on peut s'intéresser dans l'étude de durées de survie censurées à droite. Les données censurées sont matérialisées par des croix, tandis que les données de survie réellement observées le sont par des ronds.



1.3.1 Vraisemblance dans un modèle de survie censuré

Soit T^0 une durée de vie aléatoire. On suppose que la loi P_{T^0} de T^0 appartient à une famille de lois de probabilité $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ où $\Theta \subseteq \mathbb{R}^p$. La vraie loi de T^0 est ainsi notée P_{θ_0}, où $\theta_0 \in \Theta$.

Notons $f_{T^0;\theta}(.)$, $F_{T^0;\theta}(.)$, $S_{T^0;\theta}(.)$, $\lambda_{T^0;\theta}(.)$, $\Lambda_{T^0;\theta}(.)$ les densité, fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de vie T, sous la loi P_θ.

La variable de censure C est supposée indépendante de la variable T et

sa loi est supposée ne pas dépendre du paramètre θ ; on dit que la loi de la censure C est *non informative*. On note $f_C(.)$, $F_C(.)$, $S_C(.)$ les densité, fonction de répartition et fonction de répartition de la variable C.

Les observations sont donc des réalisations de $T = \min(T^0, C)$ et de l'indicatrice de censure $\Delta = \mathbb{1}_{T^0 \leq C}$. Notons $(T_i, \Delta_i)_{i\in\{1,\dots,n\}}$ un échantillon des variables (T, C). L'estimation de θ_0 à partir des observations peut être effectuée par la méthode du maximum de vraisemblance.

La vraisemblance associée à l'échantillon $(T_i, \Delta_i)_{i\in\{1,\dots,n\}}$ s'écrit sous la forme

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \left(f_{T^0;\theta}(T_i) S_C(T_i)\right)^{\Delta_i} \left(S_{T^0;\theta}(T_i) f_C(T_i)\right)^{1-\Delta_i} \\ &= \prod_{i=1}^{n} \lambda_{T^0;\theta}(T_i)^{\Delta_i} S_{T^0;\theta}(T_i) S_C(T_i)^{\Delta_i} f_C(T_i)^{1-\Delta_i}, \end{aligned}$$

en utilisant les relations liant les densité, fonction de survie et fonction de risque instantané.

Sous l'hypothèse de censure non informative, on remarque qu'il est équivalent de chercher l'estimateur du maximum de vraisemblance de θ en maximisant l'expression

$$\prod_{i=1}^{n} \lambda_{T^0;\theta}(T_i)^{\Delta_i} S_{T^0;\theta}(T_i).$$

1.3.2 Vraisemblance dans un modèle de survie censuré avec covariables

On considère un modèle de prise en compte de covariables à risque multiplicatif. On suppose que la loi conditionnelle $P_{T^0|X}$ de T^0 sachant X appartient à une famille de lois de probabilité $\mathcal{P}_X = \{P_{\theta,X}; \theta \in \Theta\}$ où $\Theta \subseteq \mathbb{R}^p$. La vraie loi de T^0 sachant X est ainsi notée $P_{\theta_0,X}$, où $\theta_0 \in \Theta$.

Notons $f_{T^0|X;\theta}(.)$, $F_{T^0|X;\theta}(.)$, $S_{T^0|X;\theta}(.)$, $\lambda_{T^0|X;\theta}(.)$, $\Lambda_{T^0|X;\theta}(.)$ les densité, fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de vie T^0, sous la loi $P_{\theta,X}$.

On suppose que la loi de X, de densité f_X, ne dépend pas du paramètre θ. De la même façon que dans le cas où les covariables n'interviennent pas, on

considère que la loi de T^0 conditionnelle à X est indépendante de la loi de C conditionnelle à X et que la loi de C est non informative pour le paramètre θ.

Avec les mêmes notations que précédemment, la vraisemblance associée à l'échantillon $(T_i, \Delta_i, X_i)_{i\in\{1,\dots,n\}}$ s'écrit, grâce à la formule de Bayes, sous la forme

$$L_n(\theta) = \prod_{i=1}^{n} \left(f_{T^0|X;\theta}(T_i) S_{C|X}(T_i)\right)^{\Delta_i} \left(S_{T^0|X;\theta}(T_i) f_{C|X}(T_i)\right)^{1-\Delta_i} f_X(X_i).$$

Les lois de X et de C conditionnellement à X ne dépendant pas du paramètre θ, l'estimateur du maximum de vraisemblance de θ peut donc être obtenu en maximisant l'expression

$$\prod_{i=1}^{n} \lambda_{T^0|X;\theta}(T_i)^{\Delta_i} S_{T^0|X;\theta}(T_i).$$

1.4 Processus de comptage et processus empiriques

Les processus de comptage et la théorie des processus empiriques fournissent des méthodes adaptées pour étudier les données de survie censurées. Cette approche a été développée par Aalen (1975) et le lecteur peut se référer aux ouvrages Fleming & Harrington (1991) et Andersen & Gill (1982) pour une étude plus complète.

Nous commençons ici par rappeler des notions sur les processus qui seront nécessaires par la suite.

1.4.1 Rappels sur les processus

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé.

Définition 1.4. *Un* processus stochastique réel *est une famille de variables aléatoires réelles $X = (X(t))_{t\in\Gamma}$ indexé par un ensemble Γ définies sur le même espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.*

L'ensemble Γ indice en général le temps et vaut habituellement $\mathbb{N}$ (processus discrets) ou $\mathbb{R}^+$ (processus continus). On s'intéressera par la suite à des processus continus.

Définition 1.5. *Pour un processus stochastique, les fonctions définies pour* $\omega \in \Omega$ *par* $X(.,\omega) : \mathbb{R}^+ \to \mathbb{R}$ *sont appelées* trajectoires de X.
Un processus sera dit continu à droite, à variation bornée, croissant, ayant des limites à droite *si l'ensemble des trajectoires ayant la propriété correspondante est de probabilité* 1.

Un processus X est donc une application de $\mathbb{R}^+ \times \Omega$ dans $\mathbb{R}$, et $X(t,\omega)$ désigne la valeur de la variable aléatoire $X(t)$ en la réalisation ω.

Définition 1.6. $\mathcal{B}(\mathbb{R}^+)$ *désigne la tribu des boréliens sur* $\mathbb{R}^+$.
Le processus X *est dit* mesurable *si* $X : (\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+)) \times (\Omega, \mathcal{F}) \to \mathbb{R}$ *est une application mesurable.*

Définition 1.7. *Un processus stochastique est dit :*
- intégrable *si* $\sup_{t\in\mathbb{R}^+} \mathbb{E}|X(t)| < +\infty$,
- de carré intégrable *si* $\sup_{t\in\mathbb{R}^+} \mathbb{E}\left(X(t)\right)^2 < +\infty$,
- borné *s'il existe une constante* $M \in \mathbb{R}^+$ *telle que*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}^+} |X(t)| \leq M\right) = 1.$$

Les définitions suivantes vont permettre d'établir une formulation rigoureuse du concept d'information s'accroissant avec le temps.

Définition 1.8.
- *Une* filtration $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *est une famille de sous-tribus de la tribu* $\mathcal{F}$ *croissante, c'est-à-dire vérifiant pour tout* $s \leq t$, $\mathcal{F}_t \subseteq \mathcal{F}_t$.
- *Si* $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *est une filtration,* $\cap_{h>0}\mathcal{F}_{t+h}$ *est une tribu et on la note* $\mathcal{F}_{t+}$. *De la même façon,* $\mathcal{F}_{t-}$ *est la tribu engendrée par* $\cup_{h>0}\mathcal{F}_{t-h}$.
- *Une filtration* $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *est dite* continue à droite *si pour tout* $t \in \mathbb{R}^+$, $\mathcal{F}_{t+} = \mathcal{F}_t$.

Les filtrations les plus naturelles sont les *histoires* de processus stochastiques, c'est-à-dire les familles $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ telles que $\mathcal{F}_t = \sigma\left(X(s); 0 \leq s \leq t\right)$ est la plus petite tribu rendant pour tout $0 \leq s \leq t$ les variables $X(s)$ mesurables. Dans ce cas, on voit que $\mathcal{F}_t$ *contient l'information* engendrée par le processus X sur $[0,t]$.

Définition 1.9. *Le processus stochastique* $(X_t)_{t\in\mathbb{R}^+}$ *est dit* adapté *à la filtration* $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *si pour tout* $t \in \mathbb{R}^+$, $X(t)$ *est* $\mathcal{F}_t$*-mesurable.*

Bien entendu, tout processus est adapté à sa filtration *histoire*.

Les processus de comptage et leurs propriétés seront particulièrement utiles dans la suite de ce travail.

Définition 1.10. *Un* processus de comptage *est un processus stochastique* $(N_t)_{t\in\mathbb{R}^+}$ *adapté à une filtration* $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *tel que* $N(0) = 0$, $N(t) < +\infty$ *p.s. et dont les trajectoires sont avec probabilité* 1 *continues à droite, constantes par morceaux avec des sauts de taille* 1.

Dans la plupart des applications, comme le terme "processus de comptage" le suggère, $N(t) - N(s)$ représentera le nombre d'événements intervenant dans l'intervalle $]s, t]$. Le processus de Poisson est un des exemples les plus classiques.

Exemple 1.4. Soient T^0 et C deux variables aléatoires positives indépendantes de lois continues. Notons $T = \min(T^0, C)$ l'observation censurée du temps de survie T^0 et $\Delta = \mathbb{1}_{T\leq C}$. Alors le processus défini pour $t \geq 0$ par

$$N(t) = \mathbb{1}_{T\leq t,\Delta=1} = \Delta\mathbb{1}_{T\leq t}$$

est un processus de comptage appelé *processus de comptage des échecs*. De la même façon, le processus défini pour $t \geq 0$ par

$$Y(t) = \mathbb{1}_{T\geq t}$$

est un processus stochastique appelé *processus à risque*.

Nous définissons maintenant l'intégrale de Lebesgue-Stieljes qui nous sera utile lorsque nous intégrerons contre un processus de comptage. Elle se base sur la bijection entre les fonctions croissantes continues à droite et la classe des mesures boréliennes sur $\mathbb{R}$.

Théorème 1.1.
- *Soit* μ *une mesure borélienne sur* $\mathbb{R}$ *et soit* G *une fonction, définie sur* $\mathbb{R}$ *à une constante additive près, par* $G(b)-G(a) = \mu(]a, b])$. *Alors* G *est continue à droite et croissante.*

- *Soit* $G : \mathbb{R} \to \mathbb{R}$ *une fonction croissante continue à droite et posons, pour tout intervalle* $]a, b]$ *de* $\mathbb{R}$ $(a < b)$, $\mu(]a, b]) = G(b) - G(a)$. *Alors il existe une unique extension de* μ *à une mesure borélienne sur* $\mathbb{R}$.

On a alors les relations suivantes, faciles à établir, entre la fonction croissante continue à droite G et sa mesure borélienne associée μ. La notation $G(t-) = \lim_{x \uparrow t} G(x)$ est utilisée.

Proposition 1.2.
- $\mu(]a,b[) = G(b-) - G(a)$,
- $\mu([a,b]) = G(b) - G(a-)$,
- $\mu([a,b[) = G(b-) - G(a-)$,
- $\mu(\{a\}) = G(a) - G(a-)$,
- *G est continue en a si et seulement si* $\mu(\{a\}) = 0$.

La définition d'intégrale de Lebesgue-Stieljes est basée sur la notion plus générale d'intégrale de Lebesgue grâce à la correspondance établie précédemment.

Définition 1.11. *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction borélienne, $G : \mathbb{R} \to \mathbb{R}$ une fonction croissante continue à droite, et μ la mesure borélienne relative à G. Pour un ensemble borélien $A \subseteq \mathbb{R}$, on définit l'*intégrale de Lebesgue-Stieljes $\int_A f \mathrm{d}G$ *par* $\int_A f \mathrm{d}\mu$.

Cette définition assez abstraite permet d'établir des formules plus explicites quand la fonction G a certaines propriétés. Notamment, quand G est une fonction en escalier, elle comporte un nombre au plus dénombrable de sauts notés ici $\{x_1, x_2, ...\}$ tels que $\Delta G(x_n) = G(x_n) - G(x_n-) > 0$. La mesure μ associée sera alors discrète et strictement positive aux points $x_1, x_2, ...$, d'où la formule, pour A un borélien de $\mathbb{R}$

$$\int_A f \mathrm{d}G = \sum_{n; x_n \in A} f(x_n) \Delta G(x_n).$$

La notation sur un intervalle de l'intégrale $\int_s^t f \mathrm{d}G$ pouvant mener à des ambiguïtés si s ou t est un point de discontinuité de G, nous utiliserons la convention

$$\int_s^t f \mathrm{d}G = \int_{]s,t]} f \mathrm{d}G.$$

Dans cette configuration, on note que l'intégrale de Lebesgue-Stieljes permet d'utiliser une notation simple pour une somme de termes dénombrables.

Si la fonction G a une dérivée g en chaque point de l'intervalle $]s,t]$, alors $\mu(]s,t]) = \int_s^t g(x) \mathrm{d}x$ et μ est absolument continue par rapport à la mesure de Lebesgue. On a alors

$$\int_s^t f(x) \mathrm{d}G(x) = \int_s^t f(x) g(x) \mathrm{d}x.$$

1.4.2 Rappels de théorie des processus empiriques

Nous rappelons ici quelques éléments de théorie des processus empiriques dont nous aurons besoin par la suite. Le lecteur peut se référer à Shorack & Wellner (1986), van der Vaart & Wellner (1996) et van der Vaart (1998) pour une étude plus complète.

Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires à valeurs dans un espace mesurable $(\mathcal{X}, \mathcal{A})$ de même loi que X notée P_X.

Définition 1.12. *La* mesure empirique $\mathbb{P}_n$ *associée à* (X_n) *est la mesure définie sur les boréliens par*

$$\mathbb{P}_n(B) = \frac{1}{n} card\left\{i \in \{1, ..., n\}; X_i \in B\right\}.$$

On la note $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$*, où* δ_a *désigne la mesure de Dirac au point* a*, c'est la mesure aléatoire qui met un poids* $1/n$ *à chaque observation* X_i*.*

Soit $\mathcal{S}$ un ensemble de fonctions mesurables $f : \mathcal{X} \to \mathbb{R}$ et P_X-intégrables, alors la mesure empirique permet de définir une application de $\mathcal{S}$ dans $\mathbb{R}$ donnée par $f \mapsto \mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$. On utilise la notation $Qf := \int f \mathrm{d}Q$ pour une fonction mesurable f et une mesure Q. Notons que si X est à valeurs réelles, pour $f_x = \mathbb{1}_{]-\infty,x]}$, $Qf_x = Q(]-\infty, x])$, et donc on retrouve pour $Q = P_X$ la fonction de répartition de X. On appelle *fonction de répartition empirique* $\mathbb{F}_n$ la fonction de répartition (aléatoire) associée à la mesure empirique, définie par $\mathbb{F}_n(x) = \mathbb{P}_n f_x = n^{-1}\sum_{i=1}^n \mathbb{1}_{X_i \leq x}$.

Définition 1.13. *Le* processus empirique $\mathbb{G}_n$ *associé à* (X_n) *et* $\mathcal{S}$ *est l'application*

$$f \mapsto \mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n - P_X)(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(f(X_i) - \mathbb{E}(f(X))\right).$$

On l'identifiera fréquemment à la mesure $\mathbb{G}_n = n^{-1/2}\sum_{i=1}^n(\delta_{Xi} - P_X)$*.*

Pour une fonction donnée f, sous l'hypothèse d'existence de $P_X f$ et l'hypothèse $P_X(f^2) < +\infty$, la loi forte des grands nombres et le théorème central limite donnent les convergences

$$\mathbb{P}_n(f) \xrightarrow{p.s.} P_X f \quad \text{et} \quad \mathbb{G}_n(f) \xrightarrow{\mathcal{F}} \mathcal{N}\left(0, P_X(f - P_X f)^2\right).$$

La théorie des processus empiriques s'intéresse au cas plus général de la convergence uniforme de ces processus sur des classes de fonctions. Les

théorèmes de Glivenko-Cantelli et Donsker donnent une première extension uniforme sur la classe de fonctions $\mathcal{I} = \{\mathbb{1}_{]-\infty,x]}, x \in \mathbb{R}\}$ pour les variables aléatoires à valeurs réelles indépendantes.

Théorème 1.3 (Glivenko-Cantelli). *Soit (X_n) une suite de variables aléatoires réelles indépendantes de même fonction de répartition F_X et de fonction de répartition empirique $\mathbb{F}_n$, alors*

$$\sup_{\mathbb{R}} |\mathbb{F}_n - F| \xrightarrow[n\to+\infty]{p.s.} 0.$$

On peut réécrire la conclusion de ce théorème sous la forme $\sup_{f\in\mathcal{I}} |\mathbb{P}_n f - P_f| \xrightarrow[n\to+\infty]{p.s.} 0$. On note $\|Qf\|_{\mathcal{S}} = \sup_{f\in\mathcal{S}} |Qf|$. Cela permet d'introduire la définition suivante.

Définition 1.14. *Une classe $\mathcal{S}$ de fonctions $f : \mathcal{X} \to \mathbb{R}$ est appelée P_X-*classe de Glivenko-Cantelli *si elle vérifie*

$$\|\mathbb{P}_n f - Pf\|_{\mathcal{S}} \xrightarrow[n\to+\infty]{p.s.} 0.$$

La classe $\mathcal{I}$ est donc un premier exemple de P_X-classe de Glivenko-Cantelli. Supposons maintenant, afin d'étendre le théorème central limite à une version uniforme, ou fonctionnelle, que

$$\text{pour tout } x \in \mathcal{X}, \quad \sup_{f\in\mathcal{F}} |f(x) - P_X f| < +\infty.$$

Sous cette condition, le processus empirique $\mathbb{G}_n$ peut être vu comme un élément de l'espace $\ell^\infty(\mathcal{F})$ des fonctions bornées de $\mathcal{F}$ dans $\mathbb{R}$.

Définition 1.15. *Une classe $\mathcal{S}$ de fonctions $f : \mathcal{X} \to \mathbb{R}$ est appelée P_X-*classe de Donsker *si elle vérifie*

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathbb{G} \quad \textit{dans } \ell^\infty(\mathcal{F}),$$

où la limite $\mathbb{G}$ est un processus gaussien tendu centré de fonction de covariance $cov(\mathbb{G}f_1, \mathbb{G}f_2) = P_X f_1 f_2 - P_X f_1 . P_X f_2$.

On remarque que le lemme de Slutsky permet de montrer que toute classe de Donsker est une classe de Glivenko-Cantelli. Réciproquement, toute classe de Glivenko-Cantelli n'est pas une classe de Donsker, mais beaucoup d'exemples peuvent se trouver parmi les classes de Glivenko-Cantelli.

Un premier exemple de classe de Donsker est obtenu grâce au théorème du même nom.

Théorème 1.4. *Soit (X_n) une suite de variables aléatoires réelles indépendantes de même fonction de répartition F_X et de fonction de répartition empirique $\mathbb{F}_n$, alors la suite des processus empiriques $\sqrt{n}(\mathbb{F}_n - F_X)$ converge en loi dans l'espace des fonctions continues à droite avec limite à gauche sur $\bar{\mathbb{R}}$ vers un processus gaussien $\mathbb{G}_{F_X}$ tendu centré de fonction de covariance au point (s,t) égale à $F_X(s \wedge t) - F_X(s)F_X(t)$.*

Ainsi, si $\mathcal{X} = \mathbb{R}$, l'ensemble $\mathcal{I} = \{\mathbb{1}_{]-\infty,x]}, x \in \mathbb{R}\}$ est une P_X-classe de Donsker. van der Vaart & Wellner (1996) et van der Vaart (1998) donnent d'autres exemples de classes de Donsker. L'ensemble des fonctions de variation uniformément bornée forme une classe de Donsker.
Soit une classe paramétrique de fonctions $\{f_t, t \in T\}$, où T est un ensemble borné de $\mathbb{R}^d$. S' il existe m fonction mesurable telle que $m(X)$ admet un moment d'ordre r et pour tout s,t, $|f_s(x) - f_t(x)| \leq m(x)\|s-t\|$, alors cet ensemble est une P_X-classe de Donsker. Les classes de Sobolev sont également des classes de Donsker.

D'autre part, certaines opérations sur les classes de fonctions permettent de préserver la propriété de Donsker. Tout d'abord, si $\mathcal{S}$ est une classe de Donsker, alors tout sous-ensemble de $\mathcal{S}$, l'adhérence et l'enveloppe convexe symétrique de $\mathcal{S}$ sont des classes de Donsker. Si $\mathcal{S}$ et $\mathcal{T}$sont des classes de Donsker telles que $\|P_X\|_{\mathcal{S}\cup\mathcal{T}} < \infty$ alors $\mathcal{S} \wedge \mathcal{T} = \{f \wedge g; f \in \mathcal{S}, g \in \mathcal{T}\}$,$\mathcal{S} \vee \mathcal{T}$ (défini de la même façon), $\mathcal{S}+\mathcal{T}$ et $\mathcal{S}\cup\mathcal{T}$ sont également des classes de Donsker. Si $\mathcal{S}$ et $\mathcal{T}$ sont des classes de Donsker uniformément bornées alors $\mathcal{S}.\mathcal{T}$ est encore une classes de Donsker. Si $\mathcal{S}$ est Donsker et vérifie $\|P_X\|_{\mathcal{S}} < \infty$ et l'existence de $\delta > 0$ tel que pour tout $f \in \mathcal{S}$, $f \geq \delta$ alors $\{1/f; f \in \mathcal{F}\}$ est une classe de Donsker. Le lecteur intéressé par des théorèmes plus généraux et d'autres exemples pourra se référer à van der Vaart & Wellner (1996).

1.4.3 Rappels sur les martingales à temps continu

Certaines méthodes utilisant la théorie des martingales permettent l'étude des propriétés des estimateurs dans le cadre de données de vie censurées. Nous rappelons donc la définition de martingale à temps continu et donnons une application classique aux données de survie. Le lecteur intéressé par une approche plus complète pourra se référer à Fleming & Harrington (1991) ou Bickel et al. (1993) par exemple.

Soit $X = (X(t))_{t\geq 0}$ un processus stochastique continu à droite avec limites à gauche et $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ une filtration.

Définition 1.16. *X est appelé* martingale adaptée à la filtration $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$ *ou* $(\mathcal{F}_t)$-martingale *si*
- *X est adapté à $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$,*
- *pour tout $t \in \mathbb{R}^+$, $\mathbb{E}|X(t)| < +\infty$,*
- *pour tous $s, t \in \mathbb{R}^+$, $\mathbb{E}\left(X(t+s)|\mathcal{F}_t\right) = X(t)$ ps.*

Si la dernière condition est remplacée par $\mathbb{E}\left(X(t+s)|\mathcal{F}_t\right) \geq X(t)$ ps, X est appelé sous-martingale.

Si la dernière condition est remplacée par $\mathbb{E}\left(X(t+s)|\mathcal{F}_t\right) \leq X(t)$ ps, X est appelé sur-martingale.

Proposition 1.5. *Soit X une martingale adaptée à $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$. Alors $\mathbb{E}\left(X(t)|\mathcal{F}_{t-}\right) = X(t-)$ ps.*

Exemple 1.5. Reprenons l'exemple 1.4. Introduisons la filtration définie par

$$\mathcal{F}_t = \sigma\{N(s), (1-\Delta)\mathbb{1}_{T^0\leq s}; 0 \leq s \leq t\}.$$

Notons λ la fonction de risque instantanée de T^0 et définissons le processus M sur $\mathbb{R}^+$ par

$$M(t) = \Delta\mathbb{1}_{T^0\leq t} - \int_0^t \mathbb{1}_{T^0\geq u}\lambda(u)\mathrm{d}u.$$

Alors M est une martingale adaptée à la filtration $(\mathcal{F}_t)$.
Ce résultat est également valable sous des hypothèses plus faibles que l'indépendance de T^0 et la censure C mais il sera suffisant pour les configurations auxquelles nous nous intéressons.

Première partie

Estimation du maximum de vraisemblance dans le modèle de Cox censuré à droite avec strates aléatoirement manquantes

Cette partie présente le travail effectué dans le cadre du modèle de Cox stratifié censuré à droite avec strates aléatoirement manquantes.

Nous commençons, dans le chapitre 2, par définir le modèle de Cox et ses généralisations, le modèle de Cox stratifié et le modèle de Cox avec strates aléatoirement manquantes. Dans le cadre de la censure à droite, nous présentons les méthodes existantes pour l'estimation du paramètre de régression et de la fonction de risque cumulé de base et expliquons les limites de ces méthodes dans le cas de strates manquantes. Nous introduisons une nouvelle méthode proposée pour remédier à cette situation.

Dans le chapitre 3, nous développons la méthode d'estimation introduite dans le chapitre 2 dans le cadre du modèle de Cox stratifié censuré à droite partiellement observé. Dans ce contexte de données manquantes, nous proposons des estimateurs du maximum de vraisemblance non paramétriques pour le paramètre de régression et des fonctions d'intensité cumulée de ce modèle. Nous étudions ensuite les propriétés asymptotiques des estimateurs obtenus.

Enfin, le chapitre 4 rappelle la démarche de l'algorithme Espérance-Maximisation et présente son application au calcul numérique des estimateurs introduits dans le chapitre 3.

Chapitre 2

Le modèle de Cox et ses généralisations

Sommaire

Ce chapitre introduit le modèle de Cox et une de ses généralisations que nous allons étudier plus en détail, le modèle de Cox stratifié avec strates aléatoirement manquantes. Le modèle de régression de Cox est un des modèles les plus utilisés pour la modélisation de l'influence de covariables sur des données de survie et il a été étudié dans de nombreux ouvrages (Cox & Oakes, 1984; Fleming & Harrington, 1991; Therneau & Grambsch, 2000; Martinussen & Scheike, 2002). Il définit une famille de lois conditionnelles de la donnée de survie, sachant un vecteur de variables explicatives. Nous commençons par introduire le modèle de Cox puis nous étendrons celui-ci au modèle de Cox stratifié et au même modèle avec strates partiellement observées.

2.1 Le modèle semiparamétrique de Cox

Nous commençons par définir les notations utilisées dans ce chapitre, donner la formulation du modèle de régression de Cox ainsi que la méthode du maximum de vraisemblance partielle permettant d'estimer le paramètre d'intérêt du modèle. Nous introduisons ensuite l'estimateur de Breslow de la fonction de risque. Enfin, nous donnons les propriétés asymptotiques des estimateurs proposés.

2.1.1 Présentation du modèle

Le modèle de régression de Cox fait partie de la famille des modèles à risques multiplicatifs définis dans la section 1.2 et a été introduit par Cox (1972). C'est un *modèle semiparamétrique* qui permet de ne spécifier que la modélisation relative à l'influence des covariables via un vecteur de régression et d'éviter ainsi le choix parfois difficile d'un modèle totalement paramétrisé. Il a été décrit originellement par la formulation de la fonction de risque instantané de la donnée de survie. Notons T^0 la variable aléatoire durée de vie du modèle et $\lambda_{T^0|X}$ sa fonction de risque instantanée connaissant X, un p-vecteur de covariables. Alors le modèle de régression de Cox est défini par la relation suivante pour tout $t \in \mathbb{R}^+$:

$$\lambda_{T^0|X}(t) = \lambda_0(t) e^{\beta_0' X},$$

où $\beta_0 \in \mathbb{R}^p$ est un paramètre de régression vectoriel et λ_0 une fonction définie sur $\mathbb{R}^+$ à valeurs positives appelée *fonction de risque instantané de base*.

Remarquons que le modèle de Cox peut également s'écrire dans le cadre de covariables dépendantes du temps t mais ce mémoire s'intéresse au cas de covariables fixes dans le temps. Dans cette configuration, le modèle de Cox appartient à la famille des *modèles à risques proportionnels*. En effet, notons X_1 et X_2 deux vecteurs de covariables indépendantes du temps, alors le rapport des risques instantanés est constant par rapport à t :

$$\frac{\lambda_{T^0|X_1}(t)}{\lambda_{T^0|X_2}(t)} = \frac{\lambda_0(t)\exp(\beta_0' X_1)}{\lambda_0(t)\exp(\beta_0' X_2)} = e^{\beta'(X_1 - X_2)}.$$

Remarquons que $e^{\beta_{0i}}$ (où β_{0i} est la i-ème coordonnée du vecteur β_0) peut s'interpréter comme le rapport des risques instantanés de deux individus pour lesquels $X_{1,i} = 1$, $X_{2,i} = 0$ et les autres composantes des deux vecteurs de covariables sont égales.

La fonction de risque instantané de base peut, quant à elle, s'interpréter comme le risque instantané de l'individu en l'absence d'influence des covariables. Ainsi, dans cette configuration, on voit que le modèle de Cox comprend différents modèles paramétriques usuels, comme le modèle exponentiel et le modèle de Weibull (obtenus avec les fonctions de risque instantané respectives $\lambda_0 \equiv \lambda$ et $\lambda_0(t) = \lambda\alpha t^{\alpha-1}$).

Définissons la *fonction de risque cumulé* conditionnelle aux covariables X pour $t \in \mathbb{R}^+$ par $\Lambda_{T^0|X}(t) = \int_0^t \lambda_{T^0|X}(s)\mathrm{d}s$ et notons qu'elle peut s'écrire dans le modèle de Cox

$$\Lambda_{T^0|X}(t) = \exp(\beta_0' X)\Lambda_0(t),$$

où $\Lambda_0(t) = \int_0^t \lambda_0$ est la *fonction de risque cumulé de base*. Ainsi, la fonction densité de la variable T^0 conditionnellement aux covariables X s'écrit

$$f_{T^0|X}(t) = \lambda_0(t)e^{\beta_0' X} \exp\left(-e^{\beta_0' X}\Lambda_0(t)\right).$$

On introduit désormais une variable de censure C à valeurs positives indépendante de la durée de vie T^0 conditionnellement à la variable explicative X. Notons $f_{C|X}$ la fonction densité et $S_{C|X}$ la fonction de survie de la variable censure conditionnelles à X. Nous nous intéressons au cas de la censure à droite : on n'observe pas T^0 mais

$$T = \min(T^0, C) \quad \text{et} \quad \Delta = \mathbb{1}_{T^0 \leq C}.$$

Le problème statistique consiste maintenant à estimer les paramètres inconnus du modèle à partir de l'observation d'un n-échantillon $(T_i, \Delta_i, X_i)_{1\leq i\leq n}$ du triplet (T, Δ, X). Le paramètre d'intérêt du modèle est le vecteur β permettant d'expliquer la nature de l'influence des covariables, tandis que le paramètre fonctionnel Λ_0 est considéré comme un paramètre de nuisance. Le problème d'estimation est développé dans la partie suivante.

2.1.2 Méthode du maximum de vraisemblance partielle

L'inférence dans le modèle à risques proportionnels nécessite des méthodes algorithmiques pour la recherche d'estimateurs des coefficients de régression et de la fonction de risque de base et l'approximation de leurs lois asymptotiques. Cox (1972) a proposé de baser l'inférence statistique sur une approche

utilisant la vraisemblance : une interprétation heuristique de la fonction de risque conditionnelle est utilisée pour construire une vraisemblance partielle, puis les techniques de vraisemblance classiques sont appliquées à la vraisemblance partielle pour l'obtention d'estimateurs.

Nous commençons par rappeler le principe de la méthode du maximum de vraisemblance partielle. Elle a été proposée de manière générale par Cox (1975) pour des modèles contenant des paramètres de nuisance de grande dimension. Dans le modèle de Cox à risques proportionnels, la fonction de risque cumulée de base Λ_0 est un paramètre de dimension infinie, de nuisance dans l'étude de l'influence des covariables sur le risque de l'individu.

Principe de la méthode du maximum de vraisemblance partielle

Soit X un vecteur aléatoire de densité $f_X(x, \theta)$, où θ est un vecteur paramètre (ϕ, β). Supposons que l'on s'intéresse à l'estimation de β, ϕ étant simplement un paramètre de nuisance. Dans certains problèmes, il suffira de maximiser la vraisemblance en $\theta = (\phi, \beta)$ conjointement et d'utiliser la partie appropriée de la matrice de covariance complète de l'estimateur du maximum de vraisemblance θ pour l'inférence sur β. Cependant, quand le paramètre ϕ est fonctionnel, ou quand la vraisemblance relative à θ est complexe, la maximisation jointe de la vraisemblance peut s'avérer difficile.

Dans certains modèles, X peut être décomposé en deux composantes V et W et la densité de X peut s'écrire comme le produit d'une marginale et d'une densité conditionnelle :

$$f_{X;\theta}(x) = f_{W|V;\theta}(w|v) f_{V;\theta}(v), \tag{2.1}$$

où $x' = (v', w')$. Même dans certains modèles compliqués, un des facteurs du membre de droite de (2.1) peut ne pas contenir ϕ et être utilisé directement pour l'inférence sur β. L'autre facteur dépendant dans la majorité des cas de ϕ et β à la fois, de l'information sera perdue lors de l'utilisation d'une seule partie de la vraisemblance, mais le gain en simplicité est susceptible de compenser une certaine perte d'efficacité. On évitera aussi les erreurs dues à l'utilisation éventuelle des méthodes standards du maximum de vraisemblance sur un paramètre de dimension infinie. Un exemple classique de décomposition telle que (2.1) est l'écriture d'un vecteur d'observations comme un vecteur de statistiques d'ordre et les rangs des observations originales.

L'inférence basée sur la méthode du maximum de vraisemblance partielle s'appuie sur cette idée de décomposition. Supposons que le vecteur d'observations puisse s'écrire comme une suite de paires $(V_1, W_1, V_2, W_2, ..., V_L, W_L)$. La vraisemblance relative à θ peut alors s'écrire

$$\begin{aligned} f_{X;\theta}(x) &= f_{V_1,W_1,...,V_L,W_L;\theta}(v_1, w_1, ..., v_L, w_L) \\ &= \prod_{i=1}^{L} f_{W_l|V_1,W_1,...,V_l;\theta}(w_l|v_1, w_1, ..., v_l) \\ &\quad \times f_{V_l|V_1,W_1,...,V_{l-1},W_{l-1};\theta}(w_l|v_1, w_1, ..., v_{l-1}, w_{l-1}) \\ &= \left(\prod_{l=1}^{L} f_{W_l|Q_l;\theta}(w_l|q_l)\right)\left(\prod_{l=1}^{L} f_{V_l|P_l;\theta}(w_l|p_l)\right), \end{aligned} \tag{2.2}$$

où $P_1 = \emptyset$, $Q_1 = V_1$ et pour $l \in \{2, ..., L\}$,

$$P_l = (V_1, W_1, ..., V_{l-1}, W_{l-1}) \quad \text{et} \quad Q_l = (V_1, W_1, ..., W_{l-1}, V_l).$$

Quand le premier terme du produit de (2.2) ne dépend que de β, il est appelé vraisemblance partielle pour β basée sur W.

Un certain nombre de questions se posent quant à l'utilisation de la vraisemblance partielle à la place de la vraisemblance totale en terme d'efficacité. Ce problème a été étudié dans le cas général par Wong (1986), et son application au modèle semiparamétrique de Cox par Tsiatis (1981) et Andersen & Gill (1982). Cette application est explicitée ici.

Nous considérons les triplets d'observations $(T_i, \Delta_i, X_i)_{1\leq i\leq n}$, répétitions indépendantes de (T, Δ, X), où T est la durée de vie éventuellement censurée issue d'un modèle de Cox décrit dans 2.1.1, Δ l'indicateur de censure, X le vecteur de covariables. La vraisemblance pour l'estimation de $\theta_0 = (\beta_0, \Lambda_0)$ relative au n-échantillon (T_i, Δ_i, X_i) est, d'après la section 1.3.2, égale à

$$\prod_{i=1}^{n} \lambda(T_i)^{\Delta_i} e^{\Delta_i \beta' X_i} \exp\left(-e^{\beta' X_i}\Lambda(T_i)\right) S_{C|X}(T_i)^{\Delta_i} f_{C|X}(T_i)^{1-\Delta_i} f_X(X_i).$$

Sous l'hypothèse supplémentaire que la censure est non informative et que la loi de X ne dépend pas du paramètre θ, on note que la vraisemblance est proportionnelle à

$$L_n(\theta) = \prod_{i=1}^{n} \lambda(T_i)^{\Delta_i} e^{\Delta_i \beta' X_i} \exp\left(-e^{\beta' X_i}\Lambda(T_i)\right). \tag{2.3}$$

L'estimation du paramètre fonctionnel Λ_0 pose problème : en effet, il n'existe pas de maximum de L_n quand Λ varie dans l'ensemble des fonctions positives strictement croissantes, s'annulant en 0 et continûment différentiables de dérivée λ_0 sur $\mathbb{R}^+$. Pour voir ceci, il suffit de considérer des fonctions Λ avec des valeurs fixes en les T_i et dont les dérivées $\lambda(T_i)$ en T_i croissent vers l'infini. Cox a donc proposé d'estimer le paramètre fini-dimensionnel β_0 à partir d'une vraisemblance partielle obtenue grâce au principe décrit précédemment.

Estimation du paramètre de régression β_0 par la méthode du maximum de vraisemblance partielle

Notons L le nombre d'observations non censurées notées $T_{(1)} < ... < T_{(L)}$ ainsi réordonnées et $(X_{(1)}, ..., X_{(L)})$ les covariables associées (on pose $T_{(0)} = 0$ et $T_{(L+1)} = +\infty$). Soit m_l le nombre de censures intervenant dans l'intervalle $[T_{(l)}, T_{(l+1)}[$. Associons à ces censures les instants d'occurrence que l'on notera $T_{(l,1)}$,...,$T_{(l,m_l)}$ et leurs covariables associées $X_{(l,1)}$,...,$X_{(l,m_l)}$. On utilise la construction donnée par la formule (2.2) dans laquelle on conditionne par les variables explicatives (X_i) et on prend pour $i \in \{0, ..., L\}$

$$V_{i+1} = \{T_{i+1}, T_{(i,j)}, (i,j); 1 \leq j \leq m_i\} \quad \text{et} \quad W_{i+1} = \{(i+1)\}.$$

Ainsi, dans la suite (P_i, Q_i) définie auparavant, P_i contient toutes les instants de censures et d'échec jusqu'à $T_{(i-1)}$ et les indices correspondants à ces instants. Q_i contient les mêmes informations additionnées de l'instant d'échec $T_{(i)}$, des instants de censure le précédant et des indices de ces censures. Cox propose d'ignorer, dans le cadre de l'hypothèse de censure non informative et de l'hypothèse sur la loi de X, le terme $\prod_{i=1}^{n} f_{V_i|P_i;\theta}(w_i|p_i)$ apportant peu d'information sur β et donc de baser l'inférence pour β sur la vraisemblance partielle

$$\prod_{l=1}^{L} \mathbb{P}\left(W_l = (l) | Q_l, (X_k)_k, \beta\right) = \prod_{i=1}^{n} \frac{e^{\beta' X_i}}{\sum_{j=1}^{n} e^{\beta' X_j} \mathbb{1}_{T_i \leq T_j}}. \tag{2.4}$$

Ce raisonnement est détaillé dans Fleming & Harrington (1991) et le lecteur intéressé pourra s'y rapporter.

L'utilisation de cette vraisemblance partielle a également été justifiée a posteriori par les propriétés asymptotiques de l'estimateur $\widehat{\beta_n}$ obtenu en maximisant la vraisemblance partielle (2.4), ce qui est développé un peu plus loin dans cette section.

2.1.3 Estimation de la fonction de risque cumulé de base Λ_0

La maximisation de la vraisemblance partielle (2.4) ne permet pas d'estimer la fonction de risque cumulé de base, puisque Λ n'apparaît pas dans la formule de la vraisemblance partielle. Plusieurs méthodes ont été proposées pour l'estimation de Λ_0, la plus souvent retenue étant celle proposée par Breslow (1972, 1974) généralisant l'estimateur de Nelson. Si le paramètre de régression β_0 a été estimé par $\widehat{\beta_n}$, Breslow propose d'estimer Λ_0 par $\widehat{\Lambda_n}$ donné par

$$\widehat{\Lambda_n}(t) = \sum_{i=1}^{n} \frac{\Delta_i \mathbb{1}_{T_i \leq t}}{\sum_{j=1}^{n} e^{\widehat{\beta_n}' X_j} \mathbb{1}_{T_i \leq T_j}} = \int_0^t \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} \exp(\widehat{\beta_n}' X_j) Y_j(s)} \mathrm{d}N_i(s). \quad (2.5)$$

Pour cela, on choisit de prendre comme estimateurs de Λ_0 les fonctions en escalier ayant des sauts aux instant T_i non censurés, c'est-à-dire tels que $\Delta_i = 1$. Il s'agit donc de remplacer dans la vraisemblance (2.3) la fonction Λ par une telle fonction et de remplacer les valeurs $\lambda(T_i)$ par les sauts de cette fonction en les T_i, quand $\Delta_i = 1$, puis de maximiser cette vraisemblance en dimension finie : les paramètres de l'estimation du modèle sont le paramètre de régression β et les valeurs des sauts de la fonction en escalier estimant Λ_0 aux instants T_i tels que $\Delta_i = 1$. Il se vérifie que les estimateurs obtenus par cette méthode sont l'estimateur du maximum de vraisemblance partielle (2.4) pour β et l'estimateur de Breslow (2.5) pour Λ. Cela est développé dans Johansen (1983), Bagdonavičius & Nikulin (2002) et Klein & Moeschberger (1997) plus complètement.

2.1.4 Propriétés asymptotiques

Nous nous intéressons ici aux propriétés asymptotiques de l'estimateur du maximum de vraisemblance partielle de β_0 et de l'estimateur de Breslow de Λ_0 définis précédemment. Andersen & Gill (1982) ont fait l'étude de la consistance et de la distribution asymptotique de ces estimateurs. Bagdonavičius & Nikulin (2002) en ont également fait un exposé détaillé.

Notons $\widehat{\beta_n}$ l'estimateur du maximum de vraisemblance partielle de β_0 et $\widehat{\Lambda_n}$ l'estimateur de Λ_0 associé. Nous supposons que l'étude prend fin au temps

τ, durée maximale d'observation. Au-delà de cette durée, l'observation est censurée. Pour $\beta \in \mathbb{R}^p$ et $t \in \mathbb{R}^+$, notons,

$$\begin{aligned} s^{(0)}(\beta,t) &= \mathbb{E}[e^{\beta' X_1}\mathbb{1}_{t\leq T_1}], \\ s^{(1)}(\beta,t) &= \mathbb{E}[X_1 e^{\beta' X_1}\mathbb{1}_{t\leq T_1}], \\ s^{(2)}(\beta,t) &= \mathbb{E}[X_1 X_1' e^{\beta' X_1}\mathbb{1}_{t\leq T_1}]. \end{aligned}$$

Remarquons que $s^{(1)}(\beta,t) = \frac{\partial s^{(0)}(\beta,t)}{\partial \beta}$ et $s^{(2)}(\beta,t) = \frac{\partial s^{(1)}(\beta,t)}{\partial \beta}$. Définissons également la matrice $\Sigma(\beta)$ par

$$\Sigma(\beta) = \int_0^\tau s^{(2)}(\beta,t) - \frac{s^{(1)}(\beta,t)\left(s^{(1)}(\beta,t)\right)'}{s^{(0)}(\beta,t)}\lambda_0(t)\mathrm{d}t.$$

Les hypothèses suivantes sont suffisantes pour l'étude asymptotique des estimateurs qui nous intéressent :

- la fonction Λ_0 vérifie $\Lambda_0(\tau) < +\infty$,
- le paramètre β appartient à un ensemble borné $\mathcal{B}$ de $\mathbb{R}^p$,
- $\mathbb{P}(T \geq \tau) > 0$,
- $\Sigma(\beta)$ est définie positive,
- la covariable X est bornée presque sûrement et sa matrice de covariance est définie positive.

Sous ces hypothèses ainsi que celles permettant de définir les estimateurs étudiés, Andersen & Gill (1982) ont montré le résultat suivant.

Théorème 2.1. *L'estimateur du maximum de vraisemblance partielle $\widehat{\beta_n}$ et l'estimateur de Breslow $\widehat{\Lambda_n}$ possèdent les propriétés suivantes :*

- *$\widehat{\beta_n}$ converge en probabilité vers β_0,*
- *$\sqrt{n}\left(\widehat{\beta_n} - \beta_0\right)$ converge en loi vers un vecteur aléatoire gaussien centré de matrice de covariance $\Sigma(\beta_0)^{-1}$,*
- *$\sqrt{n}\left(\widehat{\Lambda_n} - \Lambda_0\right)$ converge en loi vers un processus gaussien centré G de fonction de covariance*

$$\begin{aligned} cov(G(s),G(t)) &= \int_0^{\min(s,t)} \frac{\lambda_0(u)}{s^{(0)}(\beta_0,u)}\mathrm{d}u \\ &\quad + \int_0^t \frac{s^{(1)}(\beta_0,u)'}{s^{(0)}(\beta_0,u)}\lambda_0(u)\mathrm{d}u.\Sigma(\beta_0)^{-1}.\int_0^s \frac{s^{(1)}(\beta_0,u)}{s^{(0)}(\beta_0,u)}\lambda_0(u)\mathrm{d}u. \end{aligned}$$

Remarque 2.1. La dernière partie du théorème permet d'obtenir, en particulier, la loi asymptotique à t fixé de $\sqrt{n}\left(\widehat{\Lambda_n}(t) - \Lambda_0(t)\right)$. Cette variable aléatoire converge vers une gaussienne centrée de variance donnée par $cov(G(t),G(t))$.

2.2 Le modèle de Cox stratifié

Une généralisation importante du modèle de Cox à risques proportionnels est le *modèle de Cox stratifié*. Il permet de considérer une variable explicative catégorielle qui n'a pas un effet proportionnel sur le risque instantané de la variable de survie. Pour cela, on sépare selon les catégories de cette variable explicative l'échantillon en K groupes, appelés *strates*, à l'intérieur desquels l'hypothèse de risques proportionnels est vérifiée (des tests pour l'hypothèse des risques proportionnels sont proposés par exemple dans Andersen et al. (1993)). Ainsi, on établit le modèle de Cox stratifié de la façon suivante : un individu de la strate $k \in \{1, ..., K\}$ a pour fonction de risque instantanée

$$\lambda_k(t) \exp(\beta_0' X),$$

où λ_k est la fonction de risque instantanée de base spécifique à la strate k, et β_0 est le paramètre de régression commun à toutes les strates. Ce modèle est notamment étudié dans Kalbfleisch & Prentice (1980), section 4.4, dans Andersen et al. (1993), section VII.2 et dans Klein & Moeschberger (1997), section 9.3.

Si l'on note S la variable aléatoire désignant la strate de l'individu, la fonction de risque du modèle général devient $\lambda_{T^0|X,S}$ définie pour tout $t \in \mathbb{R}^+$ par

$$\begin{aligned} \lambda_{T^0|X,S}(t) &= \left(\sum_{k=1}^{K} \lambda_k(t) \mathbb{1}_{S=k} \right) \exp(\beta_0' X) \\ &= \left(\prod_{k=1}^{K} \lambda_k(t)^{\mathbb{1}_{S=k}} \right) \exp(\beta_0' X). \end{aligned} \tag{2.6}$$

On suppose ici que la variable aléatoire S désignant la strate de l'individu est observée. Alors, de la même façon que dans l'étude du modèle de Cox faite dans la section 2.1, la vraisemblance partielle au vu du n-échantillon $(T_i, \Delta_i, X_i, S_i)_{1 \leq i \leq n}$ peut être obtenue pour le modèle de Cox stratifié grâce au produit des vraisemblances partielles à l'intérieur de chaque strate :

$$L_n^{(p)}(\beta) = \prod_{k=1}^{K} \left(\prod_{i=1; S_i=k}^{n} \frac{e^{\beta' X_i}}{\sum_{j=1}^{n} e^{\beta' X_j} \mathbb{1}_{T_i \leq T_j}} \right) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{\beta' X_i}}{\sum_{j=1}^{n} e^{\beta' X_j} \mathbb{1}_{T_i \leq T_j}} \right)^{\mathbb{1}_{S_i=k}} \tag{2.7}$$

Ainsi, l'estimateur du maximum de vraisemblance partielle peut être obtenu en maximisant l'expression (2.7) ou son logarithme. Pour cela, il est nécessaire

d'établir les équations du score en différentiant

$$\ln(L_n^{(p)}(\beta)) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}_{S_i=k}\left(\beta' X_i - \ln(\sum_{j=1}^{n} e^{\beta' X_j}\mathbb{1}_{T_i\leq T_j})\right)$$

relativement aux composantes du vecteur de régression β. Si l'on note $S_n^{(p)}(\beta) = \left(\frac{\partial \ln(L_n^{(p)}(\beta))}{\partial\beta_1}, ..., \frac{\partial \ln(L_n^{(p)}(\beta))}{\partial\beta_p}\right)$ le score dérivé de la log-vraisemblance partielle, alors l'estimateur de la vraisemblance partielle $\widehat{\beta_n}$ peut s'obtenir en résolvant numériquement, par exemple grâce à un algorithme de Newton-Raphson, l'équation du score

$$S_n^{(p)}(\widehat{\beta_n}) = 0.$$

Différentes méthodes de maximisation de la vraisemblance partielle ou log-vraisemblance partielle sont détaillées dans l'appendice A de Klein & Moeschberger (1997).

Les fonctions de risque cumulé de base $(\Lambda_k)_{1\leq k\leq K}$ peuvent, tout comme dans la section 2.1, être estimées grâce à la méthode de Breslow appliquée à l'intérieur de chaque strate, après avoir obtenu l'estimateur du maximum de vraisemblance partielle $\widehat{\beta_n}$ de β_0. Les estimateurs de Breslow ainsi obtenus sont donc, pour $k \in \{1, ..., K\}$

$$\widehat{\Lambda_{k,n}}(t) = \int_0^t \sum_{i=1}^{n} \frac{\mathbb{1}_{S_i=k}}{\sum_{j=1}^{n}\exp(\widehat{\beta_n}' X_j)Y_j(s)\mathbb{1}_{S_j=k}}\mathrm{d}N_i(s).$$

Notons que la fonction en escalier $\widehat{\Lambda_{k,n}}$ estimant la fonction Λ_k de la strate k n'admet de saut en l'instant d'observation T_i que si la donnée n'est pas censurée ($\Delta_i = 1$) et l'individu i appartient bien à la strate k ($\mathbb{1}_{S_i=k} = 1$). Remarquons en effet que si le dénominateur est nul, alors le numérateur l'est aussi ($Y_j(T_i) = 1$ au moins pour $j = i$ d'où $\mathbb{1}_{S_i=k} = 0$) donc il suffit de poser $0/0 = 0$ pour que l'estimateur $\widehat{\Lambda_{k,n}}$ soit défini comme voulu.

2.3 Le modèle de Cox stratifié avec strates aléatoirement manquantes

Le problème de l'estimation dans le modèle de régression de Cox (stratifié ou non) avec covariables manquantes a été étudié intensivement ces dernières années. Le lecteur intéressé par cette problématique pourra notamment se

référer à Lin & Ying (1993), Paik (1997), Paik & Tsai (1997), Chen & Little (1999), Martinussen (1999) ou encore Pons (2002). Cependant, à notre connaissance, peu de travaux ont été effectués dans le cadre de strates partiellement manquantes. Pourtant, dans beaucoup d'applications, cette variable explicative n'est pas disponible pour tous les individus de l'échantillon. Par exemple, l'étude du stade histologique du patient, variable importante dans les études médicales portant sur le cancer ou l'hépatite, nécessite une biopsie, qui ne peut parfois pas être effectuée sur chaque individu. Dans ce cadre de strates manquantes, l'inférence du modèle des risques proportionnels stratifié, basée sur la vraisemblance partielle, ne peut être appliquée directement. En effet, la vraisemblance partielle

$$L_n^{(p)}(\beta) == \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{e^{\beta' X_i}}{\sum_{j=1}^{n} e^{\beta' X_j}\mathbb{1}_{T_i \leq T_j}}\right)^{\mathbb{1}_{S_i=k}} \tag{2.7}$$

est obtenue à partir du produit sur toutes les strates du modèle, mais l'indicatrice $\mathbb{1}_{S_i=k}$ n'est pas connue pour tous les individus i.

Récemment, ce problème d'estimation dans le modèle de Cox stratifié avec strates manquantes a été étudié par Dupuy & Leconte (2008). Leur approche est de recourir à la méthode de régression-calibration, utilisée dans l'estimation de modèles de régression quand certaines covariables sont manquantes pour des individus de l'étude. La régression-calibration consiste à remplacer une covariable non observée par son espérance conditionnelle aux covariables disponibles. Cette méthode, discutée dans Carroll et al. (1995) et Thurston et al. (2003, 2005), a été utilisée dans de nombreux cas, notamment le modèle de régression à risques proportionnels (Prentice, 1982; Tsiatis et al., 1995; Wang et al., 1997; Wang, 1999; Wang et al., 2001). Dupuy & Leconte (2008) proposent donc de remplacer dans la vraisemblance partielle l'indicatrice $\mathbb{1}_{S_i=k}$ quand elle n'est pas observée par son espérance conditionnelle $\mathbb{E}[\mathbb{1}_{S_i=k}|X_i, W_i]$, où X_i et W_i sont des covariables observées pour chaque individu, W_i fournissant une information partielle sur la strate S_i. L'article montre que l'estimateur de régression-calibration de β_0 obtenu dans le cadre de strates manquantes est en général biaisé, mais néanmoins asymptotiquement gaussien.

Dans ce mémoire, l'idée est de fournir, dans le modèle de Cox stratifié avec strates manquantes, un estimateur du paramètre de régression qui soit à la fois consistant et asymptotiquement gaussien. Il sera intéressant également d'établir des estimateurs des fonctions de risque cumulé de base ayant de bonnes propriétés asymptotiques. Nous établissons ici les notations de notre

modèle.

Notons T^0 la variable aléatoire durée de vie du modèle dont la loi dépend d'un vecteur de covariables $X \in \mathbb{R}^p$ observées et d'un indicateur de strate S. Nous supposons que la loi de T^0 conditionnelle à ces variables explicatives est donnée par

$$\lambda_{T^0|X,S}(t) = \left(\prod_{k=1}^{K} \lambda_k(t)^{\mathbb{1}_{S=k}} \right) \exp(\beta' X) = \lambda_S(t) \exp(\beta' X), \tag{2.6}$$

où β est le paramètre de régression commun à toutes les strates permettant d'expliquer l'influence des covariables X et $(\lambda_k)_{1 \leq k \leq K}$ sont les fonctions de risque instantané de base associées aux K strates. Notons $\Lambda_k = \int_0^{\cdot} \lambda_k$ la fonction de risque cumulé de base associée à λ_k. La variable T^0 est exposée à un phénomène de censure à droite : notons C la variable aléatoire positive de censure et supposons que la censure est non informative et également que l'étude prend fin au temps $\tau < +\infty$. Nous observons ainsi la durée d'intérêt éventuellement censurée $T = \min(T^0, \min(C, \tau))$ et l'indicateur de censure $\Delta = \mathbb{1}_{T^0 \leq \min(C,\tau)}$.

Notons S la variable aléatoire désignant la strate à laquelle l'individu appartient. Nous supposons qu'elle prend ses valeurs dans un ensemble fini $\{1, ..., K\}$. Cette variable n'étant pas observée pour tous les individus de l'échantillon, il est nécessaire de modéliser sa loi. Pour cela, nous supposons que la loi de S dépend d'un vecteur supplémentaire de covariables $W \in \mathbb{R}^m$ observées, pouvant éventuellement comprendre certaines des composantes de X, et que cette loi est basée sur le modèle logistique. La probabilité d'un individu d'appartenir à la strate $k \in \{1, ..., K\}$ conditionnelle aux covariables W est ainsi donnée par

$$\mathbb{P}(S = k|W) = \frac{\exp(\gamma_k' W)}{\sum_{j=1}^{k} \exp(\gamma_j' W)},$$

où $\gamma_k \in \mathbb{R}^m$ est un vecteur de régression. Dans la suite, par souci d'identifiabilité, nous posons $\gamma_K = 0$ et notons $\gamma = (\gamma_1', ..., \gamma_{K-1}')'$ le vecteur concaténé des vecteurs de régression associés à chaque strate. Notons ainsi $\pi_{k,\gamma}(W) = \mathbb{P}(S = k|W)$ pour $k \in \{1, ..., K\}$. Afin de modéliser l'information sur l'observation de la strate, nous introduisons la variable aléatoire R déterminant si S est observée. Notons

$$R = \begin{cases} 1 & \text{si } S \text{ est observée} \\ 0 & \text{si } S \text{ n'est pas observée.} \end{cases}$$

Nous supposons la variable R indépendante de (T, Δ) conditionnellement à (X, S) et indépendante de S conditionnellement à W. La loi de R est supposée non informative pour les paramètres β, γ et Λ.

Ainsi, les observations dont le statisticien dispose ici sont la durée de vie éventuellement censurée, l'indicateur de censure, les covariables X et W, l'indicateur d'observation de la strate, et la valeur de la strate si celui-ci vaut 1. Notre n-échantillon est donc composé des observations

$$\mathcal{O}_i = (T_i, \Delta_i, X_i, W_i, R_i, R_i S_i) \text{ pour tout } i \text{ dans } \{1, ..., n\},$$

répliques indépendantes et de même loi que (T, Δ, X, W, R, RS).

Le paramètre inconnu est $\theta = (\beta, \gamma, \Lambda_k; 1 \leq k \leq K)$ et la vraie valeur du modèle est notée $\theta_0 = (\beta_0, \gamma_0, \Lambda_{k,0}; 1 \leq k \leq K)$. Le paramètre d'intérêt pour l'inférence sur ce modèle est le paramètre de régression β, les paramètres γ et Λ_k $(1 \leq k \leq K)$ étant considérés comme des paramètres de nuisance.

Comme les strates ne sont pas complètement observées, nous ne pouvons pas utiliser la formule de la vraisemblance partielle pour l'estimation du paramètre d'intérêt β. Par conséquent, nous calculons la vraisemblance observée relative au paramètre θ. Nous utilisons pour cela les notations classiques f_U pour la densité de U variable ou vecteur aléatoire (P_U si U est discrète), et $f_{U|V}$ pour la densité conditionnelle de U sachant V ($P_{U|V}$ si U est discrète).

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_{T_i,\Delta_i,X_i,W_i,S_i;\theta}^{R_i} f_{T_i,\Delta_i,X_i,W_i}^{1-R_i} \mathbb{P}(R_i = 1)^{R_i} \mathbb{P}(R_i = 0)^{1-R_i} \\ &= \prod_{i=1}^{n} \left(f_{T_i,\Delta_i|X_i,W_i,S_i;\theta} P_{S_i|X_i,W_i;\theta} f_{X_i,W_i} \mathbb{P}(R_i = 1) \right)^{R_i} \qquad (2.8) \\ &\quad \times \left(\sum_{k=1}^{K} f_{T_i,\Delta_i|X_i,W_i,S_i=k;\theta} \mathbb{P}_\theta(S_i = k|X_i, W_i) f_{X_i,W_i} \mathbb{P}(R_i = 0) \right)^{1-R_i}. \end{aligned}$$

Selon notre modèle, la loi de S dépend de la covariable W et les composantes de X servant à expliquer la loi de S se trouvent également dans le vecteur W. Ainsi, il suffit de conditionner S par W pour avoir toute l'information sur la loi de S : pour tout $i \in \{1, ..., n\}$, $\mathbb{P}_{S_i|X_i,W_i;\theta} = \mathbb{P}_{S_i|W_i;\gamma}$. De plus, la connaissance de X et S suffit à déterminer la loi de (T, Δ), d'où pour tout $i \in \{1, ..., n\}$, $f_{T_i,\Delta_i|X_i,W_i,S_i;\theta} = f_{T_i,\Delta_i|X_i,S_i;\theta}$. Ainsi, la censure étant non

informative, la vraisemblance (2.8) devient

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \left(f_{T_i,\Delta_i|X_i,S_i;\theta}\pi_{S_i,\gamma}(W_i)f_{X_i,W_i}\mathbb{P}(R_i=1)\right)^{R_i} \\ &\quad \times \left(\sum_{k=1}^{K} f_{T_i,\Delta_i|X_i,S_i=k;\theta}\pi_{k,\gamma}(W_i)f_{X_i,W_i}\mathbb{P}(R_i=0)\right)^{1-R_i} \\ &= \prod_{i=1}^{n} \left(\prod_{k=1}^{K} \left(\lambda_k(T_i)^{\Delta_i}\exp(\Delta_i\beta' X_i - \Lambda_k(T_i)e^{\beta' X_i})\pi_{k,\gamma}(W_i)\right)^{\mathbb{1}_{S_i=k}} f_{X_i,W_i}\mathbb{P}(R_i=1)\right)^{R_i} \\ &\quad \times \left(\sum_{k=1}^{K} \lambda_k(T_i)^{\Delta_i}\exp(\Delta_i\beta' X_i - \Lambda_k(T_i)e^{\beta' X_i})\pi_{k,\gamma}(W_i)f_{X_i,W_i}\mathbb{P}(R_i=0)\right)^{1-R_i}. \quad (2.9) \end{aligned}$$

Par ailleurs, nous supposons dans ce modèle que la loi des covariables X et W ne dépend pas du paramètre inconnu θ, de même que la loi de la variable déterminant l'observation de la strate R. Ainsi, la vraisemblance à considérer pour l'estimation de θ (proportionnelle à celle donnée par (2.9)) est

$$\begin{aligned} L_n^{(obs)}(\theta) &= \prod_{i=1}^{n} \left(\prod_{k=1}^{K} \left(\lambda_k(T_i)^{\Delta_i}\exp(\Delta_i\beta' X_i - \Lambda_k(T_i)e^{\beta' X_i})\pi_{k,\gamma}(W_i)\right)^{\mathbb{1}_{S_i=k}}\right)^{R_i} \\ &\quad \times \left(\sum_{k=1}^{K} \lambda_k(T_i)^{\Delta_i}\exp(\Delta_i\beta' X_i - \Lambda_k(T_i)e^{\beta' X_i})\pi_{k,\gamma}(W_i)\right)^{1-R_i} \quad (2.10) \end{aligned}$$

Pour les mêmes raisons que dans la partie 2.1.3, détaillées pour ce cas dans le chapitre 3, cette vraisemblance n'admet pas de maximum sur l'ensemble des paramètres $\mathcal{B}\times\mathcal{G}\times\mathcal{A}^{\otimes K}$, où $\mathcal{A}$ est l'ensemble des fonctions positives strictement croissantes, s'annulant en 0 et continûment différentiables de dérivée λ sur $\mathbb{R}^+$ (quels que soient les ensembles d'intérieur non vide $\mathcal{B}\subseteq\mathbb{R}^p$ et $\mathcal{G}\subseteq(\mathbb{R}^m)^{K-1}$). Ainsi, nous proposons de maximiser la vraisemblance sur un espace d'estimation différent, en nous inspirant des estimateurs de Breslow pour les fonctions de risque cumulé de base.

Nous introduisons l'espace de maximisation suivant : l'estimateur de $\Lambda_{k,0}$ ($1\leq k\leq K$) est une fonction en escalier croissante continue à droite admettant des sauts aux instants T_i tels que, soit l'individu i n'est pas censuré et l'on observe sa strate qui est égale à k ($\Delta_i R_i \mathbb{1}_{S_i=k}=1$), soit l'individu i n'est pas censuré et sa strate n'est pas observée ($\Delta_i(1-R_i)=1$). Alors nous proposons de maximiser la vraisemblance $L_n^{(obs)}(\theta)$ sur l'espace $\mathcal{B}\times\mathcal{G}\times\prod_{k=1}^{K}\mathcal{E}_n^{(k)}$, où

$\mathcal{B}$ est un compact de $\mathbb{R}^p$, $\mathcal{G}$ est un compact de $(\mathbb{R}^m)^{K-1}$ et $\mathcal{E}_n^{(k)}$ est l'ensemble des fonctions définies ci-dessus permettant d'estimer $\Lambda_{k,0}$ $(1 \leq k \leq K)$. Lors de la maximisation de cette vraisemblance, l'expression $\lambda_k(T_i)$ sera remplacée par le saut de l'estimateur de Λ_k en T_i, noté $\Lambda_k\{T_i\}$. L'estimateur ainsi obtenu de θ_0 sera noté $\widehat{\theta_n}$ et appelé *estimateur semi-paramétrique* de θ_0.

Ainsi, la maximisation de la vraisemblance sur cet espace d'estimation se réduit à un problème de dimension finie, la log-vraisemblance observée étant donc

$$\begin{aligned} l_n^{(obs)}(\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} R_i \mathbb{1}_{S_i=k} \left(\Delta_i \ln \Lambda_k\{T_i\} + \Delta_i \beta' X_i - e^{\beta' X_i \Lambda_k(T_i)} + \ln \pi_{k,\gamma}(W_i) \right) \\ &\quad + \sum_{i=1}^{n} (1 - R_i) \ln \left(\sum_{k=1}^{K} \Lambda_k\{T_i\}^{\Delta_i} \exp(\Delta_i \beta' X_i - e^{\beta' X_i \Lambda_k(T_i)}) \pi_{k,\gamma}(W_i) \right). \end{aligned}$$

Etant données les difficultés posées par cette expression lors de la recherche d'un maximum explicite, nous décidons d'utiliser ici un algorithme Espérance-Maximisation pour déterminer l'estimateur du maximum de vraisemblance voulu. Cette approche est détaillée dans le chapitre 4.

Enfin, les résultats théoriques concernant l'estimateur proposé du paramètre θ_0, notamment la consistance et la normalité asymptotique de $\widehat{\theta_n}$, l'efficacité de $\widehat{\beta_n}$ et l'estimation des variances asymptotiques, sont présentés dans le chapitre 3.

Chapitre 3

Estimation du maximum de vraisemblance dans le modèle de Cox stratifié avec strates aléatoirement manquantes

Sommaire

The content of the present chapter is the one of the paper entitled *Maximum likelihood estimation in a partially observed stratified regression model with censored data* (Detais & Dupuy, 2008), submitted to Annals of the Institute of Statistical Mathematics.

Abstract

The stratified proportional intensity model generalizes Cox's proportional intensity model by allowing different groups of the population under study to have distinct baseline intensity functions. In this article, we consider the problem of estimation in this model when the variable indicating the stratum is unobserved for some individuals in the studied sample. In this setting, we construct nonparametric maximum

likelihood estimators for the parameters of the stratified model and we establish their consistency and asymptotic normality. Consistent estimators for the limiting variances are also obtained.

Keywords **: Asymptotic normality, Consistency, Missing data, Nonparametric maximum likelihood, Right-censored failure time data, Stratified proportional intensity model, Variance estimation.**

3.1 Introduction

This paper considers the problem of estimation in the stratified proportional intensity regression model for survival data, when the stratum information is missing for some sample individuals.

The stratified proportional intensity model (see Andersen et al. (1993) or Martinussen & Scheike (2006) for example) generalizes the usual Cox (1972) proportional intensity regression model for survival data, by allowing different groups -the strata- of the population under study to have distinct baseline intensity functions. More precisely, in the stratified model, the strata divide the sample individuals into K disjoint groups, each having a distinct baseline intensity function λ_k but a common value for the regression parameter.

The intensity function for the failure time T^0 of an individual in stratum k thus takes the form

$$\lambda_k(t)\exp(\beta' X), \tag{3.1}$$

where X is a p-vector of covariates, β is a p-vector of unknown regression parameters of interest, and $(\lambda_k(.))_{1\leq k\leq K}$ are K unknown baseline intensity functions defined on $\mathbb{R}^+$.

A consistent and asymptotically normal estimator of β can be obtained by maximizing the partial likelihood function (Cox, 1975). The partial likelihood for the stratified model (3.1) is the product over strata of the within-stratum partial likelihoods (we refer to Andersen et al. (1993) for a detailed treatment of maximum partial likelihood estimation in model (3.1)). In some applications, it can also be desirable to estimate the cumulative baseline intensity functions $\Lambda_k = \int \lambda_k$. The so-called Breslow (1972) estimators are commonly used for that purpose (see chapter 7 of Andersen et al. (1993) for further details on the Breslow estimator and its asymptotic properties).

One major motivation for using the stratified model is that it allows to accomodate in the analysis a predictive categorical covariate whose effect on the intensity is not proportional. To this end, the individuals under study are stratified with respect to the categories of this covariate. In many applications however, this covariate may be missing for some sample individuals (for example, histological stage determination may require biopsy and due to expensiveness, may not be performed on all the study subjects). In this case, the usual statistical inference for model (3.1), based on the product of within-stratum partial likelihoods, can not be directly applied.

In this work, we consider the problem of estimating β and the $\Lambda_k, k = 1, \ldots, K$ in model (3.1), when the covariate defining the stratum is missing for some (but not all) individuals. Equivalently said, we consider the problem of estimating model (3.1) when the stratum information is only partially available.

The problem of estimation in the (unstratified) Cox regression model $\lambda(t)\exp(\beta' X)$ with missing covariate X has been the subject of intense research over the past decade: see for example Lin & Ying (1993), Paik (1997), Paik & Tsai (1997), Chen & Little (1999), Martinussen (1999), Pons (2002), and the references therein. But to the best of our knowledge and despite its practical relevance, the problem of statistical inference in model (3.1) with partially available stratum information has not been yet extensively investigated. Recently, Dupuy & Leconte (2008) studied the asymptotic properties of a regression calibration estimator of β in this setting (regression calibration is a general method for handling missing data in regression models, see Carroll et al. (1995) for example). The authors proved that this estimator is asymptotically biased, although nevertheless asymptotically normal. No estimators of the cumulative baseline intensity functions were provided.

In this work, we aim at providing an estimator of β that is both consistent and asymptotically normal. Moreover, although the cumulative intensity functions Λ_k are usually not the primary parameters of interest, we also aim at providing consistent and asymptotically normal estimators of the values $\Lambda_k(t)$, $k = 1, \ldots, K$.

The regression calibration inferential procedure investigated by Dupuy & Leconte (2008) is essentially based on a modified version of the partial likelihood for model (3.1). In this paper, we propose an alternative method which may be viewed as a fully maximum likelihood approach. Besides assuming

that the failure intensity function for an individual in stratum k is given by model (3.1), we assume that the probability of being in stratum k conditionally on a set of observed covariates W (which may include some components of X) is of the logistic form, depending on some unknown finite-dimensional parameter γ.

A full likelihood for the collected parameter $\theta = (\beta, \gamma, \Lambda_k; 1 \leq k \leq K)$ is constructed from a sample of incompletely observed data. Based on this, we propose to estimate the finite and infinite-dimensional components of θ by using the nonparametric maximum likelihood (NPML) estimation method. We then provide asymptotic results for these estimators, including consistency, asymptotic normality, semiparametric efficiency of the NPML estimator of β, and consistent variance estimation.

Our proofs use some techniques developed by Murphy (1994, 1995) and Parner (1998) to establish the asymptotic theory for the frailty model.

The paper is organized as follows. In Section 3.2, we describe in greater detail the data structure and the model assumptions. In Section 3.3, we describe the NPML estimation method for our setting and we establish existence of the NPML estimator of θ. Section 3.4 establishes the consistency and asymptotic normality of the proposed estimator. Consistent variance estimators are also obtained for both the finite-dimensional parameter estimators and the nonparametric cumulative baseline intensity estimators. We give some concluding remarks in Section 3.5. Proofs are given in Appendix.

3.2 Data structure and model assumptions

We describe the notations and model assumptions that will be used throughout the paper.

All the random variables are defined on a probability space $(\Omega, \mathcal{C}, \mathbb{P})$. Let T^0 be a random failure time whose distribution depends on a vector of covariates $X \in \mathbb{R}^p$ and on a stratum indicator $S \in \mathcal{K} = \{1, \ldots, K\}$. We assume that conditionally on X and $S = k$ $(k \in \mathcal{K})$, the intensity function of T^0 is given by model (3.1). We suppose that T^0 may be right-censored by a positive random variable C and that the analysis is restricted to the time interval $[0, \tau]$, where $\tau < \infty$ denotes the end of the study. Thus we actually observe the potentially censored duration $T = \min(T^0, \min(C, \tau))$ and a censoring

indicator $\Delta = 1_{T^0 \leq \min(C,\tau)}$. If $t \in [0,\tau]$, we denote by $N(t) = 1_{T \leq t}\Delta$ and $Y(t) = 1_{T \geq t}$ the failure counting and at-risk processes respectively.

Let $W \in \mathbb{R}^m$ be a vector of surrogate covariates for S (W and X may share some common components). That is, W brings a partial information about S when S is missing, and it adds no information when S is observed so that the distribution of T^0 conditionally on X, S, and W does not involve the components of W that are not in X. We assume that the conditional probability that an individual belongs to the k-th stratum given his covariate vector W follows a multinomial logistic model:

$$\mathbb{P}(S = k|W) = \frac{\exp(\gamma_k' W)}{\sum_{j=1}^{K} \exp(\gamma_j' W)},$$

where $\gamma_k \in \mathbb{R}^m$ ($k \in \mathcal{K}$). Finally, we let R denote the indicator variable which is 1 if S is observed and 0 otherwise. Then, the data consist of n i.i.d. replicates

$$\mathcal{O}_i = (T_i, \Delta_i, X_i, W_i, R_i, R_i S_i), \quad i = 1, \ldots, n,$$

of $\mathcal{O} = (T, \Delta, X, W, R, RS)$. The data available for the i-th individual are therefore $(T_i, \Delta_i, X_i, W_i, S_i)$ if $R_i = 1$ and $(T_i, \Delta_i, X_i, W_i)$ if $R_i = 0$.

In the sequel, we set $\gamma_K = 0$ for model identifiability purpose and we note $\gamma = (\gamma_1', \ldots, \gamma_{K-1}')' \in (\mathbb{R}^m)^{K-1} \equiv \mathbb{R}^q$. We also note $\pi_{k,\gamma}(W) = \mathbb{P}(S = k|W)$, $k \in \mathcal{K}$. Now, let $\theta = (\beta, \gamma, \Lambda_k; k \in \mathcal{K})$ be the collected parameter and $\theta_0 = (\beta_0, \gamma_0, \Lambda_{k,0}; k \in \mathcal{K})$ denote the true parameter value. Under the true value θ_0, the expectation of random variables will be denoted P_{θ_0}. $\mathbb{P}_n$ will denote the empirical probability measure. In the sequel, the stochastic convergences will be in terms of outer measure.

We now make the following additional assumptions:

(a) The censoring time C is independent of T^0 given (S, X, W), of S given (X, W), and is non-informative. With probability 1, $\mathbb{P}(C \geq T^0 \geq \tau|S, X, W) > c_0$ for some positive constant c_0.

(b) The parameter values β_0 and γ_0 lie in the interior of known compact sets $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{G} \subset \mathbb{R}^q$ respectively. For every $k \in \mathcal{K}$, the cumulative baseline intensity function $\Lambda_{k,0}$ is a strictly increasing function on $[0,\tau]$ with $\Lambda_{k,0}(0) = 0$ and $\Lambda_{k,0}(\tau) < \infty$. Moreover, for every $k \in \mathcal{K}$, $\Lambda_{k,0}$ is continuously differentiable in $[0,\tau]$, with $\lambda_{k,0}(t) = \partial\Lambda_{k,0}(t)/\partial t$. Let $\mathcal{A}$ denote the set of functions satisfying these properties.

(c) The covariate vectors X and W are bounded (*i.e.* $\|X\| < c_1$ and $\|W\| < c_1$, for some finite positive constant c_1, where $\|\cdot\|$ denotes the Euclidean norm). Moreover, the covariance matrices of X and W are positive definite. Let $c_2 = \min_{\beta\in\mathcal{B},\|X\|<c_1} e^{\beta' X}$ and $c_3 = \max_{\beta\in\mathcal{B},\|X\|<c_1} e^{\beta' X}$.

(d) There is a constant $c_4 > 0$ such that for every $k \in \mathcal{K}$, $P_{\theta_0}(1_{S=k}Y(\tau)R) > c_4$, and the sample size n is large enough to ensure that $\sum_{i=1}^{n} 1_{S_i=k}Y_i(\tau)R_i > 0$ for every $k \in \mathcal{K}$.

(e) With probability 1, there exists a positive constant c_5 such that for every $k \in \mathcal{K}$, $P_{\theta_0}(\Delta R 1_{S=k}|T, X, W) > c_5$.

(f) R is independent of S given W, of (T, Δ) given (X, S). The distribution of S conditionally on X and W does not involve the components of X that are not in W. The distributions of R and of the covariate vectors X and W do not depend on the parameter θ.

REMARK 3.1. Conditions (b), (c), (d), and (e) are used for identifiability of the parameters and consistency of the proposed estimators. Condition (d) essentially requires that for every stratum k, some subjects are known to belong to k and are still at risk when the study ends. The first assumption in condition (f) states that S is missing at random, which is a fairly general missing data situation (we refer to chapters 6 and 7 in Tsiatis (2006) for a recent exposition of missing data mechanisms).

REMARK 3.2. We are now in position to describe our proposed approach to the problem of estimation in model (3.1) from a sample of incomplete data $\mathcal{O}_i$, $i = 1, \ldots, n$.
Let $\mathcal{S}$ denote the set of subjects with unknown stratum in this sample. The regression calibration method investigated by Dupuy & Leconte (2008) essentially allocates every subject of $\mathcal{S}$ to each of the strata, and estimates β_0 by maximizing a modified version of the partial likelihood for the stratified model, where the contribution of any individual i in $\mathcal{S}$ to the within-k-th-stratum partial likelihood is weighted by an estimate of $\pi_{k,\gamma}(W_i)$ (for every $k \in \mathcal{K}$). The asymptotic bias of the resulting estimator arises from the failure of this method to fully exploit the information carried by $(T_i, \Delta_i, X_i, W_i)$ on the unobserved stratum indicator S_i.
Therefore in this paper, we rather suggest to weight each subject i in $\mathcal{S}$ by an estimate of the conditional probability that subject i belongs to the k-th stratum given the whole observed data $(T_i, \Delta_i, X_i, W_i)$. This suggestion raises two main problems, as is described below.

REMARK 3.3. First, we should note that the suggested alternative weights depend on the unknown baseline intensity functions. Therefore, the modified partial likelihood approach considered by Dupuy & Leconte (2008) can not be used to derive an estimator for β_0. Next, the statistics to be involved in the score function for β will depend on the conditional weights and thus, this score will not be expressible as a stochastic integral of some predictable process, as is often the case in models for failure time data. This, in turn, will prevent us from using the counting process martingale theory usually associated with the theoretical developments in failure time models.

To overcome the first problem, we define our estimators from a full likelihood for the whole parameter, that is, for both the finite-dimensional -β (and γ)- and infinite-dimensional -Λ_k, $k \in \mathcal{K}$- components of θ. Empirical process theory (van der Vaart & Wellner, 1996) is used to establish asymptotics for the proposed estimators.

3.3 Maximum likelihood estimation

In the sequel, we assume that there are no ties among the observed death times (this hypothesis is made to simplify notations, but the results below can be adapted to accomodate ties). The likelihood function for observed data $\mathcal{O}_i$, $i = 1, \ldots, n$ is given by

$$L_n(\theta) = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left\{ \lambda_k(T_i)^{\Delta_i} \exp\left(\Delta_i \beta' X_i - e^{\beta' X_i} \Lambda_k(T_i) \right) \pi_{k,\gamma}(W_i) \right\}^{\mathbb{1}_{S_i=k}} \right]^{R_i} \times \left[\sum_{k=1}^{K} \lambda_k(T_i)^{\Delta_i} \exp\left(\Delta_i \beta' X_i - e^{\beta' X_i} \Lambda_k(T_i) \right) \pi_{k,\gamma}(W_i) \right]^{1-R_i} \quad (3.2)$$

It would seem natural to derive a maximum likelihood estimator of θ_0 by maximizing the likelihood (3.2). However, the maximum of this function over the parameter space $\Theta = \mathcal{B} \times \mathcal{G} \times \mathcal{A}^{\otimes K}$ does not exist. To see this, consider functions Λ_k with fixed values at the T_i, and let $(\partial \Lambda_k(t)/\partial t)|_{t=T_i} = \lambda_k(T_i)$ go to infinity for some T_i with $\Delta_i R_i \mathbb{1}_{S_i=k} = 1$ or $\Delta_i(1 - R_i) = 1$.

To overcome this problem, we introduce a modified maximization space for (3.2), by relaxing each $\Lambda_k(\cdot)$ to be an increasing right-continuous step-function on $[0, \tau]$, with jumps at the T_i's such that $\Delta_i R_i \mathbb{1}_{S_i=k} = 1$ or $\Delta_i(1 - R_i) = 1$. Estimators of $(\beta_0, \gamma_0, \Lambda_{k,0}; k \in \mathcal{K})$ will thus be derived by maximizing a modified version of (3.2), obtained by replacing $\lambda_k(T_i)$ in (3.2) with the

jump size $\Lambda_k\{T_i\}$ of Λ_k at T_i.

If they exist, these estimators will be referred to as nonparametric maximum likelihood estimators - NPMLEs - (we refer to Zeng & Lin (2007) for a review of the general principle of NPML estimation, with application to various semiparametric regression models for censored data. See also the numerous references therein). In our setting, existence of such estimators is ensured by the following theorem (proof is given in Appendix):

Theorem 3.1. *Under conditions (a)-(f), the NPMLE* $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\gamma}_n, \widehat{\Lambda}_{k,n}; k \in \mathcal{K})$ *of* θ_0 *exists and is achieved.*

The problem of maximizing L_n over the approximating space described above reduces to a finite dimensional problem, and the expectation-maximization (EM) algorithm (Dempster et al., 1977) can be used to calculate the NPMLEs. For $1 \leq i \leq n$ and $k \in \mathcal{K}$, let $w_i(k, \theta)$ be the conditional probability that the i-th individual belongs to the k-th stratum given $(T_i, \Delta_i, X_i, W_i)$ and the parameter value θ, and let $Q(\mathcal{O}_i, k, \theta)$ denote the conditional expectation of $\mathbb{1}_{S_i=k}$ given $\mathcal{O}_i$ and the parameter value θ. Then $Q(\mathcal{O}_i, k, \theta)$ has the form

$$Q(\mathcal{O}_i, k, \theta) = R_i \mathbb{1}_{S_i=k} + (1 - R_i) w_i(k, \theta).$$

In the M-step of the EM-algorithm, we solve the complete-data score equation conditional on the observed data. In particular, the following expression for the NPMLE of $\Lambda_k(\cdot)$ can be obtained by: (a) taking the derivative with respect to the jump sizes of $\Lambda_k(\cdot)$, of the conditional expectation of the complete-data log-likelihood given the observed data and the NPML estimator, (b) setting this derivative equal to 0:

Lemma 3.2. *The NPMLE* $\widehat{\theta}_n$ *satisfies the following equation for every* $k \in \mathcal{K}$*:*

$$\widehat{\Lambda}_{k,n}(t) = \int_0^t \sum_{i=1}^n \frac{Q(\mathcal{O}_i, k, \widehat{\theta}_n)}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \widehat{\theta}_n) \exp(\widehat{\beta_n}' X_j) Y_j(s)} \mathrm{d}N_i(s), \quad 0 \leq t \leq \tau.$$

The details of the calculations are omitted (note how the suggested weights $w_i(k, \theta)$ naturally arise from the M-step of the EM algorithm). We refer the interested reader to Zeng & Cai (2005) and Sugimoto & Hamasaki (2006), who recently described EM algorithms for computing NPMLEs in various other semiparametric models with censored data.

In the sequel, we shall denote the conditional expectation of the complete-data log-likelihood given the observed data and the NPML estimator by $E_n[\widetilde{l}_n(\theta)]$.

3.4 Asymptotic properties

This section states the asymptotic properties of the proposed estimators. We first obtain the following theorem, which states the strong consistency of the proposed NPMLE. The proof is given in Appendix.

Theorem 3.3. *Under conditions (a)-(f),* $\|\widehat{\beta}_n - \beta_0\|$, $\|\widehat{\gamma}_n - \gamma_0\|$, *and* $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_{k,n}(t) - \Lambda_{k,0}(t)|$ *(for every* $k \in \mathcal{K}$*) converge to 0 almost surely as* n *tends to infinity.*

To derive the asymptotic normality of the proposed estimators, we adapt the function analytic approach developed by Murphy (1995) for the frailty model (see also Chang et al. (2005), Kosorok & Song (2007), and Lu (2008), for recent examples of this approach in various other models).

Instead of calculating score equations by differentiating $E_n[\widetilde{l}_n(\theta)]$ with respect to β, γ, and the jump sizes of $\Lambda_k(\cdot)$, we consider one-dimensional submodels $\widehat{\theta}_{n,\eta}$ passing through $\widehat{\theta}_n$ and we differentiate with respect to η. Precisely, we consider submodels of the form

$$\eta \mapsto \widehat{\theta}_{n,\eta} = \left(\widehat{\beta}_n + \eta h_\beta, \widehat{\gamma}_n + \eta h_\gamma, \int_0^{\cdot} (1 + \eta h_{\Lambda_k}(s))\, \mathrm{d}\widehat{\Lambda}_{k,n}(s); k \in \mathcal{K}\right),$$

where h_β and $h_\gamma = (h'_{\gamma_1}, \ldots, h'_{\gamma_{K-1}})'$ are p- and q-dimensional vectors respectively ($h_{\gamma_j} \in \mathbb{R}^m$, $j = 1, \ldots, K-1$), and the h_{Λ_k} ($k \in \mathcal{K}$) are functions on $[0, \tau]$. Let $h = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$. To obtain the score equations, we differentiate $E_n[\widetilde{l}_n(\widehat{\theta}_{n,\eta})]$ with respect to η and we evaluate at $\eta = 0$. $\widehat{\theta}_n$ maximizes $E_n[\widetilde{l}_n(\theta)]$ and therefore satisfies $(\partial E_n[\widetilde{l}_n(\widehat{\theta}_{n,\eta})]/\partial\eta)\,|_{\eta=0} = 0$ for every h, which leads to the score equation $S_n(\widehat{\theta}_n)(h) = 0$ where $S_n(\widehat{\theta}_n)(h)$ takes the form

$$S_n(\widehat{\theta}_n)(h) = \mathbb{P}_n\left[h'_\beta S_\beta(\widehat{\theta}_n) + h'_\gamma S_\gamma(\widehat{\theta}_n) + \sum_{k=1}^{K} S_{\Lambda_k}(\widehat{\theta}_n)(h_{\Lambda_k})\right], \tag{3.3}$$

where

$$\begin{aligned}
&S_\beta(\theta) = \Delta X - \sum_{k=1}^{K} Q(\mathcal{O}, k, \theta) X \exp(\beta' X)\Lambda_k(T),\\
&S_\gamma(\theta) = (S_{\gamma_1}(\theta)', \ldots, S_{\gamma_{K-1}}(\theta)')' \text{ with } S_{\gamma_k}(\theta) = W\left[Q(\mathcal{O}, k, \theta) - \pi_{k,\gamma}(W)\right],\\
&S_{\Lambda_k}(\theta)(h_{\Lambda_k}) = Q(\mathcal{O}, k, \theta)\left[h_{\Lambda_k}(T)\Delta - \exp(\beta' X)\int_0^T h_{\Lambda_k}(s)\mathrm{d}\Lambda_k(s)\right].
\end{aligned}$$

We take the space of elements $h = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$ to be

$$\begin{aligned} H \;=\; & \{(h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K}); h_\beta \in \mathbb{R}^p, \|h_\beta\| < \infty, h_\gamma \in \mathbb{R}^q, \|h_\gamma\| < \infty, \\ & h_{\Lambda_k} \text{ is a function defined on } [0,\tau], \|h_{\Lambda_k}\|_v < \infty, \forall k \in \mathcal{K}\}, \end{aligned}$$

where $\|h_{\Lambda_k}\|_v$ denotes the total variation of h_{Λ_k} on $[0,\tau]$. We further take the functions h_{Λ_k} to be continuous from the right at 0.

Define $\theta(h) = h'_\beta\beta + h'_\gamma\gamma + \sum_{k=1}^K \int_0^\tau h_{\Lambda_k}(s)\, d\Lambda_k(s)$, where $h \in H$. From this, the parameter θ can be considered as a linear functional on H, and the parameter space Θ can be viewed as a subset of the space $l^\infty(H)$ of bounded real-valued functions on H, which we provide with the uniform norm. Moreover, the score operator S_n appears to be a random map from Θ to the space $l^\infty(H)$. Note that appropriate choices of h allow to extract all components of the original parameter θ. For example, letting $h_\gamma = 0$, $h_{\Lambda_k}(\cdot) = 0$ for every $k \in \mathcal{K}$, and h_β be the p-dimensional vector with a one at the i-th location and zeros elsewhere yields the i-th component of β. Letting $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_k}(\cdot) = 0$ for every $k \in \mathcal{K}$ except $h_{\Lambda_j}(s) = \mathbb{1}_{s \le t}$ (for some $t \in (0,\tau)$) yields $\Lambda_j(t)$.

We need some further notations to state the asymptotic normality of the NPMLE of β_0. Let us first define the linear operator $\sigma = (\sigma_\beta, \sigma_\gamma, \sigma_{\Lambda_k}; k \in \mathcal{K}) : H \to H$ by

$$\begin{aligned} \sigma_\beta(h) = {} & P_{\theta_0}\left[2X\Delta\psi(\mathcal{O},\theta_0)\sum_{k=1}^K Q(\mathcal{O},k,\theta_0)h_{\Lambda_k}(T)\right] \\ & + P_{\theta_0}\left[\psi(\mathcal{O},\theta_0)X\left\{\psi(\mathcal{O},\theta_0)X'h_\beta + S_\gamma(\theta_0)'h_\gamma\right\}\right], \\ \sigma_\gamma(h) = {} & P_{\theta_0}\left[2S_\gamma(\theta_0)\Delta\sum_{k=1}^K Q(\mathcal{O},k,\theta_0)h_{\Lambda_k}(T)\right] + P_{\theta_0}\left[S_\gamma(\theta_0)S_\gamma(\theta_0)'\right]h_\gamma \\ & + P_{\theta_0}\left[\psi(\mathcal{O},\theta_0)S_\gamma(\theta_0)X'\right]h_\beta, \\ \sigma_{\Lambda_k}(h)(u) = {} & h_{\Lambda_k}(u)P_{\theta_0}\left[Q(\mathcal{O},k,\theta_0)\phi(u,\mathcal{O},k,\theta_0)\right] \\ & + P_{\theta_0}\left[2\phi(u,\mathcal{O},k,\theta_0)\sum_{j>k}Q(\mathcal{O},j,\theta_0)\left\{h_{\Lambda_j}(u) - e^{\beta_0'X}\int_0^u h_{\Lambda_j}\, d\Lambda_{j,0}\right.\right. \\ & \left.\left. - \Delta h_{\Lambda_j}(T) + e^{\beta_0'X}\int_0^T h_{\Lambda_j}\, d\Lambda_{j,0}\right\}\right] \\ & - h'_\beta P_{\theta_0}\left[2X\psi(\mathcal{O},\theta_0)Q(\mathcal{O},k,\theta_0)e^{\beta_0'X}Y(u)\right] \\ & - h'_\gamma P_{\theta_0}\left[2S_\gamma(\theta_0)Q(\mathcal{O},k,\theta_0)e^{\beta_0'X}Y(u)\right], \end{aligned}$$

where $\phi(u, \mathcal{O}, k, \theta_0) = Y(u)Q(\mathcal{O}, k, \theta_0)e^{\beta_0' X}$ and $\psi(\mathcal{O}, \theta_0) = \Delta - \sum_{k=1}^{K} Q(\mathcal{O}, k, \theta_0)e^{\beta_0' X}\Lambda_{k,0}(T)$. This operator is continuously invertible (Lemma 3.8 in Appendix). We shall denote its inverse by $\sigma^{-1} = (\sigma_\beta^{-1}, \sigma_\gamma^{-1}, \sigma_{\Lambda_k}^{-1}; k \in \mathcal{K})$.

Next, for every $r \in \mathbb{N}^*$, the r-dimensional column vector having all its components equal to 0 will be noted by 0_r (or by 0 when no confusion may occur). Let $h = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K}) \in H$. If $h_\gamma = 0$ and h_{Λ_k} is identically equal to 0 for every $k \in \mathcal{K}$, we note $h = (h_\beta, 0, 0; k \in \mathcal{K})$. Let $\widetilde{\sigma}_\beta^{-1} : \mathbb{R}^p \to \mathbb{R}^p$ be the linear map defined by $\widetilde{\sigma}_\beta^{-1}(u) = \sigma_\beta^{-1}((u, 0, 0; k \in \mathcal{K}))$, for $u \in \mathbb{R}^p$. Let $\{e_1, \ldots, e_p\}$ be the canonical basis of $\mathbb{R}^p$.

Then the following result holds, its proof is given in Appendix.

Theorem 3.4. *Under conditions (a)-(f), $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ has an asymptotic normal distribution $N(0, \Sigma_\beta)$, where*

$$\Sigma_\beta = (\widetilde{\sigma}_\beta^{-1}(e_1), \ldots, \widetilde{\sigma}_\beta^{-1}(e_p))$$

is the efficient variance in estimating β_0.

Remark 3.4. Although γ_0 and the cumulative baseline intensity functions $\Lambda_{k,0}$ ($k \in \mathcal{K}$) are not the primary parameters of interest, we may also state an asymptotic normality result for their NMPLEs. This requires some further notations.

Define $\widetilde{\sigma}_\gamma^{-1} : \mathbb{R}^q \to \mathbb{R}^q$ by $\widetilde{\sigma}_\gamma^{-1}(u) = \sigma_\gamma^{-1}((0, u, 0; k \in \mathcal{K}))$, let $\{f_1, \ldots, f_q\}$ be the canonical basis of $\mathbb{R}^q$, and define $\Sigma_\gamma = (\widetilde{\sigma}_\gamma^{-1}(f_1), \ldots, \widetilde{\sigma}_\gamma^{-1}(f_q))$. Finally, let $h_{(j,t)} = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$ be such that $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_j}(\cdot) = 1_{(-\infty,t]}(\cdot)$ for some $t \in (0, \tau)$ and $j \in \mathcal{K}$, and $h_{\Lambda_k} = 0$ for every $k \in \mathcal{K}, k \neq j$. Then the following holds (a brief sketch of the proof is given in Appendix):

Theorem 3.5. *Assume that conditions (a)-(f) hold. Then $\sqrt{n}(\widehat{\gamma}_n - \gamma_0)$ has an asymptotic normal distribution $N(0, \Sigma_\gamma)$. Furthermore, for any $t \in (0, \tau)$ and $j \in \mathcal{K}$, $\sqrt{n}(\widehat{\Lambda}_{j,n}(t) - \Lambda_{j,0}(t))$ is asymptotically distributed as a $N(0, v_j^2(t))$, where*

$$v_j^2(t) = \int_0^t \sigma_{\Lambda_j}^{-1}(h_{(j,t)})(u)\, d\Lambda_{j,0}(u).$$

We now turn to the issue of estimating the asymptotic variances of the estimators $\widehat{\beta}_n$, $\widehat{\gamma}_n$, and $\widehat{\Lambda}_{j,n}(t)$ ($t \in (0, \tau)$, $j \in \mathcal{K}$). It turns out that the asymptotic variances Σ_β, Σ_γ, and $v_j^2(t)$ are not expressible in explicit forms,

since the inverse σ^{-1} has no closed form. However, this is not a problem if we can provide consistent estimators for them. Such estimators are defined below.

For $i = 1, \ldots, n$, let X_{ir} denote the r-th ($r = 1, \ldots, p$) component of X_i, $S_{\gamma,i}(\theta)$ be defined as in (3.3) with $\mathcal{O}$ and W replaced by $\mathcal{O}_i$ and W_i respectively, and $S_{\gamma,i,s}(\theta)$ be the s-th ($s = 1, \ldots, q$) component of $S_{\gamma,i}(\theta)$. Using these notations, we define the following block matrix

$$\mathbb{A}_n = \begin{pmatrix} A^{\beta\beta} & A^{\beta\gamma} & A^{\beta\Lambda} \\ A^{\gamma\beta} & A^{\gamma\gamma} & A^{\gamma\Lambda} \\ A^{\Lambda\beta} & A^{\Lambda\gamma} & A^{\Lambda\Lambda} \end{pmatrix} \tag{3.4}$$

where the sub-matrices $A^{\beta\beta}, A^{\gamma\gamma}, A^{\beta\gamma}$, and $A^{\gamma\beta}$ are defined as follows by their (r, s)-th component:

$$\begin{aligned}
A^{\beta\beta}_{rs} &= \frac{1}{n}\sum_{i=1}^{n}\{\psi(\mathcal{O}_i, \widehat{\theta}_n)\}^2 X_{ir}X_{is}, \quad r, s = 1, \ldots, p, \\
A^{\gamma\gamma}_{rs} &= \frac{1}{n}\sum_{i=1}^{n} S_{\gamma,i,r}(\widehat{\theta}_n)S_{\gamma,i,s}(\widehat{\theta}_n), \quad r, s = 1, \ldots, q, \\
A^{\beta\gamma}_{rs} &= \frac{1}{n}\sum_{i=1}^{n} \psi(\mathcal{O}_i, \widehat{\theta}_n)X_{ir}S_{\gamma,i,s}(\widehat{\theta}_n), \quad r = 1, \ldots, p, \quad s = 1, \ldots, q, \\
A^{\gamma\beta}_{rs} &= A^{\beta\gamma}_{sr}, \quad r = 1, \ldots, q, \quad s = 1, \ldots, p.
\end{aligned}$$

Define the block matrices $A^{\beta\Lambda} = (A^{\beta\Lambda_1}, \ldots, A^{\beta\Lambda_K})$ and $A^{\gamma\Lambda} = (A^{\gamma\Lambda_1}, \ldots, A^{\gamma\Lambda_K})$, where for every $k \in \mathcal{K}$, the sub-matrices $A^{\beta\Lambda_k}$ and $A^{\gamma\Lambda_k}$ are defined by

$$\begin{aligned}
A^{\beta\Lambda_k}_{rs} &= \frac{2}{n}X_{sr}\Delta_s\psi(\mathcal{O}_s, \widehat{\theta}_n)Q(\mathcal{O}_s, k, \widehat{\theta}_n), \quad r = 1, \ldots, p, \quad s = 1, \ldots, n, \\
A^{\gamma\Lambda_k}_{rs} &= \frac{2}{n}S_{\gamma,s,r}(\widehat{\theta}_n)\Delta_s Q(\mathcal{O}_s, k, \widehat{\theta}_n), \quad r = 1, \ldots, q, \quad s = 1, \ldots, n.
\end{aligned}$$

Define also the block matrices

$$A^{\Lambda\beta} = \begin{pmatrix} A^{\Lambda_1\beta} \\ \vdots \\ A^{\Lambda_K\beta} \end{pmatrix} \qquad A^{\Lambda\gamma} = \begin{pmatrix} A^{\Lambda_1\gamma} \\ \vdots \\ A^{\Lambda_K\gamma} \end{pmatrix} \qquad A^{\Lambda\Lambda} = \begin{pmatrix} A^{\Lambda_1\Lambda_1} & \ldots & A^{\Lambda_1\Lambda_K} \\ \vdots & & \vdots \\ A^{\Lambda_K\Lambda_1} & \ldots & A^{\Lambda_K\Lambda_K} \end{pmatrix}$$

where for every $j, k \in \mathcal{K}$,

$$A_{rs}^{\Lambda_k\beta} = -\frac{1}{n}\sum_{i=1}^{n} 2X_{is}\psi(\mathcal{O}_i,\widehat{\theta}_n)Q(\mathcal{O}_i,k,\widehat{\theta}_n)e^{\widehat{\beta_n}'X_i}Y_i(T_r), \quad r=1,\ldots,n, \quad s=1,\ldots,p,$$

$$A_{rs}^{\Lambda_k\gamma} = -\frac{1}{n}\sum_{i=1}^{n} 2S_{\gamma,i,s}(\widehat{\theta}_n)Q(\mathcal{O}_i,k,\widehat{\theta}_n)e^{\widehat{\beta_n}'X_i}Y_i(T_r), \quad r=1,\ldots,n, \quad s=1,\ldots,q,$$

$$\begin{aligned}A_{rs}^{\Lambda_k\Lambda_j} &= \mathbb{1}_{j=k}\mathbb{1}_{r=s}\frac{1}{n}\sum_{i=1}^{n} Q(\mathcal{O}_i,k,\widehat{\theta}_n)\phi(T_r,\mathcal{O}_i,k,\widehat{\theta}_n)\\ &\quad+\mathbb{1}_{j>k}\Bigg(\mathbb{1}_{r=s}\frac{1}{n}\sum_{i=1}^{n} 2\phi(T_s,\mathcal{O}_i,k,\widehat{\theta}_n)Q(\mathcal{O}_i,j,\widehat{\theta}_n)\\ &\quad+\frac{2}{n}\sum_{i=1}^{n}\phi(T_r,\mathcal{O}_i,k,\widehat{\theta}_n)Q(\mathcal{O}_i,j,\widehat{\theta}_n)e^{\widehat{\beta_n}'X_i}\widehat{\Delta\Lambda_{j,n}}(T_s)\left\{\mathbb{1}_{T_s\le T_i}-\mathbb{1}_{T_s\le T_r}\right\}\\ &\quad-\frac{2}{n}\phi(T_r,\mathcal{O}_s,k,\widehat{\theta}_n)Q(\mathcal{O}_s,j,\widehat{\theta}_n)\Delta_s\Bigg), \quad r,s=1,\ldots,n,\end{aligned}$$

and $\widehat{\Delta\Lambda_{j,n}}(T_s)$ is the jump size of $\widehat{\Lambda}_{j,n}$ at T_s that is, $\widehat{\Delta\Lambda_{j,n}}(T_s) = \widehat{\Lambda}_{j,n}(T_s) - \widehat{\Lambda}_{j,n}(T_s-)$ $(j \in \mathcal{K}, s=1,\ldots,n)$. Note that for notational simplicity, the lower (sample size) indice n has been omitted in the notations for the sub-matrices of $\mathbb{A}_n$.

Now, define

$$\begin{aligned}\widehat{\Sigma}_{\beta,n} &= \left\{A^{\beta\beta} - A^{\beta\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\beta} - (A^{\beta\Lambda} - A^{\beta\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\Lambda})\right.\\ &\qquad\left.\times(A^{\Lambda\Lambda} - A^{\Lambda\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\Lambda})^{-1}(A^{\Lambda\beta} - A^{\Lambda\gamma}(A^{\gamma\gamma})^{-1}A^{\gamma\beta})\right\}^{-1},\\ \widehat{\Sigma}_{\gamma,n} &= \left\{A^{\gamma\gamma} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma} - (A^{\gamma\Lambda} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda})\right.\\ &\qquad\left.\times(A^{\Lambda\Lambda} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda})^{-1}(A^{\Lambda\gamma} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma})\right\}^{-1},\end{aligned}$$

and

$$\begin{aligned}\widehat{\Sigma}_{\Lambda,n} &= \left\{A^{\Lambda\Lambda} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda} - (A^{\Lambda\gamma} - A^{\Lambda\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma})\right.\\ &\qquad\left.\times(A^{\gamma\gamma} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\gamma})^{-1}(A^{\gamma\Lambda} - A^{\gamma\beta}(A^{\beta\beta})^{-1}A^{\beta\Lambda})\right\}^{-1}.\end{aligned}$$

Then the following holds:

Theorem 3.6. *Under conditions (a)-(f), $\widehat{\Sigma}_{\beta,n}$ and $\widehat{\Sigma}_{\gamma,n}$ converge in probability to Σ_β and Σ_γ respectively as n tends to ∞. Moreover, for $t \in (0,\tau)$ and $j \in \mathcal{K}$, let*

$$\widehat{v}_{j,n}^2(t) = \widehat{\Xi}_{(j,t)}^{n\prime}\widehat{\Sigma}_{\Lambda,n}U_{(j,t)}^n,$$

where

$$\widehat{\Xi}^n_{(j,t)} = \left(0'_{(j-1)n}, \widehat{\Delta\Lambda_{j,n}}(T_1)\mathbb{1}_{T_1\leq t}, \ldots, \widehat{\Delta\Lambda_{j,n}}(T_n)\mathbb{1}_{T_n\leq t}, 0'_{(K-j)n}\right)'$$

and

$$U^n_{(j,t)} = (0'_{(j-1)n}, \mathbb{1}_{T_1\leq t}, \ldots, \mathbb{1}_{T_n\leq t}, 0'_{(K-j)n})'.$$

Then $\widehat{v}^2_{j,n}(t)$ *converges in probability to* $v^2_j(t)$ *as* n *tends to* ∞.

3.5 Discussion

In this paper, we have constructed consistent and asymptotically normal estimators for the stratified proportional intensity regression model when the sample stratum information is only partially available. The proposed estimator for the regression parameter of interest in this model has been shown to be semiparametrically efficient. Although computationally more challenging, these estimators improve the ones previously investigated in the literature, such as the regression calibration estimators (Dupuy & Leconte, 2008).

We have obtained explicit (and computationally fairly simple) formulas for consistent estimators of the asymptotic variances. These formulas may however require the inversion of potentially large matrices. For a large sample, this inversion may be unstable. An alternative solution relies on numerical differentiation of the profile log-likelihood (see Murphy et al. (1997) and Chen & Little (1999) for example). Note that in this latter method however, no estimator is available for the asymptotic variance of the cumulative baseline intensity estimator. Some further work is needed to evaluate the numerical performance of the proposed estimators. This is the subject for future research, and requires some extensive simulation work which falls beyond the scope of this paper.

In this paper, a multinomial logistic model (Jobson, 1992) is used for modeling the conditional stratum probabilities given covariates. This choice was mainly motivated by the fact that this model is commonly used in medical research for modeling the relationship between a categorical response and covariates. The theoretical results developed here can be extended to the case of other link functions. In addition, the covariate X in model (3.1) is assumed to be time independent, for convenience. This assumption can be relaxed to accomodate time varying covariates, provided that appropriate regularity conditions are made.

3.6 Appendix. Proofs of Theorems

A.1 Proof of Theorem 3.1

For every $k \in \mathcal{K}$, define $\mathcal{I}_k^n = \{i \in \{1,\dots,n\} | \Delta_i R_i \mathbb{1}_{S_i=k} = 1 \text{ or } \Delta_i(1-R_i) = 1\}$, and let i_k^n denote the cardinality of $\mathcal{I}_k^n$. Let $i_\bullet^n = \sum_{k=1}^K i_k^n$. Consider the set of times $\{T_i, i \in \mathcal{I}_k^n\}$. Let $t_{(k,1)} < \ldots < t_{(k,i_k^n)}$ denote the ordered failure times in this set. For any given sample size n, the NPML estimation method consists in maximizing L_n in (3.2) over the approximating parameter space

$$\Theta_n = \left\{(\beta, \gamma, \Lambda_k\{t_{(k,j)}\}) : \beta \in \mathcal{B}; \gamma \in \mathcal{G}; \Lambda_k\{t_{(k,j)}\} \in [0,\infty), j = 1,\dots,i_k^n, k \in \mathcal{K}\right\}.$$

Suppose first that $\Lambda_k\{t_{(k,j)}\} \leq M < \infty$, for $j = 1,\dots,i_k^n$ and $k \in \mathcal{K}$. Since L_n is a continuous function of β, γ, and the $\Lambda_k\{t_{(k,j)}\}$'s on the compact set $\mathcal{B} \times \mathcal{G} \times [0,M]^{i_\bullet^n}$, L_n achieves its maximum on this set.
To show that a maximum of L_n exists on $\mathcal{B} \times \mathcal{G} \times [0,\infty)^{i_\bullet^n}$, we show that there exists a finite M such that for all $\theta^M = (\beta^M, \gamma^M, (\Lambda_k^M\{t_{(k,j)}\})_{j,k}) \in (\mathcal{B}\times\mathcal{G}\times[0,\infty)^{i_\bullet^n})\backslash(\mathcal{B}\times\mathcal{G}\times[0,M]^{i_\bullet^n})$, there exists a $\theta = (\beta, \gamma, (\Lambda_k\{t_{(k,j)}\})_{j,k}) \in \mathcal{B} \times \mathcal{G} \times [0,M]^{i_\bullet^n}$ such that $L_n(\theta) > L_n(\theta^M)$. A proof by contradiction is adopted for that purpose.
Assume that for all $M < \infty$, there exists $\theta^M \in (\mathcal{B} \times \mathcal{G} \times [0,\infty)^{i_\bullet^n})\backslash(\mathcal{B} \times \mathcal{G} \times [0,M]^{i_\bullet^n})$ such that for all $\theta \in \mathcal{B} \times \mathcal{G} \times [0,M]^{i_\bullet^n}$, $L_n(\theta) \leq L_n(\theta^M)$. It can be seen that L_n is bounded above by

$$K^n \prod_{i=1}^n \left[\prod_{k=1}^K \{c_3 \Lambda_k\{T_i\}\}^{\Delta_i R_i \mathbb{1}_{S_i=k}} \exp\left(-c_2 R_i \mathbb{1}_{S_i=k} \sum_{j=1}^{i_k^n} \Lambda_k\{t_{(k,j)}\} \mathbb{1}_{t_{(k,j)} \leq T_i} \right) \right].$$

If $\theta^M \in (\mathcal{B} \times \mathcal{G} \times [0,\infty)^{i_\bullet^n})\backslash(\mathcal{B} \times \mathcal{G} \times [0,M]^{i_\bullet^n})$, then there exists $l \in \mathcal{K}$ and $p \in \{1,\dots,i_l^n\}$ such that $\Lambda_l^M\{t_{(l,p)}\} > M$. By assumption (d), there exists at least one individual with indice i_M ($i_M \in \{1,\dots,n\}$) such that $\mathbb{1}_{S_{i_M}=l} = 1$, $Y_{i_M}(\tau) = 1$ (and therefore $t_{(l,p)} \leq T_{i_M} = \tau$), and $R_{i_M} = 1$. Hence

$$R_{i_M} \mathbb{1}_{S_{i_M}=l} \sum_{j=1}^{i_l^n} \Lambda_l^M\{t_{(l,j)}\} \mathbb{1}_{t_{(l,j)} \leq T_{i_M}} \to \infty \text{ as } M \to \infty.$$

It follows that the upper bound of $L_n(\theta^M)$ (and therefore $L_n(\theta^M)$ itself) can be made as close to 0 as desired by increasing M. This is the desired contradiction. ■

A.2 Proof of Theorem 3.3

We adapt the techniques developed by Murphy (1994), in order to prove consistency of our proposed estimator $\widehat{\theta}_n$. The proof essentially consists of three steps:

(i) for every $k \in \mathcal{K}$, we show that the sequence $\widehat{\Lambda}_{k,n}(\tau)$ is almost surely bounded as n goes to infinity,

(ii) we show that every subsequence of n contains a further subsequence along which the NPMLE $\widehat{\theta}_n$ converges,

(iii) we show that the limit of every convergent subsequence of $\widehat{\theta}_n$ is θ_0.

Proof of (i). Note first that for all $s \in [0,\tau]$ and $k \in \mathcal{K}$, $\frac{1}{n}\sum_{i=1}^n Q(\mathcal{O}_i,k,\widehat{\theta}_n)e^{\widehat{\beta}_n{}' X_i}Y_i(s) \geq c_2\frac{1}{n}\sum_{i=1}^n R_i\mathbb{1}_{S_i=k}Y_i(\tau)$. Moreover, $Q(\mathcal{O}_i,k,\widehat{\theta}_n)$ is bounded by 1. It follows that for all $k \in \mathcal{K}$,

$$0 \leq \widehat{\Lambda}_{k,n}(\tau) \leq \frac{1}{c_2}\int_0^\tau \frac{\mathrm{d}\bar{N}_n(s)}{\frac{1}{n}\sum_{i=1}^n R_i\mathbb{1}_{S_i=k}Y_i(\tau)} = \frac{\frac{1}{n}\sum_{i=1}^n \Delta_i}{c_2\frac{1}{n}\sum_{i=1}^n R_i\mathbb{1}_{S_i=k}Y_i(\tau)},$$

where $\bar{N}_n(s) = n^{-1}\sum_{i=1}^n N_i(s)$. Next, $\frac{1}{n}\sum_{i=1}^n R_i\mathbb{1}_{S_i=k}Y_i(\tau)$ converges almost surely to $P_{\theta_0}[R\mathbb{1}_{S=k}Y(\tau)] > c_4 > 0$ therefore, for each $k \in \mathcal{K}$, as n goes to infinity, $\widehat{\Lambda}_{k,n}(\tau)$ is bounded above almost surely by $\frac{1}{c_2c_4}$.

Proof of (ii). If (i) holds, by Helly's theorem (see Loève (1963), p179), every subsequence of n has a further subsequence along which $\widehat{\Lambda}_{1,n}$ converges weakly to some nondecreasing right-continuous function Λ_1^*, with probability 1. By successive extractions of sub-subsequences, we can further find a subsequence (say n_j) such that $\widehat{\Lambda}_{k,n_j}$ converges weakly to some nondecreasing right-continuous function Λ_k^*, for every $k \in \mathcal{K}$, with probability 1. By the compactness of $\mathcal{B}\times\mathcal{G}$, we can further find a subsequence of n_j (we shall still denote it by n_j for simplicity of notations) such that $\widehat{\Lambda}_{k,n_j}$ converges weakly to Λ_k^* (for every $k \in \mathcal{K}$) and $(\widehat{\beta}_{n_j},\widehat{\gamma}_{n_j})$ converges to some (β^*,γ^*), with probability 1. We now show that the Λ_k^*'s must be continuous on $[0,\tau]$.

Let ψ be any nonnegative, bounded, continuous function. Then, for any

given $k \in \mathcal{K}$,

$$\begin{aligned}\int_0^\tau \psi(s)\mathrm{d}\Lambda_k^*(s) &= \int_0^\tau \psi(s)\mathrm{d}\{\Lambda_k^*(s) - \widehat{\Lambda}_{k,n_j}(s)\} \\ &+ \int_0^\tau \psi(s) \left[\frac{1}{n_j}\sum_{l=1}^{n_j} Q(\mathcal{O}_l, k, \widehat{\theta}_{n_j}) e^{\widehat{\beta_{n_j}}' X_l} Y_l(s)\right]^{-1} \frac{1}{n_j}\sum_{i=1}^{n_j} Q(\mathcal{O}_i, k, \widehat{\theta}_{n_j})\mathrm{d}N_i(s) \\ &\leq \int_0^\tau \psi(s)\mathrm{d}\{\Lambda_k^*(s) - \widehat{\Lambda}_{k,n_j}(s)\} + \int_0^\tau \psi(s) \left[\frac{c_2}{n_j}\sum_{l=1}^{n_j} R_l \mathbb{1}_{S_l=k} Y_l(s)\right]^{-1} \mathrm{d}\bar{N}_{n_j}(s).\end{aligned}$$

By the Helly-Bray lemma (see Loève (1963), p180), $\int_0^\tau \psi(s)\mathrm{d}\{\Lambda_k^*(s) - \widehat{\Lambda}_{k,n_j}(s)\} \to 0$ as $j \to \infty$. Moreover, $\bar{N}_{n_j}(\cdot)$ and $\frac{1}{n_j}\sum_{l=1}^{n_j} R_l \mathbb{1}_{S_l=k} Y_l(\cdot)$ converge almost surely in supremum norm to

$$\sum_{k=1}^{K} \int_0^\cdot P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)\right] \mathrm{d}\Lambda_{k,0}(s) \text{ and } P_{\theta_0}\left[R\mathbb{1}_{S=k} Y(\cdot)\right]$$

respectively, where the latter term is bounded away from 0 on $s \in [0, \tau]$ by assumption (d). Thus, by applying the extended version of the Helly-Bray lemma (stated by Korsholm (1998) for example) to the second term on the right-hand side of the previous inequality, we get that

$$\begin{aligned}&\int_0^\tau \psi(s)\mathrm{d}\Lambda_k^*(s) \qquad (3.5)\\ &\quad \leq c_2 \int_0^\tau \psi(s) \left\{P_{\theta_0}\left[R\mathbb{1}_{S=k} Y(s)\right]\right\}^{-1} \sum_{k=1}^{K} P_{\theta_0}[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)]\lambda_{k,0}(s)\mathrm{d}s. \\ &\quad \leq \frac{c_2 c_3}{c_4} \sum_{k=1}^{K} \int_0^\tau \psi(s)\lambda_{k,0}(s)\mathrm{d}s.\end{aligned}$$

Suppose that Λ_k^* has discontinuities, and let ψ be close to 0 except at the jump points of Λ_k^*, where it is allowed to have high and thin peaks. While the right-hand side of inequality (3.5) should be close to 0 ($\lambda_{k,0}$ is continuous by assumption (b)), its left-hand side can be made arbitrarily large, yielding a contradiction. Thus Λ_k^* must be continuous ($k \in \mathcal{K}$). A second conclusion, arising from Dini's theorem, is that $\widehat{\Lambda}_{k,n_j}$ uniformly converges to Λ_k^* ($k \in \mathcal{K}$), with probability 1. To summarize: for any given subsequence of n, we have found a further subsequence n_j and an element $(\beta^*, \gamma^*, \Lambda_k^*, k \in \mathcal{K})$ such that $\|\widehat{\beta}_{n_j} - \beta^*\|$, $\|\widehat{\gamma}_{n_j} - \gamma^*\|$, and $\sup_{t\in[0,\tau]} |\widehat{\Lambda}_{k,n_j}(t) - \Lambda_k^*(t)|$ (for every $k \in \mathcal{K}$) converge to 0 almost surely.

Proof of (iii). To prove (iii), we first define random step functions

$$\overline{\Lambda}_{k,n}(t) = \int_0^t \sum_{i=1}^n \frac{Q(\mathcal{O}_i, k, \theta_0)}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \theta_0) \exp(\beta_0' X_j) Y_j(s)} \mathrm{d}N_i(s),\ 0 \le t \le \tau, k \in \mathcal{K},$$

and we show that for every $k \in \mathcal{K}$, $\overline{\Lambda}_{k,n}$ almost surely uniformly converges to $\Lambda_{k,0}$ on $[0,\tau]$. First, note that

$$\begin{aligned}
&\sup_{t\in[0,\tau]} \left| \overline{\Lambda}_{k,n}(t) - P_{\theta_0}\left[\frac{\Delta \mathbb{1}_{T\le t} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)\right] |_{s=T}} \right] \right| \\
&\quad \le \sup_{t\in[0,\tau]} \left| \frac{1}{n} \sum_{i=1}^n \Delta_i \mathbb{1}_{T_i \le t} Q(\mathcal{O}_i, k, \theta_0) \right. \\
&\qquad \left. \times \left\{ \frac{1}{\mathbb{P}_n\left[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s)\right]} - \frac{1}{P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)\right]} \right\}_{|_{s=T_i}} \right| \\
&\qquad + \sup_{t\in[0,\tau]} \left| (\mathbb{P}_n - P_{\theta_0}) \left[\frac{\Delta \mathbb{1}_{T\le t} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)\right] |_{s=T}} \right] \right| \\
&\quad \le \sup_{s\in[0,\tau]} \left| \frac{1}{\mathbb{P}_n\left[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s)\right]} - \frac{1}{P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)\right]} \right| \\
&\qquad + \sup_{t\in[0,\tau]} \left| (\mathbb{P}_n - P_{\theta_0}) \left[\frac{\Delta \mathbb{1}_{T\le t} Q(\mathcal{O}, k, \theta_0)}{P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)\right] |_{s=T}} \right] \right| \qquad (3.6)
\end{aligned}$$

The class $\{Y(s) : s \in [0,\tau]\}$ is Donsker and $Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X}$ is a bounded measurable function, hence $\{Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s) : s \in [0,\tau]\}$ is Donsker (Corollary 9.31, Kosorok (2008)), and therefore Glivenko-Cantelli. Moreover, $P_{\theta_0}[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s)] = P_{\theta_0}[P_{\theta_0}[\mathbb{1}_{S=k}|\mathcal{O}] e^{\beta_0' X} Y(s)] = P_{\theta_0}[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)]$. Thus

$$\sup_{s\in[0,\tau]} \left| \mathbb{P}_n\left[Q(\mathcal{O}, k, \theta_0) e^{\beta_0' X} Y(s) \right] - P_{\theta_0}\left[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s) \right] \right|$$

converges to 0 a.e. Next, $P_{\theta_0}[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)]$ is larger than $c_2 . P_{\theta_0}[\mathbb{1}_{S=k} Y(\tau)]$ and thus, by assumption (d), $P_{\theta_0}[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)] > 0$. It follows that the first term on the right-hand side of inequality (3.6) converges to 0 a.e.. Similar aguments show that the class $\{\Delta \mathbb{1}_{T\le t} Q(\mathcal{O}, k, \theta_0) / P_{\theta_0}[\mathbb{1}_{S=k} e^{\beta_0' X} Y(s)] |_{s=T} : t \in [0,\tau]\}$ is also a Glivenko-Cantelli class, and therefore $\overline{\Lambda}_{k,n}$ almost surely

uniformly converges to

$$P_{\theta_0}\left[\frac{\Delta \mathbb{1}_{T\leq t} Q(\mathcal{O},k,\theta_0)}{P_{\theta_0}\left[\mathbb{1}_{S=k}e^{\beta_0' X}Y(s)\right]\mid_{s=T}}\right].$$

Now, note that $\Lambda_{k,0}(t)=\int_0^t \frac{P_{\theta_0}[\mathbb{1}_{S=k}\mathrm{d}N(s)]}{P_{\theta_0}[\mathbb{1}_{S=k}e^{\beta_0' X}Y(s)]}$, which can be reexpressed as

$$\Lambda_{k,0}(t)=\frac{P_{\theta_0}\left[\mathbb{1}_{S=k}\Delta\mathbb{1}_{T\leq t}\right]}{P_{\theta_0}\left[\mathbb{1}_{S=k}e^{\beta_0' X}Y(s)\right]\mid_{s=T}}=P_{\theta_0}\left[\frac{\Delta \mathbb{1}_{T\leq t} Q(\mathcal{O},k,\theta_0)}{P_{\theta_0}\left[\mathbb{1}_{S=k}e^{\beta_0' X}Y(s)\right]\mid_{s=T}}\right].$$

Thus $\overline{\Lambda}_{k,n}$ almost surely uniformly converges to $\Lambda_{k,0}$ on $[0,\tau]$.

Next, using somewhat standard arguments (see Parner (1998) for example), we can show that $0\leq n_j^{-1}\{\log L_{n_j}(\widehat{\theta}_{n_j})-\log L_{n_j}(\overline{\theta}_{n_j})\}$ converges to the negative Kullback-Leibler information $P_{\theta_0}[\log(L_1(\theta^*)/L_1(\theta_0))]$. Thus, the Kullback-Leibler information must be zero, and it follows that with probability 1, $L_1(\theta^*)=L_1(\theta_0)$. The proof of consistency is completed if we show that this equality implies $\theta^*=\theta_0$. For that purpose, consider $L_1(\theta^*)=L_1(\theta_0)$ under $\Delta=1$, $R=1$, and $\mathbb{1}_{S=k}=1$ (for each $k\in\mathcal{K}$ in turn). Note that this is possible by assumption (e). This yields the following equation for almost all $t\in[0,\tau]$, $\|x\|<c_1$, $\|w\|<c_1$:

$$\log\frac{\lambda_k^*(t)}{\lambda_{k,0}(t)}+(\beta^*-\beta_0)'x-\Lambda_k^*(t)e^{\beta^{*\prime}x}+\Lambda_{k,0}(t)e^{\beta_0' x}+\log\frac{\pi_k^*(w)}{\pi_{k,0}(w)}=0.$$

This equation is analogous to equation (A.2) in Chen & Little (1999). The rest of the proof of identifiability thus proceeds along the same lines as the proof of Lemma A.1.1 in Chen & Little (1999), and is omitted.
Hence, for any given subsequence of n, we have found a further subsequence n_j such that $\|\widehat{\beta}_{n_j}-\beta_0\|$, $\|\widehat{\gamma}_{n_j}-\gamma_0\|$, and $\sup_{t\in[0,\tau]}|\widehat{\Lambda}_{k,n_j}(t)-\Lambda_{k,0}(t)|$ (for every $k\in\mathcal{K}$) converge to 0 almost surely, which implies that the sequence of NPMLE $\widehat{\theta}_n$ converges almost surely to θ_0. ■

A.3 Proof of Theorem 3.4

The proof of Theorem 3.4 uses similar arguments as the proof of Theorem 3 of Fang et al. (2005), so we only highlight the parts that are different. We need a few lemmas before presenting the proof.

Lemma 3.7. *Let $h\in H$. Then the following holds:* $P_{\theta_0}\left[S_1(\theta_0)(h)\right]=P_{\theta_0}[h_\beta' S_\beta(\theta_0)+h_\gamma' S_\gamma(\theta_0)+\sum_{k=1}^K S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})]=0$.

Proof. From the properties of the conditional expectation, we first note that

$$\begin{aligned} P_{\theta_0}\left[S_\beta(\theta_0)\right] &= P_{\theta_0}\left[\Delta X - \sum_{k=1}^{K} Q(\mathcal{O},k,\theta_0) X \exp(\beta_0' X)\Lambda_{k,0}(T)\right] \\ &= P_{\theta_0}\left[\Delta X - \sum_{k=1}^{K} \mathbb{1}_{S=k} X \exp(\beta_0' X)\Lambda_{k,0}(T)\right] \\ &= P_{\theta_0}\left[XM(\tau)\right], \end{aligned}$$

where $M(t) = N(t) - \int_0^t \sum_{k=1}^K \mathbb{1}_{S=k} e^{\beta_0' X} Y(u) \mathrm{d}\Lambda_{k,0}(u)$ is the counting process martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N(u), \mathbb{1}_{C\leq u}, X, S, W : 0 \leq u \leq t\}$. X is bounded and $\mathcal{F}_t$-measurable, hence it follows that $P_{\theta_0}[S_\beta(\theta_0)] = 0$. Using similar arguments, we can verify that $P_{\theta_0}[S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})] = 0$, $k \in \mathcal{K}$. Finally, for $k = 1, \ldots, K-1$,

$$\begin{aligned} P_{\theta_0}\left[S_{\gamma_k}(\theta_0)\right] &= P_{\theta_0}\left[W\left[Q(\mathcal{O},k,\theta_0) - \pi_{k,\gamma_0}(W)\right]\right] \\ &= P_{\theta_0}\left[W P_{\theta_0}\left[\mathbb{1}_{S=k} - \pi_{k,\gamma_0}(W)|W\right]\right] \\ &= 0. \end{aligned}$$

Combining these results yields that $P_{\theta_0}\left[S_1(\theta_0)(h)\right] = 0$. ■

We now come to the continuous invertibility of the continuous linear operator σ defined in Section 3.4.

Lemma 3.8. *The operator σ is continuously invertible.*

Proof. Since H is a Banach space, to prove that σ is continuously invertible, it is sufficient to prove that σ is one-to-one and that it can be written as the sum of a bounded linear operator with a bounded inverse and a compact operator (Lemma 25.93 of van der Vaart (1998)).

Define the linear operator $A(h) = (h_\beta, h_\gamma, P_{\theta_0}\left[\mathbb{1}_{S=k}\phi(\cdot,\mathcal{O},k,\theta_0)\right] h_{\Lambda_k}(\cdot); k \in \mathcal{K})$, this is a bounded operator due to the boundedness of X. Moreover, for all $u \in [0,\tau]$ and $k \in \mathcal{K}$, $P_{\theta_0}\left[\mathbb{1}_{S=k}\phi(u,\mathcal{O},k,\theta_0)\right] \geq c_2 c_4 > 0$ by assumptions (c) and (d). This implies that A is invertible with bounded inverse $A^{-1}(h) = (h_\beta, h_\gamma, P_{\theta_0}\left[\mathbb{1}_{S=k}\phi(\cdot,\mathcal{O},k,\theta_0)\right]^{-1} h_{\Lambda_k}(\cdot); k \in \mathcal{K})$. The operator $\sigma - A$ can be shown to be compact by using the same techniques as in Lu (2008) for example.

To prove that σ is one-to-one, let $h \in H$ such that $\sigma(h) = 0$. If $\sigma(h) = 0$, $P_{\theta_0}\left[S_1(\theta_0)(h)^2\right] = 0$, and therefore $S_1(\theta_0)(h) = 0$ almost surely. Let $j \in \mathcal{K}$. By assumption (e), for almost every $t \in [0, \tau]$, $\|x\| \leq c_1$, and $\|w\| \leq c_1$, there is a non-negligible set $\Omega_{t,x,w} \subseteq \Omega$ such that $\Delta(\omega) = 1$, $R(\omega) = 1$, and $\mathbb{1}_{S(\omega)=j} = 1$ when $\omega \in \Omega_{t,x,w}$. If $S_1(\theta_0)(h) = 0$ almost surely, then in particular, for almost every $t \in [0, \tau]$, $\|x\| \leq c_1$, and $\|w\| \leq c_1$, $S_1(\theta_0)(h) = 0$ when $\omega \in \Omega_{t,x,w}$, which yields the following equation:

$$h_{\Lambda_j}(t) + h'_\beta x + w' h_{\gamma_j} - \sum_{k=1}^{K-1} w' h_{\gamma_k} \pi_{k,\gamma_0}(w) - e^{\beta'_0 x}\left[\int_0^t h_{\Lambda_j}(s)\mathrm{d}\Lambda_{j,0}(s) + h'_\beta x \Lambda_{j,0}(t)\right] = 0 \tag{3.7}$$

with $h_{\gamma_j} = 0$ when $j = K$. Then, by choosing t arbitrarily close to 0, and since $\Lambda_{j,0}$ is continuous, $\Lambda_{j,0}(0) = 0$, and h_{Λ_j} is continuous from the right at 0, we get that

$$h_{\Lambda_j}(0) + h'_\beta x + w' h_{\gamma_j} - \sum_{k=1}^{K-1} w' h_{\gamma_k} \pi_{k,\gamma_0}(w) = 0. \tag{3.8}$$

Taking the difference (3.7)-(3.8) yields that

$$h_{\Lambda_j}(t) - h_{\Lambda_j}(0) = e^{\beta'_0 x}\left[\int_0^t h_{\Lambda_j}(s)\mathrm{d}\Lambda_{j,0}(s) + h'_\beta x \Lambda_{j,0}(t)\right] \tag{3.9}$$

for almost every $t \in [0, \tau]$ and $\|x\| \leq c_1$. Since $\Lambda_{j,0}$ is increasing (by assumption (b)), for every $t > 0$, $\Lambda_{j,0}(t) > \Lambda_{j,0}(0) = 0$ and therefore (3.9) can be rewritten as

$$\frac{h_{\Lambda_j}(t) - h_{\Lambda_j}(0)}{\Lambda_{j,0}(t)} = e^{\beta'_0 x}\left[r(t) + h'_\beta x\right], \tag{3.10}$$

where $r(t) = \int_0^t h_{\Lambda_j}(s)\, d\Lambda_{j,0}(s)/\Lambda_{j,0}(t)$. Consider first the case where $\beta_0 = 0$. Since the left-hand side of (3.10) does not depend on x, h_β must equal 0. Next, consider the case where $\beta_0 \neq 0$. Let $t_1, t_2 > 0$. Then $e^{\beta'_0 x}[r(t_1) - r(t_2)]$ does not depend on x. Since the covariance matrix of X is positive definite, we can find two distinct values x_1 and x_2 of X such that $e^{\beta'_0 x_1}[r(t_1) - r(t_2)] = e^{\beta'_0 x_2}[r(t_1) - r(t_2)]$. This implies that $r(t_1) = r(t_2)$, from which we deduce that $h_{\Lambda_j}(t)$ has to be constant (say, equal to α) for almost every $t \in (0, \tau]$. From (3.10), we then deduce that $h_{\Lambda_j}(0) = \alpha$, which further implies that $h_\beta = 0$, $\alpha = 0$, and thus $h_{\Lambda_j}(t) = 0$ for almost every $t \in [0, \tau]$ ($j \in \mathcal{K}$). This, together with (3.8) implies that $h_{\gamma_j} = 0$, $j \in \mathcal{K}$.

Let $k = K$. Then $\sigma_{\Lambda_K}(h)(u) = P_{\theta_0}\left[\mathbb{1}_{S=K}\phi(u, \mathcal{O}, K, \theta_0)\right] h_{\Lambda_K}(u) = 0$ for all $u \in [0, \tau]$ since $h_\beta = 0$ and $h_\gamma = 0$. By assumptions (c) and (d), for every $u \in [0, \tau]$ and $k \in \mathcal{K}$,

$$\begin{aligned} P_{\theta_0}\left[\mathbb{1}_{S=k}\phi(u, \mathcal{O}, k, \theta_0)\right] &= P_{\theta_0}\left[\mathbb{1}_{S=k}Y(u)Q(\mathcal{O}, k, \theta_0)e^{\beta_0' X}\right] \\ &\geq P_{\theta_0}\left[\mathbb{1}_{S=k}Y(\tau)Re^{\beta_0' X}\right] > 0, \end{aligned}$$

hence we conclude that h_{Λ_K} is identically equal to 0 on $[0, \tau]$. Next, considering $\sigma_{\Lambda_{K-1}}(h)(u) = 0$ with $h_\beta = 0$, $h_\gamma = 0$ and $h_{\Lambda_K} = 0$, we conclude similarly that $h_{\Lambda_{K-1}}(u) = 0$ for every $u \in [0, \tau]$. It follows that h_{Λ_j} is identically equal to 0 on $[0, \tau]$ for every $j \in \mathcal{K}$. Therefore, σ is one-to-one. ■

We now turn to the proof of Theorem 3.4 itself. Similar to Fang et al. (2005), we get that

$$\begin{aligned} \sqrt{n}\left(h_\beta'(\widehat{\beta}_n - \beta_0) + h_\gamma'(\widehat{\gamma}_n - \gamma_0) + \sum_{k=1}^{K}\int_0^\tau h_{\Lambda_k}(s)\mathrm{d}(\widehat{\Lambda}_{k,n} - \Lambda_{k,0})(s)\right) \\ = \sqrt{n}\left(S_n(\theta_0)(\sigma^{-1}(h)) - P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}(h))\right]\right) + o_p(1), \end{aligned}$$

where S_n is given by (3.3). Consider the subset $\{(h_\beta, 0, 0; k \in \mathcal{K}) | h_\beta \in \mathbb{R}^p\} \subset H$ and let $\widetilde{h}$ be an element of this subset. Setting $h = \widetilde{h}$ in the above equation yields

$$\sqrt{n}h_\beta'(\widehat{\beta}_n - \beta_0) = \sqrt{n}\left(S_n(\theta_0)(\sigma^{-1}(\widetilde{h})) - P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))\right]\right) + o_p(1) \quad (3.11)$$

By Lemma 3.7, the central limit theorem, and Slutsky's theorem, $\sqrt{n}h_\beta'(\widehat{\beta}_n - \beta_0)$ is asymptotically normal with mean 0 and variance $P_{\theta_0}[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))^2]$. If $h \in H$, direct calculation yields

$$\begin{aligned} S_1(\theta_0)(h)^2 &= h_\beta' S_\beta(\theta_0) S_\beta(\theta_0)' h_\beta + h_\gamma' S_\gamma(\theta_0) S_\gamma(\theta_0)' h_\gamma + 2h_\beta' S_\beta(\theta_0) S_\gamma(\theta_0)' h_\gamma \\ &\quad + 2h_\beta' S_\beta(\theta_0)\left(\sum_{k=1}^{K} S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})\right) + 2h_\gamma' S_\gamma(\theta_0)\left(\sum_{k=1}^{K} S_{\Lambda_k}(\theta_0)(h_{\Lambda_k})\right) \\ &\quad + \sum_{k=1}^{K}\left(Q(\mathcal{O}, k, \theta_0)\left[h_{\Lambda_k}(T)\Delta - \exp(\beta_0' X)\int_0^T h_{\Lambda_k}(s)\mathrm{d}\Lambda_{k,0}(s)\right]\right)^2 \\ &\quad + 2\sum_{k=1}^{K}\sum_{j>k}\left(Q(\mathcal{O}, k, \theta_0)\left[h_{\Lambda_k}(T)\Delta - \exp(\beta_0' X)\int_0^T h_{\Lambda_k}(s)\mathrm{d}\Lambda_{k,0}(s)\right]\right) \\ &\quad \times \left(Q(\mathcal{O}, j, \theta_0)\left[h_{\Lambda_j}(T)\Delta - \exp(\beta_0' X)\int_0^T h_{\Lambda_j}(s)\mathrm{d}\Lambda_{j,0}(s)\right]\right). \end{aligned}$$

Taking expectation followed by some tedious algebraic manipulations and re-arrangement of terms yield that

$$P_{\theta_0}\left[S_1(\theta_0)(h)^2\right] = h'_\beta\sigma_\beta(h) + h'_\gamma\sigma_\gamma(h) + \sum_{k=1}^{K}\int_0^\tau \sigma_{\Lambda_k}(h)(u)h_{\Lambda_k}(u)\mathrm{d}\Lambda_{k,0}(u).$$

Therefore

$$\begin{aligned} P_{\theta_0}\left[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))^2\right] &= \sigma_\beta^{-1}(\widetilde{h})'\sigma_\beta(\sigma^{-1}(\widetilde{h})) + \sigma_\gamma^{-1}(\widetilde{h})'\sigma_\gamma(\sigma^{-1}(\widetilde{h})) \\ &\qquad + \sum_{k=1}^{K}\int_0^\tau \sigma_{\Lambda_k}^{-1}(\widetilde{h})(u)\sigma_{\Lambda_k}(\sigma^{-1}(\widetilde{h}))(u)\mathrm{d}\Lambda_{k,0}(u) \\ &= h'_\beta\sigma_\beta^{-1}(\widetilde{h}), \end{aligned}$$

where the last equality comes from the fact that

$$\sigma(\sigma^{-1}(\widetilde{h})) = (\sigma_\beta(\sigma^{-1}(\widetilde{h})), \sigma_\gamma(\sigma^{-1}(\widetilde{h})), \sigma_{\Lambda_k}(\sigma^{-1}(\widetilde{h})); k \in \mathcal{K}) = \widetilde{h}.$$

Now, recall that the linear map $\widetilde{\sigma}_\beta^{-1} : \mathbb{R}^p \to \mathbb{R}^p$ was defined in Section 3.4 as a restricted version of σ_β^{-1}, by setting $\widetilde{\sigma}_\beta^{-1}(h_\beta) = \sigma_\beta^{-1}(\widetilde{h})$ for any $\widetilde{h}$ of the form $(h_\beta, 0, 0; k \in \mathcal{K})$. Let $\{e_1, \ldots, e_p\}$ be the canonical basis of $\mathbb{R}^p$ and $\Sigma_\beta = (\widetilde{\sigma}_\beta^{-1}(e_1), \ldots, \widetilde{\sigma}_\beta^{-1}(e_p))$. Then for any $h_\beta \in \mathbb{R}^p$, we have $\widetilde{\sigma}_\beta^{-1}(h_\beta) = \Sigma_\beta h_\beta$ and thus $P_{\theta_0}[S_1(\theta_0)(\sigma^{-1}(\widetilde{h}))^2] = h'_\beta\Sigma_\beta h_\beta$. Hence, for every $h_\beta \in \mathbb{R}^p$, $\sqrt{n}h'_\beta(\widehat{\beta}_n - \beta_0)$ converges in distribution to $\mathcal{N}(0, h'_\beta\Sigma_\beta h_\beta)$. By the Cramér-Wold device, $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges in distribution to $\mathcal{N}(0, \Sigma_\beta)$.

Now, for $j = 1\ldots, p$, denote $\widetilde{h}_j = (e_j, 0, 0; k \in \mathcal{K})$. Letting $h = \widetilde{h}_j$ for each $j = 1\ldots, p$ in turn in (3.11) yields

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l_\beta(\mathcal{O}_i, \theta_0) + o_p(1),$$

where

$$l_\beta(\mathcal{O}, \theta_0) = \Sigma_\beta S_\beta(\theta_0) + \Xi S_\gamma(\theta_0) + \sum_{k=1}^{K} S_{\Lambda_k}(\theta_0)(\Xi^*),$$

Ξ and Ξ^* are $(p \times q)$ and $(p \times 1)$ matrices respectively defined by

$$\Xi = \begin{pmatrix} \sigma_\gamma^{-1}(\widetilde{h}_1)' \\ \vdots \\ \sigma_\gamma^{-1}(\widetilde{h}_p)' \end{pmatrix} \quad \text{and} \quad \Xi^* = \begin{pmatrix} \sigma_{\Lambda_k}^{-1}(\widetilde{h}_1) \\ \vdots \\ \sigma_{\Lambda_k}^{-1}(\widetilde{h}_p) \end{pmatrix},$$

and $S_{\Lambda_k}(\theta_0)$ is applied componentwise to Ξ^*. Thus $\widehat{\beta}_n$ is an asymptotically linear estimator for β_0, and its influence function $l_\beta(\mathcal{O}, \theta_0)$ belongs to the tangent space spanned by the score functions. It follows that $l_\beta(\mathcal{O}, \theta_0)$ is the efficient influence function for β_0, and that $\widehat{\beta}_n$ is semiparametrically efficient (see Bickel et al. (1993) or Tsiatis (2006)). ■

A.4 Proof of Theorem 3.5

The proof of asymptotic normality of $\sqrt{n}(\widehat{\gamma}_n - \gamma_0)$ proceeds along the same line as for $\sqrt{n}(\widehat{\beta}_n - \beta_0)$, and is therefore omitted.

Next, for any $t \in (0, \tau)$ and $j \in \mathcal{K}$, the asymptotic normality of $\sqrt{n}(\widehat{\Lambda}_{j,n}(t) - \Lambda_{j,0}(t))$ can be proved by using a similar argument with $\widetilde{h}$ replaced by $h_{(j,t)} = (h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$, where $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_j}(\cdot) = \mathbb{1}_{\cdot \leq t}$ $(t \in (0, \tau)$ and $j \in \mathcal{K})$, and $h_{\Lambda_k} = 0$ for every $k \in \mathcal{K}, k \neq j$. Details are omitted. ■

A.5 Proof of Theorem 3.6

The proof of Theorem 3.6 parallels the proof of Theorem 3 in Parner (1998) and thus, will be kept brief. Let $\widehat{\sigma}_n = (\widehat{\sigma}_{\beta,n}, \widehat{\sigma}_{\gamma,n}, \widehat{\sigma}_{\Lambda_k,n}; k \in \mathcal{K})$ be defined as σ with all of the θ_0 and P_{θ_0} replaced by $\widehat{\theta}_n$ and $\mathbb{P}_n$ respectively. Similar to the proof of Theorem 3 in Parner (1998), it can be shown that $\widehat{\sigma}_n$ converges in probability to σ uniformly over H and that its inverse $\widehat{\sigma}_n^{-1} = (\widehat{\sigma}_{\beta,n}^{-1}, \widehat{\sigma}_{\gamma,n}^{-1}, \widehat{\sigma}_{\Lambda_k,n}^{-1}; k \in \mathcal{K})$ is such that $\widehat{\sigma}_n^{-1}(h)$ converges to $\sigma^{-1}(h)$ in probability.

For every h_β, the asymptotic variance of $\sqrt{n}h'_\beta(\widehat{\beta}_n - \beta_0)$ is $h'_\beta \sigma_\beta^{-1}((h_\beta, 0, 0; k \in \mathcal{K}))$, which is consistently estimated by $h'_\beta \widehat{\sigma}_{\beta,n}^{-1}((h_\beta, 0, 0; k \in \mathcal{K}))$. Let $h_n = (h_{\beta,n}, h_{\gamma,n}, h_{\Lambda_k,n}; k \in \mathcal{K}) = \widehat{\sigma}_n^{-1}((h_\beta, 0, 0; k \in \mathcal{K}))$. Then $\widehat{\sigma}_n(h_n) = (h_\beta, 0, 0; k \in \mathcal{K})$, or

$$\begin{cases} \widehat{\sigma}_{\beta,n}(h_n) = h_\beta \\ \widehat{\sigma}_{\gamma,n}(h_n) = 0 \\ \widehat{\sigma}_{\Lambda_k,n}(h_n)(u) = 0, \quad k \in \mathcal{K}, \quad u \in [0, \tau]. \end{cases} \tag{3.12}$$

In particular, letting $u = T_1, \ldots, T_n$ in (3.12) yields the following system of equations:

$$\mathbb{A}_n \begin{pmatrix} h_{\beta,n} \\ h_{\gamma,n} \\ h_{\Lambda,n} \end{pmatrix} = \begin{pmatrix} h_\beta \\ 0_q \\ 0_{Kn} \end{pmatrix}, \tag{3.13}$$

where $h_{\Lambda,n} = (h_{\Lambda_1,n}(T_1), \ldots, h_{\Lambda_1,n}(T_n), \ldots, h_{\Lambda_K,n}(T_1), \ldots, h_{\Lambda_K,n}(T_n))'$, and $\mathbb{A}_n$ is defined by (3.4). Some simple algebra on (3.13) yields that $h_{\beta,n} = \widehat{\Sigma}_{\beta,n} h_\beta$ where $\widehat{\Sigma}_{\beta,n}$ is defined in Section 3.4, and therefore $h'_\beta \widehat{\Sigma}_{\beta,n} h_\beta$ is a consistent estimator of the asymptotic variance of $\sqrt{n} h'_\beta(\widehat{\beta}_n - \beta_0)$ for every h_β. We conclude that $\widehat{\Sigma}_{\beta,n}$ converges in probability to Σ_β. The consistency of $\widehat{\Sigma}_{\gamma,n}$ proceeds along the same lines and is therefore omitted.

We now turn to the estimation of the asymptotic variance of $\widehat{\Lambda}_{j,n}(t)$, for $t \in (0,\tau)$ and $j \in \mathcal{K}$. By the dominated convergence theorem and the consistency of $\widehat{\sigma}_n^{-1}$,

$$\int_0^t \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(u) \mathrm{d}\widehat{\Lambda}_{j,n}(u)$$

converges to $v_j^2(t) = \int_0^t \sigma^{-1}_{\Lambda_j}(h_{(j,t)})(u)\, d\Lambda_{j,0}(u)$, where we recall that $h_{(j,t)}$ is the element $(h_\beta, h_\gamma, h_{\Lambda_k}; k \in \mathcal{K})$ such that $h_\beta = 0$, $h_\gamma = 0$, $h_{\Lambda_j}(\cdot) = \mathbb{1}_{\cdot \le t}$ for some $t \in (0,\tau)$ and $j \in \mathcal{K}$, and $h_{\Lambda_k} = 0$ for every $k \in \mathcal{K}, k \neq j$. Letting $\widetilde{h}_n = (\widetilde{h}_{\beta,n}, \widetilde{h}_{\gamma,n}, \widetilde{h}_{\Lambda_k,n}; k \in \mathcal{K}) = \widehat{\sigma}_n^{-1}(h_{(j,t)})$, we get that $\widehat{\sigma}_n(\widetilde{h}_n) = h_{(j,t)}$ or:

$$\left\{ \begin{array}{l} \widehat{\sigma}_{\beta,n}(\widetilde{h}_n) = 0 \\ \widehat{\sigma}_{\gamma,n}(\widetilde{h}_n) = 0 \\ \widehat{\sigma}_{\Lambda_j,n}(\widetilde{h}_n)(u) = \mathbb{1}_{u \le t}, \quad u \in [0,\tau] \\ \widehat{\sigma}_{\Lambda_k,n}(\widetilde{h}_n)(u) = 0, \quad k \in \mathcal{K}, \quad k \neq j, \quad u \in [0,\tau]. \end{array} \right. \tag{3.14}$$

In particular, letting $u = T_1, \ldots, T_n$ in (3.14) yields the system of equations

$$\mathbb{A}_n \left(\begin{array}{c} \widetilde{h}_{\beta,n} \\ \widetilde{h}_{\gamma,n} \\ \widetilde{h}_{\Lambda,n} \end{array} \right) = \left(\begin{array}{c} 0_p \\ 0_q \\ U^n_{(j,t)} \end{array} \right),$$

with the notations $\widetilde{h}_{\Lambda,n} = (\widetilde{h}_{\Lambda_1,n}(T_1), \ldots, \widetilde{h}_{\Lambda_1,n}(T_n), \ldots, \widetilde{h}_{\Lambda_K,n}(T_1), \ldots, \widetilde{h}_{\Lambda_K,n}(T_n))'$ and $U^n_{(j,t)} = (0'_{(j-1)n}, \mathbb{1}_{T_1 \le t}, \ldots, \mathbb{1}_{T_n \le t}, 0'_{(K-j)n})'$. Similar algebra as above yields

$$\widetilde{h}_{\Lambda,n} = \widehat{\Sigma}_{\Lambda,n} U^n_{(j,t)},$$

where $\widehat{\Sigma}_{\Lambda,n}$ is defined in Section 3.4. Now, $\int_0^t \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(u) \mathrm{d}\widehat{\Lambda}_{j,n}(u)$ verifies

$$\begin{aligned} \int_0^t \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(u) \mathrm{d}\widehat{\Lambda}_{j,n}(u) &= \sum_{i=1}^n \widehat{\sigma}^{-1}_{\Lambda_j,n}(h_{(j,t)})(T_i) \widehat{\Delta\Lambda_{j,n}}(T_i) \mathbb{1}_{T_i \le t} \\ &= \widehat{\Xi}^{n\prime}_{(j,t)} \widetilde{h}_{\Lambda,n}, \end{aligned}$$

where $\widehat{\Xi}^n_{(j,t)} = \left(0'_{(j-1)n}, \widehat{\Delta\Lambda_{j,n}}(T_1)\mathbb{1}_{T_1\leq t}, \ldots, \widehat{\Delta\Lambda_{j,n}}(T_n)\mathbb{1}_{T_n\leq t}, 0'_{(K-j)n}\right)'$. It follows that $\widehat{\Xi}^{n'}_{(j,t)}\widehat{\Sigma}_{\Lambda,n}U^n_{(j,t)}$ is a consistent estimator of $v_j^2(t)$, which concludes the proof. ■

Chapitre 4

Utilisation de l'algorithme EM dans le modèle de Cox stratifié avec strates aléatoirement manquantes

Sommaire

Ce chapitre traite de l'algorithme Espérance-Maximisation, souvent connu sous le nom d'algorithme EM. Il s'agit d'un algorithme itératif, utilisé dans l'estimation du maximum de vraisemblance pour des problèmes de données incomplètes, qui fut introduit par Dempster et al. (1977). Nous l'appliquons ici dans le cadre du modèle de Cox stratifié avec strates aléatoirement manquantes, comme cela a été introduit dans le chapitre 2.

Nous décrivons tout d'abord la démarche générale de l'algorithme EM, puis nous présentons son application à notre modèle, afin d'illustrer les résultats obtenus dans le chapitre 3. L'algorithme de Newton-Raphson, permettant d'obtenir une estimation du paramètre de régression à chaque itération de l'algorithme EM, sera également développé.

4.1 Description de l'algorithme EM pour données manquantes

Nous rappelons dans cette section le principe de l'algorithme Espérance-Maximisation (algorithme EM).

Cet algorithme est une approche utile dans les problèmes de données incomplètes, qui permet l'implémentation itérative d'estimateurs du maximum de vraisemblance. L'idée de base de l'algorithme est d'associer au problème de *données incomplètes* un problème de *données complètes* pour lequel l'estimation du maximum de vraisemblance est plus aisé à implémenter. Supposons que l'on observe incomplètement un vecteur aléatoire X correspondant au vecteur de données complètes x. Notons $f_X(.;\theta)$ la densité du vecteur aléatoire X, où $\theta \in \Theta \subseteq \mathbb{R}^p$ est un vecteur inconnu de paramètres. Soit Y le vecteur aléatoire de densité $f_Y(.;\theta)$ correspondant aux données réellement observées y. Ainsi, x est considéré comme la concaténation des données incomplètes y et de données manquantes, que l'on peut noter z. Notons $f_{X|Y}(.|Y;\theta)$ la densité conditionnelle de X sachant Y et θ_0 la vraie valeur du paramètre θ.

Si X était observable, la log-vraisemblance complète du modèle serait donnée par

$$l^{(c)}(\theta) = \ln(f_X(x,\theta)), \tag{4.1}$$

et l'estimateur du maximum de vraisemblance de θ serait obtenu comme solution du problème de maximisation de cette expression. Mais dans le cadre de l'observation des données incomplètes Y, la log-vraisemblance observée à maximiser est donnée par

$$l^{(o)}(\theta) = \ln(f_Y(y,\theta)). \tag{4.2}$$

L'algorithme EM consiste à résoudre indirectement le problème de la maximisation de (4.2) en procédant itérativement sur la log-vraisemblance des données complètes (4.1). Comme elle n'est pas observable, elle est remplacée par son espérance conditionnée par les observations y, en utilisant la valeur courante pour θ. Plus précisément, définissons pour θ et ϑ deux valeurs de paramètres de Θ, la fonction D par

$$D(\theta,\vartheta) = \mathbb{E}_\vartheta[l^{(c)}(\theta)|Y=y],$$

alors cette espérance conditionnelle sera calculée avec une valeur courante du paramètre ϑ, et maximisée en θ. Cela conduit à envisager un algorithme

itératif qui, partant d'une valeur initiale $\theta^{(0)}$, générera une suite de valeurs $(\theta^{(q)})_{q\in\mathbb{N}}$ du paramètre θ telle que

$$\theta^{(q+1)} \in argmax_{\theta\in\Theta} D(\theta, \theta^{(q)}). \tag{4.3}$$

Ainsi, pour la $(q+1)$-ème itération de l'algorithme, la valeur courante du paramètre θ étant notée $\theta^{(q)}$, les deux étapes suivantes se succèdent :

- **Etape de calcul de l'espérance (étape E)** : calcul de l'espérance de la log-vraisemblance des données complètes conditionnée par les données réellement observées, sous la loi donnée par le paramètre courant $\theta^{(q)}$,
$$D(\theta, \theta^{(q)}).$$
- **Etape de maximisation (étape M)** : actualisation du paramètre courant comme une des valeurs de Θ maximisant la quantité calculée dans l'étape E
$$\theta^{(q+1)} \in argmax_{\theta\in\Theta} D(\theta, \theta^{(q)}).$$

Ces deux étapes, ayant donné leur nom à l'algorithme, sont répétées alternativement jusqu'à ce qu'un critère de convergence soit vérifié, par exemple jusqu'à ce que la différence $|l^{(o)}(\theta^{(q+1)}) - l^{(o)}(\theta^{(q)})|$ (ou plus simplement $||\theta^{(q+1)} - \theta^{(q)}||_{\mathbb{R}^p}$) soit arbitrairement petite.

Dempster et al. (1977) montrent que la fonction log-vraisemblance observée augmente à chaque itération EM : pour tout $q \in \mathbb{N}$,

$$l^{(o)}(\theta^{(q+1)}) \geq l^{(o)}(\theta^{(q)}),$$

ce qui permet de garantir à l'utilisateur que l'algorithme progresse dans la bonne direction des valeurs croissantes de la log-vraisemblance. Le lecteur intéressé par plus de précisions sur l'algorithme EM ainsi qu'à ses multiples applications pourra se référer à Dempster et al. (1977) ainsi qu'au traitement détaillé de McLachlan & Krishnan (1997).

4.2 Application au modèle de Cox stratifié avec strates aléatoirement manquantes

Nous appliquons maintenant l'algorithme EM à notre problème de strates aléatoirement manquantes dans le modèle de Cox stratifié afin de déterminer un estimateur du maximum de vraisemblance du paramètre de régression β.

Nous avons introduit les notations ainsi que les hypothèses de notre modèle au chapitre 2. Les données réellement observées de notre étude sont les

$$\mathcal{O} = (\mathcal{O}_i)_{i\in\{1,\dots,n\}} = (T_i, \Delta_i, X_i, W_i, R_i, R_i S_i)_{i\in\{1,\dots,n\}},$$

par opposition aux données

$$(T_i, \Delta_i, X_i, W_i, S_i)_{i\in\{1,\dots,n\}},$$

observées incomplètement.

Déterminons la log-vraisemblance associée à ces données complètes. La vraisemblance de cet échantillon est

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_{T_i,\Delta_i,X_i,W_i,S_i} \\ &= \prod_{i=1}^{n} f_{T_i,\Delta_i|X_i,S_i} P_{S_i|X_i,W_i} f_{X_i,W_i}. \end{aligned} \tag{4.4}$$

Comme la loi de S conditionnelle à (X, W) est entièrement déterminée par la donnée de W et la censure étant dans notre modèle non informative, nous obtenons

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_{T_i,\Delta_i|X_i,S_i} \pi_{S_i,\gamma}(W_i) f_{X_i,W_i} \\ &= \prod_{i=1}^{n}\prod_{k=1}^{K} \left(f_{T_i,\Delta_i|X_i,S_i=k} \pi_{k,\gamma}(W_i)\right)^{\mathbb{1}_{S_i=k}} f_{X_i,W_i} \\ &= \prod_{i=1}^{n}\prod_{k=1}^{K} \left(\lambda_k(T_i)^{\Delta_i} \exp\left(\Delta_i \beta' X_i - e^{\beta' X_i}\Lambda_k(T_i)\right) \pi_{k,\gamma}(W_i)\right)^{\mathbb{1}_{S_i=k}} f_{X_i,W_i}. \end{aligned} \tag{4.5}$$

Ainsi, la loi des covariables X et W ne dépendant pas du paramètre du modèle θ, la vraisemblance des données complètes à considérer (proportionnelle à celle donnée par (4.5)) est

$$L_n^{(c)}(\theta) = \prod_{i=1}^{n}\prod_{k=1}^{K} \left(\lambda_k(T_i)^{\Delta_i} \exp\left(\Delta_i \beta' X_i - e^{\beta' X_i}\Lambda_k(T_i)\right) \pi_{k,\gamma}(W_i)\right)^{\mathbb{1}_{S_i=k}}.$$

Par conséquent, la log-vraisemblance complète à utiliser, dans le cadre semi-paramétrique de l'estimation de θ, est

$$l_n^{(c)}(\theta) = \sum_{i=1}^{n} \left(\Delta_i \beta' X_i + \sum_{k=1}^{K} \mathbb{1}_{S_i=k} \left(\Delta_i \ln \Lambda_k\{T_i\} - e^{\beta' X_i}\Lambda_k(T_i) + \ln \pi_{k,\gamma}(W_i)\right)\right) \tag{4.6}$$

Mise en place de l'étape Espérance

Calculons maintenant la log-vraisemblance conditionnelle aux données réellement observées $\mathcal{O} = (\mathcal{O}_i)_{i\in\{1,\dots,n\}}$ sous l'hypothèse que le vrai paramètre est ϑ, utilisée dans l'algorithme EM. On l'appellera log-vraisemblance EM :

$$\begin{aligned} l_n^{EM}(\theta,\vartheta) &= D(\theta,\vartheta) \\ &= \mathbb{E}_\vartheta[l_n^{(c)}(\theta)|\mathcal{O}] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} Q(\mathcal{O}_i,k,\vartheta)\left(\Delta_i \ln \Lambda_k\{T_i\} - e^{\beta' X_i}\Lambda_k(T_i) + \ln \pi_{k,\gamma}(W_i)\right) \\ &\quad + \sum_{i=1}^{n}\Delta_i\beta' X_i, \end{aligned} \tag{4.7}$$

où $Q(\mathcal{O}_i,k,\vartheta) = R_i 1\{S_i = k\} + (1-R_i)w_i(k,\vartheta)$ est la probabilité conditionnelle que S_i soit égal à la strate k sachant $\mathcal{O}_i$ sous l'hypothèse que le modèle a pour paramètre ϑ, avec $w_i(k,\vartheta)$ la probabilité conditionnelle sous le modèle ϑ que l'individu i appartienne à la strate k sachant les variables T_i, Δ_i, X_i, W_i. Si $\vartheta = (\beta,\gamma,\Lambda_k; 1\le k\le K)$, cette probabilité conditionnelle se calcule sous la forme

$$\begin{aligned} w_i(k,\vartheta) &= \frac{f_{T_i,\Delta_i|X_i,S_i=k}(\vartheta)\pi_{k,\gamma}(W_i)}{\sum_{k=1}^{K} f_{T_i,\Delta_i|X_i,S_i=k}(\vartheta)\pi_{k,\gamma}(W_i)} \\ &= \frac{\Lambda_k\{T_i\}^{\Delta_i} e^{\Delta_i\beta' X_i}\exp\left(-e^{\beta' X_i}\sum_{j=1}^{n}\Lambda_j\{T_i\}Y_j(T_i)\right)e^{\gamma_k' W_i}}{\sum_{k=1}^{K}\Lambda_k\{T_i\}^{\Delta_i} e^{\Delta_i\beta' X_i}\exp\left(-e^{\beta' X_i}\sum_{j=1}^{n}\Lambda_j\{T_i\}Y_j(T_i)\right)e^{\gamma_k' W_i}} \end{aligned} \tag{4.8}$$

Mise en place de l'étape Maximisation

A l'étape $q+1$ de l'algorithme EM, nous cherchons à maximiser $l_n^{EM}(\theta,\widehat{\theta}_n^{(q)})$ en θ, $\widehat{\theta}_n^{(q)}$ désignant la valeur courante de paramètre θ obtenue à l'issue de l'étape q. $\theta = (\beta,\gamma,(\Lambda_k\{T_i\})_{1\le i\le n,1\le k\le K})$ varie dans $\mathcal{B}\times\mathcal{G}\times\prod_{k=1}^{K}\mathcal{E}_n^{(k)}$. Nous commençons par déterminer des relations implicites pour les estimateurs en escalier de Λ_k $(1\le k\le K)$ grâce à la proposition suivante.

Proposition 4.1. *Soit* $\widehat{\theta}_n^{(q)} = \left(\widehat{\beta}_n^{(q)}, \widehat{\gamma}_n^{(q)}, (\widehat{\Lambda_{k,n}\{T_i\}}^{(q)})_{1\le i\le n,1\le k\le K}\right)$ *l'estimateur de* θ *obtenu à l'étape* q *de l'algorithme EM. Alors si*

$$\widehat{\theta}_n^{(q+1)} = \left(\widehat{\beta}_n^{(q+1)}, \widehat{\gamma}_n^{(q+1)}, (\widehat{\Lambda_{k,n}\{T_i\}}^{(q+1)})_{1\le i\le n,1\le k\le K}\right)$$

est l'estimateur de θ obtenu à l'issue de l'étape $q+1$, la valeur du saut en T_i de l'estimateur en escalier de Λ_k à l'étape $q+1$ de l'algorithme EM vérifie pour tout k dans $\{1,...,K\}$ et n dans $\{1,...,n\}$ l'équation implicite

$$\widehat{\Lambda_{k,n}}\{T_i\}^{(q+1)} = \frac{\Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)})}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)}) \exp(X_j' \widehat{\beta_n}^{(q+1)}) Y_j(T_i)}. \tag{4.9}$$

Preuve Pour obtenir cette équation implicite, il suffit de dériver $l_n^{EM}(\theta, \widehat{\theta_n}^{(q)})$ par rapport à la composante $\Lambda_k\{T_i\}$ de θ, de vérifier que la dérivée seconde est bien négative et de trouver la valeur qui annule la dérivée première. ■

En réinjectant ces valeurs dans $l_n^{EM}(\theta, \widehat{\theta_n}^{(q)})$ et en utilisant l'expression de l'estimateur de Λ_k de la forme $\widehat{\Lambda_{k,n}}(T_i) = \sum_{j=1}^n \widehat{\Lambda_{k,n}}\{T_j\}\mathbb{1}_{T_j \leq T_i}$, on obtient une expression de la log-vraisemblance EM ne dépendant plus que des composantes β et γ de θ :

$$\begin{aligned} l_n^{EM}((\beta,\gamma), \widehat{\theta_n}^{(q)}) &= \sum_{i=1}^n \sum_{k=1}^K Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \Delta_i \ln \left(\frac{\Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)})}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)}) e^{\beta' X_j} Y_j(T_i)} \right) \\ &\quad + \sum_{i=1}^n \sum_{k=1}^K Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \ln \pi_{k,\gamma}(W_i) \\ &\quad - \sum_{i=1}^n \sum_{k=1}^K \sum_{j=1}^n \frac{Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) e^{\beta' X_i} \Delta_j Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)})}{\sum_{l=1}^n Q(\mathcal{O}_l, k, \widehat{\theta_n}^{(q)}) e^{\beta' X_l} Y_l(T_j)} \mathbb{1}_{T_j \leq T_i} \\ &\quad + \sum_{i=1}^n \Delta_i \beta' X_i \end{aligned} \tag{4.10}$$

En utilisant la commutativité des symboles de sommations et le fait que pour j dans $\{1,...,n\}$, la somme $\sum_{k=1}^K Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)})$ soit égale à 1, on obtient l'expression suivante simplifiée de (4.10) :

$$\begin{aligned} l_n^{EM}((\beta,\gamma), \widehat{\theta_n}^{(q)}) &= \sum_{i=1}^n \sum_{k=1}^K Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \Delta_i \ln \left(\frac{\Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)})}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)}) e^{\beta' X_j} Y_j(T_i)} \right) \\ &\quad + \sum_{i=1}^n \sum_{k=1}^K Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \ln \pi_{k,\gamma}(W_i) \\ &\quad + \sum_{i=1}^n \Delta_i \left(\beta' X_i - 1\right). \end{aligned} \tag{4.11}$$

Dans l'étape de maximisation de l'algorithme, nous souhaitons ensuite maximiser l'expression (4.11) en $(\beta, \gamma) \in \mathcal{B} \times \mathcal{G}$ puis réinjecter l'expression obtenue pour $\widehat{\beta_n}^{(q+1)}$ dans (4.9) pour obtenir l'ensemble des sauts des estimateurs $\widehat{\Lambda_{k,n}}^{(q+1)}$ $(1 \leq k \leq K)$. Il n'est pas possible de résoudre analytiquement les équations $\frac{\partial l_n^{EM}((\beta,\gamma),\widehat{\theta_n}^{(q)})}{\partial \beta} = 0$ et $\frac{\partial l_n^{EM}((\beta,\gamma),\widehat{\theta_n}^{(q)})}{\partial \gamma_k} = 0$ pour obtenir une expression explicite de calcul des valeurs itérées $\widehat{\beta_n}^{(q+1)}$ et $\widehat{\gamma_n}^{(q+1)}$. Nous choisissons donc de résoudre ces équations, à chaque itération de l'algorithme EM, à l'aide d'un algorithme de Newton-Raphson. Cette proposition donne les expressions des fonctions score associées aux paramètres β et γ en nous plaçant dans le cas où β varie en dimension 1 ($p = 1$) et γ en dimension K ($m = 1$) pour simplifier les notations.

Proposition 4.2. *La fonction score associée à la log-vraisemblance EM pour le paramètre β est donnée par*

$$\begin{aligned} S_\beta(\beta) &= \frac{\partial l_n^{EM}\left((\beta,\gamma),\widehat{\theta_n}^{(q)}\right)}{\partial \beta} \\ &= \sum_{i=1}^{n} \Delta_i X_i - \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \frac{\sum_{j=1}^{n} Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)}) X_j e^{\beta X_j} Y_j(T_i)}{\sum_{j=1}^{n} Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)}) e^{\beta X_j} Y_j(T_i)}. \end{aligned}$$

La fonction score associée à la log-vraisemblance EM pour le paramètre γ_k $(1 \leq k \leq K-1)$ est donnée par

$$\begin{aligned} S_{\gamma_k}(\gamma_k) &= \frac{\partial l_n^{EM}\left((\beta,\gamma),\widehat{\theta_n}^{(q)}\right)}{\partial \gamma_k} \\ &= \sum_{i=1}^{n} \left(Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) - \pi_{k,\gamma}(W_i) \right) W_i \\ &= \sum_{i=1}^{n} \left(Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) - \frac{e^{\gamma_k W_i}}{1 + e^{\gamma_k W_i}} \right) W_i \end{aligned}$$

Nous rappelons ici le principe de l'algorithme de Newton-Raphson permettant de trouver les zéros de S_β et S_{γ_k}. L'algorithme de Newton-Raphson pour la résolution numérique d'une équation $S(x) = 0$ consiste à approximer l'expression $S(x)$ par son développement de Taylor à l'ordre 1 au voisinage

de la valeur courante, que l'on notera $x^{(a)}$ pour l'itération a :

$$S(x) = S(x^{(a)}) + (x - x^{(a)})S'(x^{(a)}).$$

Ainsi, la nouvelle valeur de x à l'itération $a+1$ est obtenue en résolvant l'équation où le second membre de l'équation précédente est pris égal à 0 :

$$x^{(a+1)} = x^{(a)} + S'(x^{(a)})^{-1} S(x^{(a)}).$$

Cette étape est répétée jusqu'à ce qu'un critère de convergence soit vérifié, par exemple $|x^{(a+1)} - x^{(a)}|$ arbitrairement petit.

Appliquons maintenant cet algorithme pour le calcul de $\widehat{\beta_n}^{(q+1)}$ et $\widehat{\gamma_n}^{(q+1)}$. Nous initialisons l'algorithme de Newton-Raphson avec les valeurs obtenues à l'étape q de l'algorithme EM, $\widehat{\beta_n}^{(q)}$ et $\widehat{\gamma_n}^{(q)}$. Notons la valeur courante de ces paramètres dans l'algorithme de Newton-Raphson $\beta^{(a)}$ et $\gamma_k^{(a)}$ $(1 \leq k \leq K-1)$, en éludant l'indice n représentant la taille de l'échantillon et l'étape $q+1$ de l'algorithme EM, fixes lors de l'application de la méthode de Newton-Raphson. La proposition suivante donne les expressions dérivées des fonctions score utilisées dans l'implémentation de cet algorithme.

Proposition 4.3. *La dérivée de S_β est donnée par*

$$\begin{aligned} S'_\beta(\beta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \left(\frac{\sum_{j=1}^{n} Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) X_j e^{\beta X_j} Y_j(T_i)}{\sum_{j=1}^{n} Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) e^{\beta X_j} Y_j(T_i)} \right)^2 \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) \frac{\sum_{j=1}^{n} Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) X_j^2 e^{\beta X_j} Y_j(T_i)}{\sum_{j=1}^{n} Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)}) e^{\beta X_j} Y_j(T_i)}. \end{aligned}$$

Quel que soit k dans $\{1, ..., K-1\}$, la dérivée de S_{γ_k} est donnée par

$$S_{\gamma_k}(\gamma_k) = -\sum_{i=1}^{n} \frac{W_i^2 e^{\gamma_k W_i}}{(1 + e^{\gamma_k W_i})^2}.$$

Résumons maintenant les étapes de l'implémentation de l'itération $q+1$ de l'algorithme EM, où l'on dispose au départ de la valeur des estimateurs obtenue à l'issue de l'étape q

$$\widehat{\theta_n}^{(q)} = \left(\widehat{\beta_n}^{(q)}, \widehat{\gamma_n}^{(q)}, (\widehat{\Lambda_{k,n}\{T_i\}}^{(q)})_{1 \leq i \leq n, 1 \leq k \leq K} \right),$$

où $\widehat{\gamma_n}^{(q)} = (\widehat{\gamma_{1,n}}^{(q)}, ..., \widehat{\gamma_{K-1,n}}^{(q)}, 0)$.

Etape $q+1$ de l'algorithme EM

(I) **Etape Espérance** Calcul de $w_i(k, \widehat{\theta_n}^{(q)})$ grâce à la formule (4.8), puis de $Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)})$, pour tout $k \in \{1, ..., K\}$ et $i \in \{1, ..., n\}$.

(II) **Etape Maximisation**

- **Etape 1** : Initialisation des paramètres

$$\beta^{(0)} = \widehat{\beta_n}^{(q)}, \gamma^{(0)} = \widehat{\gamma_n}^{(q)} \quad \text{et} \quad \Lambda_k^{(0)} = (\widehat{\Lambda_{k,n}}\{T_i\}^{(q)})_{1 \leq i \leq n}$$

  pour tout k dans $\{1, ..., K\}$.
- **Etape 2** : Exécution des K algorithmes de Newton-Raphson,
  - répétition des calculs de $S_\beta(\beta^{(a)})$, $S'_\beta(\beta^{(a)})$ et

$$\beta^{(a+1)} = \beta^{(a)} - S'_\beta(\beta^{(a)})^{-1} S_\beta(\beta^{(a)})$$

    jusqu'à ce que la condition $|\beta^{(a+1)} - \beta^{(a)}| < 10^{-5}$ soit vérifiée,
  - répétition pour chaque $k \in \{1, ..., K\}$ des calculs de $S_{\gamma_k}(\gamma_k^{(a)})$, $S'_{\gamma_k}(\gamma_k^{(a)})$ et

$$\gamma_k^{(a)} - S'_{\gamma_k}(\gamma_k^{(a)})^{-1} S_{\gamma_k}(\gamma_k^{(a)})$$

    jusqu'à ce que la condition $|\gamma_k^{(a+1)} - \gamma_k^{(a)}| < 10^{-5}$ soit vérifiée.
- **Etape 3** : Assignation des valeurs

$$\begin{aligned} \widehat{\beta_n}^{(q)} &= \beta^{(a)}, \\ \widehat{\gamma_{k,n}}^{(q)} &= \gamma_k^{(a)}, \forall 1 \leq k \leq K-1, \\ \widehat{\gamma_{K,n}}^{(q)} &= 0. \end{aligned}$$

- **Etape 4** : Calcul des sauts des estimateurs des Λ_k $(1 \leq k \leq K)$

$$\widehat{\Lambda_{k,n}}\{T_i\}^{(q+1)} = \frac{\Delta_i Q(\mathcal{O}_i, k, \widehat{\theta_n}^{(q)})}{\sum_{j=1}^n Q(\mathcal{O}_j, k, \widehat{\theta_n}^{(q)}) \exp(X'_j \widehat{\beta_n}^{(q+1)}) Y_j(T_i)},$$

  pour tout $1 \leq k \leq K$ et $1 \leq i \leq n$.
- **Etape 5** : Obtention de

$$\widehat{\theta_n}^{(q+1)} = \left(\widehat{\beta_n}^{(q+1)}, \widehat{\gamma_n}^{(q+1)}, (\widehat{\Lambda_{k,n}}\{T_i\}^{(q+1)})_{1 \leq i \leq n, 1 \leq k \leq K} \right).$$

Deuxième partie

Estimation par moindres carrés pénalisés dans le modèle de régression linéaire censuré à droite

Cette partie s'intéresse au travail effectué dans le cadre du modèle de régression linéaire censuré à droite.

Nous commençons, dans le chapitre 5, par introduire le modèle de durée de vie à temps accéléré permettant de comprendre l'utilisation du modèle de régression linéaire dans un cadre de durées de vie censurées. Nous rappelons les résultats asymptotiques obtenus sur les estimateurs de type Kaplan-Meier dans un cadre général de durées de vie censurée à droite. Enfin, nous développons l'estimation par moindres carrés pénalisés dans le modèle de régression linéaire et donnons les résultats existants sur ce type d'estimateurs.

Dans le chapitre 6, nous introduisons une méthode d'estimation par moindres carrés pénalisés dans le modèle de régression linéaire censuré à droite. Nous proposons un nouvel estimateur du coefficient de régression minimisant un critère des moindres carrés pondéré par des poids de Kaplan-Meier et contraint par une pénalité bridge. Des résultats de consistance et normalité asymptotique sont obtenus et une simulation permet de donner des résultats numériques pour des échantillons à taille fixée et de comparer l'estimateur proposé à celui obtenu par la méthode simple consistant à supprimer les données censurées.

Enfin, le chapitre 7 rappelle la démarche de l'algorithme bootstrap et présente son application à l'estimation de l'écart-type de l'estimateur obtenu dans le chapitre 6.

Chapitre 5

Modèle de durée de vie à temps accéléré et approche pénalisée

Sommaire

Dans ce chapitre, nous commençons par motiver l'utilisation du modèle de régression linéaire pour des données de survie en expliquant qu'il correspond à un modèle de durée de vie à temps accéléré. Nous donnons l'interprétation de ce type de modèle en terme des fonctions usuelles caractérisant la loi. Nous introduisons ensuite dans le cadre général de durée de survie censurées à droite les estimateurs de type Kaplan-Meier, initiés par Kaplan & Meier (1958), et les résultats asymptotiques obtenus par Stute (1993, 1996). Nous détaillons les applications à l'estimateur des moindres carrés dans le modèle de régression linéaire censuré. Enfin, nous présentons les méthodes de réduction de dimension par approche pénalisée de type ridge, LASSO et bridge dans le modèle de régression linéaire ainsi que les résultats asymptotiques obtenus par Knight & Fu (2000) dans le cas d'un design fixe. Notre objectif sera, dans le chapitre suivant, de proposer ce type d'estimateurs dans le modèle de régression linéaire censuré à droite.

5.1 Modèle de durée de vie à temps accéléré

Le modèle de temps accéléré pour des données de survie est une approche utilisant le modèle de régression linéaire classique. Dans ce modèle, les covariables $X \in \mathbb{R}^p$ agissent en accroissant ou contractant le temps d'un facteur $\exp(-\beta' X)$, où β est un p-vecteur de paramètres. En effet, le logarithme népérien $Y = \ln(T)$ de la durée de vie T est modélisé, de manière à transformer une variable T prenant ses valeurs dans les réels positifs en une variable réelle Y, et cette variable Y est supposée suivre un modèle de régression linéaire

$$Y = \beta' X + \varepsilon, \tag{5.1}$$

où ε est une variable aléatoire centrée représentant l'erreur. Les choix classiques de distribution pour ε comprennent la loi gaussienne, conduisant pour T à un modèle de régression log-normal, la loi des valeurs extrêmes, conduisant à un modèle de Weibull ou encore la loi logistique, conduisant à un modèle log-logistique. Notons S_0 la fonction de survie de T quand les covariables sont fixées égales à 0 et $S_{T|X=x}$ celle de T quand les covariables sont fixées égales à $x \in \mathbb{R}^p$, alors

$$\begin{aligned} S_{T|X=x}(t) &= \mathbb{P}\left(T > t | X = x\right) \\ &= \mathbb{P}\left(\ln(T) > \ln(t) | X = x\right) \\ &= \mathbb{P}\left(\varepsilon > \ln(t) - \beta' X | X = x\right) \\ &= \mathbb{P}\left(\exp(\varepsilon) > t \exp(-\beta' x)\right) \\ &= \mathbb{P}\left(T > t \exp(-\beta' x) | X = 0\right) \\ &= S_0(t \exp(-\beta' x)). \end{aligned}$$

De la même façon, si on note λ_0, Λ_0 les fonctions de risques instantané et cumulé de T quand les covariables sont fixées égales à 0 et $\Lambda_{T|X}$, $\lambda_{T|X}$ celle de T connaissant les covariables $X \in \mathbb{R}^p$ alors

$$\begin{aligned} \Lambda_{T|X}(t) &= \Lambda_0(t \exp(-\beta' X)), \\ \lambda_{T|X}(t) &= \lambda_0(t \exp(-\beta' X)) \exp(-\beta' X). \end{aligned}$$

La densité $f_{T|X}$ de T peut alors s'écrire en fonction de la fonction de risque de base

$$f_{T|X}(t) = \lambda_0(te^{-\beta' X}) e^{-\beta' X} \exp\left(-\Lambda_0(te^{-\beta' X})\right).$$

Ce modèle d'accélération ou décélération du temps est couramment utilisé dans l'industrie, où des échelles de temps multiplicatives sont communes.

Par exemple, supposons qu'un photocopieur ait une fonction de risque dépendant du nombre total de copies faites, mais que les données des temps de panne aient été répertoriées en fonction du temps calendaire. Les covariables liées au nombre de copies par jour devraient alors donner de très bons résultats dans le modèle de survie accéléré. Cette approche peut également être utilisée dans le domaine médical.

Le lecteur intéressé par un développement plus détaillé sur le modèle de durée de vie à temps accéléré pourra se référer à Bagdonavičius & Nikulin (2002). Par ailleurs, le modèle de régression linéaire jouant un rôle central dans la statistique moderne, il est développé dans de nombreux ouvrages, par exemple Jørgensen (1993), Rao & Toutenburg (1995), Searle (1997), ou encore Rencher & Schaalje (2008).

5.2 Estimateurs de type Kaplan-Meier sous censure aléatoire

5.2.1 Définitions et notations

Supposons que l'on dispose d'un n-échantillon $(Y_1, ..., Y_n)$ de répétitions indépendantes de Y, variable aléatoire réelle de fonction de répartition F. Alors un estimateur non-paramétrique et efficace de F est donné par la fonction de répartition empirique F_n définie par

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \leq y}.$$

Nous nous intéressons dans cette partie à des données censurées aléatoirement à droite. Ce phénomène, commun pour des données de survie, est introduit dans la partie 1.3. Nous introduisons donc une variable aléatoire réelle de censure C de fonction de répartition G indépendante de Y et supposons que les données observées sont les covariables X, la variable d'intérêt éventuellement censurée $Z = \min(Y, C)$ et l'indicateur de censure $\Delta = \mathbb{1}_{Y \leq C}$. Nous disposons ainsi d'un n-échantillon $(X_i, Z_i, \Delta_i)_{1 \leq i \leq n}$ de répétitions indépendantes de (X, Z, Δ).

Introduisons les notations suivantes. Posons $Z_{1:n} \leq Z_{2:n} \leq ... \leq Z_{n:n}$ les valeurs réordonnées dans l'ordre croissant de $(Z_1, ..., Z_n)$ et $(\Delta_{[1:n]}, \Delta_{[2:n]}, ..., \Delta_{[n:n]})$,

$(X_{[1:n]}, X_{[2:n]}, ..., X_{[n:n]})$ les valeurs de Δ et X associées aux $(Z_{i:n})$.L'analogue non-paramétrique de F_n, lors de l'observation du n-uplet $(X_i, Z_i, \Delta_i)_{1\leq i\leq n}$ devient alors l'estimateur de Kaplan & Meier (1958) $\widehat{F_n}$ défini par

$$\begin{aligned}\widehat{F_n}(y) &= 1-\prod_{i=1}^{n}\left(1-\frac{\Delta_{[i:n]}}{n-i+1}\right)^{\mathbb{1}_{Z_{i:n}\leq y}}\\ &= \sum_{i=1}^{n} W_{in}\mathbb{1}_{Z_{i:n}\leq y},\end{aligned}$$

où les poids W_{in} sont donnés par

$$W_{in}=\frac{\Delta_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\Delta_{[j:n]}}.$$

Gill (1981) montre la convergence uniforme de l'estimateur de Kaplan & Meier (1958) $\widehat{F_n}$ vers F dans le cas de variables positives. Stute, dans les années '90, s'intéresse dans un cadre très général à l'estimation de la fonction de répartition bivariée $F^0 = F_{X,Y}$, comme extension de l'estimation de la fonction de répartition univariée $\widehat{F_n}$ de Y. L'estimateur $\widehat{F_n^0}$ de F^0 devrait vérifier la propriété : pour tout $y \in \mathbb{R}$, $\widehat{F_n}(y) = \widehat{F_n^0}(+\infty, y)$. Seules les deux hypothèses suivantes sont posées sur le modèle

(i) $\mathbb{P}(Y \leq C|X, Y) = \mathbb{P}(Y \leq C|Y)$,

(ii) F et G n'ont pas de sauts en commun,

Pour cela, il introduit les estimateurs de la forme générale

$$S_n^{\varphi} = \sum_{i=1}^{n} W_{in}\varphi(X_{[i:n]}, Z_{i:n}).$$

REMARQUE 5.1. En choisissant $\varphi(x, y) = \mathbb{1}_{]-\infty,x]\times]-\infty,y]}$, S_n^{φ} devient l'estimateur de la fonction de répartition bivariée proposé par Stute

$$\widehat{F_n^0} = \sum_{i=1}^{n} W_{in}\mathbb{1}_{]-\infty,X_{[i:n]}]\times]-\infty,Z_{i:n}]}.$$

REMARQUE 5.2. Supposons que X soit univariée. En posant

$$\varphi_1(x, y) = yx\ ,\ \varphi_2(x, y) = y\ ,\ \varphi_3(x, y) = x\ ,\ \varphi_4(x, y) = y^2\ ,\ \varphi_5(x, y) = x^2,$$

et en notant S_n^i $(1 \leq i \leq 5)$ les quantités correspondantes, des combinaisons de ces estimateurs permettent d'obtenir des estimations de la covariance et corrélation de (X, Y).

Afin d'étudier les propriétés des estimateurs de la forme S_n^φ proposés, introduisons quelques notations. Notons H la fonction de répartition de la variable effectivement observée Z et posons

$$\tau_H = \inf\{x \in \mathbb{R}; H(x) = 1\}$$

la borne supérieure du support de H. Nous utiliserons la même notation τ_F, τ_G pour les fonctions de répartition F et G. Remarquons que $\tau_H = \min(\tau_F, \tau_G)$ en raison de l'indépendance de Y et C. Dans la suite de ce chapitre, nous remplacerons l'hypothèse (ii) par l'hypothèse plus restrictive

(iii) F et G sont continues sur $\mathbb{R}$,

qui permettra de simplifier les notations et qui correspond au cadre dans lequel notre étude a été faite.

Nous présentons maintenant différents résultats obtenus par Stute (1993, 1996) sur les variables aléatoires S_n^φ.

5.2.2 Résultats de consistance

Nous énonçons ici des résultats de consistance sur les estimateurs de type S_n^φ introduits dans la partie 5.2.1. Ils sont développés plus en détail dans Stute (1993).

Théorème 5.1. *Sous les hypothèses* (i),(iii), *si* $\varphi(X, Y)$ *est intégrable, alors presque sûrement*

$$\lim_{n\to\infty} S_n^\varphi = \int_{Y \le \tau_H} \varphi(X, Y) \mathrm{d}\mathbb{P}.$$

La démonstration de ce théorème se trouve dans Stute (1993).

Remarque 5.3. Notons que si $\tau_F \le \tau_G$ alors S_n^φ est un estimateur consistant de $\int \varphi(X, Y)\mathrm{d}\mathbb{P}$.

Ce théorème permet de conclure sur la convergence uniforme de l'extension bivariée de l'estimateur de Kaplan-Meier.

Corollaire 5.2. *Sous les hypothèses* (i),(iii), *si* $\tau_F \le \tau_G$, *alors*

$$\sup_{x\in\mathbb{R}^p, y\in\mathbb{R}} |\widehat{F_n^0}(x, y) - F^0(x, y)| \xrightarrow[n\to\infty]{p.s.} 0.$$

Supposons maintenant que la variable aléatoire Y soit issue du modèle de régression linéaire (5.1) introduit dans (5.1) où le vrai paramètre de régression sera noté β_0. Ajoutons l'hypothèse que $\mathbb{E}[\varepsilon|X] = 0$. Il est alors possible de proposer un nouvel estimateur de β coïncidant avec l'estimateur des moindres carrés en l'absence de censure, et ayant la propriété de consistance. Il est pour cela nécessaire d'introduire les matrices M_{1n} et M_{2n} suivantes pour $1 \leq i, j \leq p$, et $1 \leq s \leq n$.

$$\begin{aligned} M_{1n}(i,s) &= W_{sn} X^i_{[s:n]}, \\ M_{2n}(i,j) &= \sum_{k=1}^{n} W_{kn} X^i_{[k:n]} X^j_{[k:n]}, \end{aligned}$$

où $X^i_{[k:n]}$ désigne la i-ème coordonnée des covariables $X_{[k:n]}$ associées à la donnée réordonnée $Z_{k:n}$. Notons $\widehat{\beta_n}$ l'estimateur minimisant la somme des moindres carrés suivante :

$$\widehat{\beta_n} = argmin_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} W_{in} \left(Z_{i:n} - \beta' X_{[i:n]} \right)^2 .$$

En utilisant la notation $\tilde{Z_n} = (Z_{1:n}, ..., Z_{n:n})'$, nous remarquons que

$$\widehat{\beta_n} = M_{2n}^{-1} M_{1n} \tilde{Z_n}. \tag{5.2}$$

Le théorème (5.1) permet alors d'établir la propriété de consistance forte de ce nouvel estimateur.

Corollaire 5.3. *Sous les hypothèses* (i), (iii), *si* $\tau_F \leq \tau_G$ *et* $\mathbb{E}[XX']$ *existe et est définie positive, alors*

$$\widehat{\beta_n} \xrightarrow[n\to\infty]{p.s.} \beta_0$$

Preuve Comme l'existence de $\mathbb{E}[XX']$ est supposée, d'après le théorème 5.1, on obtient la convergence presque sûre $M_{2n} \underset{n\to\infty}{\longrightarrow} \mathbb{E}[XX']$. De plus,

$$M_{1n} \tilde{Z_n} = \sum_{k=1}^{n} W_{kn} Z_{k:n} X_{[k:n]}.$$

En appliquant à nouveau le théorème (5.1), et comme $\tau_F \leq \tau_G$, on obtient la convergence presque sûre $M_{1n} \tilde{Z_n} \underset{n\to\infty}{\longrightarrow} \mathbb{E}[ZX] = \mathbb{E}[XX']\beta_0$, ce qui permet de conclure sur la consistance de $\widehat{\beta_n}$. ■

Cet estimateur, facile a implémenter, a été comparé numériquement aux estimateurs de Miller (1976) et Buckley & James (1979) dans le modèle linéaire dans Stute (1993) et donne de meilleurs résultats dans l'ensemble.

5.2.3 Normalité asymptotique

Stute (1996) s'intéresse par la suite à la normalité asymptotique des estimateurs de la forme S_n^φ. Il est nécessaire d'introduire pour cette étude quelques notations supplémentaires. Notons $\tilde{H}$ la fonction définie pour $x \in \mathbb{R}^p$ et $y \in \mathbb{R}$ par

$$\tilde{H}(x,y) = \mathbb{P}\left(X \leq x, Z \leq y, \Delta = 1\right),$$

où l'inégalité $X \leq x$ est prise coordonnée par coordonnée. Nous définissons également pour $j \in \{1,...,p\}$ les fonctions Φ_1^j et Φ_2^j sur $\mathbb{R}$ par :

$$\begin{aligned}
\Phi_1^\varphi(z) &= \frac{1}{1-H(z)} \int \mathbb{1}_{z<y} \varphi(x,y) \exp\left(\frac{G(y)}{1-G(y)}\right) \mathrm{d}\tilde{H}(x,y) \\
\Phi_2^\varphi(z) &= \iint \frac{\mathbb{1}_{u<z, u<y} \varphi(x,y) \exp\left(\frac{G(y)}{1-G(y)}\right)}{(1-F(u))(1-G(u))^2} \mathrm{d}G(u) \mathrm{d}\tilde{H}(x,y).
\end{aligned}$$

Le théorème établissant la normalité asymptotique des estimateurs de type S_n^φ sera vérifié sous les hypothèses suivantes.

(iv) $\int \left(\varphi(X,Z) \exp\left(\frac{G(Z)}{1-G(Z)}\right) \Delta\right)^2 \mathrm{d}\mathbb{P} < +\infty$,

(v) $\int |\varphi(X,Y)| \sqrt{C(Y)} \mathrm{d}\mathbb{P} < +\infty$,

où

$$C(y) = \int_0^y \frac{\mathrm{d}G(v)}{(1-H(v))(1-G(v))}.$$

Nous énonçons maintenant le théorème établi par Stute (1996).

Théorème 5.4. *Sous les hypothèses* (i), (iii), (iv), (v) *et si* $\tau_F \leq \tau_G$ *alors*

$$\sqrt{n}\left(S_n^\varphi - \mathbb{E}[\varphi(X,Y)]\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2(\varphi)),$$

où

$$\sigma^2(\varphi) = \mathrm{var}\left(\varphi(X,Z) \exp\left(\frac{G(Z)}{1-G(Z)}\right) \Delta + \Phi_1^\varphi(Z)(1-\Delta) - \Phi_2^\varphi(Z)\right).$$

REMARQUE 5.4. La variance asymptotique dépendant des fonctions inconnues Φ_1^φ et Φ_2^φ, il sera difficile de déterminer un estimateur consistant de $\sigma^2(\varphi)$.

Pour de nombreuses applications statistiques, il est intéressant d'avoir une version multidimensionnelle du théorème (5.4). Posons donc $\varphi = (\varphi_1, ..., \varphi_k)$ une fonction mesurable définie sur $\mathbb{R}^{p+1}$ à valeurs dans $\mathbb{R}^k$. Définissons pour tout $j \in \{1, ..., k\}$ la fonction

$$\psi_j = \varphi_j(X, Z) \exp\left(\frac{G(Z)}{1 - G(Z)}\right) \Delta + \Phi_1^{\varphi_j}(Z)(1 - \Delta) - \Phi_2^{\varphi_j}(Z)$$

et posons

$$\sigma_{ij} = cov(\psi_i, \psi_j).$$

Etablissons alors le théorème pour la fonction vectorielle

$$S_n^\varphi = (S_n^{\varphi_1}, ..., S_n^{\varphi_k})',$$

et notons

$$S^\varphi = (\mathbb{E}[\varphi_1(X, Y)], ..., \mathbb{E}[\varphi_k(X, Y)])'.$$

THÉORÈME 5.5. *Sous les hypothèses* (i), (iii), *si* (iv), (v) *sont vérifiées pour tout* $j \in \{1, ..., k\}$, *et si* $\tau_F \leq \tau_G$ *alors*

$$\sqrt{n}\,(S_n^\varphi - S^\varphi) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma(\varphi)),$$

où

$$\Sigma(\varphi) = (\sigma_{ij})_{1 \leq i,j \leq k}.$$

Preuve Les théorèmes (5.4) et (5.5) sont démontrés dans l'article de Stute (1996). ■

Appliquons maintenant ce résultat au modèle de régression linéaire (5.1) avec le vrai paramètre du modèle noté β_0, sous les hypothèses $\mathbb{E}[\varepsilon|X] = 0$ et $\Sigma_0 = \mathbb{E}[XX']$ existe et est définie positive. Nous utiliserons les fonctions $(\varphi_j)_{1 \leq j \leq p}$ définies sur $\mathbb{R}^{p+1}$ par

$$\varphi_j(x, z) = \varphi_j(x^1, ..., x^p, z) = x^j(z - \beta_0' x).$$

Le corollaire suivant donne la distribution asymptotique de $\widehat{\beta_n}$ défini par (5.2).

Corollaire 5.6. *Sous les hypothèses* (i), (iii) *et* $\tau_F \leq \tau_G$, *si les hypothèses* (iv) *et* (v) *sont vérifiées par* φ_j *pour tout* $j \in \{1, ..., p\}$, *alors*

$$\sqrt{n}\left(\widehat{\beta_n} - \beta_0\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_0^{-1}\Sigma(\varphi)\Sigma_0^{-1}),$$

où $\Sigma(\varphi)$ *est la matrice définie dans* (5.5).

Preuve Pour commencer, calculons

$$\begin{aligned} S^{\varphi} &= (\mathbb{E}[\varphi_1(X,Y)], ..., \mathbb{E}[\varphi_p(X,Y)]) \\ &= \mathbb{E}[(Y - \beta_0' X)X] \\ &= \mathbb{E}[\varepsilon X] \\ &= (0, ..., 0)', \end{aligned}$$

d'après l'hypothèse $\mathbb{E}[\varepsilon|X] = 0$.
Nous appliquons ainsi le théorème 5.5 pour établir la convergence en loi de $\sqrt{n}S_n^{\varphi}$. Les hypothèses étant vérifiées, nous obtenons

$$\sqrt{n}S_n^{\varphi} = \sum_{i=1}^{n} W_{in}(Z_{i:n} - \beta_0' X_{[i:n]})X_{[i:n]} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma(\varphi)).$$

Or d'après la preuve du corollaire (5.3), la convergence

$$M_{2n} \xrightarrow[n\to\infty]{p.s.} \mathbb{E}[XX'] = \Sigma_0$$

est établie, d'où d'après le théorème de Slutsky

$$\sqrt{n}M_{2n}^{-1}S_n^{\varphi} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_0^{-1}\Sigma(\varphi)\Sigma_0^{-1}).$$

Pour conclure, remarquons que

$$\begin{aligned} \widehat{\beta_n} - \beta_0 &= M_{2n}^{-1}M_{1n}\tilde{Z}_n - \beta_0 \\ &= M_{2n}^{-1}\left(M_{1n}\tilde{Z}_n - M_{2n}\beta_0\right) \\ &= M_{2n}^{-1}\left(\sum_{k=1}^{n} W_{kn}Z_{k:n}X_{[k:n]} - \sum_{k=1}^{n} W_{kn}\beta_0' X_{[k:n]}X_{[k:n]}\right) \\ &= M_{2n}^{-1}\sum_{k=1}^{n} W_{kn}\left(Z_{k:n} - \beta_0' X_{[k:n]}\right)X_{[k:n]} \\ &= M_{2n}^{-1}S_n^{\varphi}. \end{aligned}$$

■

Les corollaires 5.3 et 5.6 permettent d'établir d'intéressantes propriétés sur l'estimateur des moindres carrés pondéré par des poids de Kaplan-Meier dans le modèle de régression linéaire censuré aléatoirement à droite. Nous nous intéressons par la suite à l'estimation pénalisée dans ce modèle de régression.

5.3 Estimation par moindres carrés pénalisés

La recherche sur l'estimation pénalisée dans le modèle de régression linéaire s'est beaucoup développée lors des dernières décennies. En effet, en raison de l'émergence de données en très grande dimension issues notamment du séquençage génétique, de nombreux chercheurs se sont intéressés au problème de la sélection d'un petit nombre de covariables ayant un fort effet sur la variable réponse Y parmi un grand nombre de prédicteurs potentiels. L'avantage, en sélection de variables, de retenir un sous-ensemble plus restreint de prédicteurs est l'interprétabilité du modèle ainsi qu'une possibilité d'obtenir une erreur de prédiction plus faible qu'avec le modèle complet. Cependant, comme la sélection de variables est un processus discret (les variables sont soit retenues, soit écartées), les estimateurs obtenus présentent souvent une variance élevée. Les méthodes dites *de réduction de dimension* se sont développées dans cette perspective de sélection et ont l'avantage d'être continues, ce qui leur confère une variabilité moindre.

5.3.1 Régression ridge

Par exemple, dans le modèle de régression linéaire, la méthode de régression dite *ridge regression* réduit les coefficients de régression en imposant une pénalité sur leur norme. L'estimateur ridge minimise une somme de carrés pondérée par une norme de type ℓ^2 :

$$\widehat{\beta_n}^{ridge} = argmin_\beta \left(\sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right). \tag{5.3}$$

Ici, le paramètre $\lambda \leq 0$ est un paramètre de complexité qui contrôle la réduction des coefficients de régression : plus la valeur de λ est grande, plus les valeurs des coefficients sont proches de zéro. Une manière équivalente de

réécrire le problème ridge est la suivante :

$$\begin{aligned} \widehat{\beta_n}^{ridge} &= argmin_\beta \sum_{i=1}^{n} (Y_i - \beta' X_i)^2 \\ &\quad \text{sous la contrainte } \sum_{j=1}^{p} \beta_j^2 \leq s, \end{aligned} \tag{5.4}$$

ce qui explicite la contrainte de norme sur les paramètres de régression. Une correspondance bijective est établie entre les paramètres λ de (5.3) et s de (5.4).

Quand plusieurs covariables sont corrélées dans le modèle de régression linéaire, les coefficients associés peuvent être mal déterminés et posséder une variance élevée. Par ailleurs, l'effet d'un coefficient positif très grand sur une variable peut être compensé par un coefficient négatif du même ordre associé à la variable qui lui est corrélée. En imposant la contrainte (5.4) aux coefficients, on empêche l'émergence de ce cas de figure.

En réécrivant le critère à minimiser de l'équation (5.3) sous forme matricielle

$$RSS_\lambda(\beta) = \left(\sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right).$$

on obtient l'écriture de l'estimateur ridge sous forme d'une fonction linéaire de Y :

$$\widehat{\beta_n}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda I_p)^{-1} \mathbf{X}'\mathbf{Y},$$

où $\mathbf{X}$ désigne la matrice de dimension $n \times p$ possédant le vecteur de covariables X_i sur sa i-ème ligne et $\mathbf{Y}$ le vecteur colonne composé des n variables réponses (Y_i). La solution diffère de l'estimateur des moindres carrés classique par l'ajout d'une constante positive λ à la diagonale de la matrice $\mathbf{X}'\mathbf{X}$, ce qui la rend inversible. Cela était la principale motivation de l'introduction de l'estimation ridge (cf Hoerl & Kennard (1970)).

5.3.2 Régression LASSO

La régression de type LASSO (pour Least Absolute Shrinkage and Selection Operator), introduite par Tibshirani (1996), est également une méthode

de réduction de dimension, mais elle est différente de la régression ridge : l'estimateur lasso de β est défini par

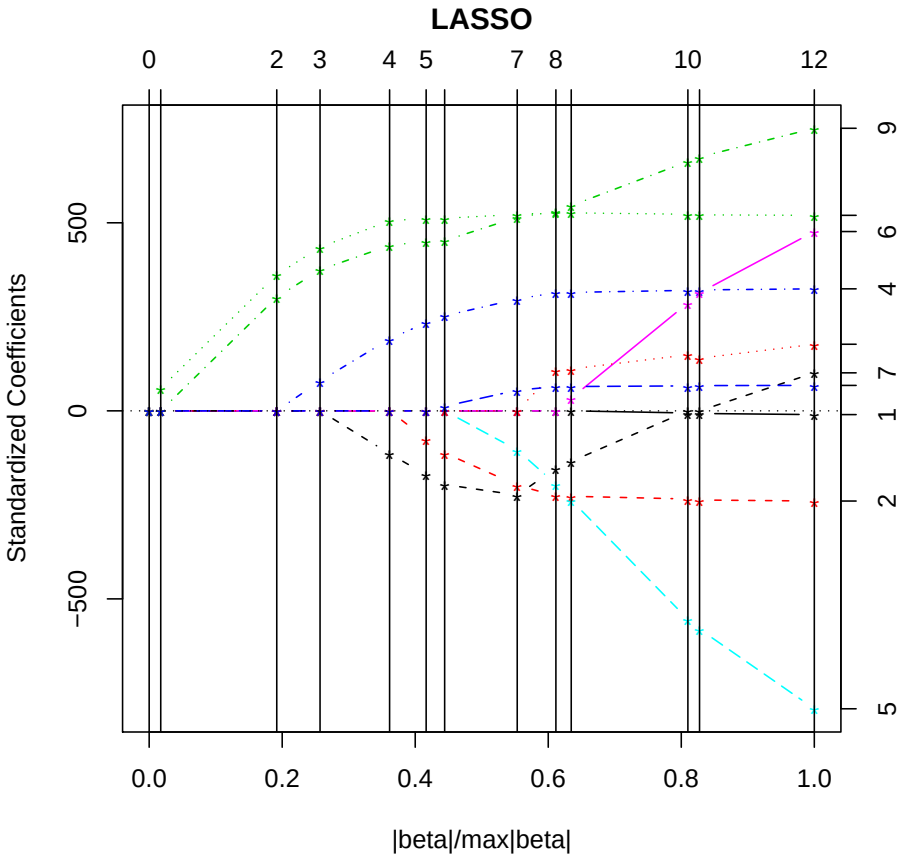
$$\widehat{\beta_n}^{lasso} = argmin_\beta \left(\sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{5.5}$$

Tout comme en régression ridge, il est équivalent de réécrire le problème lasso sous la forme

$$\begin{aligned} \widehat{\beta_n}^{lasso} &= argmin_\beta \sum_{i=1}^{n} (Y_i - \beta' X_i)^2 \\ &\quad \text{sous la contrainte } \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{5.6}$$

Il est important de remarquer la similarité entre le problème de régression ridge (5.3) et celui de la régression lasso : la pénalité ridge de type ℓ^2 $\sum_{j=1}^{p} \beta_j^2$ est remplacée par la pénalité ℓ^1 lasso $\sum_{j=1}^{p} |\beta_j|$. En raison de cette dernière contrainte, la solution $\widehat{\beta_n}^{lasso}$ ne peut être linéaire en $\mathbf{Y}$ et un algorithme sera nécessaire pour la déterminer. D'autre part, remarquons que le choix d'une valeur faible pour t contraindra certains des coefficients à être exactement nuls. Ainsi, le lasso effectue une sorte de sélection de variables continue.

Notons que si t est choisi supérieur ou égal à la norme ℓ^1 de l'estimateur des moindres carrés noté $\widehat{\beta}^{ls}$, alors l'estimateur lasso obtenu est égal à l'estimateur des moindres carrés. Comme en sélection de variables, la valeur de t doit être choisie de manière adaptative afin de minimiser une erreur de prédiction. Cela peut-être mis en place grâce à une procédure de validation croisée. Cependant, la nature de la réduction de dimension n'est pas évidente, donc pour faciliter l'interprétation, nous représentons dans la figure suivante un exemple d'utilisation de la régression lasso, issu des données *diabetes* du package **lars** du logiciel R. Les coefficients de l'estimateur lasso du vecteur de régression sont représentés en fonction du paramètre standardisé $s = t / \sum_{j=1}^{p} |\widehat{\beta}_j^{ls}|$. Chaque courbe correspond à un des coefficients du modèle (dont l'étiquette est donnée sur la droite de la figure) en fonction du paramètre de régularisation s. Remarquons que quand $s = 1$, les coefficients sont ceux de l'estimateur des moindres carrés, et qu'ils convergent vers 0 quand $s \to 0$.



5.3.3 Régression bridge

Les méthodes de régression ridge et lasso peuvent être généralisées en considérant l'estimateur défini par

$$\widehat{\beta_n}^{bridge} = argmin_\beta \left(\sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma \right), \tag{5.7}$$

pour $\lambda \geq 0$ et $\gamma \geq 0$. Ces estimateurs, introduits par Frank & Friedman (1993), sont appelés estimateurs *bridge*. La valeur $\gamma = 0$ correspond à la sélection de variables ; $\gamma = 1$ à la régression lasso, tandis que $\gamma = 2$ équivaut à la régression ridge. Par ailleurs, si $\lambda = 0$, on trouve l'estimateur des moindres carrés. Le cas $\gamma = 1$ (lasso) est le paramètre minimal pour lequel la région de contrainte est convexe ; une région non convexe rend le problème d'optimisation plus difficile. Nous nous intéresserons par la suite au cas $\gamma \geq 1$.

Les propriétés asymptotiques des estimateurs bridge dans le modèle de régression linéaire (5.1) *avec variables explicatives fixes* sont étudiées dans Knight & Fu (2000). Le modèle s'écrit donc, pour le i-ème échantillon,

$$Y_i = \beta^t x_i + \varepsilon_i, \tag{5.8}$$

où Y_i est la variable réponse réelle, x_i le p-vecteur de covariables fixes, et ε_i une variable aléatoire d'erreur centrée de variance σ^2. Les conditions de régularité suivantes sont supposées :

(a) $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i x_i^t \underset{n\to\infty}{\longrightarrow} C$, où C est une matrice définie positive,

(b) $\dfrac{1}{n}\displaystyle\max_{1\le i\le n} x_i x_i^t \underset{n\to\infty}{\longrightarrow} 0.$

Les propriétés asymptotiques de l'estimateur bridge $\widehat{\beta_n}^{bridge}$ peuvent être déterminées en étudiant le comportement de la fonction à minimiser $\Pi_n^{\lambda,\gamma}$, définie sur $\mathbb{R}^p$, par

$$\Pi_n^{\lambda,\gamma}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - x_i^t\beta\right)^2 + \frac{\lambda_n}{n}\sum_{j=1}^{p}|\beta_j|^\gamma.$$

Le théorème suivant montre la consistence de $\widehat{\beta_n}^{bridge}$ sous la condition $\lambda_n = o(n)$.

Théorème 5.7. *Si les conditions* (a) *et* (b) *sont vérifiées et si* $\lambda_n/n \to \lambda_0 \ge 0$ *alors*

$$\widehat{\beta_n}^{bridge} \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} \operatorname{argmin}_{\mathbb{R}^p}\Pi^\gamma,$$

où

$$\Pi^\gamma(\beta) = (\beta-\beta_0)^t C(\beta-\beta_0) + \lambda_0\sum_{j=1}^{p}|\beta_j|^\gamma.$$

Ainsi, si $\lambda_n = o(n)$ *(*$\lambda_0 = 0$*),* $\operatorname{argmin}\Pi^\gamma = \beta_0$ *et donc* $\widehat{\beta_n}^{bridge}$ *est consistant.*

Tandis que la condition $\lambda_n = o(n)$ est suffisante pour la consistance de l'estimateur bridge, il est nécessaire que λ_n ait une vitesse plus faible pour sa normalité asymptotique : la condition $\lambda_n = O(\sqrt{n})$ doit être vérifiée.

Théorème 5.8. *Si la condition* (a) *est vérifiée et si* $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ *alors*

$$\sqrt{n}\left(\widehat{\beta_n}^{bridge} - \beta_0\right) \xrightarrow[n\to\infty]{\mathcal{L}} \text{argmin}_{\mathbb{R}^p} M^\gamma,$$

où

$$M^\gamma(\beta) = 2\beta^t W + \beta^t C\beta + \lambda_0 \Psi^\gamma_{\beta_0}(\beta)$$

avec W *un vecteur aléatoire gaussien centré de matrice de covariance* $\sigma^2 C$ *et* Ψ^γ_β *donné par*

$$\Psi^\gamma_\beta(h) = \begin{cases} \gamma \sum_{j=1}^p h_j \text{sgn}(\beta_j)|\beta_j|^{\gamma-1} & si\ \gamma > 1 \\ \sum_{j=1}^p h_j \text{sgn}(\beta_j)\mathbb{1}_{\beta_j\neq 0} + |h_j|\mathbb{1}_{\beta_j=0} & si\ \gamma = 1. \end{cases}$$

Preuve Les théorèmes (5.7) et (5.8) sont démontrés dans l'article de Knight & Fu (2000). ■

Chapitre 6

Estimation de type LASSO dans le modèle de régression linéaire censuré

Sommaire

The content of the present chapter is the one of the paper untitled *LASSO-type estimation in the censored linear regression model* (Detais, 2008), ready to be submitted.

Abstract

Assume that (X_i, Y_i) is a n-sample of random vectors, where X_i is a p-vector of observable covariables and Y_i is the response variable. We suppose Y_i is real at risk of being randomly right-censored. We study the linear regression model $Y = X\beta^t + \varepsilon$, where ε is an error. We introduce a new estimator of the unknown parameter β using a Kaplan-Meier-weighted LASSO-penalized least squares criteria. Results of consistency and asymptotic distribution are obtained. A simulation study is also conducted to investigate the small sample properties of this weighted LASSO-type estimator.

Keywords **: linear regression, censored data, LASSO estimator, bridge estimator, Kaplan-Meier weights, bootstrap estimation.**

6.1 Introduction

The linear regression model is one of the cornerstones of statistical analysis. Its applications include engineering, economics, physical sciences, life and social sciences, management, among many others. Consider the familiar linear regression model

$$Y = X^t\beta + \varepsilon, \tag{6.1}$$

where Y is a one-dimensional response variable, X is an observable p-dimensional vector of covariates, ε is a random error term, $\beta = (\beta_1, \ldots, \beta_p)^t$ is a p-vector of unknown parameters, and t denotes the transpose sign. One problem of interest consists in estimating β from n independent and identically distributed replicates $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). The least squares method is the common approach to estimation in model (6.1) (e.g., Rao et al. (2008)).

In the past three decades, there has been an increasing interest in extending the methods of linear regression analysis to the case of a randomly right-censored response. The response Y is said to be randomly right-censored when one only observes the minimum $Z = \min(Y, C)$ of Y and a random variable C. An obvious example arises when $Y = g(T)$, where T denotes some random failure time and g is a real-valued function (when $g(\cdot) = \ln(\cdot)$, model (6.1) is called the accelerated failure time model; e.g., Bagdonavičius & Nikulin (2002)). In this case, censoring may occur when a patient enrolled in a clinical trial is lost to follow-up, for example. An other example is given by Huang & Harrington (2004), who report a study of HIV infection where the change in HIV viral load (between two consecutive time points) is right-censored when the level of viral load falls below the limit of quantification. Several methods have been proposed to handle right-censored data in model (6.1). For example, the Buckley-James approach (Buckley & James, 1979) replaces a censored observation Y by an estimator of its conditional expectation given the censored value Z and the covariates X. The large sample theory for the resulting estimator of β, and variants of it, has been investigated by Ritov (1990), Lai & Ying (1991), and Jin et al. (2006) among others. One alternative approach relies on weighted least squares (see, among others, Zhou (1992), Stute (1993, 1996)). In this approach, censoring is accounted for by introducing suitable weights in the least squares criterion. In particular, Stute (1993, 1996) proposes and investigates a weighted least squares

estimator, where the weights are given by the jumps of the Kaplan-Meier estimator (Kaplan & Meier, 1958) of the distribution function of Y. Lastly, the synthetic data approach (Koul et al., 1981) is based on a transformation of $Z = \min(Y, C)$, which has the same conditional expectation as Y.

Over the past few years, a large amount of work has also been devoted to penalized estimation in the linear regression model (6.1). Motivated by the new challenge of statistical inference with high dimensional data (arising from microarray gene expression data for example), many investigators have considered the problem of selecting a small subset of covariates with strong effect on the response Y, among a large number of potential predictors. So-called shrinkage methods have been developed for that purpose. For example, the ridge method shrinks the regression coefficients by imposing a ℓ^2-norm penalty on their size, or, equivalently, by minimizing a ℓ^2-norm penalized residual sum of squares. The lasso method (for least absolute shrinkage and selection operator, Tibshirani (1996)) replaces the ℓ^2-norm ridge penalty $\sum_{j=1}^{p} \beta_j^2$ by the ℓ^1-norm penalty $\sum_{j=1}^{p} |\beta_j|$. The so-called bridge regression assumes a penalty of the form $\sum_{j=1}^{p} |\beta_j|^\gamma$ ($\gamma \geq 1$), and includes the lasso and ridge regressions as special cases. We refer to Fu (1998) and Hastie et al. (2001) for a detailed exposition of the relative merits and differences of these methods. Knight & Fu (2000) establish the consistency and weak convergence of the bridge estimator in model (6.1).

In this work, we consider the problem of bridge regression in model (6.1), when the response Y is randomly right-censored. Precisely, we investigate a bridge-penalized weighted least squares approach, where the weights are derived from the Kaplan-Meier estimator of the distribution function of Y. Unlike Stute's (unpenalized) weighted least squares estimator (Stute, 1993, 1996), the bridge-penalized weighted least squares estimator cannot be written in an explicit form when $\gamma \neq 2$. Therefore, the methodology developed by Stute for investigating the asymptotic properties of the Kaplan-Meier weighted least squares estimator cannot be directly applied to the general bridge estimator. However, by combining Stute's techniques and results from the M-estimation theory, we establish the consistency and weak convergence of the proposed estimator. A simulation study is also conducted to investigate the small-sample properties of this estimator.

The paper is organized as follows. In Section 6.2, we describe the model assumptions and we propose a bridge-penalized weighted least squares estimator in model (6.1), when the response variable is right-censored. Section 6.3 establishes the consistency and weak convergence of this estimator. Section 6.4 reports the results of the simulation study. We give some concluding remarks in Section 6.5. Technical proofs are given in Appendix, along with

some useful results from the M-estimation theory.

6.2 Data structure and the bridge estimator

We describe the notations and model assumptions that will be used throughout the paper.

All the random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let Y be a real random variable whose distribution depends on a p-dimensional vector of covariates X. We assume that Y is distributed according to the model (6.1):

$$Y = X^t\beta + \varepsilon,$$

where ε is a random error term and $\beta = (\beta_1, \ldots, \beta_p)^t$ is a p-dimensional vector of unknown parameters to be estimated. In this paper, we consider and investigate the bridge-penalized estimation approach when the response Y is randomly right-censored. Precisely, assume that instead of Y, we observe the random pair (Z, Δ), where $Z = \min(Y, C)$, C is a real random variable, and $\Delta = \mathbb{1}_{Y \leq C}$ is a censoring indicator, which indicates whether Y is complete ($\Delta = 1$) or censored ($\Delta = 0$). Thus, the data consist of n independent and identically distributed replicates (X_i, Z_i, Δ_i), $i = 1, ..., n$, of the triple (X, Z, Δ).

In the uncensored case, Tibshirani (1996) proposed the so-called lasso method for estimation and model selection in model (6.1). The lasso estimator of β minimizes a penalized least square criterion, where the penalty term has the form $\sum_{j=1}^{p} |\beta_j|$. The lasso method permits to retain the good features of both subset selection (which provides interpretable models) and ridge regression (ridge regression uses a penalty of the form $\sum_{j=1}^{p} \beta_j^2$, and ensures a continuous, and thus stable, process). The lasso and ridge estimators are special cases of the bridge estimator, which minimizes the penalized least squares criterion

$$\sum_{i=1}^{n} \left(Y_i - X_i^t\beta\right)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^\gamma, \tag{6.2}$$

where γ is a positive constant and $(\lambda_n)_{n\geq 1}$ is a nonnegative sequence. Bridge estimators were introduced by Frank & Friedman (1993). The asymptotic properties of the bridge estimator of β in model (6.1) were studied by Knight

& Fu (2000). Fu (1998) compared numerically the bridge ($\gamma > 1$) and lasso ($\gamma = 1$) estimators derived from (6.2).

In the censored data situation described above, we propose and investigate a bridge estimator for β. We construct this estimator as a bridge-penalized version of Stute's weighted least squares estimator (Stute, 1993, 1996). Precisely, let $Z_{[1:n]} \leq \ldots \leq Z_{[n:n]}$ denote the ordered values of $Z_1, \ldots, Z_n$, and let $\Delta_{[1:n]}, \ldots, \Delta_{[n:n]}$ (respectively $X_{[1:n]}, \ldots, X_{[n:n]}$) be the corresponding values of Δ (respectively X). When the response variable Y is randomly right-censored, Stute (1993, 1996) suggests estimating β in model (6.1) by minimizing the weighted sum of squares $\sum_{i=1}^{n} W_{in}(Z_{[i:n]} - X_{[i:n]}^t\beta)^2$, where the

$$W_{in} := \frac{\Delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\Delta_{[j:n]}}$$

are the increments of the Kaplan-Meier estimator of the distribution function of Y. The resulting estimator is consistent and asymptotically normal, which essentially follows from the convergence in probability and distribution of rescaled versions of empirical quantities of the form $\sum_{i=1}^{n} W_{in}\Phi(X_{[i:n]}, Z_{i:n})$, where Φ is a real-valued function defined on $\mathbb{R}^{p+1}$.

Mimicking Stute (1993, 1996), we propose the following Kaplan-Meier weighted bridge (KMWB for short, in the sequel) estimator of β:

$$\widehat{\beta_n^{\lambda,\gamma}} = \text{argmin}_{\beta \in \mathbb{R}^p} \Pi_n^{\lambda,\gamma}(\beta), \tag{6.3}$$

where $\Pi_n^{\lambda,\gamma} : \mathbb{R}^p \to \mathbb{R}^+$ is the random function defined as

$$\Pi_n^{\lambda,\gamma}(\beta) = \sum_{i=1}^{n} W_{in} \left(Z_{[i:n]} - X_{[i:n]}^t\beta \right)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^\gamma,$$

where $\lambda := (\lambda_n)_{n \geq 1}$ is a nonnegative sequence, and γ is a positive constant.

In the following, we shall note $\|\beta\|_{\ell^\gamma} = \left(\sum_{j=1}^{p} |\beta_j|^\gamma \right)^{1/\gamma}$ and $\| \cdot \| = \| \cdot \|_{\ell^2}$. We shall also use the notation "sgn" for the sign function. We denote by F (respectively F^0, G) the cumulative distribution function of Y (respectively (X, Y), C). Let β be the collected parameter and β_0 denote the true parameter value. We now make the following additional assumptions :

(a) the error variable ε verifies $\mathbb{E}[\varepsilon|X] = 0$ and $\sigma^2(X) := \mathbb{E}[\varepsilon^2|X] < +\infty$ almost surely,

(b) the matrix $\mathbb{E}[XX^t]$ exists and is positive definite,

(c) Y and C are independent,

(d) Y and C have got continuous distributions and $\inf\{x \in \mathbb{R} | F(x) = 1\} \leq \inf\{x \in \mathbb{R} | G(x) = 1\}$,

(e) Δ and X are independent conditionally on Y.

The next remark states the existence of the proposed estimator.

REMARK 6.1. The KMWB estimator $\widehat{\beta_n^{\lambda,\gamma}}$ of β_0 exists and is achieved. Indeed, we notice that for all $\beta \in \mathbb{R}^p$, $\Pi_n^{\lambda,\gamma}(\beta) \geq \lambda_n \|\beta\|_{\ell^\gamma}^\gamma \underset{\|\beta\| \to +\infty}{\longrightarrow} +\infty$. Hence as $\Pi_n^{\lambda,\gamma}$ is continuous on $\mathbb{R}^p$, the existence of a minimum is stated.

6.3 Asymptotic properties

This section states the asymptotic properties of the proposed estimators. Stute's results (1993) use an explicit expression of the estimator, which cannot be obtained for $\gamma \neq 2$, so we combine these results and the theory of M-estimation developed by van der Vaart & Wellner (1996) to study the asymptotics of the Kaplan-Meier-weighted bridge estimator.

We first determine the limiting behaviour of the KMWBE $\widehat{\beta_n^\lambda}$ by studying the asymptotic behaviour of the random function $\Pi_n^{\lambda,\gamma}$. We obtain the following theorem, which states the consistency of the proposed KMWB estimator in the case $\gamma \geq 1$ and $\lambda_n = o(1)$.

THEOREM 6.1. *Under conditions (a)–(e), if $\gamma \geq 1$ and $\lambda_n \underset{n \to +\infty}{\longrightarrow} \lambda_0 \geq 0$, then almost surely,*

$$\widehat{\beta_n^{\lambda,\gamma}} \underset{n \to +\infty}{\longrightarrow} \operatorname{argmin}_{\mathbb{R}^p} \Pi^\gamma,$$

where

$$\Pi^\gamma(\beta) = (\beta - \beta_0)^t \mathbb{E}(XX^t)(\beta - \beta_0) + \lambda_0 \sum_{j=1}^{p} |\beta_j|^\gamma.$$

Thus if $\lambda_n = o(1)$, $\operatorname{argmin}(\Pi^\gamma) = \beta_0$ and so $\widehat{\beta_n^{\lambda,\gamma}}$ is strongly consistent.

To derive the asymptotic normality of the proposed estimator, we study the convergence in law of the random function M_n defined on $\mathbb{R}^p$ by

$$M_n^{\lambda,\gamma}(h) = n\left(\Pi_n^{\lambda,\gamma}(\beta_0 + \frac{h}{\sqrt{n}}) - \Pi_n^{\lambda,\gamma}(\beta_0)\right),$$

which is minimized at $\sqrt{n}\left(\widehat{\beta_n^{\lambda,\gamma}} - \beta_0\right)$.

Under random censoring, the limit variance becomes much more complicated. We need to introduce the following integrability assumptions ($j \in \{1, ..., p\}$), under which theorem 6.2 will be valid :

$$\int |\varepsilon| \|X\| \sqrt{C(Y)} \mathrm{d}\mathbb{P} < +\infty, \tag{6.4}$$

$$\int \left(\exp\left(\frac{G(Y)}{1-G(Y)}\right)(Z - X^t\beta_0)\right)^2 \Delta X_j \mathrm{d}\mathbb{P} < +\infty \tag{6.5}$$

where

$$C(y) = \int_{-\infty}^{y} \frac{\mathrm{d}G(v)}{(1-F(v))(1-G(v))^2}.$$

Let define the following sub-distribution function :

$$\tilde{H}(x, y) = \mathbb{P}(X \leq x, Z \leq y, \Delta = 1),$$

and introduce for $j \in \{1, ..., p\}$ the functions φ_1^j and φ_2^j defined on $\mathbb{R}$ by :

$$\varphi_1^j(z) = \frac{1}{(1-F(z))(1-G(z))} \int \mathbb{1}_{z<y}(y - x^t\beta_0) x_j \exp\left(\frac{G(y)}{1-G(y)}\right) \mathrm{d}\tilde{H}(x, y)$$

$$\varphi_2^j(z) = \iint \frac{\mathbb{1}_{u<z, u<y}(y - x^t\beta_0) x_j \exp\left(\frac{G(y)}{1-G(y)}\right)}{(1-F(u))(1-G(u))^2} \mathrm{d}G(u)\mathrm{d}\tilde{H}(x, y).$$

Put for $j \in \{1, ..., p\}$

$$\psi_j = (Z - X^t\beta_0) X_j \exp\left(\frac{G(Z)}{1-G(Z)}\right)\Delta + \varphi_1^j(Z)(1-\Delta) - \varphi_2^j(Z)$$

and set

$$\sigma_{i,j} = \mathrm{cov}(\psi_i, \psi_j).$$

Theorem 6.2. *Under conditions (a)–(e), if hypotheses* (6.4), (6.5) *are verified for all* $j \in \{1, ..., p\}$, *if* $\gamma \geq 1$ *and* $\sqrt{n}\lambda_n \underset{n\to+\infty}{\longrightarrow} \lambda_0 \geq 0$, *then*

$$\sqrt{n}(\widehat{\beta_n^{\lambda,\gamma}} - \beta_0) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathrm{argmin}_{\mathbb{R}^p} M^\gamma$$

where

$$M^\gamma(h) = 2h^tW + h^t\mathbb{E}(XX^t)h + \lambda_0\Psi_{\beta_0}^\gamma(h),$$

W has a $\mathcal{N}(0, \Sigma)$ distribution, $\Sigma = (\sigma_{i,j})_{1\leq i,j\leq k}$ and $\Psi_{\beta_0}^\gamma$ is given by

$$\Psi_\beta^\gamma(h) = \begin{cases} \gamma \sum_{j=1}^p h_j \mathrm{sgn}(\beta_j)|\beta_j|^{\gamma-1} & \text{if } \gamma > 1 \\ \sum_{j=1}^p h_j \mathrm{sgn}(\beta_j)\mathbb{1}_{\beta_j\neq 0} + |h_j|\mathbb{1}_{\beta_j=0} & \text{if } \gamma = 1. \end{cases} \tag{6.6}$$

6.4 A simulation study

In order to complete our investigations of the performances of the Kaplan-Meier-weighted bridge estimator in the context of censored data deriving from a linear regression model, we carried out a simulation study under various scenarii.

We consider the linear regression model of equation (6.1) with the same kind of settings as those in Tibshirani (1996). We choose $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $\varepsilon \sim \mathcal{N}(0,1)$, $X \sim \mathcal{N}(0, \Sigma)$, the pairwise correlation between two predictors X_i and X_j being $\Sigma_{i,j} = 0.5^{|i-j|}$. Censoring times are generated from the normal distribution of variance 1 and mean chosen to yield the wanted censoring percentages. Four percentages of censoring (20%, 50%, 70% and 80%) and four sample sizes are considered ($n = 50$, 100, 200 and 400). For each combination of the simulation parameters, 1000 data sets are generated. In these examples, we are interested in the performances of the weighted lasso (WL) estimator $\widehat{\beta_n^\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} \left(\sum_{i=1}^n W_{in} \left(Z_{i:n} - X_{[i:n]}^t \beta \right)^2 + \lambda_n \sum_{j=1}^p |\beta_j| \right)$.

In order to estimate the regularization parameter λ_n in each setting, we use fivefold cross-validation (leave out 20% of the data), as described and studied in Breiman & Spector (1992). The procedure randomly splits the data set into 5 equal-sized parts. For the kth part, for each $\lambda_n \geq 0$, we get a Kaplan-Meier lasso estimate $\widehat{\beta_n^\lambda}$ of β using the other 4 parts as a learning sample. Let denote by E_k the subset of $\{1, ..., n\}$ containing the indexes of the individuals belonging to this k-th part. We use the k-th part as a test sample and calculate the prediction error of the fitted model when predicting the non-censored data of the k-th part :

$$e_k(\lambda_n) = \sum_{i \in E_k} \Delta_i \left(\widehat{Y_{i,k}^{\lambda_n}} - Y_i \right)^2,$$

where $\widehat{Y_{i,k}^{\lambda_n}}$ is the prediction of Y_i using the estimator $\widehat{\beta_n^\lambda}$ obtained from the learning sample. The chosen value of λ_n is the one which minimizes the estimate of the test error

$$e(\lambda_n) = \frac{1}{5} \sum_{k=1}^{5} e_k(\lambda_n).$$

We notice that this procedure prevents us from studying small samples ($n = 50$) with high censoring probability (70 or 80%), some test samples not containing any non censored data.

For comparison, the method which applies a usual lasso estimation procedure to the subset of complete cases (CC) only is also evaluated. The results of a full-data analysis (FD) using the actual values of the response variable Y as if they were known are also obtained. These latter results indeed provide a natural benchmark for evaluating the performance of the Kaplan-Meier-weighted lasso estimator in the setting considered in this paper.

Attaching standard error estimates to weighted lasso estimates is clearly non trivial, watching theorem 6.2. An alternative approach to obtaining standard error estimates is to use the bootstrap. For each data set, we draw 50 independent bootstrap samples and apply our procedure to these samples. The covariance matrix of the 50 obtained bootstrap estimates gives an estimation of the covariance matrix of the weighted lasso estimator (Efron & Tibshirani, 1993).

Tables 1 and 2 summarize the key results of these simulations. For each component of the vector β_0 (corresponding to the 8 rows of one simulation), "bias" is the estimator of the bias constructed from the average mean of the 1000 differences $\widehat{\beta_n^\lambda} - \beta_0$, "mean($\widehat{se}_B$)" denotes the average of the 1000 standard error estimates obtained from the bootstrap procedure and "emp.var" is the sample variance of the 1000 estimates. Finally, "prob" gives, according to the true value of the component of β_0, the estimation of the power or the level of the 5% Wald test for testing nullity of each component, based on the weighted lasso estimate. It is the estimated probability of rejecting the hypothesis of nullity.

Tables 1 and 2 confirm the result of Theorem 6.1, which states the consistency of the weighted lasso estimator. We observe that the ℓ^1-norm of the vector of bias is in all cases smaller than the one of the lasso estimator when all censored data are deleted (CC). The bias can even be divided par more than 15 for the coefficient 3 when the censoring rate is high (80%), or by 5 for the smaller coefficient 1.5. We notice that for high censoring rates, the bias of the weighted lasso estimator is comparable to the one that would be obtained if the censored data were actually observed (FD), which is a very attractive feature. In terms of standard errors, the estimation given by bootstrapping shows that the weighted lasso results are very comparable to the ones obtained by deleting all censored data (CC). In the example of a small sample of 50 data with 50%-censoring, it is interesting to notice that in addition to a lower bias, the weighted lasso estimator may have a standard error divided by 3 compared to the CC method. Finally, the level of the Wald test is comparable for both weighted lasso and complete case methods, whereas the power function of the test based on the weighted lasso estimator is better for high censoring rates (70 or 80%) and as well in other cases.

Table 1. Simulation study with sample sizes $n = 50$ and 100 and various censoring rates (%cens).

	%		Full Data				Weighted Lasso				Complete Case			
n	*cens*	β_0	bias	mean($\widehat{se}_B$)	emp.var	prob	bias	mean($\widehat{se}_B$)	emp.var	prob	bias	mean($\widehat{se}_B$)	emp.var	prob
50	20	3	-0,0739	0,0364	0,0336	1	-0,1935	0,2476	0,0561	1	-0,2090	0,2516	0,0543	1
		1,5	-0,0804	0,0422	0,0396	1	-0,1397	0,2503	0,0531	0,997	-0,1466	0,2519	0,0502	0,997
		0	0,0322	0,0292	0,0164	0,017	0,0323	0,1998	0,0220	0,017	0,0307	0,2023	0,0221	0,017
		0	0,0195	0,0286	0,0170	0,024	0,0244	0,2006	0,0212	0,016	0,0237	0,2029	0,0203	0,015
		2	-0,1237	0,0429	0,0358	1	-0,2237	0,2591	0,0510	1	-0,2275	0,2636	0,0519	1
		0	0,0164	0,0284	0,0158	0,017	0,0274	0,1948	0,0206	0,018	0,0239	0,1974	0,0191	0,014
		0	0,0039	0,0277	0,0139	0,013	0,0048	0,1938	0,0184	0,015	0,0079	0,1922	0,0166	0,013
		0	0,0033	0,0228	0,0125	0,017	0,0036	0,1777	0,0170	0,018	0,0011	0,1775	0,0164	0,012
50	50	3	-0,0702	0,0369	0,0358	1	-0,2809	0,6920	0,1144	0,98	-0,4509	2,0682	0,1680	0,947
		1,5	-0,0809	0,0423	0,0366	1	-0,2150	0,6664	0,1097	0,88	-0,2852	2,1661	0,1152	0,762
		0	0,0182	0,0289	0,0155	0,017	0,0328	0,5580	0,0331	0,017	0,0324	3,1949	0,0389	0,015
		0	0,0300	0,0285	0,0139	0,015	0,0376	0,5800	0,0350	0,01	0,0320	0,7363	0,0408	0,014
		2	-0,1356	0,0423	0,0341	1	-0,3404	0,7413	0,1079	0,93	-0,4329	1,9955	0,1377	0,846
		0	0,0197	0,0278	0,0142	0,02	0,0315	0,9199	0,0352	0,015	0,0308	2,6132	0,0433	0,012
		0	0,0012	0,0279	0,0126	0,012	0,0116	0,8820	0,0288	0,009	0,0107	1,0367	0,0304	0,011
		0	0,0010	0,0225	0,0117	0,018	-0,0034	0,6241	0,0207	0,009	0,0008	2,3545	0,0272	0,014
100	20	3	-0,0539	0,0149	0,0147	1	-0,1579	0,1563	0,0290	1	-0,1715	0,1544	0,0239	1
		1,5	-0,0453	0,0180	0,0151	1	-0,0994	0,1592	0,0245	1	-0,1066	0,1562	0,0216	1
		0	0,0156	0,0125	0,0071	0,015	0,0170	0,1270	0,0117	0,022	0,0184	0,1268	0,0100	0,025
		0	0,0212	0,0125	0,0063	0,02	0,0211	0,1256	0,0101	0,023	0,0203	0,1259	0,0087	0,02
		2	-0,0890	0,0184	0,0159	1	-0,1715	0,1625	0,0253	1	-0,1712	0,1598	0,0235	1
		0	0,0046	0,0124	0,0068	0,019	0,0082	0,1247	0,0096	0,021	0,0037	0,1253	0,0099	0,022
		0	0,0085	0,0116	0,0064	0,02	0,0120	0,1225	0,0088	0,018	0,0095	0,1228	0,0083	0,017
		0	-0,0033	0,0099	0,0057	0,019	-0,0017	0,1113	0,0078	0,021	-0,0024	0,1112	0,0072	0,021
100	50	3	-0,0472	0,0154	0,0153	1	-0,2033	0,2089	0,0471	1	-0,3259	0,2387	0,0486	1
		1,5	-0,0535	0,0176	0,0159	1	-0,1468	0,2149	0,0465	0,998	-0,2028	0,2179	0,0409	0,999
		0	0,0144	0,0122	0,0073	0,019	0,0296	0,1675	0,0166	0,022	0,0256	0,1702	0,0159	0,021
		0	0,0229	0,0122	0,0071	0,02	0,0278	0,1643	0,0178	0,025	0,0250	0,1681	0,0170	0,021
		2	-0,0854	0,0179	0,0154	1	-0,2234	0,2186	0,0457	0,999	-0,2889	0,2314	0,0447	1
		0	0,0111	0,0124	0,0075	0,015	0,0216	0,1653	0,0207	0,035	0,0176	0,1684	0,0167	0,019
		0	0,0073	0,0119	0,0061	0,012	0,0001	0,1639	0,0191	0,02	0,0040	0,1690	0,0153	0,014
		0	0,0008	0,0098	0,0049	0,009	0,0082	0,1473	0,0152	0,021	0,0063	0,1500	0,0122	0,016
100	70	3	-0,0470	0,0156	0,0159	1	-0,1180	0,4256	0,0853	0,998	-0,5382	0,5443	0,1351	0,98
		1,5	-0,0591	0,0180	0,0157	1	-0,1512	0,8466	0,0880	0,955	-0,3522	0,4295	0,0884	0,855
		0	0,0188	0,0126	0,0067	0,016	0,0484	0,3997	0,0353	0,031	0,0402	0,3129	0,0341	0,017
		0	0,0197	0,0127	0,0064	0,016	0,0448	0,9252	0,0338	0,015	0,0290	0,2853	0,0315	0,014
		2	-0,0919	0,0182	0,0137	1	-0,2217	0,4305	0,0805	0,982	-0,4726	0,4894	0,1047	0,926
		0	0,0172	0,0124	0,0065	0,019	0,0382	0,5383	0,0370	0,029	0,0354	0,3166	0,0345	0,019
		0	0,0019	0,0120	0,0064	0,014	0,0050	0,7010	0,0292	0,016	0,0028	0,2957	0,0267	0,008
		0	0,0033	0,0098	0,0054	0,016	0,0060	0,3669	0,0240	0,017	-0,0006	0,2570	0,0226	0,008

NOTE : *Method Full Data shows the results that would be obtained if the response variable were not censored. Method Weighted Lasso refers to the Kaplan-Meier-weighted lasso estimator studied in this paper. Method Complete Case deletes all subjects with censored response variable.*

Table 2. Simulation study with sample sizes $n = 200$ *and* 400 *and various censoring rates (%cens).*

			Full Data				Weighted Lasso				Complete Case			
n	%cens	β_0	bias	mean($\widehat{se}_B$)	emp.var	prob	bias	mean($\widehat{se}_B$)	emp.var	prob	bias	mean($\widehat{se}_B$)	emp.var	prob
200	20	3	-0,0379	0,0073	0,0074	1	-0,1291	0,1105	0,0139	1	-0,1463	0,1050	0,0108	1
		1,5	-0,0335	0,0086	0,0083	1	-0,0872	0,1113	0,0136	1	-0,0884	0,1057	0,0107	1
		0	0,0131	0,0060	0,0034	0,015	0,0147	0,0874	0,0052	0,026	0,0108	0,0867	0,0044	0,02
		0	0,0130	0,0061	0,0039	0,027	0,0185	0,0876	0,0056	0,028	0,0156	0,0869	0,0047	0,023
		2	-0,0598	0,0086	0,0076	1	-0,1354	0,1130	0,0130	1	-0,1380	0,1091	0,0105	1
		0	0,0112	0,0058	0,0033	0,018	0,0113	0,0852	0,0043	0,017	0,0121	0,0853	0,0037	0,011
		0	0,0006	0,0057	0,0028	0,01	0,0006	0,0835	0,0037	0,013	0,0018	0,0844	0,0037	0,016
		0	0,0005	0,0047	0,0028	0,017	0,0029	0,0776	0,0040	0,02	0,0019	0,0775	0,0037	0,015
200	50	3	-0,0278	0,0072	0,0072	1	-0,1782	0,1462	0,0247	1	-0,2785	0,1533	0,0225	1
		1,5	-0,0395	0,0085	0,0074	1	-0,1260	0,1495	0,0237	0,999	-0,1711	0,1399	0,0174	1
		0	0,0095	0,0059	0,0035	0,018	0,0179	0,1159	0,0095	0,017	0,0145	0,1120	0,0079	0,019
		0	0,0116	0,0057	0,0035	0,017	0,0207	0,1149	0,0112	0,035	0,0121	0,1096	0,0076	0,023
		2	-0,0528	0,0084	0,0071	1	-0,1776	0,1492	0,0229	1	-0,2322	0,1470	0,0185	1
		0	0,0090	0,0059	0,0033	0,016	0,0162	0,1148	0,0101	0,025	0,0164	0,1105	0,0075	0,017
		0	0,0004	0,0057	0,0028	0,011	0,0012	0,1121	0,0087	0,017	0,0024	0,1096	0,0067	0,013
		0	0,0008	0,0046	0,0023	0,015	-0,0019	0,1042	0,0075	0,017	-0,0002	0,0994	0,0061	0,015
200	80	3	-0,0329	0,0072	0,0075	1	-0,0294	0,2289	0,0543	1	-0,5552	0,3457	0,0872	0,999
		1,5	-0,0311	0,0084	0,0081	1	-0,0653	0,2648	0,0707	0,992	-0,3159	0,2814	0,0628	0,977
		0	0,0086	0,0060	0,0037	0,02	0,0359	0,2120	0,0313	0,023	0,0225	0,2008	0,0206	0,019
		0	0,0135	0,0058	0,0032	0,013	0,0244	0,2111	0,0268	0,028	0,0188	0,2015	0,0219	0,015
		2	-0,0557	0,0086	0,0076	1	-0,1072	0,2535	0,0600	1	-0,4411	0,3121	0,0684	0,997
		0	0,0100	0,0058	0,0034	0,022	0,0352	0,2059	0,0254	0,026	0,0240	0,1975	0,0185	0,012
		0	0,0006	0,0057	0,0031	0,017	0,0036	0,2011	0,0241	0,014	0,0065	0,1965	0,0182	0,013
		0	0,0018	0,0047	0,0026	0,011	0,0021	0,1858	0,0202	0,018	0,0007	0,1781	0,0185	0,022
400	20	3	-0,0241	0,0035	0,0037	1	-0,1090	0,0804	0,0079	1	-0,1228	0,0731	0,0050	1
		1,5	-0,0217	0,0042	0,0040	1	-0,0697	0,0818	0,0075	1	-0,0726	0,0736	0,0051	1
		0	0,0095	0,0030	0,0018	0,026	0,0137	0,0608	0,0024	0,02	0,0092	0,0608	0,0023	0,025
		0	0,0097	0,0030	0,0017	0,019	0,0118	0,0606	0,0024	0,022	0,0097	0,0608	0,0022	0,015
		2	-0,0376	0,0042	0,0033	1	-0,1060	0,0804	0,0059	1	-0,1076	0,0757	0,0045	1
		0	0,0046	0,0029	0,0018	0,016	0,0075	0,0606	0,0025	0,022	0,0040	0,0606	0,0023	0,016
		0	0,0021	0,0028	0,0014	0,013	0,0019	0,0582	0,0020	0,012	0,0030	0,0593	0,0018	0,016
		0	-0,0002	0,0024	0,0013	0,012	0,0010	0,0538	0,0017	0,01	-0,0004	0,0549	0,0017	0,014
400	50	3	-0,0194	0,0035	0,0035	1	-0,1677	0,1076	0,0152	1	-0,2549	0,1026	0,0103	1
		1,5	-0,0268	0,0041	0,0039	1	-0,1081	0,1108	0,0151	1	-0,1489	0,0946	0,0085	1
		0	0,0094	0,0030	0,0017	0,014	0,0176	0,0856	0,0070	0,027	0,0111	0,0757	0,0036	0,023
		0	0,0082	0,0029	0,0018	0,018	0,0189	0,0859	0,0074	0,039	0,0116	0,0758	0,0038	0,025
		2	-0,0418	0,0041	0,0037	1	-0,1583	0,1114	0,0148	1	-0,2063	0,0996	0,0086	1
		0	0,0074	0,0029	0,0018	0,018	0,0139	0,0850	0,0066	0,034	0,0106	0,0753	0,0032	0,021
		0	0,0038	0,0028	0,0014	0,009	0,0045	0,0818	0,0061	0,025	0,0041	0,0747	0,0032	0,016
		0	-0,0006	0,0023	0,0013	0,016	-0,0013	0,0738	0,0053	0,037	0,0002	0,0678	0,0029	0,024
400	80	3	-0,0244	0,0035	0,0035	1	-0,0262	0,1581	0,0350	1	-0,4841	0,1945	0,0367	1
		1,5	-0,0235	0,0041	0,0041	1	-0,0443	0,1831	0,0422	1	-0,2616	0,1654	0,0261	1
		0	0,0071	0,0030	0,0018	0,016	0,0325	0,1458	0,0211	0,041	0,0158	0,1248	0,0095	0,027
		0	0,0082	0,0030	0,0018	0,017	0,0306	0,1481	0,0213	0,034	0,0175	0,1267	0,0106	0,028
		2	-0,0383	0,0042	0,0037	1	-0,0806	0,1762	0,0385	1	-0,3693	0,1792	0,0299	1
		0	0,0056	0,0029	0,0019	0,021	0,0201	0,1468	0,0195	0,029	0,0128	0,1240	0,0105	0,018
		0	0,0016	0,0029	0,0015	0,013	0,0116	0,1405	0,0178	0,026	0,0001	0,1224	0,0102	0,029
		0	0,0015	0,0024	0,0012	0,013	0,0025	0,1299	0,0159	0,038	0,0075	0,1106	0,0083	0,02

NOTE : *Method Full Data shows the results that would be obtained if*

the response variable were not censored. Method Weighted Lasso refers to the Kaplan-Meier-weighted lasso estimator studied in this paper. Method Complete Case deletes all subjects with censored response variable.

6.5 Discussion

In this paper, we have adapted bridge regression in the linear model to right-censored response data, considering a Kaplan-Meier-weighted bridge-penalized least squares approach. The proposed estimator for the regression parameter in this model has been shown to be consistent and its convergence in distribution has been studied in the case $\gamma \geq 1$. Computations for the Kaplan-Meier-weighted lasso estimator are conclusive, first because the time to calculate the estimator for one data set does not exceed a few seconds, second because they show that it is a very interesting alternative to the inefficient complete-case analysis. It illustrates that the lasso method has, in the case of right-censored data also, the nice features of prediction accuracy and interpretation (since it shrinks some parameters estimators to 0).

The case of the Kaplan-Meier-weighted bridge-penalized estimator in the case $\gamma < 1$ still has to be studied. The main difficulty remains in the non convexity of the function to be minimized. It is a subject for future work.

6.6 Appendix. Technical Proofs

Proof of Theorem 6.1 This proof draws on a theorem given for instance in Rockafellar (1970) that we state here.

Theorem 6.3. *Let E be an open convex subset of $\mathbb{R}^p$ and let $(f_n)_{n\geq 1}$ be a sequence of real random convex functions on E such that for all $x \in E$, $f_n(x) \to f(x)$ almost surely as $n \to +\infty$ where f is some real function on E. Then f is also convex and if f has a unique minimum on E,* $\operatorname{argmin}_E(f_n) \to \operatorname{argmin}_E(f)$ *almost surely as $n \to +\infty$.*

Set $\gamma \geq 1$. We first show that for all $n \in \mathbb{N}^*$, $\Pi_n^{\lambda,\gamma}$ is a convex function. Let f be the function defined on $\mathbb{R}^p$ by $f(\beta) = \sum_{i=1}^n W_{in} \left(Z_{i:n} - X_{[i:n]}^t \beta\right)^2$. Then the gradient of f at $\beta \in \mathbb{R}^p$ is $\nabla_\beta(f) = -2\sum_{i=1}^n W_{in} \left(Z_{i:n} - X_{[i:n]}^t \beta\right) X_{[i:n]}$. Hence, for all $\alpha, \beta \in \mathbb{R}^p$,

$$\langle \nabla_\beta(f) - \nabla_\alpha(f), \beta - \alpha \rangle = 2\sum_{i=1}^n W_{in} \left(X_{[i:n]}^t(\alpha - \beta)\right)^2 \geq 0.$$

Therefore, as f is a convex function, λ_n a nonnegative real and $\gamma \geq 1$ ($\|.\|_{\ell^\gamma}$ is a norm hence convex and $x \mapsto x^\gamma$ a nondecreasing convex function), $\Pi_n^{\lambda,\gamma}$ is a convex function.

Applying results from Stute (1993), we now get the pointwise almost sure convergence of $\Pi_n^{\lambda,\gamma}$ to the function $\Pi^\gamma + \sigma^2$ under the hypotheses (a)–(e) :

$$\begin{aligned}
\mathbb{E}\left((Y - X^t\beta)^2\right) &= \mathbb{E}\left(\varepsilon^2 + 2\varepsilon X^t(\beta - \beta_0) + \left(X^t(\beta - \beta_0)\right)^2\right) \\
&= \sigma^2 + 2\mathbb{E}\left(\mathbb{E}(\varepsilon|X)X^t(\beta - \beta_0)\right) + \mathbb{E}\left(\left(X^t(\beta - \beta_0)\right)^2\right) \\
&= \sigma^2 + (\beta - \beta_0)^t\mathbb{E}(XX^t)(\beta - \beta_0).
\end{aligned}$$

We now show that Π^γ has a unique minimum to state the existence of $\operatorname{argmin}_{\mathbb{R}^p}\Pi^\gamma$. First, thanks to the hypothesis (b), we get that Π^γ is a strictly convex function on $\mathbb{R}^p$. Then, denoting by $\lambda_{min} > 0$ the smaller eigenvalue of $\mathbb{E}(XX^t)$, we get (with K a positive constant whose existence comes from the equivalence of the norms $\|.\|_{\ell^\gamma}$ and $\|.\|$ in $\mathbb{R}^p$) :

$$\begin{aligned}
\Pi^\gamma(\beta) &= (\beta - \beta_0)^t\mathbb{E}(XX^t)(\beta - \beta_0) + \lambda_0\|\beta\|_{\ell^\gamma}^\gamma \\
&\geq \lambda_{min}\|\beta - \beta_0\|^2 + \lambda_0 K\|\beta\|^\gamma \\
&\geq \|\beta\|\left(\lambda_{min}\|\beta\| + \lambda_0 K\|\beta\|^{\gamma-1} - 2\lambda_{min}\|\beta_0\|\right) + \lambda_{min}\|\beta_0\|^2.
\end{aligned}$$

Hence $\Pi^\gamma(\beta) \underset{\|\beta\|\to+\infty}{\longrightarrow} +\infty$ and Π^γ has a unique minimum on $\mathbb{R}^p$.

Theorem 6.3 enables us to give the conclusion. ■

Proof of Theorem 6.2 Let rewrite for $h \in \mathbb{R}^p$

$$\begin{aligned} M_n^{\lambda,\gamma}(h) &= n\sum_{i=1}^{n} W_{in}\left(\left(Z_{i:n} - X_{[i:n]}^t(\beta_0 + \frac{h}{\sqrt{n}})\right)^2 - \left(Z_{i:n} - X_{[i:n]}^t\beta_0\right)^2\right) \\ &\quad + n\lambda_n\left(\sum_{j=1}^{p}|\beta_{0,j} + \frac{h_j}{\sqrt{n}}|^\gamma - |\beta_{0,j}|^\gamma\right) \\ &= -2h^t\sqrt{n}\sum_{i=1}^{n} W_{in}\left(Z_{i:n} - X_{[i:n]}^t\beta_0\right)X_{[i:n]} + \sum_{i=1}^{n} W_{in}(X_{[i:n]}^t h)^2 \\ &\quad + n\lambda_n\left(\sum_{j=1}^{p}|\beta_{0,j} + \frac{h_j}{\sqrt{n}}|^\gamma - |\beta_{0,j}|^\gamma\right) \\ &= -2h^t\sqrt{n}\sum_{i=1}^{n} W_{in}\varepsilon_{[i:n]}X_{[i:n]} + h^t\left(\sum_{i=1}^{n} W_{in}X_{[i:n]}X_{[i:n]}^t\right)h \\ &\quad + n\lambda_n\left(\sum_{j=1}^{p}|\beta_{0,j} + \frac{h_j}{\sqrt{n}}|^\gamma - |\beta_{0,j}|^\gamma\right) \end{aligned}$$

Let study the limit when $n \to +\infty$ of each term of this sum. First, the determinist term $\sum_{j=1}^{p}|\beta_{0,j} + \frac{h_j}{\sqrt{n}}|^\gamma - |\beta_{0,j}|^\gamma$ is equivalent to

$$\sum_{j=1}^{p}\frac{\gamma \mathrm{sgn}(\beta_j)|\beta_j|^{\gamma-1}h_j}{\sqrt{n}}\mathbb{1}_{\beta_{0,j}\neq 0} + \frac{|h_j|^\gamma}{n^{\gamma/2}}\mathbb{1}_{\beta_{0,j}=0}.$$

Therefore, let introduce the function Ψ_β^γ defined on $\mathbb{R}^p$ by

$$\Psi_\beta^\gamma(h) = \begin{cases} \sum_{j=1}^{p}\gamma h_j \mathrm{sgn}(\beta_j)|\beta_j|^{\gamma-1} & \text{if } \gamma > 1 \\ \sum_{j=1}^{p} h_j \mathrm{sgn}(\beta_j)\mathbb{1}_{\beta_j\neq 0} + |h_j|\mathbb{1}_{\beta_j=0} & \text{if } \gamma = 1. \end{cases} \tag{6.7}$$

Then, we get under the hypothesis $\sqrt{n}\lambda_n \to \lambda_0$ the convergence

$$n\lambda_n\left(\sum_{j=1}^{p}|\beta_{0,j} + \frac{h_j}{\sqrt{n}}|^\gamma - |\beta_{0,j}|^\gamma\right) \underset{n\to+\infty}{\longrightarrow} \lambda_0\Psi_{\beta_0}^\gamma(h).$$

Second, thanks to hypotheses (b), (c) and (e), the theorem of Stute (1993) gives

$$\sum_{i=1}^{n} W_{in}X_{[i:n]}X_{[i:n]}^t \underset{n\to+\infty}{\overset{a.s.}{\longrightarrow}} \mathbb{E}(XX^t).$$

Finally the convergence of $\sqrt{n}\sum_{i=1}^{n} W_{in}\varepsilon_{[i:n]}X_{[i:n]}$ has to be studied. Under hypotheses (a)–(e), (6.4), (6.5) verified for all $j \in \{1, ..., p\}$, theorem 1.2 of Stute (1996) gives for all $h \in \mathbb{R}^p$

$$\sqrt{n}\sum_{i=1}^{n} W_{in}\varepsilon_{[i:n]}X_{[i:n]} \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \Sigma).$$

Hence, we get for all $h \in \mathbb{R}^p$

$$M_n^{\lambda,\gamma}(h) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} M^{\gamma}(h).$$

The finite-dimensional convergence is holding trivially since the vector h can be factorized in each component of $M_n^{\lambda,\gamma}(h)$.

Next, the convexity of function $M_n^{\lambda,\gamma}$ comes from the convexity of $\Pi_n^{\lambda,\gamma}$ showed in the proof of theorem 6.1.

Finally, M^{γ} is shown to have a unique minimum on R^p. Hypothesis (b) actually gives the strict convexity of M^{γ} : it yields that $h \mapsto h^t\mathbb{E}(XX^t)h$ is strictly convex and M^{γ} is the sum of this function, a linear function and $\Psi_{\beta_0}^{\gamma}$ which is convex (if $\gamma \geq 1$, it is linear, if $\gamma = 1$ it is convex due to the convexity of $|.|$). Furthermore, denoting by $|h|$ the vector composed of the absolute values of the components of h and $\lambda_{min} > 0$ the smallest eigenvalue of $\mathbb{E}(XX^t)$, we have for $\gamma = 1$

$$\begin{aligned} |M^{\gamma}(h)| &\geq |h^t\mathbb{E}(XX^t)h| - |\langle h, 2W + (\mathrm{sgn}(\beta_{0,j})\mathbb{1}_{\beta_{0,j}\neq 0})_j\rangle + \langle |h|, (\mathbb{1}_{\beta_{0,j}=0})_j\rangle| \\ &\geq \lambda_{min}\|h\|^2 - \|h\|.\|2W + (\mathrm{sgn}(\beta_{0,j})\mathbb{1}_{\beta_{0,j}\neq 0})_j\| - \|h\|.\|(\mathbb{1}_{\beta_{0,j}=0})_j\| \underset{\|h\|\to+\infty}{\xrightarrow{a.s.}} +\infty. \end{aligned}$$

The proof for $\gamma > 1$ is very similar. So the minimum of the continuous function M^{γ} is almost surely achieved at a unique point.

We now state the corollary of a theorem given in Geyer (1996) which will enable to give the conclusion of theorem (6.2).

Theorem 6.4. *Suppose $(f_n)_{n\geq 1}$ is a sequence of random continuous convex functions on $\mathbb{R}^p$ and f is another such function. If for each finite subset $(x_1, ..., x_k)$ of $\mathbb{R}^p$, the random vector $(f_n(x_1), ..., f_n(x_k))$ converges in law to the random vector $(f(x_1), ..., f(x_k))$ and if with probability one f is finite and has a unique minimizer, then* $\mathrm{argmin}_{\mathbb{R}^p}(f_n)$ *converges in law to* $\mathrm{argmin}_{\mathbb{R}^p}(f)$.

Since $M_n^{\lambda,\gamma}$ is convex continuous, M^γ is another such function with a unique minimizer and the finite-dimensional convergence in law was shown, it follows from Theroem 6.4 that

$$\sqrt{n}(\widehat{\beta_n^{\lambda,\gamma}} - \beta_0) = \text{argmin}_{R^p}(M_n^{\lambda,\gamma}) \xrightarrow[n\to+\infty]{\mathcal{L}} \text{argmin}_{R^p}(M^\gamma).$$

■

Chapitre 7

Estimation bootstrap de l'écart-type

Sommaire

La méthode *bootstrap*, introduite par Efron (1979), a pour objectifs premiers l'estimation de paramètres d'intérêt et l'évaluation de la précision d'un estimateur via l'estimation de son écart-type ou la détermination d'intervalles de confiance. La généralité du boostrap a par la suite permis son application à de nombreux autres types de problèmes, comme la sélection de modèles en régression, la régression non linéaire ou encore l'analyse de séries chronologiques. Nous souhaitons ici appliquer cette méthode pour évaluer l'écart-type et la puissance du test de Wald associés à l'estimateur de type LASSO introduit dans le chapitre 6.

Nous commençons par expliquer les mécanismes de la démarche bootstrap. Supposons, en toute généralité, que nous disposons d'un échantillon de taille n et que nous souhaitons estimer un paramètre ou déterminer son écart-type. Considérons la distribution empirique associée à l'échantillon, qui associe la probabilité $1/n$ à chaque valeur de l'échantillon. L'idée de base du bootstrap est simplement de remplacer la distribution inconnue de l'échantillon par cette distribution empirique qui estime simplement la distribution de l'échantillon et qui, elle, est connue. Les propriétés de l'estimateur, comme son écart-type, peuvent alors être déterminées sur la base de la distribution empirique.

7.1 Estimation plug-in

Expliquons maintenant plus formellement la démarche d'obtention de l'estimateur bootstrap de l'écart-type et introduisons pour cela la démarche du plug-in. Soit $(X_1, ..., X_n)$ un n-uplet de vecteurs aléatoires indépendant et identiquement distribués de loi P_X. Notons $\mathbb{P}_n$ la distribution empirique associée à l'échantillon $(X_i)_{1 \leq i \leq n}$, définie précédemment dans le chapitre 1. C'est un estimateur de la distribution P_X ayant de bonnes propriétés de convergence (théorème de Glivenko-Cantelli). Le principe du *plug-in* est une méthode simple pour l'estimation de paramètres à partir d'échantillons.

Définition 7.1. *L'*estimateur plug-in *d'un paramètre $\theta = t(P_X)$ fonction de la distribution P_X est défini par*

$$\hat{\theta} = t(\mathbb{P}_n).$$

Il estime la fonction $\theta = t(P_X)$ de la loi de probabilité P_X par la même fonction de la distribution empirique $\mathbb{P}_n$, $\hat{\theta} = t(\mathbb{P}_n)$.

Les exemples de base d'application du principe plug-in sont bien connus de tout statisticien.

Exemple 7.1. L'estimateur plug-in de l'espérance $E_{P_X}(X)$ d'une variable aléatoire réelle X, obtenu à partir d'un échantillon $(X_i)_{1 \leq i \leq n}$ issu de la distribution P_X, est $E_{\mathbb{P}_n}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$.

Exemple 7.2. L'estimateur plug-in de l'écart-type
$\sigma_{P_X}(X) = ((E_{P_X}(X - E_{P_X}(X))^2))^{1/2}$ d'une variable aléatoire réelle X, obtenu à partir de l'échantillon $(X_i)_{1 \leq i \leq n}$, est $\sigma_{\mathbb{P}_n}(X) = \hat{\sigma} = \left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2\right)^{1/2}$.

Ces exemples permettent notamment d'évaluer la précision de $\bar{X}$ comme estimateur de l'espérance de la distribution des données. En effet, nous savons que si X a pour moyenne μ et variance σ^2 alors la moyenne empirique $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ a pour moyenne μ et variance $\frac{\sigma^2}{n}$ et donc pour écart-type $se_{P_X}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. D'après les exemples précédents, un estimateur de l'écart-type de la moyenne empirique est

$$se_{\mathbb{P}_n}(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{1}{n} \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^{1/2}.$$

On reconnaît, à une constante multiplicative près (tendant vers 1), l'estimateur usuel de l'écart-type de la moyenne empirique. Pour la plupart des

objectifs, $\frac{\hat{\sigma}}{\sqrt{n}}$ donne d'aussi bons résultats dans l'estimation de $se_{P_X}(X)$. Dans cette application, le principe du plug-in a été utilisé deux fois : d'abord pour l'estimation de la moyenne μ par $\bar{X}$, puis dans l'estimation de l'écart-type $se_{P_X}(X)$.

L'estimation bootstrap de l'écart-type, sujet de la section suivante, revient à utiliser le principe plug-in pour l'estimation de l'écart-type d'une statistique arbitraire $\hat{\theta}$.

7.2 Estimation bootstrap de l'écart-type

Supposons que nous disposons d'un échantillon $\underline{X} = (X_1, ..., X_n)$ issu d'une loi de probabilité P_X inconnue et que nous souhaitons estimer un paramètre d'intérêt $\theta = t(P_X)$ en nous basant sur l'échantillon $\underline{X}$. Dans cette optique, nous calculons un estimateur $\hat{\theta} = s(\underline{X})$, qui peut être, ou ne pas être, l'estimateur plug-in $t(\mathbb{P}_n)$. Nous cherchons à connaître la précision de cet estimateur. Le *bootstrap*, introduit par Efron (1979), est une méthode basée sur ordinateur pour l'estimation de l'écart-type de $\hat{\theta}$. L'estimateur bootstrap de cet écart-type ne nécessite aucun calcul théorique et est disponible quelle que soit la complexité de l'écriture de l'estimateur $\hat{\theta} = s(\underline{X})$.

La méthode bootstrap s'appuie sur la notion d'*échantillon bootstrap*. Soit $\mathbb{P}_n$ la distribution empirique associée à l'échantillon $\underline{X}$, affectant la probabilité $1/n$ à chacune des valeurs observées X_i $(1 \leq i \leq n)$.

Définition 7.2. *Un* échantillon bootstrap *est défini comme un échantillon de taille n issu de la loi $\mathbb{P}_n$. On le note $\underline{X^*} = (X_1^*, ..., X_n^*)$, la notation étoile $*$ indiquant que $\underline{X^*}$ n'est pas le véritable échantillon $\underline{X}$ mais une version rééchantillonnée de $\underline{X}$.*

On peut résumer la situation de cette façon :

Échantillon		Échantillon Bootstrap
$\underline{X} = (X_1, ..., X_n)$ généré par P_X de loi empirique $\mathbb{P}_n$	$\rightsquigarrow$	$\underline{X^*} = (X_1^*, ..., X_n^*)$ généré par $\mathbb{P}_n$ de loi empirique $\mathbb{P}_n^*$.

Autrement dit, l'échantillon bootstrap $\underline{X^*} = (X_1^*, ..., X_n^*)$ est un échantillon de taille n tiré équiprobablement et avec remise dans la population des n objets $(X_1, ..., X_n)$. Il se peut donc par exemple qu'on obtienne $X_1^* = X_7$, $X_2^* = X_3$, $X_3^* = X_3$, $X_4^* = X_{22}$, ..., $X_n^* = X_7$. L'échantillon bootstrap

$(X_1^*, ..., X_n^*)$ est constitué d'éléments de l'échantillon de départ $(X_1, ..., X_n)$, certains éléments n'apparaissant pas, d'autres une fois, d'autres deux fois etc.

Définition 7.3. *Une* réplication bootstrap *de* $\hat{\theta} = s(\underline{X})$ *est défini par l'application de la même fonction* s *à l'échantillon bootstrap* $\underline{X^*}$

$$\hat{\theta}^* = s(\underline{X^*}).$$

L'estimateur bootstrap idéal de l'écart-type $se_{P_X}(\hat{\theta})$ de la statistique $\hat{\theta}$ est l'estimateur plug-in qui remplace la loi inconnue P_X par la distribution empirique $\mathbb{P}_n$. Il est donc défini par

$$se_{\mathbb{P}_n}(\hat{\theta}^*).$$

C'est l'écart-type de $\hat{\theta}$ pour des échantillons de taille n issus de la distribution $\mathbb{P}_n$. Il est appelé estimateur bootstrap *idéal* de l'écart-type de $\hat{\theta}$ car pour un estimateur quelconque $\hat{\theta}$, différent de la moyenne empirique, on ne dispose pas a priori de formule permettant de calculer une valeur numérique de $se_{\mathbb{P}_n}(\hat{\theta}^*)$.

Il est cependant aisé d'implémenter le rééchantillonnage bootstrap sur ordinateur et l'algorithme bootstrap fonctionne en construisant plusieurs échantillons de type bootstrap, en évaluant la réplication bootstrap associée à chacun, et en estimant l'écart-type de $\hat{\theta}$ par l'écart-type empirique des réplications.

Définition 7.4. *L'*estimateur bootstrap de l'écart-type *de* $\hat{\theta}$ *est défini par*

$$\widehat{se}_B = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}_{boot} \right)^2 \right)^{1/2},$$

où $\hat{\theta}_b^*$ *est la réplication bootstrap de* $\hat{\theta}$ *associée au* b*-ème l'échantillon bootstrap* ($b \in \{1, ..., B\}$) *et* $\hat{\theta}_{boot}$ *est la moyenne des réplications bootstrap de* $\hat{\theta}$ *:*

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*.$$

Ainsi, voici l'algorithme permettant l'estimation bootstrap de l'écart-type d'une statistique.

Algorithme bootstrap pour l'estimation de l'écart-type de $s(\underline{X})$

- **Étape 1** : Sélection de B échantillons bootstrap indépendants $\underline{X}_1^*$, $\underline{X}_2^*$, ..., $\underline{X}_B^*$ issus de la loi $\mathbb{P}_n$.

- **Étape 2** : Evaluation de la réplication bootstrap associée à chacun des échantillons bootstrap

$$\hat{\theta}_b^* = s(\underline{X}_b^*) \quad (1 \leq b \leq B).$$

- **Étape 3** : Estimation de l'écart-type $se_{P_X}(\hat{\theta})$ par l'écart-type empirique des B réplications

$$\widehat{se}_B = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}_{boot} \right)^2 \right)^{1/2},$$

où $\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$.

La limite de $\widehat{se}_B$ quand B tend vers l'infini est l'estimateur bootstrap idéal de $se_{P_X}(\hat{\theta})$:

$$\lim_{b \to +\infty} \widehat{se}_B = se_{\mathbb{P}_n}(\hat{\theta}^*).$$

Le nombre de réplications bootstrap B sera habituellement situé entre 25 et 200. Il a été montré expérimentalement que même 25 réplications bootstrap donnent suffisamment d'information pour obtenir un bon estimateur de $se_{P_X}(\hat{\theta})$.

Un développement plus détaillé sur le principe du bootstrap pourra être trouvé par le lecteur intéressé dans Efron & Tibshirani (1986) et Chernick (1999).

7.3 Application à l'estimateur de type LASSO dans le modèle de régression linéaire censuré

Nous souhaitons estimer la précision de l'estimateur de type LASSO introduit dans le chapitre 6. En effet, un théorème de consistance (5.1) ainsi qu'un théorème de normalité asymptotique (5.4) ont été obtenus mais nous ne disposons pas de formule explicite permettant d'estimer la précision de

cet estimateur $\widehat{\beta_n^{\lambda,\gamma}}$. Nous décidons donc d'estimer l'écart-type de $\widehat{\beta_n^{\lambda,\gamma}}$ grâce à la méthode bootstrap.

Les simulations ont été effectuées sur 1000 échantillons indépendants afin que les moyennes soient significatives. Nous avons, pour chacun de ces échantillons, construit 50 échantillons bootstrap pour lesquels l'estimateur associé $\widehat{\beta_n^{\lambda,\gamma}}^*_b$ a été construit. Ces estimateurs bootstrap ont permis, grâce à l'utilisation de l'algorithme donné dans la partie précédente, de donner une estimation de l'écart-type de $\widehat{\beta_n^{\lambda,\gamma}}$, notée $\widehat{se}_B$. Celle-ci a par ailleurs été utilisée pour construire le test de Wald de nullité de chaque composante et établir son niveau ou sa puissance (suivant la vraie valeur de la composante). Les estimations de l'écart-type $\widehat{se}_B$ ont été moyennées sur les 1000 échantillons afin d'obtenir les colonnes intitulées mean($\widehat{se}_B$) dans les tableaux *Table 1* et *Table 2* du chapitre 6.

Les programmes permettant la mise en oeuvre du calcul des estimateurs proposés dans le chapitre 6 $\widehat{\beta_n^{\lambda,\gamma}}$, avec $\gamma = 1$ pour l'application, sont donnés dans l'addenda.

Conclusion et perspectives

Motivés par les nombreuses applications possibles de l'analyse de durées de vie censurées, notre travail a consisté à proposer de nouveaux estimateurs et étudier leurs propriétés dans deux types de modèles différents.

Le modèle de Cox stratifié permet de s'intéresser à une population dont certaines strates possèdent une fonction d'intensité de base différente mais de paramètre de régression commun à tous les individus. Nous avons estimé ce paramètre de régression lorsque l'indicateur de strate est aléatoirement manquant. Dans ce modèle, nous avons proposé un estimateur du maximum de vraisemblance non paramétrique et montré ses consistance, normalité et efficacité asymptotiques. De plus, la méthode espérance-maximisation proposée et explicitée dans les détails permet d'implémenter cet estimateur.

Il serait maintenant intéressant de l'appliquer à des données simulées ou données réelles afin de comparer ses résultats à ceux d'autres méthodes d'estimation, à taille d'échantillon fixée. Il est envisageable de comparer les résultats de cet estimateur de maximum de vraisemblance non paramétrique à l'estimateur régression-calibration ou à l'estimateur du maximum de vraisemblance obtenu en utilisant l'échantillon nettoyé de tous les individus dont la strate n'est pas observée.

Dans ce travail sur le modèle à risques proportionnels stratifié, la loi conditionnelle de la strate est supposée être une loi multinômiale logistique, couramment utilisée en recherche médicale. Les résultats théoriques obtenus dans le cadre de cette loi devraient pouvoir être étendus à d'autres fonctions de lien. Une autre perspective serait de s'intéresser à des covariables dépendant du temps. Des résultats du même type devraient être obtenus sous de bonnes conditions de régularité. Cela ouvrirait des possibilités d'application encore plus larges, notamment en médecine.

Dans un deuxième temps, nous nous sommes intéressés au modèle de

régression linéaire censuré aléatoirement à droite. Nous avons introduit un estimateur du paramètre de régression de type LASSO, motivé par l'émergence de données en très grande dimension. Nous avons montré que l'estimateur proposé, de type Kaplan-Meier ℓ^γ-pénalisé (avec $\gamma \geq 1$), est fortement consistant et asymptotiquement normal. Une étude de simulation est proposée pour l'estimateur LASSO pondéré Kaplan-Meier et donne des résultats très concluants en comparaison avec l'estimateur LASSO calculé sur l'échantillon nettoyé des données effectivement censurées.

Il serait intéressant d'étendre ces simulations à l'estimateur de type Kaplan-Meier ℓ^γ-pénalisé ($\gamma > 1$), ce qui demande l'élaboration de programmes informatiques plus complexes. Enfin, la perspective d'étude de l'estimateur pondéré Kaplan-Meier et ℓ^γ-pénalisé avec $\gamma < 1$ est également très attrayante, la difficulté principale résidant dans le fait que la fonction à minimiser pour l'obtention de l'estimateur n'est pas convexe.

Addenda

Dans cet addenda, nous présentons les principales fonctions programmées dans le langage **R** pour la mise en oeuvre du calcul des estimateurs proposés dans le chapitre 6 et permettant d'obtenir les tableaux *Table 1* et *Table 2*.

La fonction `completvar`

Description : La fonction `completvar` permet de comparer l'estimateur LASSO des moindres carrés pondérés par des poids de Kaplan-Meier à l'estimateur LASSO obtenu en supprimant toutes les données censurées. L'estimateur LASSO qui aurait été obtenu si la censure avait été absente est également considéré pour établir une référence.
Cette fonction nécessite l'installation des packages de **R** appelés **boot** et **lars**.

Arguments :

`N`	nombre d'échantillons	`N` $= 1000$
`n`	taille de chacun des échantillons	`n` $= 50, 100, 200$ ou 400
`beta`	le paramètre de régression	`beta` $= (3, 1.5, 0, 0, 2, 0, 0, 0)$
`sigma`	écart-type de l'erreur	`sigma` $= 1$
`mu`	moyenne de la censure	`mu` $= 0$ pour 50% de censure
`tau`	écart-type de la censure	`tau` $= 1$
`m`	moyenne des covariables	`m` $= 0_{\mathbb{R}^p}$ (p dimension de `beta`)
`rho`	paramètre de la matrice de covariance des covariables X	`rho` $= 0.5$
`K`	nombre de sous-échantillons dans la validation croisée	`K` $= 5$
`R`	nombre d'échantillons bootstrap	`R` $= 50$

Pour la valeur `beta` $= (3, 1.5, 0, 0, 2, 0, 0, 0)$, on choisit `mu` $= -4.06$ pour 80% de censure, -2.53 pour 70% de censure, 0 pour 50% de censure, 4.06 pour 20% de censure.

Valeurs de sortie : Une liste contenant les composantes (beta,betmoyFD,biasFD,semoyFD,betvarFD,puisFD,errFD,seerrFD,zerosFD, betmoyLC,biasLC,semoyLC,betvarLC,puisLC,errLC,seerrLC,zerosLC, betmoyCC,biasCC,semoyCC,betvarCC,puisCC,errCC,seerrCC,zerosCC), où

`beta`	vecteur de régression
`betmoyXX`	la moyenne des estimateurs de `beta` donnés par la méthode `XX`
`biasXX`	le biais moyen des estimateurs de `beta` donnés par la méthode `XX`
`semoyXX`	écart-type estimé grâce au bootstrap de chacune des composantes de l'estimateur de `beta` donné par la méthode `XX`,
`betvarXX`	variance empirique de chacune des composantes de l'estimateur de `beta` donné par la méthode `XX`,
`puisXX`	puissance (ou niveau) du test de Wald testant la nullité de chaque composante à partir de l'estimateur de `beta` donné par la méthode `XX`,
`errXX`	moyenne de l'erreur ℓ^2 au carré de l'estimateur de `beta` donné par la méthode `XX`,
`seerrXX`	écart-type empirique de l'erreur ℓ^2 au carré de l'estimateur de `beta` donné par la méthode `XX`,
`zerosXX`	moyenne du nombre de composantes nulles de l'estimateur de `beta` donné par la méthode `XX`,

Code :

```
completvar<-function(N=1000,n=100,beta,sigma,mu,tau,
m=rep(0,length(beta)),rho=0.5,K=5,R=50) {

p=length(beta)

D=donneesLASSO(N,n,beta,mu,tau,m,sigma,rho)

betachapLC=matrix(0,p,N) betachapCC=matrix(0,p,N)
betachapFD=matrix(0,p,N)

betachapvarLC=array(data = NA, dim = c(N,p,p))
betachapvarCC=array(data = NA, dim = c(N,p,p))
betachapvarFD=array(data = NA, dim = c(N,p,p))

for (j in 1:N)
{
Z=D$Z[,j] Y=D$Y[,j]
del=D$del[,j]
X=D$X[,(1+(j-1)*n):(j*n)]
data=matrix(c(Z,del,t(X)),nrow=n,ncol=(p+2))

fraction = seq(from = 0, to = 1, length = 100)

boutLC=censboot(data, statistic=censfun, R=R, sim="ordinary", K=K,
fraction=fraction, plot.it=FALSE, trace=FALSE)

boutCC=censboot(data, statistic=censfuncomp, R=R, sim="ordinary",
K=K, fraction=fraction, plot.it=FALSE, trace=FALSE)
```

```
boutFD=censboot(matrix(c(Y,rep(1,n),t(X)),nrow=n,ncol=(p+2)),
statistic=censfuncomp, R=R, sim="ordinary", K=K,
fraction=fraction, plot.it=FALSE, trace=FALSE)

estboutLC=boutLC$t estboutCC=boutCC$t estboutFD=boutFD$t
betachapvarLC[j,,]=var(estboutLC)
betachapvarCC[j,,]=var(estboutCC)
betachapvarFD[j,,]=var(estboutFD) betachapLC[,j]=boutLC$t0
betachapCC[,j]=boutCC$t0 betachapFD[,j]=boutFD$t0
}

errLC=mean(apply((betachapLC-beta)^2,2,sum))
seerrLC=sqrt(var(apply((betachapLC-beta)^2,2,sum)))
zerosLC=mean(apply(betachapLC==0,2,sum))
betmoyLC=apply(betachapLC,1,mean) betvarLC=apply(betachapLC,1,var)
betcovLC=apply(betachapvarLC,2:3,mean)
semoyLC=sqrt(diag(betcovLC))
seLC=sqrt(apply(betachapvarLC,1,diag))
puisLC=1/N*apply(matrix(as.numeric(abs(betachapLC)/seLC>1.96),
nrow=p),1,sum)

errCC=mean(apply((betachapCC-beta)^2,2,sum))
seerrCC=sqrt(var(apply((betachapCC-beta)^2,2,sum)))
zerosCC=mean(apply(betachapCC==0,2,sum))
betmoyCC=apply(betachapCC,1,mean) betvarCC=apply(betachapCC,1,var)
betcovCC=apply(betachapvarCC,2:3,mean)
semoyCC=sqrt(diag(betcovCC))
seCC=sqrt(apply(betachapvarCC,1,diag))
puisCC=1/N*apply(matrix(as.numeric(abs(betachapCC)/seCC>1.96),
nrow=p),1,sum)

errFD=mean(apply((betachapFD-beta)^2,2,sum))
seerrFD=sqrt(var(apply((betachapFD-beta)^2,2,sum)))
zerosFD=mean(apply(betachapFD==0,2,sum))
betmoyFD=apply(betachapFD,1,mean) betvarFD=apply(betachapFD,1,var)
betcovFD=apply(betachapvarFD,2:3,mean) semoyFD=diag(betcovFD)
seFD=sqrt(apply(betachapvarFD,1,diag))
puisFD=1/N*apply(matrix(as.numeric(abs(betachapFD)/seFD>1.96),
nrow=p),1,sum)

resul=c(beta,betmoyFD,betmoyFD-beta,semoyFD,betvarFD,puisFD,
betmoyLC,betmoyLC-beta,semoyLC,betvarLC,puisLC,
betmoyCC,betmoyCC-beta,semoyCC,betvarCC,puisCC) dim(resul)=c(p,16)
dimnames(resul)=list(c("beta1","beta2","beta3","beta4",
"beta5","beta6","beta7","beta8"),
c("beta_0","est","bias","mean(se*)","seemp","power",
"est","bias","mean(se*)","seemp","power",
"est","bias","mean(se*)","seemp","power"))

resulerr=c(errFD,errLC,errCC,seerrFD,seerrLC,seerrCC,
zerosFD,zerosLC,zerosCC) dim(resulerr)=c(3,3)
dimnames(resulerr)=list(c("FD","LC","CC"),c("erreur L2","se erreur
```

```
L2","nb of 0"))
print(paste("N",N,sep="=")) print(paste("n",n,sep="="))
print(paste("proba moyenne de
censure",round(D$censure*100)/100,sep="=")) print(resul)
print(resulerr)
list(beta=beta,betmoyFD=betmoyFD,biasFD=betmoyFD-beta,
semoyFD=semoyFD,betvarFD=betvarFD,puisFD=puisFD,errFD=errFD,
seerrFD=seerrFD,zerosFD=zerosFD,betmoyLC=betmoyLC,biasLC=betmoyLC-beta,
semoyLC=semoyLC,betvarLC=betvarLC,puisLC=puisLC,errLC=errLC,
seerrLC=seerrLC,zerosLC=zerosLCbetmoyCC=betmoyCC,biasCC=betmoyCC-beta,
semoyCC=semoyCC,betvarCC=betvarCC,puisCC=puisCC,errCC=errCC,
seerrCC=seerrCC,zerosCC=zerosCC)
}
```

La fonction `donneesLASSO`

Description : La fonction `donneesLASSO` simule des durées de survie gaussiennes censurées par des variables de censure gaussienne.

Arguments :

`N`	nombre d'échantillons	`N` $= 1000$
`n`	taille de chacun des échantillons	`n` $= 50, 100, 200$ ou 400
`beta`	le paramètre de régression	`beta` $= (3, 1.5, 0, 0, 2, 0, 0, 0)$
`mu`	moyenne de la censure	`mu` $= 0$ pour 50% de censure
`tau`	écart-type de la censure	`tau` $= 1$
`m`	moyenne des covariables	`m` $= 0_{\mathbb{R}^p}$ (p dimension de `beta`)
`sigma`	écart-type de l'erreur	`sigma` $= 1$
`rho`	paramètre de la matrice de covariance des covariables X	`rho` $= 0.5$

Pour la valeur `beta` $= (3, 1.5, 0, 0, 2, 0, 0, 0)$, on choisit `mu` $= -4.06$ pour 80% de censure, -2.53 pour 70% de censure, 0 pour 50% de censure, 4.06 pour 20% de censure.

Valeurs de sortie : Une liste contenant les composantes (`X`,`Z`,`del`,`Y`,`censure`), où

`X` matrice de taille $p \times (N \times n)$ où chaque colonne est un p-vecteur de covariables de moyenne `m` et matrice de covariance Σ (cf partie 6.4), les n premières colonnes correspondent à $N = 1$ etc,

`Z` matrice de taille $n \times N$ de durées de vie éventuellement censurées, par une gaussienne de moyenne `mu` et écart-type `tau`,

`del` matrice de taille $n \times N$ des indicateurs de censure,

`Y` matrice de taille $n \times N$ des durées de vie non censurées, de loi `beta`$'$`X` + `sigma`ε, où ε gaussienne centrée réduite,

`censure` pourcentage moyen de censure sur les $n \times N$ données.

Code :

```
donneesLASSO <- function(N,n,beta,mu,tau,m,sigma,rho)
{
p=length(beta);
cov=matrix(0,p,p)
for (i in 1:p)
    {
    for (j in 1:p)
        {
        cov[i,j]=rho^abs(i-j)
        }
    }
R=t(chol(cov))
epsilon=rnorm(N*n,0,1)
X=matrix(0,p,N*n)
Y=rep(0,N*n)
for (i in 1:(n*N))
    {
    X[,i]=m+R%*%rnorm(p,0,1)
    Y[i]=t(X[,i])%*%beta+sigma*epsilon[i]
    }
dim(Y)=c(n,N)
dim(epsilon)=c(n,N)
C=matrix(rnorm(N*n,mu,tau),ncol=N)
Z=pmin(Y,C)
del=matrix(as.numeric(Y<=C),ncol=N)
censure=mean(1-apply(del,2,mean))
print(paste("proba moyenne de
censure",censure,sep="="))
list(X=X,Z=Z,del=del,Y=Y,censure=censure)
}
```

La fonction `censfun`

Description : La fonction `censfun` détermine l'estimateur LASSO pondéré Kaplan-Meier du paramètre de régression où le paramètre de la pénalité

LASSO est choisi par validation croisée.

Arguments :

`data`	données de travail	`data` = (`Z`,`del`,`X`)
`K`	nombre de sous-échantillons pour la validation croisée	`K` = 10
`fraction`	vecteur de discrétisation du paramètre de pénalisation	`fraction` = $(\frac{k}{100})_{k=0}^{100}$

Valeurs de sortie : Un vecteur `betachap` estimateur LASSO pondéré Kaplan-Meier du paramètre de régression, où le choix du coefficient de pénalité LASSO est fait par validation croisée.

Code :

```
censfun<-function(data, K = 10, fraction, trace = FALSE , plot.it
= FALSE , se = TRUE)
{
Z=data[,1]
del=data[,2]
X=t(data[,3:ncol(data)])
all.folds <- cv.folds(length(Z), K)
residmat <- matrix(0, length(fraction), K)
for (i in seq(K))
    {
    omit <- all.folds[[i]]
    Zapp=Z[-omit]
    napp=length(Zapp)
    Xapp=X[,-omit]
    delapp=del[-omit]
    DR=poids(Zapp,Xapp,delapp)
    Zlars=sqrt(DR$W*napp)*DR$Zreord
    Xlars=sqrt(DR$W*napp)*t(DR$Xreord)
    fit <- lars(Xlars, Zlars, trace, type="lasso")
    Xtest=X[,omit]
    deltest=del[omit]
    if (sum(deltest)==0)
        {
        var=0
        residmat[, i]<-NA
        }
    else
        {
        Xtest=Xtest[,deltest==1]
        Ytest=Z[omit][deltest==1]
        Ychap <- predict(fit, t(Xtest), type="fit",
            mode = "fraction", s = fraction)$fit
        if (length(omit) == 1)
            fit <- matrix(fit, nrow = 1)
```

```
        if (length(Ytest)==1)
            residmat[, i] <- mean((Ytest - Ychap)^2)
        else
            {residmat[, i] <- apply((Ytest - Ychap)^2, 2, mean)}
        }
    if (trace)
           cat("\n CV Fold", i, "\n\n")
    }
cv <- apply(residmat, 1, mean, na.rm=TRUE)
cv.error<-sqrt(apply(residmat, 1, var, na.rm=TRUE)/K)
object<-list(fraction = fraction, cv = cv, cv.error = cv.error)
if(plot.it)
    plotCVLars(object, se)
schap=fraction[which.min(cv)] n=length(Z) DR=poids(Z,X,del)
Zlars=sqrt(DR$W*n)*DR$Zreord Xlars=sqrt(DR$W*n)*t(DR$Xreord)
fittot=lars(Xlars, Zlars, trace, type="lasso")
betachap=predict(fittot , s = schap , mode= "fraction",
type="coefficients")$coefficients
}
```

La fonction `censfuncomp`

Description : La fonction `censfuncomp` détermine l'estimateur LASSO du paramètre de régression en n'utilisant que les données non censurées de l'échantillon. Le paramètre de la pénalité LASSO est choisi par validation croisée.

Arguments :

`data`	données de travail	`data` = (`Z`,`del`,`X`)
`K`	nombre de sous-échantillons pour la validation croisée	`K` = 10
`fraction`	vecteur de discrétisation du paramètre de pénalisation	`fraction` = $(\frac{k}{100})_{k=0}^{100}$

Valeurs de sortie : Un vecteur `betachap` estimateur LASSO du paramètre de régression sur l'échantillon des données non censurées, où le choix du coefficient de pénalité LASSO est fait par validation croisée.

Code :

```
censfuncomp<-function(data, K = 10, fraction, trace = FALSE ,
plot.it = FALSE , se = TRUE)
{
Z=data[,1]
```

```
del=data[,2]
X=t(data[,3:ncol(data)])
Zcomp=Z[del==1] Xcomp=t(X[,del==1])
all.folds <- cv.folds(length(Zcomp), K)
crossvallars=cvlars(x=Xcomp, y=Zcomp, K=K, fraction=fraction,
trace=trace, plot.it=plot.it, se,type="lasso",all.folds=all.folds)
fraction=crossvallars$fraction
cv=crossvallars$cv
cv.error=crossvallars$cv.error
scomp=fraction[which.min(cv)]
fittot=lars(x=Xcomp, y=Zcomp,type="lasso", trace)
betachap=predict(object=fittot, s=scomp,type= "coefficients",
mode= "fraction")$coefficients
}
```

La fonction `poids`

Description : La fonction `poids` calcule les poids de Kaplan-Meier associés aux données de survie.

Arguments :

`Z`	vecteur comprenant un n-échantillon de durées de survie
`X`	matrice de taille $p \times n$ de covariables
`del`	n-vecteur d'indicateurs de censure

Valeurs de sortie : Une liste contenant les composantes (`Zreord`,`Xreord`,`delreord`,W), où

`Zreord`	vecteur des durées de survie réordonné par ordre croissant,
`Xreord`	matrice des covariables réordonnées dans le même ordre que `Zreord`,
`delreord`	vecteur des indicateurs de censure réordonnés dans le même ordre que `Zreord`,
`W`	vecteur des poids associés aux durées de survie réordonnées `Zreord`.

Code :

```
poids <- function(Z,X,del)
{
n=length(Z)
perm=order(Z)
Zperm=Z[perm] Xperm=X[,perm] deltap=del[perm] W=deltap[1]/n P=1
for (i in 2:n)
```

```
    {
    P=P*((n-i+1)/(n-i+2))^(deltap[i-1])
    W=c(W,(deltap[i]/(n-i+1))*P)
    }
list(Zreord=Zperm,Xreord=Xperm,delreord=deltap,W=W)
}
```

Bibliographie

Aalen, O.O. *Statistical inference for a family of counting processes.* PhD thesis, University of California, Berkeley., 1975.

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

Andersen, P.K. & R.D. Gill. "Cox's regression model for counting processes : a large sample study." *Ann. Statist.* 10 (1982) : 1100–1120.

Bagdonavičius, V. & M.S. Nikulin. *Accelerated Life Models : Modeling and Statistical Analysis.* CRC Press, 2002.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov, & J.A. Wellner. *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD : Johns Hopkins University Press, 1993.

Breiman, L. & P. Spector. "Submodel selection and evaluation in regression. The X-random case." *Int. Statist. Rev.* 60 (1992) : 291–319.

Breslow, N. "Covariance Analysis of Censored Survival Data." *Biometrics* 30 (1974) : 89–99.

Breslow, N.E. "Contribution à la discussion sur l'article de D.R. Cox, Regression models and life-tables." *J. Roy. Statist. Soc. Ser. B* 34 (1972) : 216–217.

Buckley, J. & I. James. "Linear regression with censored data." *Biometrika* 66 (1979) : 429–436.

Buckley, J. D. "Additive and multiplicative models for relative survival rates." *Biometrics* 40 (1984) : 51–62.

Carroll, R. J., D. Ruppert, & L. A. Stefanski. *Measurement error in nonlinear models.* Volume 63 of Monographs on Statistics and Applied Probability. London : Chapman & Hall, 1995.

Chang, I.-S., C. A. Hsuing, M.-C. Wang, & C.-C. Wen. "An asymptotic theory for the nonparametric maximum likelihood estimator in the Cox gene model." *Bernoulli* 11 (2005) : 863–892.

Chen, H. Y. & R. J. A. Little. "Proportional hazards regression with missing covariates." *J. Amer. Statist. Assoc.* 94 (1999) : 896–908.

Chernick, M.R. *Bootstrap methods.* A practitioner's guide, A Wiley-Interscience Publication. Wiley Series in Probability and Statistics : Applied Probability and Statistics. New York : John Wiley & Sons Inc., 1999.

Cox, D. R. "Regression models and life-tables." Avec discussion de 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. *J. Roy. Statist. Soc. Ser. B* 34 (1972) : 187–220.

Cox, D. R. "Partial likelihood." *Biometrika* 62 (1975) : 269–276.

Cox, D.R. & D. Oakes. *Analysis of Survival Data.* Chapman & Hall, 1984.

Dempster, A.P., N.M. Laird, & D.B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." With discussion. *J. Roy. Statist. Soc. Ser. B* 39 (1977) : 1–38.

Detais, A. "LASSO-type estimation in the censored linear regression model." *En cours de soumission* (2008).

Detais, A. & J.-F. Dupuy. "Maximum likelihood estimation in a partially observed stratified regression model with censored data." *Soumis à Ann. Inst. Statist. Math.* (2008).

Dupuy, J.-F. & E. Leconte. "A study of regression calibration in a partially observed stratified Cox model." *Journal of Statistical Planning and Inference* A paraître (2008).

Efron, B. "Bootstrap methods : another look at the jackknife." *Ann. Statist.* 7 (1979) : 1–26.

Efron, B. & R. Tibshirani. "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." Avec un commentaire de J. A. Hartigan et une réponse des auteurs. *Statist. Sci.* 1 (1986) : 54–77.

Efron, B. & R.J. Tibshirani. *An introduction to the bootstrap.* Volume 57 of Monographs on Statistics and Applied Probability. New York : Chapman and Hall, 1993.

Fang, H.-B., G. Li, & J. Sun. "Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model." *Scand. J. Statist.* 32 (2005) : 59–75.

Fleming, T.R. & D.P. Harrington. *Counting processes and survival analysis.* Wiley Series in Probability and Statistics. John Wiley & Sons, 1991.

Frank, I.E. & J.H. Friedman. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics* 35 (1993) : 109–135.

Fu, W. J. "Penalized regressions : the bridge versus the lasso." *J. Comput. Graph. Statist.* 7 (1998) : 397–416.

Geyer, C.J. "On the Asymptotics of Convex Stochastic Optimization." *Unpublished manuscript* (1996).

Gill, R.D. "Testing with replacement and the product limit estimator." *Ann. Statist.* 9 (1981) : 853–860.

Hastie, T., R. Tibshirani, & J. Friedman. *The elements of statistical learning.* Data mining, inference, and prediction. Springer Series in Statistics. New York : Springer-Verlag, 2001.

Hoerl, A.E. & R.W. Kennard. "Ridge regression : biased estimation for nonorthogonal prob." *Technometrics* 12 (1970) : 55–67.

Huang, J. & D. Harrington. "Dimension reduction in the linear model for right-censored data : predicting the change of HIV-I RNA levels using clinical and protease gene mutation data." *Lifetime Data Anal.* 10 (2004) : 425–443 (2005).

Jin, Z., D.Y. Lin, & Z. Ying. "On least-squares regression with censored data." *Biometrika* 93 (2006) : 147–161.

Jobson, J.D. *Applied multivariate data analysis. Volume II : Categorical and multivariate methods.* Springer Texts in Statistics. New York : Springer-Verlag, 1992.

Johansen, S. "An extension of Cox's regression model." *Internat. Statist. Rev.* 51 (1983) : 165–174.

Jørgensen, B. *The theory of linear models.* New York : Chapman & Hall, 1993.

Kalbfleisch, J.D. & R.L. Prentice. *The statistical analysis of failure time data.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York-Chichester-Brisbane, 1980.

Kaplan, E.L. & P. Meier. "Nonparametric estimation from incomplete observations." *J. Amer. Statist. Assoc.* 53 (1958) : 457–481.

Klein, J. P. & M. L. Moeschberger. *Survival Analysis : Methods for Censored and Truncated Data.* Statistics for Biology and Health. New York : Springer, 1997.

Knight, K. & W. Fu. "Asymptotics for lasso-type estimators." *Ann. Statist.* 28 (2000) : 1356–1378.

Korsholm, L. Likelihood Ratio Test in the Correlated Gamma-Frailty Model. Technical Report 98-11, Centre for Labour Market and Social Research - University of Aarhus School of Business, 1998.

Kosorok, M.R. *Introduction to empirical processes and semiparametric inference.* Springer Series in Statistics. Springer-Verlag, 2008.

Kosorok, M.R. & R. Song. "Inference under right censoring for transformation models with a change-point based on a covariate threshold." *Ann. Statist.* 35 (2007) : 957–989.

Koul, H., V. Susarla, & J. Van Ryzin. "Regression analysis with randomly right-censored data." *Ann. Statist.* 9 (1981) : 1276–1288.

Lai, T.L. & Z. Ying. "Large sample theory of a modified Buckley-James estimator for regression analysis with censored data." *Ann. Statist.* 19 (1991) : 1370–1402.

Lin, D.Y., D. Oakes, & Z. Ying. "Additive hazards regression with current status data." *Biometrika* 85 (1998) : 289–298.

Lin, D.Y. & Z. Ying. "Cox regression with incomplete covariate measurements." *J. Amer. Statist. Assoc.* 88 (1993) : 1341–1349.

Loève, M. *Probability theory.* Third edition. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1963.

Lu, W. "Maximum likelihood estimation in the proportional hazards cure model." *Ann. Inst. Statist. Math.* A paraître (2008).

Martinussen, T. "Cox regression with incomplete covariate measurements using the EM-algorithm." *Scand. J. Statist.* 26 (1999) : 479–491.

Martinussen, T. & T.H. Scheike. "Efficient estimation in additive hazards regression with current status data." *Biometrika* 89 (2002) : 649–658.

Martinussen, T. & T.H. Scheike. *Dynamic regression models for survival data.* Statistics for Biology and Health. New York : Springer, 2006.

McLachlan, G.J. & T. Krishnan. *The EM algorithm and extensions.* A Wiley-Interscience Publication. Wiley Series in Probability and Statistics : Applied Probability and Statistics. New York : John Wiley & Sons Inc., 1997.

Miller, R.G. "Least squares regression with censored data." *Biometrika* 63 (1976) : 449–464.

Murphy, S.A. "Consistency in a proportional hazards model incorporating a random effect." *Ann. Statist.* 22 (1994) : 712–731.

Murphy, S.A. "Asymptotic theory for the frailty model." *Ann. Statist.* 23 (1995) : 182–198.

Murphy, S.A., A.J. Rossini, & A.W. van der Vaart. "Maximum likelihood estimation in the proportional odds model." *J. Amer. Statist. Assoc.* 92 (1997) : 968–976.

Paik, M.C. "Multiple imputation for the Cox proportional hazards model with missing covariates." *Lifetime Data Anal.* 3 (1997) : 289–298.

Paik, M.C. & W.-Y. Tsai. "On using the Cox proportional hazards model with missing covariates." *Biometrika* 84 (1997) : 579–593.

Parner, E. "Asymptotic theory for the correlated gamma-frailty model." *Ann. Statist.* 26 (1998) : 183–214.

Pons, O. "Estimation in the Cox model with missing covariate data." *J. Nonparametr. Stat.* 14 (2002) : 223–247.

Prentice, R.L. "Covariate measurement errors and parameter estimation in a failure time regression model." *Biometrika* 69 (1982) : 331–342.

Rao, C.R. & H. Toutenburg. *Linear models.* Least squares and alternatives. Springer Series in Statistics. New York : Springer-Verlag, 1995.

Rao, C.R., H. Toutenburg, Shalabh, & C. Heumann. *Linear models and generalizations.* Least squares and alternatives, With contributions by Michael Schomaker. extended edition. Springer Series in Statistics. Berlin : Springer, 2008.

Rencher, A.C. & G.B. Schaalje. *Linear models in statistics.* Second edition. Hoboken, NJ : Wiley-Interscience [John Wiley & Sons], 2008.

Ritov, Y. "Estimation in a linear regression model with censored data." *Ann. Statist.* 18 (1990) : 303–328.

Rockafellar, R.T. *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton, N.J. : Princeton University Press, 1970.

Searle, S.R. *Linear models.* Reprint of the 1971 original. Wiley Classics Library. New York : John Wiley & Sons Inc., 1997.

Shorack, G.R. & J.A. Wellner. *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. New York : John Wiley & Sons Inc., 1986.

Stute, W. "Consistent estimation under random censorship when covariables are present." *J. Multivariate Anal.* 45 (1993) : 89–103.

Stute, W. "Distributional convergence under random censorship when covariables are present." *Scand. J. Statist.* 23 (1996) : 461–471.

Sugimoto, T. & T. Hamasaki. "Properties of estimators of baseline hazard functions in a semiparametric cure model." *Ann. Inst. Statist. Math.* 58 (2006) : 647–674.

Therneau, T.M. & P.M. Grambsch. *Modeling survival data : extending the Cox model.* Statistics for Biology and Health. New York : Springer-Verlag, 2000.

Thurston, S.W., D. Spiegelman, & D. Ruppert. "Equivalence of regression calibration methods in main study/external validation study designs." *J. Statist. Plann. Inference* 113 (2003) : 527–539.

Thurston, S.W., P.L. Williams, R. Hauser, H. Hu, M. Hernandez-Avila, & D. Spiegelman. "A comparison of regression calibration approaches for designs with internal validation data." *J. Statist. Plann. Inference* 131 (2005) : 175–190.

Tibshirani, R. "Regression shrinkage and selection via the lasso." *J. Roy. Statist. Soc. Ser. B* 58 (1996) : 267–288.

Tsiatis, A.A. "A large sample study of Cox's regression model." *Ann. Statist.* 9 (1981) : 93–108.

Tsiatis, A.A. *Semiparametric theory and missing data.* Springer Series in Statistics. New York : Springer, 2006.

Tsiatis, A.A., V. Degruttola, & M.S. Wulfsohn. "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS.." *J. Amer. Statist. Assoc.* 90 (1995).

Vaart, A.W.van der . *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge : Cambridge University Press, 1998.

Vaart, A.W.van der & J.A. Wellner. *Weak convergence and empirical processes.* With applications to statistics. Springer Series in Statistics. New York : Springer-Verlag, 1996.

Wang, C.Y. "Robust sandwich covariance estimation for regression calibration estimator in Cox regression with measurement error." *Statist. Probab. Lett.* 45 (1999) : 371–378.

Wang, C.Y., L. Hsu, Z.D. Feng, & R.L. Prentice. "Regression calibration in failure time regression." *Biometrics* 53 (1997) : 131–145.

Wang, C.Y., S.X. Xie, & R.L. Prentice. "Recalibration based on an approximate relative risk estimator in Cox regression with missing covariates." *Statist. Sinica* 11 (2001) : 1081–1104.

Wong, W.H. "Theory of partial likelihood." *Ann. Statist.* 14 (1986) : 88–123.

Zeng, D. & J. Cai. "Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time." *Ann. Statist.* 33 (2005) : 2132–2163.

Zeng, D. & D.Y. Lin. "Maximum likelihood estimation in semiparametric regression models with censored data." *J. R. Stat. Soc. Ser. B Stat. Methodol.* 69 (2007) : 507–564.

Zhou, M. "M-estimation in censored linear models." *Biometrika* 79 (1992) : 837–841.

Maximum likelihood and penalized least squares in censored survival data models

Life data analysis is used in various application fields. Different methods have been proposed for modeling such data. In this thesis, we are interested in two distinct modelisation types, the stratified Cox model with randomly missing strata indicators and the right-censored linear regression model. We propose methods for estimating the parameters and establish the asymptotic properties of the obtained estimators in each of these models.

First, we consider a generalization of the Cox model, allowing different groups, named strata, of the population to have distinct baseline intensity functions, whereas the regression parameter is shared by all the strata. In this stratified proportional intensity model, we are interested in the parameters estimation when the strata indicator is missing for some of the population individuals. Nonparametric maximum likelihood estimators are proposed for the model parameters and their consistency and asymptotic normality are established. We show the asymptotic efficiency of the regression parameter and obtain consistant estimators of its variance. The Expectation-Maximization algorithm is proposed and developed for the evaluation of the estimators of the model parameters.

Second, we are interested in the regression linear model when the response data is randomly right-censored. We introduce a new estimator of the regression parameter, which minimizes a Kaplan-Meier-weighted penalized least squares criterion. Results of consistency and asymptotic normality are obtained and a simulation study is conducted in order to investigate the small sample properties of this LASSO-type estimator. The bootstrap method is used for the estimation of the asymptotic variance.

Résumé

L'analyse de durées de vie censurées est utilisée dans des domaines d'application variés et différentes possibilités ont été proposées pour la modélisation de telles données. Nous nous intéressons dans cette thèse à deux types de modélisation différents, le modèle de Cox stratifié avec indicateurs de strates aléatoirement manquants et le modèle de régression linéaire censuré à droite. Nous proposons des méthodes d'estimation des paramètres et établissons les propriétés asymptotiques des estimateurs obtenus dans chacun de ces modèles.

Dans un premier temps, nous considérons une généralisation du modèle de Cox qui permet à différents groupes de la population, appelés strates, de posséder des fonctions d'intensité de base différentes tandis que la valeur du paramètre de régression est commune. Dans ce modèle à intensité proportionnelle stratifié, nous nous intéressons à l'estimation des paramètres lorsque l'indicateur de strate est manquant pour certains individus de la population. Des estimateurs du maximum de vraisemblance non paramétrique pour les paramètres du modèle sont proposés et nous montrons leurs consistance et normalité asymptotique. L'efficacité du paramètre de régression est établie et des estimateurs consistants de sa variance asymptotique sont également obtenus. Pour l'évaluation des estimateurs du modèle, nous proposons l'utilisation de l'algorithme Espérance-Maximisation et le développons dans ce cas particulier.

Dans un second temps, nous nous intéressons au modèle de régression linéaire lorsque la donnée réponse est censurée aléatoirement à droite. Nous introduisons un nouvel estimateur du paramètre de régression minimisant un critère des moindres carrés pénalisé et pondéré par des poids de Kaplan-Meier. Des résultats de consistance et normalité asymptotique sont obtenus et une étude de simulations est effectuée pour illustrer les propriétés de cet estimateur de type LASSO. La méthode bootstrap est utilisée pour l'estimation de la variance asymptotique.

Mots clés

Durées de vie censurées à droite, régression linéaire, modèle à intensité proportionnelle stratifié, estimateur LASSO, estimateur bridge, estimateur du maximum de vraisemblance non paramétrique, poids de Kaplan-Meier, estimation bootstrap, consistance, normalité asymptotique, données manquantes, estimation de la variance.