Inférence pour les modèles statistiques mal spécifiés, application à une étude sur les facteurs pronostiques dans le cancer du sein

Roxane Duroux

Université Pierre et Marie Curie

École Doctorale de Sciences Mathématiques de Paris Centre

Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistiques

présentée par

Roxane Duroux

Inférence pour les modèles statistiques mal spécifiés, application à une étude sur les facteurs pronostiques dans le cancer du sein.

---

dirigée par John O'Quigley

Soutenue le 21/09/2016 devant le jury composé de :

M. Jacques Bénichou	Université de Rouen	rapporteur
M. Michel Broniatowski	Université Paris 6	examinateur
M. Philippe Flandre	Inserm	rapporteur
M. John O'Quigley	Université Paris 6	directeur de thèse

Laboratoire de Statistique Théorique et Appliquée (LSTA)
Université Pierre et Marie Curie
Boîte 158, Tours 15-25, 2ème étage
4, place Jussieu
75 252 Paris Cedex 05

A mes proches.

Remerciements

Mes premiers remerciements vont à mon directeur de thèse. John, je vous remercie pour l'autonomie de recherche que vous m'avez accordée, notamment à travers mes diverses collaborations. Merci de m'avoir fait partager vos connaissances, qu'elles soient statistiques ou sur le monde de la recherche en général.

Je remercie l'ensemble des membres du jury. Merci à Michel Broniatowski d'avoir accepté de participer à ma soutenance. Merci à Philippe Flandre et Jacques Bénichou d'avoir accepté de rapporter ma thèse, et de m'avoir fourni de très intéressantes remarques sur mon travail.

Je tiens également à remercier l'ensemble de l'équipe du LSTA grâce à laquelle j'ai pu travailler dans un cadre particulièrement agréable pendant ces trois années. Merci notamment à Louise pour sa disponibilité et sa gentillesse. Merci aussi aux permanents du laboratoire pour leur écoute et leur aide en de nombreuses circonstances. Je remercie également Corentin Lacombe pour son aide et ses conseils pour l'organisation administrative de ma soutenance.

Impossible de ne pas remercier l'ensemble des doctorants du LSTA. Pour leur disponibilité, leur écoute, leur aide, leur joie de vivre, les discussions passionnantes (et pas toujours mathématiques) autour de la machine à café et leur immense soutien, merci donc à Baptiste, Diaa, Dimbihery, Emilie, Lucie, Mokhtar, Nazih, Quyen, Sarah, Simon, Thibault et Yohann. Assia, merci pour ta bonne humeur constante et nos fous rires. Les TD de l'ISUP n'auraient pas été les même sans toi ! Cécile, que ce soit par tes connaissances en survie, en R, ton soutien sans faille, ta grande gentillesse, ou les conseils que tu continues à me donner, tu as su me faire partager toutes les ficelles du métier. Merci pour tout ! Matthieu, vivement que tu reviennes de Los Angeles ! Merci tout d'abord pour notre collaboration. C'est un plaisir de travailler avec toi. Merci pour ta grande disponibilité, pour être toujours prêt à aider les autres. Et puis merci pour nos longues discussions culinaires, culturelles, mais surtout politiques. Je ne doute pas que nous reprendrons à ton retour ces conversations engagées. Enfin Erwan, les doctorants ont changé chaque année, mais tu as réussi à me supporter trois années entières ! Évidemment, merci pour nos collaborations, que ce soit pour les forêts ou pour le GTT. Qu'il est facile de travailler avec toi ! Merci pour ton soutien à toute épreuve, ta disponibilité (j'avoue que la concordance de nos emplois du temps peu matinaux, c'est pratique), tes relectures attentives et le large panel de discussions abordées aux pauses cafés.

D'un point de vue plus personnel, je tiens à remercier deux amies présentes à mes côtés ces trois années. Élodie, merci de m'avoir toujours épaulée, surtout lors des périodes un peu plus compliquées de ma thèse. Les conférences de stats sont toujours plus intéressantes

quand je t'y croise ! Céline, bien sûr que je ne t'oublie pas ! Quoiqu'il advienne, tu restes d'un soutien indéfectible. Merci pour ta relecture et tes conseils, mais surtout pour ton oreille attentive et nos nombreuses heures au téléphone. Malgré la distance, je te sais toujours aussi proche.

Je remercie mes parents et mes soeurs de leurs relectures de mon manuscrit (ce qui est loin d'être trivial quand on ne fait pas de stats, voire pas de maths !), de m'avoir invariablement soutenue et d'avoir toujours cru en ma capacité d'aller plus loin. Quelle fraîcheur de parler avec vous tous ! Arriver jusqu'ici, c'était impossible sans vous.

Quentin, tes relectures, tes avis éclairés sur mon travail, ton soutien au quotidien, ta grande diplomatie... tu fais partie de chacun de mes accomplissements. Merci d'être toujours là. C'est chaque jour un bonheur de partager ta vie.

Résumé

Cette thèse est consacrée à l'inférence de certains modèles statistiques mal spécifiés. Chaque résultat obtenu trouve son application dans une étude sur les facteurs pronostiques dans le cancer du sein, grâce à des données collectées par l'Institut Curie. Dans un premier temps, nous nous intéressons au modèle à risques non proportionnels, et exploitons la connaissance de la survie marginale du temps de décès. Ce modèle autorise la variation dans le temps du coefficient de régression, généralisant ainsi le modèle à hasards proportionnels. Dans un deuxième temps, nous étudions un modèle à hasards non proportionnels ayant un coefficient de régression constant par morceaux. Nous proposons une méthode d'inférence pour un modèle à un unique point de rupture, et une méthode d'estimation pour un modèle à plusieurs points de rupture. Dans un troisième temps, nous étudions l'influence du sous-échantillonnage sur la performance des forêts médianes et essayons de généraliser les résultats obtenus aux forêts aléatoires de survie à travers une application. Enfin, nous présentons un travail indépendant où nous développons une nouvelle méthode de recherche de doses, dans le cadre des essais cliniques de phase I à ordre partiel.

Mots-clefs

Survie, modèle à risques non proportionnels, modèle avec changepoints, estimation non paramétrique, forêts aléatoires, essais cliniques de phase I, coupure maximale tolérée.

Inference for statistical misspecified models, application to a prognostic factors study for breast cancer

Abstract

The thesis focuses on inference of statistical misspecified models. Every result finds its application in a prognostic factors study for breast cancer, thanks to the data collection of Institut Curie. We consider first non-proportional hazards models, and make use of the marginal survival of the failure time. This model allows a time-varying regression coefficient, and therefore generalizes the proportional hazards model. On a second time, we study step regression models. We propose an inference method for the changepoint of a two-step regression model, and an estimation method for a multiple-step regression model. Then, we study the influence of the subsampling rate on the performance of median forests and try to extend the results to random survival forests through an application. Finally, we present a new dose-finding method for phase I clinical trials, in case of partial ordering.

Keywords

Survival, Non-proportional hazards model, step regression model, non-parametric estimation, random forests, phase I clinical trials, maximum tolerated cut.

Table des matières

Plan détaillé de la thèse

Chapitre 1 : Introduction

Dans ce premier chapitre, nous proposons un état de l'art sur les notions clés de cette thèse. La première partie concerne l'analyse de survie. Nous commençons par des définitions et notations de base qui nous serviront tout au long du document. Nous rappelons notamment la définition de l'estimateur de Kaplan and Meier (1958) qui permet une estimation de la fonction de survie du temps de décès en présence de censure. Nous introduisons la définition de modèles à hasards proportionnels et nous nous attardons sur le modèle de Cox (1972) qui est largement utilisé en analyse de survie. Ce modèle suppose un effet constant des covariables sur le temps de décès. Nous citons quelques modèles qui relâchent un peu cette hypothèse, tout en restant des modèles semi-paramétriques. Nous présentons ensuite plusieurs méthodes d'estimation non paramétriques de la fonction de risque instantané dans le cadre de données censurées. Nous détaillons dans une deuxième partie le cas particulier des forêts aléatoires. Dans cette partie, nous nous plaçons d'abord dans le cas de données complètes pour présenter les forêts de Breiman (2001) et développons ensuite les forêts de survie (Ishwaran et al., 2008). Enfin, nous concluons avec une troisième partie qui sert d'introduction au Chapitre 5. Nous décrivons le contexte des essais cliniques de phase I, ainsi que les problèmes qui y sont soulevés. Nous passons ensuite en revue des méthodes de recherche de doses, en particulier une méthode proposée par O'Quigley et al. (1990).

Chapitre 2 : Modèle à hasards non proportionnels et Survie marginale

Dans le Chapitre 2, nous introduisons un estimateur $\tilde{\beta}$, qui exploite la connaissance de la survie marginale. En effet, il est intéressant pour certaines études de prendre en considération de l'information, connue au préalable, sur la loi du temps de décès. C'est le cas par exemple en survie relative où l'étude du taux de hasard de la population d'intérêt, celui d'une population de patients atteints d'une maladie particulière par exemple, s'effectue grâce à celui de la population générale, d'un pays par exemple. Ce dernier est souvent obtenu grâce à des tables de mortalité de population. Nous cherchons ici à savoir dans quelle mesure la connaissance de la survie marginale améliore, ou non, l'estimation de l'effet des covariables sur le temps de décès.

Sous l'hypothèse de hasards proportionnels, l'estimateur est consistant pour le vrai paramètre de régression. Cet estimateur est aussi simple à calculer en pratique. Nous étudions ses propriétés asymptotiques ; en particulier, il converge en probabilité vers un réel noté β^* qui peut être vu comme un effet moyen de population. Sous un modèle à hasards proportionnels, l'estimateur de la vraisemblance partielle (Cox, 1972) est connu pour être optimal dans le sens où son efficacité est maximale. C'est pourquoi nous cherchons à

comparer son efficacité par rapport à celle de l'estimateur $\tilde{\beta}$ afin d'avoir une idée des performances de ce dernier sous le modèle de Cox. Nous fournissons des simulations dans le cas de modèles à hasards proportionnels et non-proportionnels. Les dernières parties de ce chapitre montrent, à l'aide d'exemples, comment $\tilde{\beta}$ peut être utilisé en pratique, notamment en survie relative.

Ce chapitre fait l'objet d'un article soumis.

Chapitre 3 : Détection de points de rupture

Le Chapitre 3 s'intéresse à l'étude de modèles de survie généralisant le modèle de Cox (1972). Pour être plus précis, on considère un modèle de Cox pour lequel le coefficient de régression ne serait pas constant mais constant par morceaux. D'après les travaux de Andersen and Gill (1982), il est évident que, si nous connaissions à l'avance la position des points de discontinuité de la fonction de régression, il serait alors aisé d'estimer les coefficients de part et d'autre des discontinuités. En effet, entre chaque point de discontinuité, on se retrouve dans le cadre d'un modèle de Cox à coefficient de régression constant. On peut alors utiliser le maximum de vraisemblance partielle sur chacun de ces segments. La difficulté se situe donc dans l'estimation de ces points de discontinuité, appelés dans ce cadre "points de rupture" ou "changepoints".

Nous commençons par rappeler la méthode d'estimation proposée par Anderson and Senthilselvan (1982) dans le cas d'un modèle à un unique changepoint et proposons une méthode d'inférence de ce dernier à l'aide des travaux de Davies (1977). Nous introduisons ensuite une nouvelle méthode d'estimation basée sur le processus du score standardisé Chauvel and O'quigley (2014) dans le cas d'un modèle à plusieurs changepoints. Dans ce cadre, l'estimation ne se fait pas par vraisemblance mais par minimisation des résidus quadratiques. Nous effectuons des simulations pour le modèle à un changepoint afin d'observer le niveau empirique de l'intervalle de confiance proposé. Pour le modèle à plusieurs changepoints, nous utilisons le package `strucchange` (Kleiber et al., 2002) du logiciel R pour la méthode des moindres carrés. Nous finissons par une application de la méthode des moindres carrés à des données de cancer du sein de l'Institut Curie.

Ce chapitre fait l'objet d'un article soumis.

Chapitre 4 : Une approche non paramétrique : les forêts aléatoires de survie

Ce chapitre est le résultat d'un travail réalisé en collaboration avec M. Erwan Scornet et fait l'objet d'un article soumis.

Nous souhaitons, dans ce chapitre, étudier les propriétés des forêts aléatoires de survie (Ishwaran et al., 2008). Nous sommes alors confrontés au caractère censuré des données auxquelles s'appliquent ces forêts aléatoires. C'est pourquoi nous nous plaçons pour commencer dans le cas de données complètes, et plus particulièrement dans le cadre de la régression. Dans ce cadre, les forêts les plus utilisées sont les forêts aléatoires de Breiman (2001). Ces dernières trouvent leur place dans de nombreuses applications, que ce soit pour la régression ou la classification. On peut citer par exemple les problèmes de reconnaissance de forme (Rogez et al., 2008).

Nous commençons donc par introduire les notations utiles à l'étude de ces forêts aléatoires et rappelons l'algorithme des forêts de Breiman. Nous souhaitons connaître l'influence du taux de sous-échantillonnage sur la performance de ces forêts. Pour cela, nous

nous basons sur des résultats théoriques que nous établissons sur les forêts médianes. Nous présentons une majoration de la vitesse de convergence des forêts médianes et montrons que la performance des forêts médianes ne dépend pas du taux de sous-échantillonnage, mais plutôt du niveau de chaque arbre dans la forêt. Les simulations présentées montrent que ces résultats s'adaptent aux forêts de Breiman. Nous revenons ensuite au problème initial des forêts aléatoires de survie et des données censurées à travers une application. Les preuves sont détaillées en Annexe A.

Chapitre 5 : Méthode semi-paramétrique en essais cliniques de phase I dans les cas d'ordre partiel

Ce chapitre est issu d'un article en cours de rédaction réalisé en collaboration avec M. Mathieu Clertant, et peut être trouvé en partie dans sa thèse (Clertant, 2015).

Dans le cadre des essais cliniques de phase I, on souhaite tester simultanément deux substances (traitements, agents cytotoxiques, etc.). En supposant que la toxicité d'une substance augmente quand sa quantité augmente, les couples de doses à tester sont donc ordonnés partiellement en terme de toxicité. En effet, prenons deux couples de doses (d_1, d_2) et (d'_1, d'_2), alors on distingue trois cas. Si $d_1 \leq d'_1$ et $d_2 \leq d'_2$, alors le couple (d_1, d_2) est moins toxique que le couple (d'_1, d'_2). De la même manière, si $d_1 \geq d'_1$ et $d_2 \geq d'_2$, alors le couple (d_1, d_2) est plus toxique que le couple (d'_1, d'_2). Mais si $d_1 \leq d'_1$ et $d_2 \geq d'_2$, alors on ne peut pas ordonner les couples (d_1, d_2) et (d'_1, d'_2). On propose dans ce chapitre une extension, dans ce cas d'ordre partiel, de la Méthode Semi-Paramétrique (SPM) introduite par Clertant (2015) qu'on notera simplement "poSPM", pour Partial Ordering Semi Parametric Method.

Dans les essais cliniques de phase I, en général, on cherche l'unique dose ayant la toxicité la plus proche d'un certain seuil fixé à l'avance. On appelle celle-ci la dose maximale tolérée (MTD). Pour le cas d'une combinaison de deux substances, il semble raisonnable de supposer qu'une telle dose n'est pas unique. On s'intéresse donc à la notion de contour introduite par Mander and Sweeting (2015). Nous proposons deux paramétrisations de la poSPM. La première visant une recherche de la MTD, et la seconde une recherche de contour. Cette dernière peut être vue comme une extension du modèle PIPE (Mander and Sweeting, 2015). On obtient des propriétés asymptotiques de convergence pour ces deux paramétrisations de la poSPM, et donc également pour la méthode PIPE en tant que cas particulier de la poSPM.

Annexe

L'Annexe présente les preuves de différents théorèmes, lemmes et propriétés énoncés dans les chapitres précédents. Plus précisément, l'Annexe A présente les preuves associées aux résultats du Chapitre 4, et l'Annexe B présente celles associées au Chapitre 5.

Chapitre 1

Introduction

1.1 Introduction à l'analyse de survie

L'analyse statistique des durées de vie, ou analyse de survie, est l'étude du délai de la survenue d'un événement. En biostatistiques, on nomme l'événement étudié "décès", et il peut être un temps de guérison, un temps avant rechute ou avant décès, par exemple. Dans la plupart des études ou essais cliniques nécessitant l'analyse de données de survie, on dispose, en plus des temps de décès, de variables explicatives individuelles appelées covariables. Celles-ci peuvent être fixes comme le sexe, le type de traitement administré ou encore le lieu de résidence. Elles peuvent aussi dépendre du temps, c'est le cas pour l'âge, le stade de développement d'un cancer ou encore des mesures répétées d'une quantité biologique. L'étude des données de survie permet entre autres de relier certaines de ces covariables à la durée de vie, et ainsi limiter les facteurs de risque d'une maladie.

L'analyse des durées de vie pose des problèmes spécifiques dus au fait que certaines durées de vie ne sont pas observées totalement. Il s'agit du phénomène de censure. C'est par exemple le cas lorsqu'un patient quitte une étude avant que celle-ci se termine. On a alors accès à la date de son départ, mais pas à son temps de décès. Sa durée de vie est alors dite censurée. Cette censure empêche l'utilisation immédiate de nombreux résultats de statistiques classiques, tels que les résultats sur le maximum de vraisemblance. Nous commençons par définir les notations d'analyse de survie utilisées tout au long de la thèse.

1.1.1 Définitions et notations

Quelques définitions

Nous nous intéressons à la variable aléatoire T symbolisant le temps de décès. La fonction F désigne la fonction de répartition de T et f sa densité. On note S la fonction de survie de T définie par

$$S : \left\{ \begin{array}{rcl} \mathbb{R}_+ & \to & [0,1] \\ t & \mapsto & P(T > t) = 1 - F(t). \end{array} \right.$$

$S(t)$ représente donc la probabilité d'être encore en vie à l'instant t. Si la loi de T est absolument continue par rapport à la mesure de Lebesgue, on introduit la fonction de risque instantané.

Définition 1.1 (Risque instantané)**.** Le risque instantané, ou taux de hasard, de la va-

riable aléatoire positive X, est la fonction $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ définie par

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq X < t + h | X \geq t)}{h}, \ t \geq 0.$$

Le réel $\lambda(t)$ représente donc la probabilité de mourir dans un petit intervalle de temps après t, sachant que l'on a survécu jusque t. C'est donc la probabilité de mort instantanée pour ceux qui ont survécu. On peut remarquer que la fonction de survie S et le risque instantané λ sont liés par la relation suivante. Pour tout $t \in \mathbb{R}_+$,

$$S(t) = \exp\left(- \int_0^t \lambda(s) ds \right).$$

Il suffit donc de connaître l'une des trois fonctions S, F ou λ pour en déduire les autres. Généralement, la fonction λ est la plus intéressante, puisqu'elle donne une description du futur immédiat d'un sujet encore observé et permet la prise en compte de différences entre modèles qui seraient difficiles à formuler pour la fonction de survie ou de répartition. La plupart du temps, on établit un modèle de données de survie à partir de cette fonction de risque instantané. C'est ce que nous effectuerons dans la Section 1.2.

Quelques notations

En plus de la variable T, nous introduisons la variable aléatoire C représentant le temps de censure et $Z(.) \in \mathbb{R}^d$ un vecteur de covariables. Nous considérons que C est indépendant de T conditionnellement à $Z(.)$ et qu'il existe $\tau > 0$ tel que le segment $[0, \tau]$ est le support de T et C. On note $(T_i, C_i, Z_i(.))_{i \in \{1,\dots,n\}}$ un n-échantillon de loi $(T, C, Z(.))$.

Nous faisons l'hypothèse de censure à droite, *i.e.*, nous considérons que, pour tout $i \in \{1, \dots, n\}$, nous observons la variable aléatoire $X_i = \min(T_i, C_i)$. Nous observons également un indicateur de décès $\Delta_i = \mathbb{1}_{T_i \leq C_i}$. Cet indicateur prend la valeur 1 si l'observation X_i est un temps de décès, et 0 si c'est un temps de censure. Le n-échantillon auquel nous avons accès est alors $(X_i, \Delta_i, Z_i(.))_{i \in \{1,\dots,n\}}$. Pour tout $i \in \{1, \dots, n\}$, nous notons $Y_i(t) = \mathbb{1}_{X_i \geq t}$. Le processus $(Y_i(t))_{t \in \mathbb{R}_+}$ indique si l'individu i est encore à risque au temps t, c'est-à-dire si l'individu i est encore vivant au temps t. On note

$$N_i(t) = \mathbb{1}_{\{X_i \leq t, \ T_i \leq C_i\}},$$

le processus de comptage valant 1 à partir de X_i et 0 avant. Ce processus est identiquement nul si l'individu i est censuré. La somme de ces processus de comptage individuels permet de définir le processus de comptage

$$\bar{N}(t) = \sum_{i=1}^{n} N_i(t).$$

Ce processus possède un saut de taille 1 à chaque temps de décès. On suppose dans l'ensemble de la thèse qu'il n'y a pas d'égalité entre temps de décès : chaque temps de décès correspond au décès d'un seul individu.

1.1.2 Estimateur de Kaplan-Meier

Nous affirmions précédemment que la censure empêchait l'utilisation de résultats classiques en statistiques. Nous pouvons en faire la démonstration avec l'estimation de la fonction de survie S de T. En effet, une première idée pour l'estimation non paramétrique

de la fonction de survie S est d'utiliser la fonction de répartition empirique F_n de T définie de la façon suivante : pour tout $t \in [0, \tau]$,

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

En effet, le théorème de Glivenko-Cantelli nous assure de la convergence uniforme de F_n vers F. Cependant, en présence de données censurées, nous n'avons pas accès à l'échantillon $(T_1, \ldots, T_n)$ mais seulement à $(X_1, \ldots, X_n)$. Nous ne pouvons donc évaluer la fonction F_n. On pourrait alors penser à utiliser la fonction de répartition empirique de X,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} = 1 - \frac{1}{n} \sum_{i=1}^{n} Y_i(t).$$

Cependant cette estimation de F est biaisée en présence de censure. En effet, son espérance au temps t vaut $P(T \leq t, T \leq C)$, ce qui est une sous-estimation de $P(T \leq t)$. Un estimateur consistant de F, en présence de censure, est introduit pour la première fois par Kaplan and Meier (1958).

Définition 1.2 (Estimateur de Kaplan-Meier)**.** L'estimateur de Kaplan-Meier de la fonction de répartition F, noté $\hat{F}$, est défini, pour tout $t \in [0, \tau]$, par

$$\hat{F}(t) = 1 - \prod_{\substack{i \in \{1,\ldots,n\} \\ X_i \leq t}} \left(1 - \frac{\Delta_i}{\sum_{j=1}^{n} Y_j(X_i)}\right). \tag{1.1}$$

Ce processus est constant par morceaux, continu à droite avec limite à gauche en tout point. Il possède de plus un saut à chaque temps de décès. On peut remarquer que l'estimateur $\hat{F}$ coïncide avec F_n dans le cas où il n'y a pas de censure. On définit alors naturellement l'estimateur de Kaplan-Meier de la fonction de survie S par $\hat{S} = 1 - \hat{F}$. Illustrons la construction de l'estimateur de Kaplan-Meier de la fonction de survie S sur un exemple.

Exemple 1.3. Supposons qu'on ait accès aux données de la Table 1.1

Table 1.1 – Exemple de jeu de données

X	2	3	7	9	12	15	16	20	21	24
Δ	1	1	0	1	0	1	1	1	1	0

Les temps $\{2, 3, 9, 15, 16, 20, 21\}$ sont donc des temps de décès et les temps $\{7, 12, 24\}$ des temps de censure. Pour $i \in \{1, \ldots, 10\}$, la quantité $\sum_{j=1}^{10} Y_j(X_i)$ correspond au nombre d'individus à risque au temps X_i. On a donc

$$\left(\sum_{j=1}^{10} Y_j(X_1), \ldots, \sum_{j=1}^{10} Y_j(X_{10})\right) = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1).$$

En utilisant la définition de l'estimateur de Kaplan-Meier (1.1), on obtient l'estimateur représenté en Figure 1.1. On constate bien que les sauts de l'estimateur de Kaplan-Meier, $\hat{S}$, s'effectuent sur les temps de décès. De plus, on remarque qu'il ne s'annule pas à la fin du jeu de données, comme on pourrait l'imaginer pour un estimateur d'une fonction de survie. Ceci est dû au fait que le dernier temps de décès est un temps censuré.

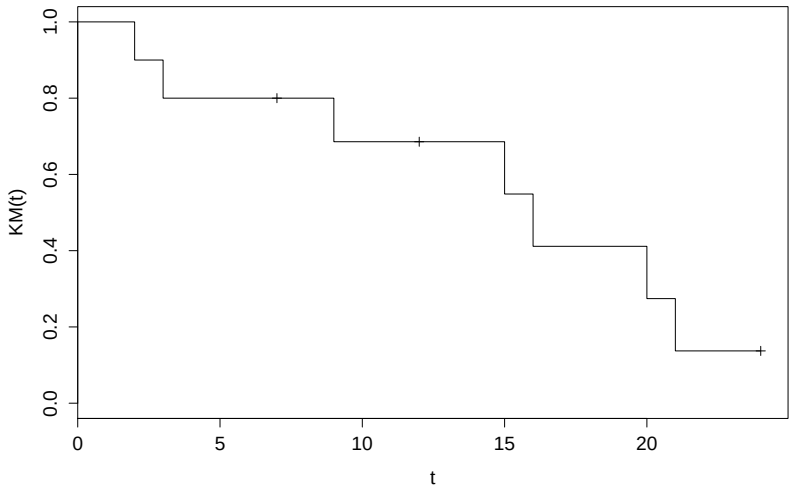


FIGURE 1.1 – Exemple d'estimateur de Kaplan-Meier en fonction du temps

On peut montrer que l'estimateur de Kaplan–Meier est uniformément convergent et asymptotiquement normal. Ces propriétés sont intéressantes pour l'utilisation de cet estimateur dans différents cadres. Par exemple, les incréments de l'estimateur de Kaplan–Meier peuvent être utilisés pour pondérer des variables dépendantes du temps intégrées par rapport à $\bar{N}(t)$. Les intégrales obtenues, dites intégrales de Kaplan-Meier, convergent alors vers l'espérance de fonctions de T et ne sont pas biaisées en présence de censure. De plus, ces intégrales sont asymptotiquement normales sous de bonnes conditions (Stute, 1995). Satten and Datta (2001) ont montré qu'il est possible d'exprimer l'estimateur de Kaplan–Meier de F comme une somme pondérée, avec des poids dépendants de l'estimateur de Kaplan–Meier de la fonction de répartition de la censure C.

De nombreux modèles ont été développés pour l'analyse de données de survie. La Section 1.2 présente un état de l'art de ces différents modèles en fonction de leur degré de paramétrisation, *i.e.*, les modèles sont classés en deux catégories : semi-paramétriques et non paramétriques.

1.2 Du semi-paramétrique au non paramétrique

1.2.1 Modèles de survie semi-paramétriques

Modèles à hasards proportionnels et modèle de Cox

Les modèles à hasards proportionnels servent à exprimer un effet multiplicatif des covariables, que l'on note β, sur le taux de hasard. Ce dernier prend donc la forme suivante, pour tout $t \in [0, \tau]$,

$$\lambda(t|Z) = \lambda_0(t)h(\beta, Z), \tag{1.2}$$

où Z est toujours un vecteur de covariables, λ_0 est appelé taux de hasard de base, β est le paramètre d'intérêt (exprimant l'effet des covariables Z) et h est une fonction positive. On remarque que le taux de hasard se décompose ici en une fonction ne dépendant que

de t et une autre n'en dépendant pas. Considérons la relation suivante : pour tout i et j,

$$\frac{\lambda(t|Z_i)}{\lambda(t|Z_j)} = \frac{h(\beta, Z_i)}{h(\beta, Z_j)}.$$

On remarque que pour deux individus distincts, au temps $t \in [0, \tau]$, les taux de hasard qui leur sont associés sont proportionnels, d'où la dénomination de "modèles à hasards proportionnels". On peut noter que le modèle (1.2) est paramétrique si λ_0 l'est. Par exemple, si $\lambda_0(t) = \lambda$ pour tout $t \geq 0$, alors on retrouve le modèle paramétrique exponentiel. Le modèle (1.2) est semi-paramétrique si on ne suppose aucune forme sur λ_0. On considère ici des modèles semi-paramétriques.

Nous nous intéressons maintenant à un modèle à hasards proportionnels particulier introduit par Cox (1972).

Définition 1.4 (Modèle de Cox)**.** Le modèle de Cox est défini par la fonction de hasard

$$\lambda(t|Z) = \lambda_0(t) \exp(\beta^T Z), \ \forall t \geq 0, \tag{1.3}$$

où λ_0 est un risque instantané de base et β^T la transposée du vecteur $\beta \in \mathbb{R}^d$.

Cox (1972) introduisit le modèle (1.3) dans le cas de covariables Z ne dépendant pas du temps. Le passage à des covariables dépendantes du temps a été étudié, entre autres, par Kalbfleisch and Prentice (1980); Huber (2000); Therneau and Grambsch (2000); Klein and Moeschberger (2003); Lawless (2011). Faisons le point maintenant sur l'estimation de β dans le cas du modèle de Cox (1.3). Rappelons nous d'abord que nous avons supposé que deux décès ne peuvent arriver simultanément. Ainsi la probabilité qu'il y ait un décès en X_i est

$$\sum_{j=1}^{n} Y_j(X_i)\lambda_0(X_i) \exp(\beta^T Z_j).$$

Donc la probabilité que ce soit l'individu i qui décède en X_i, sachant qu'un décès a eu lieu en X_i est

$$\frac{Y_i(X_i)\lambda_0(X_i) \exp(\beta^T Z_i)}{\sum_{j=1}^{n} Y_j(X_i)\lambda_0(X_i) \exp(\beta^T Z_j)} = \frac{\exp(\beta^T Z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta^T Z_j)}. \tag{1.4}$$

Les contributions à la vraisemblance s'effectuent à chaque temps de décès. On définit ainsi la vraisemblance partielle de Cox.

Définition 1.5 (Vraisemblance partielle de Cox)**.** La vraisemblance partielle de Cox, notée L_n, est une fonction du coefficient de régression $\beta \in \mathbb{R}^d$ et s'écrit

$$L_n(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta^T Z_i)}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta^T Z_j)} \right]^{\Delta_i}. \tag{1.5}$$

On voit alors tout de suite l'intérêt du modèle de Cox (1.3) pour l'estimation de β. En effet, l'expression de la vraisemblance partielle de Cox (1.5) ne fait pas intervenir la fonction de hasard λ_0 et permet donc l'estimation directe de β sans avoir à estimer λ_0. La fonction λ_0 est ici considérée comme un paramètre de nuisance, et β est le seul paramètre d'intérêt. Cette vraisemblance partielle n'est pas une vraisemblance au sens strict du terme, mais elle se comporte comme telle. Il est alors naturel de chercher des

résultats asymptotiques la concernant. Avant de les énoncer, nous nous plaçons dans le cas où les covariables Z dépendent du temps. La vraisemblance partielle devient alors

$$L_n(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta^T Z_i(X_i))}{\sum_{j=1}^{n} Y_j(X_i) \exp(\beta^T Z_j(X_i))} \right\}^{\Delta_i}. \tag{1.6}$$

Remarquons que l'expression (1.6) nécessite que la covariable de l'individu $i \in \{1, \dots, n\}$, $Z_i(.)$, soit observée pour $t > X_i$, ce qui n'est pas le cas en pratique. Une première manière de pallier à ce problème est de supposer que $Z_i(t) = 0$ pour $t > X_i$ (Arjas, 1988). Nous choisissons ici de poser $Z_i(t) = X_i$ pour tout $t > X_i$, *i.e.*, remplacer $Z_i(t)$ par $Z_i(\min(t, X_i))$ pour tout $t \in [0, \tau]$. Pour éviter de surcharger les équations, nous ne l'écrirons cependant pas explicitement. Ce choix implique que les valeurs covariables soient connues à tous les temps de décès jusqu'à ce qu'ils soient censurés ou décèdent à leur tour. Cette hypothèse est courante pour traiter des covariables dépendantes du temps (Andersen and Gill, 1982). Prenons le logarithme de (1.6), nous obtenons la log-vraisemblance partielle qui est la fonction à maximiser pour l'estimation du coefficient de régression β.

$$\begin{aligned} l_n(\beta) = \log(L_n(\beta)) &= \sum_{i=1}^{n} \Delta_i \left\{ \beta^T Z_i(X_i) - \log \left[\sum_{j=1}^{n} Y_j(X_i) \exp(\beta^T Z_j(X_i)) \right] \right\} \\ &= \sum_{i=1}^{n} \int_0^\tau \left\{ \beta^T Z_i(t) - \log \left[\sum_{j=1}^{n} Y_j(t) \exp(\beta^T Z_j(t)) \right] \right\} dN_i(t). \end{aligned}$$

La fonction l_n est concave. La maximiser est donc équivalent à annuler sa dérivée, appelée fonction score et notée U.

$$U(\beta) = \frac{\partial l_n}{\partial \beta}(\beta) = \sum_{i=1}^{n} \int_0^\tau \{Z_i(t) - \mathcal{E}_\beta(Z|t)\} \, dN_i(t),$$

où

$$\mathcal{E}_\beta(Z|t) = \sum_{i=1}^{n} Z_i(t) \pi_i(\beta, t), \;\; \mathcal{V}_\beta(Z|t) = \sum_{i=1}^{n} Z_i^2(t) \pi_i(\beta, t) - \mathcal{E}_\beta(Z|t)^2, \tag{1.7}$$

avec

$$\pi_i(\beta, t) = \frac{Y_i(t) \exp(\beta Z_i(t))}{\sum_{j=1}^{n} Y_j(t) \exp(\beta Z_j(t))}. \tag{1.8}$$

On reconnaît dans l'équation (1.8) la probabilité (1.4) adaptée au cas de covariables dépendantes du temps. Les quantités $\mathcal{E}_\beta(Z|t)$ et $\mathcal{V}_\beta(Z|t)$ définies en (1.7) correspondent donc respectivement à l'espérance et la variance de Z par rapport à la famille de probabilités $(\pi_i(\beta, t))_{i \in \{1,\dots,n\}}$. Un estimateur du paramètre de régression β, dans le cas du modèle de Cox (1.3), est donc la solution de l'équation $U(\beta) = 0$ et est notée $\hat{\beta}_{PL}$ pour spécifier que cet estimateur est obtenu en maximisant la vraisemblance partielle (ou "Partial Likelihood" en anglais).

Pour effectuer de l'inférence sur β, on peut utiliser les résultats présentés par Andersen and Gill (1982). En effet, ils montrent que $\hat{\beta}_{PL}$ est un estimateur consistant de β et même que $\sqrt{n}(\hat{\beta}_{PL} - \beta)$ converge en loi vers une variable aléatoire gaussienne centrée de variance estimable par $I(\hat{\beta}_{PL})^{-1}$, où

$$I(\beta) = -\frac{\partial U}{\partial \beta}(\beta) = \sum_{i=1}^{n} \Delta_i \mathcal{V}_\beta(Z|X_i).$$

On peut trouver plus de détails sur l'estimation de β dans Kalbfleisch and Prentice (2011), par exemple.

Le modèle de Cox est le modèle à risques proportionnels le plus utilisé pour l'analyse de survie. D'autres modèles existent néanmoins, comme celui de Marubini and Valsecchi (2004) :

$$\lambda(Z|t) = \lambda_0(t)\exp(g(Z(s),\ 0 \leq s \leq t; \beta),$$

où $g(.;.)$ est une fonction de la famille $\{Z(s), 0 \leq s \leq t\}$ et du paramètre de régression β définie par

$$g(Z(s), 0 \leq s \leq t; \beta) = \beta^T(Z(t) - Z(t-h)),\ h > 0.$$

Ce choix de g permet d'évaluer l'influence de l'évolution de la covariable Z entre les instants $t-h$ et t sur le risque de survenue d'un décès au temps t.

Les modèles à risques proportionnels, et en particulier le modèle de Cox, sont très utilisés en pratique pour leur simplicité d'interprétation. En effet, prenons le cas d'une covariable Z binaire, par exemple $Z_i = 1$ si le patient i reçoit un traitement et $Z_i = 0$ s'il reçoit un placebo. Alors, l'hypothèse des risques proportionnels revient à dire que l'effet entre les groupes est constant au cours du temps. Or il est courant que l'effet d'un médicament diminue à cause de l'accoutumance. Cette hypothèse est donc très forte et peut mener à des erreurs d'interprétation si l'effet des covariables, β, dépend du temps. C'est pourquoi nous présentons maintenant des alternatives au modèle de Cox prenant en compte un coefficient de régression qui varie dans le temps, c'est-à-dire nous nous intéressons au modèle

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta(t)^T Z),\ \forall t \geq 0. \tag{1.9}$$

Alternatives semi-paramétriques au modèle de Cox

Puisque nous considérons toujours dans cette partie des modèles semi-paramétriques, il s'agit donc de remplacer la fonction $\beta(t)$ dans (1.9) par un vecteur de paramètres à estimer en maximisant la vraisemblance partielle. Autrement dit, nous supposons que la fonction β prend la forme

$$\beta_i(t) = \beta_{0i}h_i(t),\ \forall i \in \{1, \ldots, d\},$$

où les β_{0i} sont des constantes et les h_i des fonctions déterministes. On peut citer quelques exemples avec Cox (1972) pour des fonctions h_i linéaires, Brown (1975) pour des fonctions constantes par morceaux, ou encore Stablein et al. (1981) pour des fonctions polynomiales.

Modèles constants par morceaux Deux modèles nous intéresserons particulièrement dans le Chapitre 3. Il s'agit du modèle introduit par Anderson and Senthilselvan (1982) :

$$\beta(t) = \beta_1 \mathbb{1}_{t \leq \gamma} + \beta_2 \mathbb{1}_{t > \gamma},\ \forall t \in [0, \tau], \tag{1.10}$$

et de sa généralisation, présentée par exemple dans O'Quigley (2008),

$$\beta(t) = \beta_1 \mathbb{1}_{t \leq \gamma_1} + \beta_2 \mathbb{1}_{\gamma_1 < t \leq \gamma_2} + \cdots + \beta_K \mathbb{1}_{t > \gamma_{K-1}},\ t \in [0, \tau]. \tag{1.11}$$

La fonction de régression β est ici supposée constante par morceaux (avec seulement deux morceaux dans le cas du modèle de Anderson and Senthilselvan (1982)). Les points de

discontinuité de β, les γ_i, sont appelés points de rupture. Un cas d'intérêt particulier du modèle (1.10) est le cas $\beta_1 = -\beta_2$ (O'Quigley and Pessione, 1989). En effet, lors d'un traitement chirurgical par exemple, l'effet peut être négatif au début de l'étude puisque le risque de décès est élevé lors de l'opération. Et par la suite, les patients qui ont survécu à l'opération bénéficient de ce traitement, l'effet devient donc positif.

La fonction β peut aussi prendre la forme $\beta(t) = \beta_0 + G(t)$ (Moreau et al., 1985), où G est une fonction constante par morceaux avec des temps de rupture prédéfinis. Grambsch and Therneau (1994) ont considéré un effet de la forme $\beta(t) = \beta_0 + \theta g(t)$, où g est une fonction déterministe ou aléatoire mais prévisible.

Modèles de Cox stratifiés Dans le cas où une variable ne vérifie pas l'hypothèse de hasards proportionnels, on peut considérer un modèle de Cox stratifié (Kalbfleisch and Prentice, 1973). Pour expliquer ce modèle, prenons l'exemple d'une variable Y à valeurs dans $\{0, 1\}$. On considère alors que le risque de base est différent dans les deux strates mais les covariables Z agissent de la même manière sur le taux de hasard. Plus précisément,

$$\lambda(t|Z, Y = 0) = \lambda_0(t) \exp(\beta^T Z)$$
$$\lambda(t|Z, Y = 1) = \lambda_1(t) \exp(\beta^T Z).$$

On constate que l'effet des covariables est le même dans chaque strate. On peut obtenir des estimations des différents paramètres $\lambda_0, \lambda_1, \beta$ de la même manière que pour le modèle de Cox, *i.e.*, par vraisemblance partielle. La vraisemblance totale du modèle étant le produit des vraisemblances de chaque strate. Therneau and Grambsch (2000) montrent qu'il peut être judicieux d'effectuer cette stratification selon les modalités d'une variable catégorielle qui ne respecte pas l'hypothèse de hasards proportionnels.

Modèles de survie accélérée Les modèles de survie accélérée sont souvent utilisés en fiabilité. En effet, si on considère un temps de décès comme le temps passé avant qu'une machine tombe en panne, il est raisonnable de penser que plus le temps passe, plus on a de chance de voir une machine tomber en panne, *i.e.*, la survie décroît. Une représentation des modèles de survie accélérée se fait par la fonction de survie accélérée

$$S(t|Z) = S_0\left(t \exp\left(\beta^T Z\right)\right),$$

où Z est toujours le vecteur de covariables et β le vecteur des coefficients de régression. On obtient alors le risque instantané suivant

$$\lambda(t|Z) = \exp\left(\beta^T Z\right) \lambda_0\left(t \exp\left(\beta^T Z\right)\right).$$

On constate alors que la quantité $\exp(\beta^T Z)$ est un facteur d'accélération puisqu'un changement dans les covariables Z implique un changement d'échelle du temps. Plus de détails peuvent être trouvés dans Bagdonavičius and Nikulin (1997, 2000).

Modèles de fragilité Les modèles de fragilité permettent la prise en compte d'une hétérogénéité dans les observations. C'est le cas lorsque des covariables importantes ne sont pas observables ou inconnues (des facteurs environnementaux par exemple). Prenons par exemple une nouvelle covariable Y. On suppose que l'effet de cette covariable est exprimé par la quantité $\exp(\beta_Y Y)$, comme dans le cas d'un modèle de Cox. On obtient alors le taux de hasard suivant

$$\lambda(t|Z, Y) = \lambda_0(t) \exp(\beta_Y Y) \exp(\beta^T Z).$$

On note $\omega = \exp(\beta_Y Y)$. La variable aléatoire réelle positive ω est appelée "fragilité". Comme ω est une variable aléatoire, on travaille ensuite avec la fonction de survie moyennée sur ω.

On utilise souvent ces modèles pour prendre en compte la dépendance entre individus, par exemple des individus d'une même ville ou se faisant soigner dans un même hôpital. La fragilité est alors commune à tous les individus d'un même groupe, mais différente entre chacun des groupes. On parle alors de modèles à fragilités partagées. On peut trouver des détails sur ces modèles dans Klein et al. (2013).

Modèle de Yang et Prentice Le modèle de Yang and Prentice (2005) a été développé dans le cas d'une covariable binaire (à valeurs dans $\{0, 1\}$) pour prendre en compte des fonctions de survie conditionnelles aux covariables, $S(t|Z = 0)$ et $S(t|Z = 1)$ qui se croisent. Plus précisément, les auteurs définissent le taux de hasard suivant

$$\lambda(t|Z = 1) = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)S(t|Z = 0)}\lambda(t|Z = 0), \tag{1.12}$$

où θ_1 et θ_2 sont des constantes strictement positives et $S(t|Z = 0) = \exp(-\int_0^t \lambda(s|Z = 0)ds)$. On constate bien que les risques du modèle (1.12) ne sont pas proportionnels, à cause de l'utilisation de la fonction de survie de groupe $Z = 0$:

$$\frac{\lambda(t|Z = 1)}{\lambda(t|Z = 0)} = \frac{\theta_1\theta_2}{\theta_1 + (\theta_2 - \theta_1)S(t|Z = 0)}.$$

Pour comprendre l'interprétation des paramètres θ_1 et θ_2, on peut remarquer que

$$\theta_1 = \lim_{t\to 0} \frac{\lambda(t|Z = 1)}{\lambda(t|Z = 0)}, \ \theta_2 = \lim_{t\to \tau} \frac{\lambda(t|Z = 1)}{\lambda(t|Z = 0)}.$$

Le paramètre θ_1 peut alors être vu comme le risque relatif à court terme et, de la même manière, le paramètre θ_2 peut être vu comme le risque relatif à long terme. Pour l'estimation de ces paramètres, ils définissent une pseudo vraisemblance spécifique au modèle (1.12) et la maximisent.

1.2.2 Modèles de survie non-paramétriques

Les méthodes non-paramétriques en analyse de survie se focalisent généralement sur l'estimation du risque instantané et non sur la fonction $\beta(.)$. Nous citons ici quelques méthodes d'estimation possibles. Nous ne nous attardons cependant pas sur les forêts aléatoires qui seront plus amplement détaillées à la Section 1.3.

Modèle additif généralisé

Le modèle à hasards proportionnels peut être vu comme un modèle à hasards multiplicatifs. Rien n'empêche de s'intéresser également aux modèles additifs. Le modèle à hasards additifs a été introduit par Aalen (1980) et est défini par

$$\lambda(t|Z) = \beta_0(t) + \beta(t)^T Z(t).$$

Sous ce modèle, il est aisé d'estimer les coefficients de régression cumulés $\int_0^t \beta(s)ds$ par moindres carrés. Certains auteurs se sont penchés sur des combinaisons de modèles additifs

et multiplicatifs. On peut citer Lin and Ying (1995) qui proposent un risque instantané de la forme

$$\lambda(t|Z^{(1)}, Z^{(2)}) = Y(t)g\{\beta^T Z^{(1)}(t)\} + \lambda_0(t)h\{\gamma^T Z^{(2)}(t)\},$$

où $Y(t)$ représente toujours l'indicateur d'être à risque à l'instant t, $Z^{(1)} \in \mathbb{R}^p$ et $Z^{(2)} \in \mathbb{R}^q$ sont des vecteurs de covariables, $\beta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}^q$, et g et h sont des fonctions de lien. Scheike and Zhang (2002) ont, quant à eux, proposé le modèle suivant

$$\lambda(t|Z^{(1)}, Z^{(2)}) = Y(t)\beta(t)^T Z^{(1)}(t) \exp\left(\gamma^T Z^{(2)}(t)\right),$$

où cette fois β est bien une fonction de régression. McKeague and Utikal (1990) proposent d'utiliser une méthode à noyaux pour estimer la fonction α dans le modèle

$$\lambda(t|Z) = Y(t)\alpha(t, Z(t)).$$

Ce dernier article s'inscrit dans la lignée de nombreux travaux proposant l'étude des estimateurs de type Beran (1981), comme Stute (1986); Dabrowska (1989); Lin and Ying (1995).

Méthodes de lissage

Une autre manière d'estimer non paramétriquement la fonction β, est de la projeter sur une base de fonctions connues. On remplace ainsi cette fonction par un vecteurs de paramètres à estimer.

Dans le cas de la faible dimension, on peut toujours estimer grâce à la maximisation de la vraisemblance partielle. On trouve quelques exemples de cette démarche avec une base de splines cubiques (Hess, 1994), une base de splines quadratiques (Abrahamowicz et al., 1996), ou encore une base d'histogrammes (Murphy and Sen, 1991). Dans ce dernier cas, l'estimateur est alors constant par morceaux sur des intervalles de temps prédéfinis. On note donc la différence de ce dernier avec le modèle (1.11) qui ne définit pas les intervalles de temps sur lesquels la fonction β est constante.

Cai and Sun (2003) proposent un lissage de la log-vraisemblance par un noyau pour des coefficients $\beta(t)$ localement linéaires. On obtient ainsi une estimation non paramétrique locale de β autour des temps de décès.

Castellan and Letué (2000) étendent les travaux de Murphy and Sen (1991) au cas de la grande dimension. Les auteurs commencent par projeter β sur une base d'histogrammes. Pour cela on note $\mathcal{M}$ une partition de $[0, 1]$, dont les intervalles sont notés $(I_m, m \in \mathcal{M})$. On cherche alors à estimer β par une fonction $\hat{\beta}$ de la forme

$$\hat{\beta}(t) = \sum_{m \in \mathcal{M}} \hat{\beta}_m \mathbb{1}_{I_m},$$

où les $\hat{\beta}_m$ sont les estimateurs des projections de β sur chaque intervalle I_m. On souhaite ensuite choisir la bonne partition $\mathcal{M}$. En effet, pour bien approcher la fonction β, on aimerait que la partition $\mathcal{M}$ comporte un grand nombre d'intervalles. Cependant, en augmentant ce nombre, on augmente celui des coefficients à estimer et on perd ainsi en qualité d'estimation. Les auteurs optent donc pour une sélection de modèle, selon la méthode proposée par Barron et al. (1999).

Dans le cas d'un modèle additif généralisé, Hastie and Tibshirani (1990) et Gray (1992) ont choisi pour $\mathcal{B}$ une base de splines cubiques et une pénalisation de la forme

$$\text{pen}(\beta(t)) = \frac{1}{2} \sum_{i=1}^{p} \lambda_i \beta_i''(s) ds,$$

où les λ_i sont des constantes positives de lissage. Pour les projections sur des splines cubiques sous un modèle à hasards non proportionnels, on peut citer plus récemment LeBlanc and Crowley (1999).

On peut noter en marge l'article de Antoniadis et al. (1999) qui estime le taux de hasard, non conditionnel aux covariables, par une base d'ondelettes.

Minimisation de contraste

On s'intéresse dans cette partie à deux contrastes en particulier : la vraisemblance partielle et les moindres carrés. On trouvera plus de détails sur la minimisation de contraste dans Birgé and Massart (1997).

Vraisemblance partielle On s'intéresse dans ce paragraphe à des estimations de la fonction de régression β par log-vraisemblance pénalisée. Le problème d'estimation s'écrit alors

$$\hat{\beta}(t) = \max_{\beta \in \mathcal{B}} \{l_n(\beta(t)) - \text{pen}(\beta(t))\},$$

où l_n est la log-vraisemblance partielle de Cox, l'ensemble $\mathcal{B}$ et la pénalisation changent selon les cas.

Zucker and Karr (1990) ont considéré pour l'ensemble $\mathcal{B}$ l'ensemble des fonctions k-différentiables par morceaux pour $k \geq 3$ et une pénalisation

$$\text{pen}(\beta) = \frac{1}{2}\alpha_n \int_0^1 \left\{\beta^{(k)}(t)\right\}^2 dt,$$

où α_n est un nombre strictement positif choisi par le statisticien. Les auteurs ont montré que $\hat{\beta}$ est consistant, et asymptotiquement normal pour $k \geq 4$.

O'Sullivan (1993) propose pour l'ensemble $\mathcal{B}$ des conditions de régularité à travers l'appartenance à des espaces de Sobolev. Pour la pénalisation, ils suivent la méthode développée par Wahba (1990), c'est-à-dire

$$\text{pen}(\beta(t)) = \frac{1}{n}\sum_{i=1}^{n} \beta(t) dN_i(t),$$

où N_i est le processus de comptage valant 1 à partir de X_i et 0 avant.

Verweij and Houwelingen (1995) proposent plutôt de définir la fonction β uniquement sur les temps définis par les observations et définissent une pénalisation en variation totale pour la norme $\mathbb{L}^2$ pondérée, c'est-à-dire

$$\text{pen}(\beta(t)) = \frac{1}{2}\sum_{k=1}^{d} \lambda_k \sum_{j=1}^{m-1} w_{jk} \left\{\beta_k(X_{j+1}) - \beta_k(X_j)\right\}^2,$$

où les λ_k sont des paramètres de lissage, les w_{jk} sont des poids strictement positifs et m le nombre de décès. Avec cette pénalisation, la fonction β est lissée dans le sens où on force les valeurs consécutives $(\beta_k(X_j))_{j\in\{1,\ldots,m\}}$ à être proches les unes des autres pour tout $k \in \{1, \ldots, d\}$.

Moindres carrés Comte et al. (2011) étudient directement l'estimation non paramétrique du risque instantané conditionnel aux covariables. Pour cela, les auteurs considèrent un contraste des moindres carrés adapté au problème de l'estimation d'intensité :

$$c_n(h) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau h^2(Z_i,t)Y_i(t)dt - \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau h(Z_i,t)dN_i(t), \tag{1.13}$$

où $h \in \mathbb{L}^2 \cap \mathbb{L}^\infty([0,1]\times[0,\tau])$. Une piste pour comprendre le lien entre la minimisation de c_n (1.13) et l'estimation de $\lambda(.|Z)$ vient du calcul de l'espérance de c_n :

$$E\left[c_n(h)\right] = \|h - \lambda(.|Z)\|_\mu^2 - \|\lambda(.|Z)\|_\mu^2,$$

où $\|.\|_\mu$ est une norme dépendant de la loi de Z. On voit ainsi à travers cette espérance l'intuition derrière la minimisation de c_n.

Dans le cas d'une estimation du risque instantané non conditionnel aux covariables, on peut citer par exemple Reynaud-Bouret (2006) qui projette l'intensité sur une base d'histogrammes avant de pénaliser par la norme $\mathbb{L}^2$.

1.3 Introduction aux forêts aléatoires

Lors des dernières décennies sont apparus des jeux de données de très grandes dimensions. En conséquence, de nouvelles méthodes d'apprentissage ont vu le jour dans l'objectif de traiter cette quantité de données massive, tout en gardant une efficacité statistique raisonnable. Nous nous intéressons ici à l'une des méthodes développées pour ce type de données : les forêts aléatoires. Les forêts aléatoires ont été introduites pour la première fois par Breiman (2001) et sont maintenant parmi les méthodes les plus efficaces pour la grande dimension, tant en régression qu'en classification. On leur trouve donc naturellement de nombreuses applications en reconnaissance de forme (Rogez et al., 2008), chimio-informatique (Svetnik et al., 2003) ou en génomique (Qi, 2012) par exemple.

Le principe de base des forêts aléatoires est de faire pousser un grand nombre d'arbres, puis de les agréger pour former une prédiction. Pour obtenir des arbres différents, on introduit de l'aléatoire dans le processus de construction. Il existe un grand nombre de façon de construire des arbres (CART, uniforme, centré...). On comprend donc qu'il existe une grande variété de forêt aléatoire, car elles dépendent à la fois de la manière d'introduire de l'aléatoire dans la construction de l'arbre et du processus de construction lui-même. Nous nous intéressons ici à l'une des forêts les plus utilisées en pratique.

Les forêts de Breiman (2001) sont sans doute le type de forêts le plus utilisé. Cependant, les raisons de ses bonnes performances sont encore mal comprises. Pour obtenir des résultats théoriques, on remplace le critère de coupure CART (Breiman et al., 1984), dépendant de tout l'échantillon, par un critère plus simple à analyser car ne dépendant pas des étiquettes des données. Nous présentons dans cette section les modèles de forêt qui seront utiles dans la suite.

1.3.1 Notations

Nous introduisons ici les notations utiles à cette section et au Chapitre 4. On note $\mathbf{X} = (\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(d)}) \in [0,1]^d$ le vecteur des variables explicatives et $Y \in \mathbb{R}$ la variable réponse. On cherche à estimer la fonction de régression $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Pour cela, on introduit un échantillon d'apprentissage $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$. Les $(\mathbf{X}_i, Y_i)$ sont des couples de variables indépendants et identiquement distribués selon la loi du couple de variables aléatoires $(\mathbf{X}, Y)$, où $\mathbb{E}[Y^2] < \infty$. A l'aide de ces observations, on souhaite construire un estimateur $m_n : [0,1]^d \to \mathbb{R}$ de m.

Les forêts aléatoires sont des méthodes de classification et de régression basées sur une collection de plusieurs arbres aléatoires. Considérons maintenant une forêt à M arbres. On note $m_n(\mathbf{x}, \Theta_j, \mathcal{D}_n)$ la valeur prédite du j-ème arbre, au point $\mathbf{x}$. Les $\Theta_1, \dots, \Theta_M$ sont des variables aléatoires indépendantes, distribuées selon une variable aléatoire générique Θ, indépendante de l'échantillon $\mathcal{D}_n$. En pratique, la variable Θ peut être utilisée pour échantillonner le jeu d'apprentissage ou pour sélectionner les directions ou positions candidates à une coupure. Les prédictions de M arbres aléatoires sont ensuite moyennées pour donner la prédiction finale, *i.e.*,

$$m_{M,n}(\mathbf{x}, \Theta_1, \dots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m, \mathcal{D}_n). \tag{1.14}$$

D'après la loi forte des grands nombres, à $\mathbf{x}$ fixé, conditionnellement à $\mathcal{D}_n$, l'estimateur de la forêt finie (1.14) converge vers l'estimateur de la forêt infinie

$$m_{\infty,n}(\mathbf{x}, \mathcal{D}_n) = \mathbb{E}_\Theta \left[m_n(\mathbf{x}, \Theta, \mathcal{D}_n) \right].$$

Par la suite, on omet la dépendance en l'échantillon $\mathcal{D}_n$ dans les notations. On notera par exemple $m_{\infty,n}(\mathbf{x})$ pour $m_{\infty,n}(\mathbf{x}, \mathcal{D}_n)$. On dit que l'estimateur de la forêt infinie $m_{\infty,n} = m_n$ est consistant si

$$\mathbb{E}\left[(m_n(\mathbf{X}) - m(\mathbf{X}))^2\right] \underset{n\to\infty}{\longrightarrow} 0.$$

Nous présentons dans la suite les forêts aléatoires centrées, médianes, de Breiman et de survie. La part d'aléatoire dans la construction des forêts va donc croître. En effet, les forêts centrées font partie de la famille des forêts purement aléatoire, c'est-à-dire dont la construction ne dépend pas des données. La construction des forêts médianes dépend uniquement des $\mathbf{X}$, celle des forêts de Breiman à la fois des $\mathbf{X}$ et Y. Quand aux forêts de survie, elles permettent la prise en compte de données censurées.

1.3.2 Forêts purement aléatoires

On s'intéresse ici à des forêts aléatoires simplifiées, introduites pour approfondir la compréhension des forêts de Breiman. Les forêts dont nous parlons ici ont la particularité d'être construites indépendamment des données. C'est pourquoi on les appelle des forêts purement aléatoires. Nous nous focalisons plus particulièrement sur les forêts centrées introduites par Breiman (2004). Détaillons tout d'abord la construction de ces forêts :

(1) à chaque noeud de chaque arbre, on choisit uniformément une seule coordonnée parmi $\{1, \ldots, d\}$;

(2) on effectue ensuite la coupure au centre de la cellule selon la coordonnée choisie en (1) ;

(3) on recommence les étapes (1) et (2) k fois, où k est un paramètre fixé par l'utilisateur. On obtient ainsi un arbre binaire de niveau k, c'est-à-dire avec 2^k feuilles.

La Figure 1.2 illustre, en dimension 2, le résultat de la construction d'un arbre centré de niveau 2. Le paramètre k règle la profondeur des arbres. Plus il est grand, plus l'arbre est développé et plus l'erreur d'approximation est faible. Plus il est petit, plus les cellules contiennent de points et plus l'erreur d'estimation est faible. Biau (2012) montre que les forêts centrées sont consistantes si $k \to \infty$ et $n/2^k \to \infty$. L'hypothèse $k \to \infty$ permet le contrôle de l'erreur d'approximation en imposant des arbres suffisamment développés. L'hypothèse $n/2^k \to \infty$ permet le contrôle de l'erreur d'estimation. En effet, si $\mathbf{X}$ est uniformément distribué sur $[0,1]^d$, $n/2^k$ représente le nombre moyen de points dans une cellule au niveau k de l'arbre. S'assurer que ce nombre est grand, c'est s'assurer qu'on garde un nombre de points suffisant dans chaque feuille.

Breiman (2004) et Biau (2012) ont déterminé une vitesse de convergence pour les forêts centrées. Entrons un peu dans les détails. Ils considèrent que la fonction de régression m, qui dépend a priori des d variables $(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)})$, ne dépend en fait que de certaines d'entre elles. On note $\mathcal{S}$ (pour $\mathcal{S}$trong) le sous-ensemble de $\{1, \ldots, d\}$ défini par ces variables importantes. On a alors

$$m(\mathbf{x}) = \mathbb{E}\left[Y | X_{\mathcal{S}}\right],$$

où $X_{\mathcal{S}} = (\mathbf{X}^{(j)},\ j \in \mathcal{S})$. Alors, si m est lipschitzienne, et si les arbres sont construits en n'utilisant que les variables $\mathbf{X}_{\mathcal{S}}$, alors

$$\mathbb{E}\left[(m_n(\mathbf{X}) - m(\mathbf{X}))^2\right] = O\left(n^{\frac{-0.75}{|\mathcal{S}|\log 2 + 0.75}}\right), \tag{1.15}$$

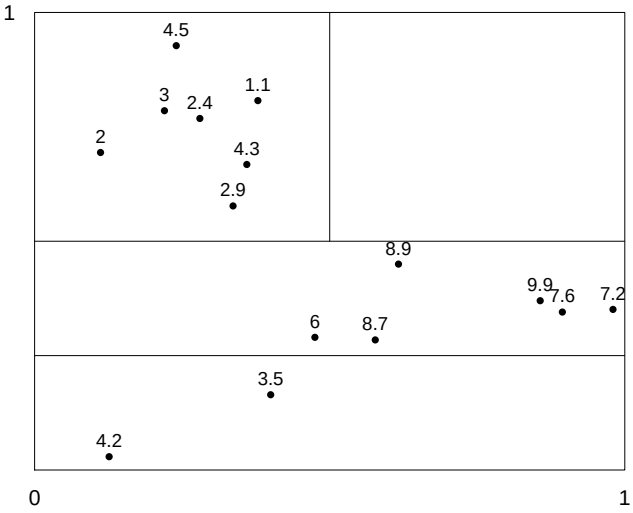


Figure 1.2 – Un arbre centré de niveau 2

où $|\mathcal{S}|$ est le cardinal de $\mathcal{S}$, c'est-à-dire le nombre de variables importantes. La convergence (1.15) montre que la vitesse de convergence de m_n vers m ne dépend que des variables importantes du modèle, et non de l'ensemble des d variables explicatives. Ce résultat pourrait expliquer pourquoi les forêts aléatoires ne sur-ajustent pas, même en présence d'un grand nombre de variables.

1.3.3 Forêts médianes

On s'intéresse dans cette partie à des forêts aléatoires un peu moins simplifiées que les forêts purement aléatoires. Celles dont nous parlons ici ont la particularité d'avoir un processus de construction ne dépendant que des variables explicatives $\mathbf{X}$. Ces forêts sont donc un bon compromis entre la complexité des forêts de Breiman et la simplicité des forêts purement aléatoires. Nous nous focalisons plus particulièrement sur les forêts médianes. Détaillons la construction de ces forêts :

(1) pour chaque arbre, on sélectionne uniformément a_n observations parmi n sans remise ;

(2) à chaque noeud de chaque arbre, on choisit uniformément une seule coordonnée parmi $\{1, \ldots, d\}$;

(3) on effectue ensuite la coupure à la médiane des observations de la cellule selon la coordonnée choisie en (1) ;

(4) on recommence les étapes (2) et (3) jusqu'à ce qu'il ne reste qu'un point dans chaque feuille.

La Figure 1.3 illustre, en dimension 2, le résultat de la construction d'un arbre médian de niveau 2.

Scornet (2014) montre que les forêts médianes, et les forêts quantiles en général, sont consistantes si $a_n \to \infty$ et $a_n/n \to 0$. L'hypothèse $a_n/n \to \infty$ permet de garantir que chaque couple d'observations $(\mathbf{X}_i, Y_i)$ est utilisé dans la construction du m-ième arbre avec une probabilité qui tend vers 0. Ceci force donc le point d'intérêt $\mathbf{x}$ a être déconnecté de

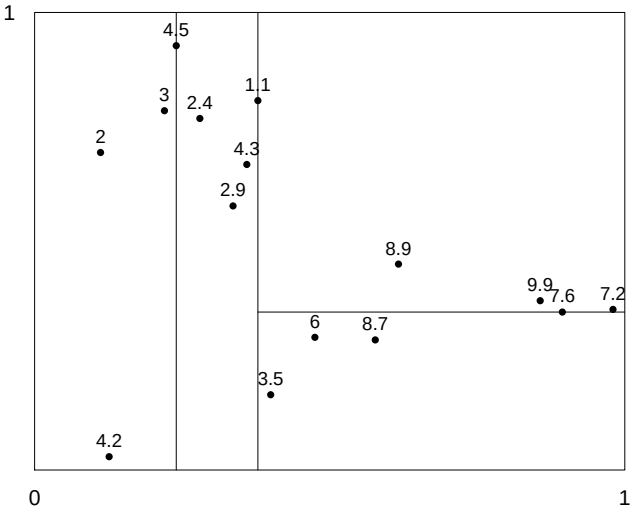


FIGURE 1.3 – Un arbre médian de niveau 2

$(\mathbf{X}_i, Y_i)$ dans un grand nombre d'arbres. Si ce n'était pas le cas, un couple $(\mathbf{X}_i, Y_i)$ pourrait sur-influencer la prédiction en $\mathbf{x}$. Cette hypothèse implique donc que les partitions, formées par les feuilles de chaque arbre, sont suffisamment différentes et elle contrôle ainsi l'erreur d'estimation. Une borne supérieure de la vitesse de convergence des forêts médianes est proposée au Chapitre 4.

1.3.4 Forêts de Breiman

On s'intéresse maintenant aux forêts de Breiman (2001). Leur algorithme fait apparaître trois paramètres importants : $a_n \in \{1, \dots, n\}$ est le nombre de points échantillonnés pour chaque arbre ; $m_{try} \in \{1, \dots, d\}$ est le nombre de coordonnées que l'on peut sélectionner pour couper, à chaque noeud de chaque arbre ; `nodesize` $\in \{1, \dots, n\}$ est le nombre de points minimal dans une cellule, *i.e.*, on ne peut pas couper une cellule qui contient `nodesize` points.

Détaillons la construction de ces forêts :

(1) pour chaque arbre, on sélectionne uniformément a_n observations parmi n avec remise ;

(2) à chaque noeud de chaque arbre, on choisit uniformément, sans remise, m_{try} coordonnées parmi $\{1, \dots, d\}$. On note cet ensemble $\mathcal{M}_{try}$;

(3) on effectue ensuite la coupure selon l'une des m_{try} coordonnées en optimisant le critère CART (Breiman et al., 1984) ;

(4) on recommence les étapes (2) et (3) jusqu'à ce qu'il reste moins de `nodesize` points dans chaque feuille.

Expliquons maintenant ce qu'est le critère de coupure CART. Pour cela, on note A une cellule et $N_n(A)$ le nombre de points tombant dans cette cellule A. Une coupure dans la cellule A est un couple (j, z), où j est une coordonnée de $\{1, \dots, d\}$ et z est la position de la coupure selon la j-ème coordonnée. On note $\mathcal{C}_A$ l'ensemble des coupures possibles pour

la cellule A. Alors le critère de coupure CART s'écrit

$$\mathcal{L}_n(j,z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_A\right)^2 \mathbb{1}_{\mathbf{X}_i \in A} - \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z}\right)^2 \mathbb{1}_{\mathbf{X}_i \in A},$$

où $A_L = \{\mathbf{x} \in A \ / \ \mathbf{x}^{(j)} < z\}$, $A_R = \{\mathbf{x} \in A \ / \ \mathbf{x}^{(j)} \geq z\}$ et $\bar{Y}_A$ (resp. $\bar{Y}_{A_L}$, $\bar{Y}_{A_R}$) est la moyenne des Y_i tombant dans la cellule A (resp. A_l, A_R), avec la convention $0/0 = 0$. La meilleure coupure est ensuite choisie en maximisant $\mathcal{L}_n(j,z)$ sur $\mathcal{M}_{try}$ et $\mathcal{C}_A$, c'est-à-dire

$$(j_n^*, z_n^*) \in \underset{j \in \mathcal{M}_{try},\ (j,z) \in \mathcal{C}_A}{\arg\max} \mathcal{L}_n(j,z).$$

Ce critère mesure, dans la cellule A, la différence entre la variance empirique avant et après coupure. La Figure 1.4 illustre, en dimension 2, le résultat de la construction d'un arbre de Breiman de niveau 2.

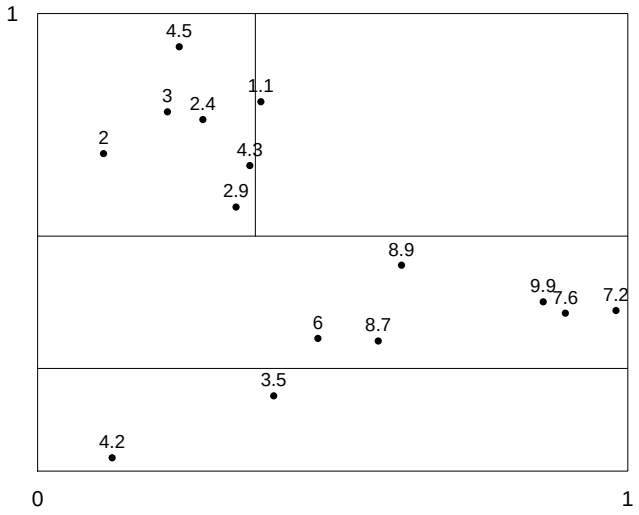


Figure 1.4 – Un arbre de Breiman de niveau 2

Mentch and Hooker (2014) montrent un résultat de normalité asymptotique de la méthode des forêts aléatoires de Breiman finies. Scornet et al. (2014) ont, quant à eux, prouvé un résultat de convergence dans le cas de modèles de régression additifs pour les forêts de Breiman élaguées. Plus précisément, si $\mathbf{X}$ est uniformément distribuée sur $[0,1]^d$, $a_n \to \infty$ et $t_n(\log a_n)^9 a_n \to 0$, où t_n est le nombre total de feuilles dans chaque arbre, alors la forêt de Breiman est consistante.

1.3.5 Forêts de survie aléatoires

Test du log-rank

Le test du log-rank a été introduit par Mantel (1966). Ce test a pour objectif de tester la présence d'un effet des covariables sur la survie. Il s'applique au cas de covariables

catégorielles. Expliquons un peu plus le principe de ce test. Pour cela, on considère une covariable Z à valeurs dans $\{0,1\}$, représentant par exemple un groupe de patients recevant un traitement ($Z=1$) et un groupe de patient recevant un placebo ($Z=0$). On souhaite tester l'hypothèse nulle H_0 contre l'alternative H_1 suivantes.

$$\begin{aligned} H_0 &: \forall t,\ S_1(t) = S_0(t); \\ H_1 &: \exists t_0,\ S_1(t_0) \neq S_0(t_0), \end{aligned}$$

où S_0 est la fonction de survie du groupe $Z=0$, et S_1 la fonction de survie du groupe $Z=1$. On note $D_{1,t}$ le nombre d'individus décédés au temps t dans le groupe $Z=1$, $N_{1,t}$ le nombre d'individus à risque au temps t dans le groupe $Z=1$ et N_t le nombre total d'individus à risque dans les deux groupes au temps t. On définit de la même manière $D_{0,t}$ et $N_{0,t}$. Sous l'hypothèse nulle H_0, la variable $D_{1,t}$ suit une loi géométrique de paramètres $N_{1,t}, D_{1,t}/N_{1,t}, N_t$. Elle admet donc l'espérance et la variance suivante.

$$\begin{aligned} E\left[D_{1,t}\right] &= \frac{D_{1,t}}{N_{1,t}} N_{1,t}, \\ \mathrm{Var}\left(D_{1,t}\right) &= \frac{D_{1,t} N_{0,t} \left(N_{1,t} - D_{1,t}\right)}{N_{1,t}\left(N_t - 1\right)}. \end{aligned}$$

On peut alors considérer la statistique de test Z définie par

$$Z = \frac{\left\{\sum_{i=1}^{n} \Delta_i \left(D_{1,X_i} - E\left[D_{1,X_i}\right]\right)\right\}^2}{\sum_{i=1}^{n} \Delta_i \mathrm{Var}\left(D_{1,X_i}\right)}.$$

Cette statistique converge en loi vers une loi du χ^2 à 1 degré de liberté sous l'hypothèse H_0. C'est cette convergence qui construit le test du log-rank.

Ce test est très utilisé en pratique. La légitimité de son utilisation est renforcée par Peto and Peto (1972) qui prouvent que ce test est le plus puissant sous un modèle à hasards proportionnels. Cependant, cette puissance peut fortement diminuer sous un modèle à hasards non proportionnels (Leurgans, 1983, 1984). Pour pallier à cette perte de puissance, on peut par exemple utiliser des tests du log-rank pondéré (Peto and Peto, 1972; Fleming and Harrington, 2011).

Description des forêts de survie

Les forêts dont nous parlons ici ont la particularité de traiter des données censurées. C'est pourquoi on les appelle des forêts de survie aléatoires. Leur formalisme a été introduit par Ishwaran et al. (2008). Détaillons la construction de ces forêts :

(1) pour chaque arbre, on sélectionne uniformément n observations parmi $\mathcal{D}_n$ avec remise ;

(2) à chaque noeud de chaque arbre, on choisit uniformément, sans remise, m_{try} coordonnées parmi $\{1,\ldots,d\}$. On note cet ensemble $\mathcal{M}_{try}$;

(3) on effectue ensuite la coupure selon l'une des m_{try} coordonnées en maximisant la statistique de test du log-rank (Segal, 1988; Leblanc and Crowley, 1993) ;

(4) on recommence les étapes (2) et (3) jusqu'à ce qu'il reste moins de `nodesize` temps de décès dans chaque feuille.

D'autres critères peuvent être utilisés pour l'étape (3) :

— le critère de coupure de conservation des événements (Naftel et al., 1985) ;

— le critère du score du log-rank (Hothorn and Lausen, 2003), qui coupe en utilisant la statistique du log-rank standardisé ;

— le critère de coupure log-rank aléatoire, pour lequel on choisit une direction de coupure aléatoirement parmi $\mathcal{M}_{try}$, puis on maximise la statistique du log-rank pour couper.

La Figure 1.5 illustre, en dimension 2, le résultat de la construction d'un arbre de survie de niveau 2. Ishwaran and Kogalur (2010) montrent que les forêts de survie sont consistantes si toutes les variables explicatives sont catégorielles et si le temps de décès est à support borné. La preuve de cette consistance repose en fait sur la consistance de chacun des arbres de la forêt. Yang et al. (2010) ont amélioré les performances des forêts de survie dans de nombreuses situations en y introduisant des fonctions noyaux.

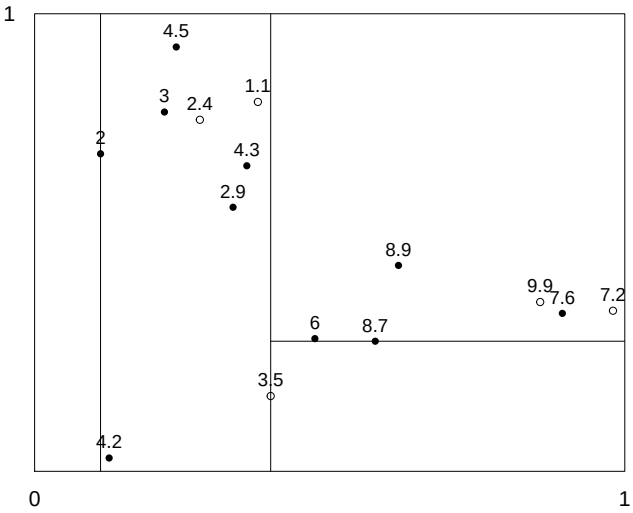


Figure 1.5 – Un arbre de survie de niveau 2. • : individus décédés. ○ : individus censurés.

1.4 Introduction aux essais cliniques de phase I

1.4.1 Contexte

Un essai clinique est une étude scientifique réalisée pour évaluer l'efficacité et la tolérance d'un traitement avant sa mise sur le marché. Ces études sont souvent effectuées après des études expérimentales non-cliniques (sur des modèles animaux ou cellulaires) pour confirmer leur pertinence et leur sécurité. En fonction du type d'étude et du stade du développement du médicament, les cliniciens enrôlent des volontaires sains ou des patients. Un essai clinique comporte généralement quatre phases.

— **Phase I** : elle évalue la toxicité et la tolérance de l'homme à un traitement. Parfois ces essais peuvent être proposés à des patients en impasse thérapeutique, pour lesquels le traitement étudié représente la seule chance de survie. Cette phase permet également d'étudier la cinétique et le métabolisme chez l'homme de la substance étudiée. Les groupes étudiés sont le plus souvent de petite taille (20 à 80 participants).

— **Phase II** : elle consiste à déterminer la dose optimale du médicament et ses éventuels effets indésirables. Elle est subdivisée en deux phases : les phases IIa et IIb. La phase IIa estime l'efficacité de la molécule sur un nombre limité (de 100 à 200) de malades, alors que la phase IIb détermine la dose thérapeutique de la molécule sur une plus grande échelle (de 100 à plus de 300 malades).

— **Phase III** : elle étudie l'efficacité comparative du traitement. Elle compare le traitement soit à un placebo, soit à un traitement de référence. Les groupes sont de taille importante, souvent plusieurs milliers de participants.

— **Phase IV** : elle s'intéresse au suivi du traitement à long terme alors qu'il est autorisé sur le marché. Elle doit permettre de dépister des effets secondaires rares ou des complications tardives.

Cette section, comme le Chapitre 5, est centrée sur la phase I de tels essais dans le cadre de traitements contre le cancer. Ces derniers se révèlent très toxiques, même à faible dose. C'est pourquoi on ne cherchera pas une dose à administrer non toxique, mais on s'autorisera une certaine probabilité de toxicité. Le fait que la toxicité recherchée soit non nulle implique que cette phase soit réalisée avec des patients malades, parfois en impasse curative, qui peuvent tirer des bénéfices de ce traitement. La partie suivante est consacrée à l'introduction des notations utiles à cette section et au Chapitre 5, ainsi qu'aux critères auxquels nous nous intéressons.

1.4.2 Notations et critères d'intérêts

Notons (i, j) la combinaison de la i-ème dose du traitement 1 avec la j-ème dose du traitement 2, où $i \in \{1, \cdots, I\}$ et $j \in \{1, \cdots, J\}$. On appelle D l'ensemble de toutes les combinaisons de doses possibles, c'est-à-dire $D = \{1, \cdots, I\} \times \{1, \cdots, J\}$. Nous nous intéressons ici à l'estimation de la racine de la fonction de régression représentant le lien dose / toxicité. Pour cela, nous avons accès à la suite d'observations $(X_n, Y_n)_{n\in\mathbb{N}}$. A l'étape n, *i.e.* quand le n-ème patient est enrôlé dans l'essai, la variable X_n représente la dose administrée à ce patient parmi les IJ doses testées ; et la variable Y_n représente la réponse de ce patient. Cette variable prend les valeurs 0 et 1 : 1 si on observe un signe de toxicité (Dose Limiting Toxicity, DLT) et 0 sinon). La loi conditionnelle de Y_n sachant X_n suit une loi de Bernoulli de paramètre $\beta_{(i,j)}$. Ceci peut s'écrire de la manière suivante.

$$\forall n \in \mathbb{N}, \quad \mathbb{P}(Y_n = 1 | X_n = (i, j)) = \beta_{(i,j)}.$$

Un seuil de toxicité α est fixé par les cliniciens. On souhaite ensuite trouver la dose ayant la probabilité de toxicité la plus proche de ce seuil, ce qu'on effectue en estimant la racine de la fonction de régression. On appelle cette dose recherchée la Dose Maximale Tolérée (MTD) et on la note d^*. On suppose que les gammes de doses $\{1, \cdots, I\}$ et $\{1, \cdots, J\}$ sont ordonnées en terme de probabilité de toxicité. Cela induit un ordre naturel sur l'ensemble D. Cet ordre est partiel et est défini sur l'ensemble des doses D par $(i, j) \leq (r, s)$ si et seulement si $i \leq r$ et $j \leq s$. On suppose que cet ordre partiel est respecté, c'est-à-dire

$$(i, j) < (r, s) \Rightarrow \beta_{(i,j)} < \beta_{(r,s)}. \tag{1.16}$$

On émet de plus l'hypothèse que tout l'échantillon $(X_1^n, Y_1^n) = ((X_1, Y_1), \cdots, (X_n, Y_n))$ peut être utilisé pour la recommendation de la dose X_{n+1}. Plus précisément, pour toutes les méthodes présentées dans cette section, X_{n+1} est complètement déterminée par l'historique (X_1^n, Y_1^n), c'est-à-dire l'estimateur associé à la méthode $\mathcal{M}$ satisfait $\mathcal{M}(X_1^n, Y_1^n) \in \sigma(X_1^n, Y_1^n)$, où $\sigma(X_1^n, Y_1^n)$ est la sigma-algèbre générée par l'échantillon (X_1^n, Y_1^n).

Dans les essais cliniques de Phase I, on s'intéresse en particulier à trois critères. Les deux premiers sont détaillés par Azriel et al. (2011). Le premier est appelé le principe de traitement : on souhaite recommender la MTD à un maximum de patients. Le second est appelé principe d'expérimentation : on veut obtenir le meilleur estimateur possible pour la MTD à la fin de l'étude. Malheureusement, on sait depuis Azriel et al. (2011) que si le principe de traitement est vérifié, alors le principe d'expérimentation ne peut pas l'être. Plus précisément, on ne peut pas avoir une convergence presque sûre vers la MTD. Il faut donc faire un compromis entre ces deux principes, mais idéalement on souhaite trouver la méthode qui sera la meilleure pour ces deux critères à la fois. Le dernier critère est appelé cohérence, et a été introduit Cheung (2005). Il peut s'expliquer de la manière suivante : si on observe une DLT chez le n-ème patient, alors on ne veut pas recommander une dose strictement supérieure à la sienne au patient $(n+1)$; de même, si on n'observe aucune DLT chez le n-ème patient, alors on ne veut pas recommander une dose strictement inférieure à la sienne au $(n+1)$-ème patient. Ce dernier critère est essentiel si l'on veut avoir confiance en un design dans le cadre des essais cliniques de Phase I.

Nous présentons dans la suite certains des designs existant de recherche de doses dans le cadre des phases I dans le cas d'ordre partiel. Nous commençons par des extensions de la CRM (O'Quigley et al., 1990) que nous rappelons. Nous détaillons ensuite deux designs paramétriques et enfin un design non paramétrique : PIPE (Mander and Sweeting, 2015).

1.4.3 Extensions de la CRM

De nombreux designs de recherche de doses, dans le cas d'une combinaison de deux traitements, découlent directement d'un design développé pour le cas d'un seul traitement, la CRM. Nous commençons par détailler cette méthode, puis étudions ses généralisations.

Présentation de la CRM

La CRM, ou Continual Reassessment Method, est introduit pour la première fois par O'Quigley et al. (1990). C'est un design de recherche de doses paramétrique dans le cas d'un ordre total sur les doses. Son principe est d'estimer à chaque étape n la MTD d'un traitement grâce aux informations de toutes les patients précédents. Détaillons ce modèle. Nous voulons trouver la MTD pour un unique traitement parmi les doses $\{d_1, ..., d_K\}$. Pour cela les cliniciens précise un seuil α. Il est généralement autour de $20 - 30\%$ et notre but est de trouver la dose d_0 associée à cette toxicité, ou possédant la toxicité la plus proche

possible de ce seuil. Nous considérons par la suite que $d_0 \in \{d_1, ..., d_K\}$ pour simplifier, mais il revient au même de chercher ν, la dose de $\{d_1, ..., d_K\}$ ayant la probabilité de toxicité la plus proche de α. En notant R la fonction représentant la vraie relation dose / toxicité, ceci revient à dire que $R(d_0) = \theta$. On choisit de modéliser cette relation par un modèle à un paramètre $\psi(., a)$.

On note X la variable aléatoire à valeurs dans $\{d_1, ..., d_K\}$ représentant la dose recommandée au patient, et Y une variable aléatoire à valeurs dans $\{0, 1\}$. $Y = 1$ si on observe un signe de toxicité et $Y = 0$ sinon. Alors

$$R(x) = \mathbb{E}[Y|X = x] = \mathrm{P}(Y = 1|X = x) \doteq \psi(x, a).$$

On note x_1, ..., x_n les doses recommandées aux n premiers patients et y_1, ..., y_n les réponses observées.
La vraisemblance du modèle après n patients est alors

$$L(a|\Omega_n) = \prod_{i=1}^{n} \psi(x_i, a)^{y_i}(1 - \psi(x_i, a))^{1-y_i}.$$

On dispose de plus d'un a priori f sur le paramètre a. D'après la formule de Bayes, on obtient

$$\hat{a}_n = \frac{\int_{\mathcal{A}} a f(a) L(a|\Omega_n) da}{\int_{\mathcal{A}} f(a) L(a|\Omega_n) da}.$$

On définit x_{n+1} par

$$x_{n+1} = \mathrm{argmin}|\psi(d_k, \hat{a}_n) - \theta|, \tag{1.17}$$

où l'argmin est pris sur l'ensemble $\{d_1, ..., d_K\}$. L'algorithme de la CRM est implémenté dans la librairie `dfcrm` du logiciel R.

Cette méthode est consistante sous des conditions de régularité sur la fonction ψ et pourvu que le vrai modèle sous-jacent aux données ne soit pas trop éloigné de la famille $(\psi(., a))_{a\in\mathcal{A}}$. Plus précisément, x_n converge presque sûrement vers d_0. Elle est de plus cohérente par construction. Pour plus de détails, on peut se référer à Shen and O'Quigley (1996); Cheung (2011).

Extensions adaptées à l'ordre partiel

CRM généralisée (Braun and Jia, 2013) Pour ce modèle, les auteurs proposent un modèle logit pour décrire le lien entre le traitement 1 et le traitement 2, c'est-à-dire

$$\mathrm{logit}\left(\beta_{(i,j)}\right) = a_j + b * d_i^1,$$

où d_i^1 représente la dose i du traitement 1. Le paramètre b est supposé suivre une loi gamma d'espérance μ_b et de variance σ_b^2, et le paramètre a_1 une loi gaussienne d'espérance μ_a et de variance σ_a^2. Les auteurs modélisent ensuite la différence entre les intercepts, c'est-à-dire entre deux doses du traitement 2, par une loi gaussienne d'espérance δ_j et de variance σ_a^2. De cette manière, la loi jointe des intercepts est une gaussienne multivariée. On obtient alors un modèle sur les $\beta_{(i,j)}$ dépendant des deux paramètre a et b. Ce modèle joue le rôle de ψ dans le cas de la CRM classique. On procède ensuite en suivant les étapes de la CRM : ré-estimation bayésienne de a et b, puis minimisation de (1.17).

Design à deux dimensions (Wang and Ivanova, 2005) Les auteurs proposent ici un modèle à 3 paramètres :

$$\psi((i,j),a,b,\gamma) = 1 - \left(1-d_i^1\right)^a \left(1-d_j^2\right)^{b+\gamma\log(1-d_i^1)},$$

où $0 \leq d_1^1 < \cdots < d_I^1 < 1$ et $0 \leq d_1^2 < \cdots < d_J^2 < 1$ sont des constantes remplaçant respectivement les doses 1 à I du traitement 1 et les doses 1 à J du traitement 2. Le paramètre γ définit ici la dépendance entre les deux traitements. Si les deux traitements sont indépendants, le modèle est réduit à

$$\psi((i,j),a,b,\gamma) = 1 - \left(1-d_i^1\right)^a \left(1-d_j^2\right)^b.$$

On procède ensuite comme pour la CRM classique. L'Equation (1.17) est résolue en déterminant d'abord toutes les MTD i_j^* optimales pour le traitement 1 à j fixé (à une dose pour le traitement 2 fixé). Puis la combinaison optimale $(i_{j^*}^*, j^*)$ est choisie comme la combinaison ayant la toxicité la plus proche du seuil α parmi les combinaisons $(i_j^*, j)_{j\in\{1,\ldots,J\}}$.

po-CRM (Wages et al., 2011) Dans l'ensemble D de toutes les combinaisons possibles, nous savons qu'il existe un ordre partiel. On peut donc considérer tous les ordres totaux possibles satisfaisant la relation d'ordre partiel (1.16). On sélectionne alors certains ordres parmi les M ordres possibles entre les IJ doses. Disons qu'on choisit les ordres $\{1,...,M_0\}$ (quitte à renuméroter les ordres). Pour chaque ordre $m \in \{1,...,M_0\}$, on modélise la probabilité de toxicité par $R(i,j)$ via

$$R(i,j) = \mathbb{P}(Y=1|X=(i,j)) = \mathbb{E}[Y|(i,j)] \doteq \psi_m((i,j),a),$$

avec un modèle dose / toxicité $\psi_m((i,j),a)$ et $a \in \mathcal{A}$. Après l'inclusion du n-ème patient dans l'essai, on obtient des données sous la forme $\Omega_n = \{x_1, y_1, ..., x_n, y_n\}$. La vraisemblance sous le modèle m est alors

$$L_m(a|\Omega_n) = \prod_{k=1}^{n} \psi_m^{y_k}(x_k,a)(1-\psi_m(x_k,a))^{1-y_k}.$$

La CRM à ordre partiel (poCRM) se définit par les étapes suivantes. Quand on enrôle un n-ème patient, on choisit un ordre particulier h et je note m l'ordre précédemment utilisé. L'ordre h est choisi comme étant le plus vraisemblable parmi les M_0 ordres possibles. On trouve ensuite un estimateur $\hat{a}_h$ qui est l'estimateur du maximum de vraisemblance sous le modèle h. Enfin on choisit la dose x_{n+1} en minimisant

$$|\psi_h((i,j),\hat{a}_h) - \theta|.$$

On peut remarquer la spécificité de ce modèle : la dépendance entre le traitement 1 et le traitement 2 n'est pas modélisée.

1.4.4 Autres designs

Nous présentons maintenant des designs de recherche de doses adaptés au cas de l'ordre partiel et ne provenant pas d'une généralisation de la CRM. Le premier (Yin and Yuan, 2009b) est une méthode paramétrique qui repose sur la modélisation du copule représentant la dépendance entre le traitement 1 et le traitement 2. Le second (Mander and Sweeting, 2015) est un des rares modèles non paramétriques.

Paramétrisation de copules

Yin and Yuan (2009b) supposent ici que les toxicités marginales des traitements 1 et 2 sont connues, c'est-à-dire que l'on considère comme fixes les probabilités de toxicité $p_1, \ldots, p_I$ du traitement 1 et les probabilités de toxicité $q_1, \ldots, q_J$ du traitement 2. Pour prendre en compte l'incertitude dans le calcul de ces probabilités, les auteurs introduisent les paramètres a et b inconnus d'a priori centré en 1. On ne considère alors plus les (p_i, q_j), mais (p_i^a, q_j^b).

En se basant sur le modèle de copule de Clayton, on peut proposer le modèle de régression de type copule suivant pour la probabilité jointe $\beta_{(i,j)}$

$$\beta_{(i,j)} = 1 - \left\{ (1 - p_i^a)^{-\gamma} + \left(1 - q_j^b\right)^{-\gamma} - 1 \right\}^{-1/\gamma},$$

où le paramètre γ représente la dépendance entre le traitement 1 et le traitement 2. Pour plus de simplicité, les a priori sur les paramètres a, b et γ sont prises indépendantes. On peut choisir par exemple pour les paramètre a et b une loi Gamma de paramètres 2 et 2. Ces paramètres sont alors centrés en 1 et de faible variance. La loi sur γ, quant à elle, doit être moins informative. On peut encore choisir une loi Gamma, mais cette fois avec grande variance pour que l'information vienne essentiellement des données. Les paramètres sont ensuite ré-évalués par la formule de Bayes, puis on choisit la combinaison (i, j) minimisant la distance $|\beta_{(i,j)} - \alpha|$. La différence avec la CRM se situe ici dans la connaissance que l'on a a priori sur les probabilités de toxicité marginales.

Méthode non paramétrique : PIPE

L'objectif de Mander and Sweeting (2015) est différent de celui de toutes les méthodes citées précédemment. En effet, les auteurs ne s'intéressent pas à déterminer une unique MTD, mais plutôt un contour maximal toléré (MTC) qui se définit comme la frontière entre les doses de toxicité inférieure au seuil α et les doses de toxicité supérieure à α. Il est effectivement raisonnable, dans le cas d'une combinaison de traitements, de chercher à exhiber plusieurs combinaisons de doses à toxicités proches afin de laisser aux cliniciens la possibilité de choisir parmi celles-ci, par exemple une combinaison avec une toxicité un peu moins proche du seuil mais plus efficace.

Quant au modèle proprement dit, Mander and Sweeting (2015) propose un produit de lois Beta indépendantes :

$$\beta_{(i,j)} | a_{(i,j)}, b_{(i,j)} \sim Beta(a_{(i,j)}, b_{(i,j)}), \forall i \in \{1, \ldots, I\},\ j \in \{1, \ldots, J\},$$

où $a_{(i,j)}$ et $b_{(i,j)}$ sont des paramètres fixés. Supposons maintenant, qu'après le n-ème patient, on ait observé $n^1_{(i,j)}$ DLTs à la dose (i, j) sur les $n_{(i,j)}$ patients traités à cette dose. Alors, par conjugaison, la loi a posteriori de $\beta_{(i,j)}$ est la loi Beta

$$\beta_{(i,j)} | \mathcal{D}_n \sim Beta\left(a_{(i,j)} + n^1_{(i,j)}, b_{(i,j)} + n_{(i,j)} - n^1_{(i,j)}\right),$$

où $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ est l'historique de l'essai jusqu'au patient n. On peut alors déterminer la probabilité que la MTC soit égale à contour particulier $\mathcal{C}$. Pour cela, on représente le contour $\mathcal{C}$ par une matrice C dont les coefficients sous le contour sont nuls et les autres égaux à 1. On obtient

$$\mathbb{P}\left(MTC = \mathcal{C} | \mathcal{D}_n\right) = \prod_{(i,j)} \left\{1 - \mathbb{P}\left(\beta_{(i,j)} \leq \alpha | \mathcal{D}_n\right)\right\}^{C[i,j]} \mathbb{P}\left(\beta_{(i,j)} \leq \alpha | \mathcal{D}_n\right)^{1 - C[i,j]},$$

où $C[i,j]$ est l'élément de la matrice C à la ligne i et la colonne j. On choisit alors le MTC $\mathcal{C}^*$ tel que

$$\mathcal{C}^* = \arg\max_{\mathcal{C}} \mathbb{P}\left(MTC = \mathcal{C}|\mathcal{D}_n\right).$$

Chapitre 2

Modèle à hasards non proportionnels et Survie marginale

2.1 Introduction

Avant les travaux sur le modèle à hasards proportionnels de Cox (Cox, 1972), les modèles paramétriques remportaient un franc succès pour effectuer des régressions sur des données de survie censurées. En effet, même un simple modèle exponentiel peut se montrer fiable et valable dans une gamme assez large de situations. Il l'est par exemple dans les cas où l'hypothèse de perte de mémoire est vérifiée ou est une bonne approximation. Néanmoins ces modèles paramétriques ont rapidement été presque totalement éclipsés par l'arrivée du modèle de Cox. Les avantages de la régression de Cox se situent à au moins deux niveaux. D'une part, l'utilisateur n'est pas obligé d'imposer une forme à la fonction de hasard initiale, *i.e.* la loi conditionnelle du temps de décès, T, par rapport à un vecteur de covariable Z est non paramétrique. D'autre part, le modèle de Cox s'étend immédiatement à des situations plus complexes comme le cas de covariables dépendantes du temps, les processus multi-états et les événements récurrents. L'inférence est aussi invariante par transformations croissantes en T et son efficacité reste élevée. En d'autres termes, on perd assez peu d'information en utilisant ce modèle. L'inférence sur les paramètres de risque relatif en est facilitée.

Supposons cependant que l'on souhaite inclure dans notre analyse la connaissance de la fonction de survie marginale de T, notée S. Il peut exister de nombreux cas pour lesquels on possède de l'information sur la loi marginale de T, avant d'avoir effectué une analyse de régression de T sur Z. On peut avoir un estimateur précis de S, en utilisant par exemple des données de registre, comme dans un contexte de survie relative. On peut vouloir se calibrer sur une autre étude dans laquelle, par hypothèse, le mécanisme gouvernant la génération de la variable aléatoire T est le même. Enfin, on peut décider d'ajuster un modèle marginal à la loi de T. Les paramètres inconnus de la loi marginale sont alors remplacés par des estimateurs basés sur l'échantillon ou un échantillon indépendant.

On se demande dans ce chapitre dans quelle mesure cette connaissance sur la fonction de survie peut être transférée à la régression. A la lumière de précédents travaux concernant l'efficacité des estimateurs de vraisemblance partielle, nous n'anticipons pas d'améliorations significatives (Struthers and Kalbfleisch, 1986; Lin, 1991; Xu and O'Quigley, 2000). Dans le cas du modèle de Cox, un résultat important est que l'on peut mal spécifier le modèle pour S, ne pas modéliser correctement la loi de T, et toujours estimer de manière consistante les paramètres de risque relatif. Dans un tel cas, il y a une perte d'efficacité mais il se trouve qu'elle est assez faible. Mais supposons maintenant qu'à la

place d'un log-risque relatif constant, β_0, comme supposé dans le modèle à hasards proportionnels, les observations sont générées par une fonction $t \mapsto \beta(t)$ qui varie en fonction du temps. L'estimateur de la vraisemblance partielle va alors converger vers une quantité qui dépend fortement du mécanisme de censure, qui est inconnu, et cela même si ce dernier est indépendant du temps de décès et des covariables (Struthers and Kalbfleisch, 1986). Cette forte dépendance a été montrée par un grand nombre d'auteurs. En situations de hasards non-proportionnels, dans les cas particuliers d'un changement lisse dans le temps ; par exemple des effets diminuant de manière continue avec le temps, on aimerait savoir si on est capable d'estimer $E[\beta_0(T)]$ en utilisant nos connaissances sur la survie marginale. En fait, c'est tout à fait possible en conditionnant sur l'estimateur de la survie marginale observée de F.

Le modèle à hasards proportionnels de Cox nous permet de faire de l'inférence sur les coefficients de régression du modèle de risque relatif en gardant non-spécifié le taux de hasard de base, λ_0. Un modèle à hasards non-proportionnels, pour lequel le modèle de Cox serait un cas particulier, peut s'écrire,

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_0(t)Z(t)\},$$

et, dans tous les cas, le taux de hasard de base λ_0 peut être interprété comme le hasard $\lambda(t|Z = 0)$. Ce modèle a été étudié par Moreau et al. (1985); O'Quigley and Pessione (1989, 1991); Liang et al. (1990); Zucker and Karr (1990); Murphy and Sen (1991); Gray (1992); Hastie and Tibshirani (1993); Verweij and Houwelingen (1995); Lausen and Schumacher (1996); Marzec and Marzec (1997), notamment. Le principal point d'intérêt est d'estimer l'effet de régression β_0 en tant que fonction de t. Estimer β_0 représente un défi important puisque, en règle générale, β_0 est de dimension infinie. Un objectif moins ambitieux mais plus facilement atteignable est d'estimer un effet moyen $E[\beta(T)]$. Il se trouve qu'une solution à ce problème est de conditionner par rapport à la survie marginale. L'estimation d'un effet moyen peut être utilisée en étude préliminaire d'un jeu de données dépendantes du temps. On pourrait penser de prime abord que l'estimateur du maximum de vraisemblance partielle dans les cas où β_0 varie avec t, correspond à une moyenne par rapport à la variable T. Cependant la dépendance par rapport au mécanisme de censure rend cette assertion fausse (Xu and O'Quigley, 2000).

En Section 2.2.3 et 2.2.4, nous déduisons de nos connaissances sur la survie marginale de T deux estimateurs du paramètre de régression β_0, l'un dans le cas fréquentiste, l'autre dans le cas bayésien. Quand l'hypothèse de hasards proportionnels est vérifiée, ces estimateurs sont consistants pour β_0. Nous étudions aussi les propriétés asymptotiques dans le cas de modèles à hasards non proportionnels. La Section 2.2.5 présente un résultat asymptotique pour l'estimation de la survie conditionnellement aux covariables. Des simulations sont fournies en Section 2.3. L'efficacité relative de l'estimateur fréquentiste par rapport à l'estimateur de la vraisemblance partielle est étudiée en Section 2.4. Les Sections 2.5, 2.6 et 2.7 présentent des applications des méthodes d'estimation introduites sur des jeux de données réels.

2.2 Propriétés asymptotiques

2.2.1 Notations

Dans ce Chapitre 2, je noterai T la variable aléatoire de fonction de répartition F et de fonction de survie $S = 1 - F$ représentant le temps de décès. C désigne le temps de censure et $Z(.) \in \mathbb{R}^d$ un vecteur de covariables. On considérera que C est indépendant de T conditionnellement à $Z(.)$. Je suppose qu'il existe $\tau > 0$ tel que T et C ont pour support le segment $[0, \tau]$. Je supposerai de plus que T et Z suivent le modèle à risques non-proportionnels

$$\lambda(t|Z) = \lambda_0(t) \exp\{\beta_0(t)Z\}, \tag{2.1}$$

où $\lambda_0(t)$ le risque instantané de base. Il est défini pour tout $t \in [0, \tau]$ par la formule

$$\lambda_0(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq T < t + h | T \leq t)}{h} = -\ln(S)'(t).$$

On notera $(T_i, C_i, Z_i(.))$ avec $i \in \{1, \cdots, n\}$ une suite de variables aléatoires i.i.d. de même loi que $(T, C, Z(.))$.

Définissons pour tout $i \in \{1, \cdots, n\}$, $X_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{T_i \leq C_i}$. X_i est alors le temps observé de l'individu i et Δ_i est le statut associé à cet individu : "décédé" $(= 1)$ ou "censuré" $(= 0)$. La covariable du patient i est observée jusqu'au temps X_i, mais nous étendons sa définition sur $[0, \tau]$ en notant $Z_i(s) = Z_i(X_i)$ pour tout $s \in [X_i, \tau]$. Nous définissons ensuite $\mathcal{D}$ l'ensemble des fonctions discontinues de $[0, \tau]$ dans $\mathbb{R}^d$, et supposons que $\beta_0 \in \mathcal{D}$. Pour tout $i \in \{1, \cdots, n\}$ et tout $t \in [0, \tau]$, on notera $Y_i(t) = \mathbb{1}_{X_i \geq t}$. $Y_i(t)$ indique si l'individu i est encore à risque au temps t $(= 1)$, ou s'il ne l'est plus $(= 0)$. Enfin, je note F_n la distribution empirique de T, *i.e.* $F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{T_i \leq t}$, pour tout $t \in [0, \tau]$.

Après ces quelques notations, la partie suivante rappellera quelques résultats sur les équations estimatrices.

2.2.2 Equations estimatrices

Les notations utilisées ici sont essentiellement celles introduites par Andersen and Gill (1982).

Pour tout $r \in \{0, 1, 2\}$ et $\beta \in \mathcal{D}$, définissons

$$S^{(r)}(\beta, t) = \frac{1}{n} \sum_{j=1}^n Y_j(t) \exp\left\{\beta(t)' Z_j(t)\right\} Z_j(t)^{\otimes r}, \tag{2.2}$$

où, pour un vecteur colonne v, $v^{\otimes 2}$ représente la matrice vv', $v^{\otimes 1}$ le vecteur v et $v^{\otimes 0}$ le scalaire 1. On peut alors écrire la vraisemblance partielle (Cox, 1972) sous la forme suivante

$$l(\beta) = \sum_{i=1}^n \Delta_i \left[\beta' Z_i(X_i) - \log\left\{S^{(0)}(\beta, X_i)\right\}\right].$$

et on définit

$$E(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}, \quad V(\beta, t) = \frac{S^{(2)}(\beta, t)}{S^{(0)}(\beta, t)} - E(\beta, t)^{\otimes 2}.$$

On note maintenant le score U la fonction du paramètre β (appartenant à $\mathcal{D}$) définie par

$$U(\beta) = \sum_{i=1}^n \Delta_i \left\{Z_i(X_i) - E(\beta, X_i)\right\} = \int \left\{\mathcal{Z}(t) - E(\beta, t)\right\} d\bar{N}(t), \tag{2.3}$$

où $\bar{N}(t) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t,\ \Delta_i = 1\}}$ et $\mathcal{Z}(.)$ est une fonction en escaliers, continue à gauche, discontinue aux points X_i où elle prend la valeur $Z_i(X_i)$. On peut remarquer qu'en absence de censure,

$$U(\beta) = \int \{\mathcal{Z}(t) - E(\beta, t)\}\, d\bar{N}(t) = \int \{\mathcal{Z}(t) - E(\beta, t)\}\, dF_n(t). \tag{2.4}$$

Pour maximiser la vraisemblance partielle du modèle (2.1) sous l'hypothèse que la fonction β_0 est constante, on résout l'équation estimatrice $U(\beta) = 0$. On obtient ainsi un estimateur réel $\hat{\beta}_{PL}$ de la fonction β_0. Si β_0 est effectivement une fonction constante, *i.e.* $\beta_0(t) = \beta_0$ pour tout $t \in [0, \tau]$, Andersen and Gill (1982) ont montré que $\sqrt{n}(\hat{\beta}_{PL} - \beta_0)$ est asymptotiquement normal d'espérance nulle et de variance pouvant être estimée de manière consistante par $\mathcal{I}^{-1}(\hat{\beta}_{PL})$, où $\mathcal{I}(\hat{\beta}_{PL}) = n^{-1} \sum_{i=1}^{n} \Delta_i V(\hat{\beta}_{PL}, X_i)$.

2.2.3 Inférence fréquentiste

Si l'hypothèse β_0 constante dans le modèle 2.1, c'est-à-dire l'hypothèse de hasards proportionnels, n'est pas vérifiée, nous aimerions trouver un nombre réel qui caractérise la fonction β_0 dans son ensemble et étudier comment estimer cette quantité. Notons que dans la fonction score (2.3), tous les termes de la somme sont affectés du même poids. Un choix naturel est alors de leur assigner des poids différents. C'est ce qui est fait par exemple dans l'utilisation du test du log-rank pondéré (Gehan, 1965).

L'idée est de remplacer F_n dans (2.4) par $\tilde{F}_n$ qui est un estimateur consistant de F. Nous définissons alors la fonction score pondérée suivante

$$U_W(\beta) = \sum_{i=1}^{n} \Delta_i W(X_i) \{Z_i(X_i) - E(\beta, X_i)\}, \tag{2.5}$$

où $W(.)$ est un processus stochastique réel $(\mathcal{F}_t)_{t>0}$ - prévisible avec

$$\mathcal{F}_t = \sigma((X_i, \Delta_i, Z_i(s));\ i : X_i \leq s,\ 0 \leq s \leq t) \text{ pour } t \in [0, \tau].$$

En d'autres termes, $(\mathcal{F}_t)_{t>0}$ est la filtration incluant toute l'information avant le temps t. Avec un choix spécifique de W, nous voulons trouver

$$U_W(\beta) = \int \{\mathcal{Z}(t) - E(\beta, t)\}\, d\tilde{F}(t). \tag{2.6}$$

Notons $\hat{\beta}_W$ la solution de l'équation $U_W(\beta) = 0$ pour une fonction de pondération W quelconque. Nous constatons qu'en choisissant W constant égal à 1, alors on retrouve la fonction score (2.3) définie par la vraisemblance partielle usuelle.

Pour étudier le comportement asymptotique de $\hat{\beta}_W$ pour n'importe quelle fonction de pondération W, on considère les hypothèses ci-dessous. Les quatre premières ont été introduites par Andersen and Gill (1982), la dernière dans Xu and O'Quigley (2000) ;

Hypothèses. *(A) (Intervalle fini)* $\int_0^\tau \lambda_0(t)dt < \infty$.

(B) (Stabilité asymptotique) Il existe un voisinage $\mathcal{B}$ *de* β_0 *dans* $\mathcal{D}([0, \tau])$ *tel que la fonction nulle et* β_0 *sont à l'intérieur de* $\mathcal{B}$*, et pour tout* $r \in \{0, 1, 2\}$ *il existe des fonctions* $s^{(r)}(\beta, t)$ *définies sur* $\mathcal{B} \times [0, \tau]$ *telles que*

$$\sup_{\beta \in \mathcal{B}, t \in [0,\tau]} \|S^{(r)}(\beta, t) - s^{(r)}(\beta, t)\| \xrightarrow[n \to \infty]{P} 0.$$

(C) (Régularité asymptotique) Pour tout $r \in \{0,1,2\}$, *les fonctions* $s^{(r)}(\beta,t)$ *sont uniformément continues en* $t \in [0,\tau]$, *continues en* $\beta \in \mathcal{B}$ *et bornées sur* $\mathcal{B} \times [0,\tau]$; $s^{(0)}(\beta,t)$ *est bornée et sa valeur absolue est minorée par une constante strictement positive.*

(D) (Stabilité asymptotique de W*) Il existe une fonction positive bornée* w *définie sur* $[0,\tau]$ *telle que*

$$\sup_{t\in[0,\tau]} |nW(t) - w(t)| \xrightarrow[n\to\infty]{P} 0.$$

(E) (Homoscédasticité) Si on pose

$$\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)} = E[Z(t)|T=t],$$

et

$$v(\beta_0,t) = \frac{\partial}{\partial\beta}\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\bigg|_{\beta=\beta_0} = Var[Z(t)|T=t],$$

on suppose que $v(\beta_0,t)$ *est constant par rapport au temps.*

Introduisons quelques notations supplémentaires : on définit β_w^* comme l'unique solution de l'équation

$$h_w(\beta) = \int_0^\tau w(t)\left\{\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right\} s^{(0)}(\beta_0,t)dt = 0, \tag{2.7}$$

et $A_w(\beta)$ par

$$A_w(\beta) = \int_0^\tau w(t)\left\{\frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} - \left(\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)}\right)^{\otimes 2}\right\} s^{(0)}(\beta_0,t)dt.$$

On peut trouver dans Lin (1991) la propriété suivante

Propriété 2.1. *Sous le modèle (2.1) de paramètre* β_0 *et les hypothèses* (A), (B), (C) *et* (D), $\hat{\beta}_W \xrightarrow[n\to\infty]{P} \beta_w^*$ *si* $A_w(\beta_w^*)$ *est définie positive.*

On peut remarquer en particulier qu'en choisissant $w(t) = 1$, on retrouve la convergence de $\hat{\beta}_{PL}$ vers la solution de l'équation

$$\int_0^\tau \left\{\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right\} s^{(0)}(\beta_0,t)\lambda_0(t)dt = 0. \tag{2.8}$$

En général, cette solution $\hat{\beta}_{PL}$ dépend de la censure à travers le terme $s^{(0)}(\beta_0,t)$ (Struthers and Kalbfleisch, 1986). C'est pourquoi les résultats d'estimations provenant de l'équation (2.5) doivent être lus avec attention et leur interprétation n'est pas immédiate en général. On comprend alors que le choix de la fonction de pondération W a son importance.

Rappelons que l'on note S la fonction de survie du temps de décès T. Nous noterons de plus $\hat{S}$ la version continue à gauche de l'estimateur de Kaplan and Meier (1958) de S. $\hat{S}$ est un estimateur consistant de S (Xu and O'Quigley, 2000). Comme expliqué dans la Section 2.1, l'objectif est de bien choisir la fonction de pondération W. Commençons par

regarder le cas connu de l'estimateur de Kaplan-Meier. Plus précisément, prenons pour tout $t \in [0, \tau]$,

$$W_{KM}(t) = \frac{\hat{S}(t)}{\sum_{i=1}^{n} Y_i(t)} = \frac{\hat{S}(t)}{nS^{(0)}(0,t)}.$$

On peut alors écrire $\hat{S}$ sous la forme $\hat{S}(t) = \sum_{i:X_i \leq t} \delta_i W_{KM}(X_i)$, pour tout $t \in [0, \tau]$. Cette écriture montre que les poids $\delta_i W_{KM}(X_i)$ sont les incréments de la fonction $\hat{S}$. On sait de plus que nW_{KM} converge uniformément vers w_{KM} où $w_{KM}(t) = S(t)/s^{(0)}(0,t)$, pour tout $t \in [0, \tau]$.

Revenons maintenant à la Propriété 2.1. Grâce à celle-ci, on obtient le résultat suivant, dû à Xu and O'Quigley (2000),

Proposition 2.2. *Sous le modèle (2.1) de paramètre β_0 et sous les hypothèses (A), (B) et (C), si $A_{w_{KM}}(\beta^*)$ est définie positive, on a $\hat{\beta}_{KM} \xrightarrow[n\to\infty]{P} \beta^*_{w_{KM}}$.*

Pour plus de simplicité, je noterai dans la suite du chapitre $\beta^* = \beta^*_{KM}$. β^* est alors par définition l'unique solution de l'équation $h_{w_{KM}}(\gamma) = 0$ *i.e.* de l'équation

$$\int_0^\tau \left\{ \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \right\} dF(t) = 0. \tag{2.9}$$

On remarque que contrairement à l'équation (2.8), la censure n'intervient pas dans l'équation (2.9). C'est pourquoi sa solution β^* est elle aussi indépendante de la censure.

L'indépendance de β^* est une première caractéristique intéressante. Montrons maintenant que β^*, solution de (2.9), peut être vu comme l'effet moyen de $\beta_0(.)$ sous certaines conditions. Commençons par appliquer une approximation de Taylor à l'équation (2.9). On obtient

$$\int_0^\tau \mathrm{Var}[Z(t)|T = t]\{\beta^* - \beta_0(t)\}dF(t) \approx 0.$$

On peut directement écrire cette équation en fonction de β^* de la manière suivante

$$\beta^* = \frac{\int_0^\tau \mathrm{Var}[Z(t)|T = t]\beta_0(t)dF(t)}{\int_0^\tau \mathrm{Var}[Z(t)|T = t]dF(t)} = \frac{\int_0^\tau v(t)\beta_0(t)dF(t)}{\int_0^\tau v(t)dF(t)}.$$

β^* peut alors être vu comme un effet moyen pondéré du coefficient de régression β_0. Pour être plus précis, β^* est l'effet moyen de β_0 pondéré par des poids proportionnels à $v(.)$. Maintenant, supposons que v varie peu dans le temps. Alors on obtient

Propriété 2.3. *Sous l'hypothèse (E),*

$$\beta^* \approx \frac{\int_0^\tau \beta_0(t)dF(t)}{F(\tau)} = \int_0^\tau \beta_0(t)dF(t) = E[\beta_0(T)]. \tag{2.10}$$

On rappelle que $F(\tau) = 1$ car $[0, \tau]$ est le support de T. L'hypothèse (E) dans la Proposition 2.10 est suffisante mais pas nécessaire. En effet, dans le cas des modèles introduits par Harrington and Fleming (1982), l'équation (2.10) est exacte. Pour rentrer un peu plus dans les détails, considérons la relation entre β^* et un vecteur α mesurant la différence de groupe dans des modèles de transformation de k-échantillon quand l'erreur aléatoire appartient à la famille G^ρ de Harrington and Fleming (1982). On rappelle qu'un tel modèle de transformation peut s'écrire $g(T) = \alpha' Z + \varepsilon$ où g est une fonction strictement croissante, Z un vecteur à valeurs dans $\{0, 1\}$ et ε, l'erreur aléatoire, appartient à la famille G^ρ, *i.e.* admet une fonction de survie définie par

$$\begin{cases} H_0(t) = \exp(-e^t) & (\rho = 0) \\ H_\rho(t) = (1 + \rho e^t)^{1/\rho} & (\rho > 0) \end{cases}$$

Le modèle à hasards non-proportionnels et le modèle de transformation ci-dessus sont équivalents pour $g(t) = \log \Lambda_0(t)$ où Λ_0 est la fonction de hasard cumulative de base et $\alpha = \beta_0(0)$. Xu and Harrington (2001) ont montré que, dans le cas d'un 2-échantillon et avec la paramétrisation suivante :

— $\rho = 1$,

— $g = \log$,

— $P(Z = 1) = 1 - P(Z = 0) = 1/2$,

on a

$$E[\beta_0(T)] = -\frac{\alpha}{2} = \beta^*,$$

où la première égalité est un calcul explicite et la seconde utilise le résultat

$$E[Z|T = t] = \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)}.$$

Finalement, la Proposition 2.10 ou encore les modèles d'Harrington et Fleming nous autorisent à considérer β^*, en première approximation, comme un effet moyen du coefficient de régression fonctionnel β_0.

Revenons sur les propriétés de l'estimateur $\hat{\beta}_{KM}$. Le travail de simulation de Xu and O'Quigley (2000) montre que cet estimateur présente en général de meilleures performances que les estimateurs $\hat{\beta}_{PL}$ et $\hat{\beta}_W$ en terme d'estimation. De plus, le fait que cet estimateur ne dépende pas de la censure rend l'interprétation de β^* plus claire et donc incite à son utilisation en pratique. Cependant les sauts de l'estimateur de Kaplan-Meier $\hat{\beta}_{KM}$ sont grands à la fin de la collection du jeu de données. Ceci est dû à la diminution de la taille de l'ensemble des individus à risque au cours du temps. De cette manière, on met plus de poids sur les observations tardives dans l'équation estimatrice. De plus, ces derniers incréments sont très variables (Stute and Wang, 1993). Pour éviter ces problèmes, une idée est d'utiliser une fonction de pondération W lisse dans l'équation estimatrice (2.5). Ceci peut être effectué par exemple en prenant en compte un modèle paramétrique pour T. C'est ce qui est détaillé ci-après.

Définissons d'abord le modèle paramétrique $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$ où $\Theta \subset \mathbb{R}^p$ et P_θ est une fonction de répartition continue pour tout θ. Supposons que T suit la loi P_{θ_0}, $\theta_0 \in \Theta$. Alors on note $S = S(.; \theta_0)$ sa fonction de survie. Elle est donc continue par rapport à $t \in [0, \tau]$. Nous allons supposer de plus qu'elle est uniformément continue par rapport au couple $(t; \theta) \in [0, \tau] \times \Theta$. On note $\hat{\theta}_n$ un estimateur consistant de θ_0, *i.e.* on a la convergence

$$\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_0. \tag{2.11}$$

On note T_n une variable aléatoire de loi $P_{\hat{\theta}_n}$ et $S(.; \hat{\theta}_n)$ sa fonction de survie. Nous définissons alors une fonction de pondération d'après le modèle paramétrique

$$W_{par}(t) = \frac{S(t; \hat{\theta}_n)}{nS^{(0)}(0, t)},$$

où *par* indique la dépendance au modèle paramétrique. On obtient alors le résultat suivant.

Théorème 2.4. *Sous le modèle (2.1) de paramètre β_0 et sous les hypothèses (A), (B) et (C), si $A_w(\beta^*)$ est définie positive, on a $\tilde{\beta} \xrightarrow[n\to\infty]{P} \beta^*$, où $\tilde{\beta}$ est la solution de l'équation $U_{W_{par}}(\beta) = 0$.*

Preuve. Il suffit de montrer que W_{par} satisfait l'hypothèse (D) pour obtenir le résultat en utilisant la Propriété 2.1.

Comme w est continue, elle est uniformément continue sur $[0,\tau]$. Ensuite nous allons montrer que

$$\sup_{t\in[0,\tau]} |nW_{\text{par}}(t) - w(t)| \xrightarrow[n\to\infty]{P} 0,$$

S est uniformément continue par rapport à $(t;\theta)$, c'est-à-dire que pour tout $\varepsilon_0 > 0$, il existe $\delta > 0$ tel que pour tout $(t_1,\theta_1),(t_2,\theta_2) \in [0,\tau]\times\Theta$,

$$\|(t_1,\theta_1) - (t_2,\theta_2)\| \leq \delta \Longrightarrow |S(t_1,\theta_1) - S(t_2,\theta_2)| \leq \varepsilon_0.$$

En appliquant l'uniforme continuité pour $t_1 = t_2$ avec $\|.\| = \|.\|_1$, on obtient que pour tout $t \in [0,\tau]$ et tout $(\theta_1,\theta_2) \in \Theta^2$,

$$|\theta_1 - \theta_2| \leq \delta \Longrightarrow |S(t,\theta_1) - S(t,\theta_2)| \leq \varepsilon_0. \tag{2.12}$$

Soit maintenant $\varepsilon > 0$ et $\varepsilon_0 > 0$. D'après l'équation (2.11), il existe $N \in \mathbb{N}$ tel que pour tout $n \geq N$, $P(|\hat{\theta}_n - \theta_0| \leq \delta) \geq 1-\varepsilon$. Donc, pour tout $n \geq N$,

$$\begin{aligned} P(\sup_t |S(t;\hat{\theta}_n) - S(t;\theta_0)| \leq \varepsilon_0) &= P(\forall t,\ |S(t;\hat{\theta}_n) - S(t;\theta_0)| \leq \varepsilon_0) \\ &\geq P(|\hat{\theta}_n - \theta_0| \leq \delta) \\ &\geq 1-\varepsilon, \end{aligned}$$

où la première inégalité est due à l'équation (2.12). On a donc

$$\sup_{t\in[0,\tau]} |S(t;\hat{\theta}_n) - S(t;\theta_0)| \xrightarrow[n\to\infty]{P} 0. \tag{2.13}$$

De plus, l'hypothèse (B) implique que

$$\sup_{t\in[0,\tau]} |S^{(0)}(0,t) - s^{(0)}(0,t)| \xrightarrow[n\to\infty]{P} 0. \tag{2.14}$$

Revenons à W_{par}.

$$\begin{aligned} \sup_t |nW_{par}(t) - w(t)| &= \sup_t \left| \frac{S(t;\hat{\theta}_n)}{S^{(0)}(0,t)} - \frac{S(t;\theta_0)}{s^{(0)}(0,t)} \right| \\ &\leq \sup_t \left| \frac{1}{S^{(0)}(0,t)} \right| \sup_t |S(t;\hat{\theta}_n) - S(t;\theta_0)| \\ &\quad + \sup_t |S(t;\theta_0)| \sup_t \left| \frac{1}{S^{(0)}(0,t)} - \frac{1}{s^{(0)}(0,t)} \right|. \end{aligned}$$

$\sup_t \left|\frac{1}{S^{(0)}(0,t)}\right|$ est borné à partir d'un certain rang d'après la condition de bornitude (C) de $s^{(0)}(0,t)$ et l'équation (2.14). $\sup_t |S(t;\theta_0)|$ est borné car $S(.\theta_0)$ est continue sur le compact $[0,\tau]$.

$$\sup_t \left| \frac{1}{S^{(0)}(0,t)} - \frac{1}{s^{(0)}(0,t)} \right| \leq \varepsilon,$$

d'après l'hypothèse (C), l'équation (2.14) et le fait que la fonction inverse est continue sur tout intervalle de la forme $[a;+\infty[$. En utilisant l'équation (2.13), on obtient finalement que W_{par} satisfait (D) et on peut utiliser la Propriété 2.1, ce qui clôt la preuve du Théorème 2.4. □

Dans l'équation (2.5), nous avons estimé β^* avec des poids différents de W_{KM}. Cette démarche a été réalisée pour prendre en compte des connaissances sur la survie marginale. Les intérêts de β^* sont son indépendance par rapport au mécanisme de censure et sa valeur proche de $E[\beta_0(T)]$, qui est un bon résumé du paramètre de régression fonctionnel β_0.

On peut voir que l'estimateur du maximum de vraisemblance ou la méthode des moments satisfont l'équation (2.11) si la loi marginale de T est bien définie. Donc le paramètre θ_0 peut être estimé en maximisant la vraisemblance partielle ou par la méthode des moments avant de passer au problème de l'estimation consistante de β^*, ce qui est suggéré dans le Théorème 2.4. Il peut être intéressant de considérer un estimateur lisse de la fonction de survie S de T, qui n'est pas trop variable pour des temps tardifs dans le jeu de données. De plus, on peut avoir une meilleure interprétation de T en utilisant sa forme paramétrique. Enfin, supposons que l'on possède de l'information sur la loi de T grâce à un grand échantillon d'une part ; et que, d'autre part, l'échantillon que l'on souhaite étudier est plus petit et constitué du même type de données. Alors il est raisonnable d'utiliser l'information sur le grand échantillon pour l'estimation sur le petit. C'est le cas par exemple en survie relative. On trouvera une illustration de ce cas en Section 2.7. Le Théorème 2.4 nous assure que nous pouvons ajuster les poids dans la fonction score (2.5) avec W_{par}. Alors l'estimateur obtenu convergera vers β^*, proche de $E[\beta_0(T)]$.

2.2.4 Inférence bayésienne

Nous pouvons aussi voir l'estimation paramétrique de β_0 d'un point de vue bayésien. Soit g la densité associée à la loi a priori sur θ et supposons que Θ est un ensemble compact de $\mathbb{R}^p$. On note $\hat{\theta}_n$ la n-ème estimation bayésienne de θ (en utilisant la vraisemblance) et g_n sa densité associée. Pour chaque θ fixé, on peut considérer que T suit la loi $F(t;\theta)$ et on définit

$$U_n(\theta,\beta) = \int \{\mathcal{Z}(t) - E(t,\beta)\}\, dF_n(t,\theta) \tag{2.15}$$

$$= \sum_{i=1}^{n} \Delta_i \frac{S(X_i,\theta)}{nS^{(0)}(0,X_i)} \{Z_i(X_i) - E(\beta,X_i)\}\,, \tag{2.16}$$

où $S(t;\theta) = 1-F(t;\theta)$. On suppose que $S(t;.)$ est continûment différentiable et minorée par une constante strictement positive. On obtient le théorème suivant

Théorème 2.5. *Si la fonction $v(\beta,t)$ est minorée par une constante strictement positive sur $\mathcal{B}\times[0,\tau]$ et sous les hypothèses (A), (B), (C) et (D), alors $\hat{\beta}_b \xrightarrow[n\to\infty]{P} \beta^*$, où $\hat{\beta}_b$ est la solution de l'équation $U_n(\hat{\theta}_n,\beta)=0$.*

Démonstration. La preuve du Théorème 2.5 repose sur le Lemme 2.6 ci-dessous dont la preuve peut être trouvée en Annexe.

Lemme 2.6. *Théorème des fonctions implicites probabiliste*
Soit $\mathcal{U}$ un ouvert de $\mathbb{R}^m\times\mathbb{R}^p$, $k\in\mathbb{N}^*$ et X un processus presque sûrement $\mathcal{C}^k$ défini sur $\mathcal{U}$ et à valeurs dans $\mathbb{R}^p$. Soit $(a,b)\in\mathbb{R}^m\times\mathbb{R}^p$ tel que $X(a,b)=0$ et $D_tX(a,b)$ inversible presque sûrement.

Alors il existe un voisinage $\mathcal{V}$ de a dans $\mathbb{R}^m$, un voisinage $\mathcal{W}$ de b dans $\mathbb{R}^p$ et un processus $\phi:\mathcal{V}\longrightarrow\mathcal{W}$ presque sûrement $\mathcal{C}^k$ tel que $\mathcal{V}\times\mathcal{W}\subset\mathcal{U}$ et

$$\forall s\in\mathcal{V},\ \forall t\in\mathcal{W},\ \textit{presque sûrement},\ X(s,t)=0\Leftrightarrow t=\phi(s)$$

De plus, presque sûrement,

$$\forall s\in\mathcal{V},\ d\phi(s) = -D_tX(s,\phi(s))^{-1}\circ D_sX(s,\phi(s)). \tag{2.17}$$

Revenons maintenant à la preuve du Théorème 2.5. Pour chaque θ, on peut appliquer la Propriété 2.1 et donc définir la suite $(\hat{\beta}_{n,\theta})_n$ telle que pour tout $n\in\mathbb{N}$, $U_n(\theta,\hat{\beta}_{n,\theta})=0$ et $\hat{\beta}_{n,\theta}\xrightarrow[n\to\infty]{P}\beta_\theta$, où β_θ est la solution de l'équation

$$h_\theta(\beta) = \int_0^\tau w_\theta(t)\left\{\frac{s^{(1)}(\beta_0,t)}{s^{(0)}(\beta_0,t)} - \frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right\} s^{(0)}(\beta_0,t)dt = 0, \tag{2.18}$$

avec $w_\theta(t) = S(t;\theta)/s^{(0)}(0,t)$. D'après le Lemme 2.6, cela nous autorise à considérer le processus $\left(\hat{\beta}_n(\theta)\right)_n$ presque sûrement $\mathcal{C}^1$ défini sur Θ tel que, pour tout $\theta\in\Theta$, $U_n(\theta,\hat{\beta}_n(\theta))=0$ et $\hat{\beta}_n(\theta)\xrightarrow[n\to\infty]{P}\beta(\theta)$. De plus, en appliquant (2.17), on obtient

$$\hat{\beta}_n'(\theta) = -\left(\frac{\partial U_n}{\partial\beta}(\theta,\hat{\beta}_n(\theta))\right)^{-1}\frac{\partial U_n}{\partial\theta}(\theta,\hat{\beta}_n(\theta)). \tag{2.19}$$

Etudions maintenant les deux dérivées partielles présentes dans (2.19). D'une part,

$$\left|\frac{\partial U_n}{\partial \theta}(\theta, \hat{\beta}_n(\theta))\right| = \left|\sum_{i=1}^{n} \Delta_i \frac{\frac{\partial S}{\partial \theta}(X_i;\theta)}{nS^{(0)}(0,X_i)} \left\{Z_i(X_i) - E(\hat{\beta}_n(\theta), X_i)\right\}\right|$$
$$\leq \sum_{i=1}^{n} \frac{\left|\frac{\partial S}{\partial \theta}(X_i,\theta)\right|}{n|S^{(0)}(0,X_i)|} |Z_i(X_i) - E(\hat{\beta}_n(\theta), X_i)|.$$

— $S(t;.)$ est $\mathcal{C}^1$-différentiable et Θ est compact. Donc il existe une constante M_1 indépendante de n telle que pour tout $i \in \{1, \cdots, n\}$,

$$\left|\frac{\partial S}{\partial \theta}(X_i, \theta)\right| \leq M_1.$$

— Les hypothèses (B) et (C) impliquent qu'à partir d'un certain rang, il existe $m_0 > 0$ indépendant de n tel que pour tout $i \in \{1, \cdots, n\}$, $|S^{(0)}(0, X_i)| \geq m_0$.

— $E(\hat{\beta}_n(\theta), X_i) = S^{(1)}(\hat{\beta}_n(\theta), X_i)/S^{(0)}(\hat{\beta}_n(\theta), X_i)$. Donc en utilisant les hypothèses (B) et (C), il existe une constante M_2 indépendante de n telle que pour tout $i \in \{1, \cdots, n\}$, $E(\hat{\beta}_n(\theta), X_i) \leq M_2$. D'où $|Z_i(X_i) - E(\hat{\beta}_n(\theta), X_i)| \leq M_3 = max(Z) + M_2$.

Finalement, au moins pour n assez grand, $\frac{\partial U_n}{\partial \theta}(\theta, \beta_n(\theta))$ est bornée par une constante indépendante de n (égale à $M_1 M_3/m_0$). D'autre part,

$$\left|\frac{\partial U_n}{\partial \beta}(\theta, \beta_n(\theta))\right| = \sum_{i=1}^{n} \Delta_i \frac{S(X_i;\theta)}{nS^{(0)}(0,X_i)} V(\beta_n(\theta), X_i).$$

— Pour tout t fixé, $S(t;.)$ est minorée par une constante strictement positive. Donc, comme il n'y a qu'un nombre fini de X_i, il existe une constante m_1 strictement positive et indépendante de n telle que pour tout $i \in \{1, \cdots, n\}$, $S(X_i;\theta) \geq m_1$.

— Les hypothèses (B) et (C) impliquent qu'à partir d'un certain rang, il existe une constante M_0 indépendante de n telle que $|S^{(0)}(0, X_i)| \leq M_0$.

— Le fait que la fonction $v(\beta, t)$ soit minorée par une constante strictement positive et l'hypothèse (B) impliquent qu'à partir d'un certain rang, il existe une constante m_2 strictement positive et indépendante de n telle que, pour tout $i \in \{1, \cdots, n\}$, $V(\hat{\beta}_n(\theta), X_i) \geq m_2$.

Donc

$$\left|\frac{\partial U_n}{\partial \beta}(\theta, \beta_n(\theta))\right| \geq \frac{m_1 m_2}{M_0} \frac{1}{n} \sum_{i=1}^{n} \Delta_i.$$

Or si on applique la loi forte des grands nombres à la suite de variables i.i.d. $(\Delta_i)_{i \in \{1,\cdots,n\}}$, on obtient qu'à partir d'un certain rang, presque sûrement,

$$\frac{1}{n} \sum_{i=1}^{n} \Delta_i \geq m_3 = P(\Delta_1 = 1) - \varepsilon, \ \forall \varepsilon > 0.$$

Finalement, au moins pour n assez grand, $\frac{\partial U_n}{\partial \beta}(\theta, \hat{\beta}_n(\theta))$ est bornée et majorée par une constante K_2 strictement positive indépendante de n. Donc, avec l'équation (2.19), $\hat{\beta}_n'(.)$ est presque sûrement bornée par une constante $K = K_1/K_2 > 0$ indépendante de n. On en déduit que $\hat{\beta}_n$ est presque sûrement uniformément Lipschitz et donc

$$\|\hat{\beta}_n - \beta\|_\infty \xrightarrow[n\to\infty]{P} 0. \tag{2.20}$$

Par définition de $\hat{\theta}_n$, on a

$$\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta_0. \tag{2.21}$$

Soit maintenant $\varepsilon > 0$ et $\varepsilon_0 > 0$.

$$\begin{aligned}
P(|\hat{\beta}_b - \beta^*| \leq \varepsilon_0) &= P(|\hat{\beta}_n(\hat{\theta}_n) - \beta(\theta_0)| \leq \varepsilon_0) \\
&\geq P(|\hat{\beta}_n(\hat{\theta}_n) - \beta(\hat{\theta}_n)| + |\beta(\hat{\theta}_n) - \beta(\theta_0)| \leq \varepsilon_0) \\
&\geq P\left(\left\{\|\hat{\beta}_n - \beta\|_\infty \leq \frac{\varepsilon_0}{2}\right\} \cap \left\{|\beta(\hat{\theta}_n) - \beta(\theta_0)| \leq \frac{\varepsilon_0}{2}\right\}\right) \\
&\geq P\left(\left\{|\hat{\beta}_n - \beta\|_\infty \leq \frac{\varepsilon_0}{2}\right\}\right) + P\left(\left\{|\beta(\hat{\theta}_n) - \beta(\theta_0)| \leq \frac{\varepsilon_0}{2}\right\}\right) - 1.
\end{aligned}$$

Or, d'après l'équation (2.20), on sait que $P\left(\left\{|\hat{\beta}_n - \beta\|_\infty \leq \frac{\varepsilon_0}{2}\right\}\right) \geq 1 - \varepsilon/2$. De plus, d'après l'équation (2.21) et la continuité de la fonction β, on a, à partir d'un certain rang, $P\left(\left\{|\beta(\hat{\theta}_n) - \beta(\theta_0)| \leq \frac{\varepsilon_0}{2}\right\}\right) \geq 1 - \varepsilon/2$. D'où $P(|\hat{\beta}_b - \beta^*| \leq \varepsilon_0) \geq 1 - \varepsilon$, c'est-à-dire

$$\hat{\beta}_b \xrightarrow[n\to\infty]{P} \beta^*.$$

□

Une première remarque est que le Théorème 2.5 nous permet de considérer une inférence bayésienne pour l'étude de l'effet moyen de la fonction β_0. En pratique, à l'étape n, on commence par calculer l'estimation bayésienne de θ, notée $\hat{\theta}_n$. Ensuite on estime β par $\hat{\beta}_b$ solution de l'équation $U_n(\hat{\theta}_n, \beta) = 0$ où

$$U_n(\hat{\theta}_n, \beta) = \sum_{i=1}^{n} \Delta_i \frac{S(X_i; \hat{\theta}_n)}{nS^{(0)}(0, X_i)}\{Z_i(X_i) - E(\beta, X_i)\}.$$

On peut résoudre cette équation avec la méthode de Newton-Raphson par exemple. Le Théorème 2.5 nous assure que cet algorithme converge vers $\beta^* \approx E[\beta(T)]$.

2.2.5 Survie conditionnelle aux covariables

Dans cette section, on se place dans le cadre de modèles de survie à hasards proportionnels, *i.e.* $\lambda(t|Z) = \lambda_0(t)\exp(\beta_0 Z(t))$, et une censure indépendante. On définit la probabilité suivante

$$\tilde{P}(Z \in H|t) = \frac{\sum_H Y_j(t)\exp(\tilde{\beta}Z_j(t))}{\sum_{j=1}^n Y_j(t)\exp(\tilde{\beta}Z_j(t))}.$$

On peut trouver dans O'Quigley (2003) le résultat suivant

Proposition 2.7. *Sous le modèle (2.1) et une censure indépendante, en supposant que la fonction β est connue, la loi conditionnelle de $Z(.)$ sachant $T = t$ est estimée de manière consistante par*

$$\hat{P}\{Z(t) \leq z|T = t\} = \sum_{j=1}^n \pi_j(\beta(t), t)\mathbb{1}_{\{Z_j(t)\leq z\}},$$

où, pour tout $j \in \{1, \ldots, n\}$,

$$\pi_j(\beta(t), t) = \frac{Y_j(t)\exp(\beta(t)Z_j(t))}{\sum_{i=1}^n Y_i(t)\exp(\beta(t)Z_i(t))}.$$

En associant ce résultat avec le Théorème 2.4 et le lemme de Slutsky, on obtient

Proposition 2.8. *Sous un modèle à hasards proportionnels et une censure indépendante, $P(Z \in H|t)$ est estimée de manière consistante par $\tilde{P}(Z \in H|t)$.*

Un corollaire direct de la Proprosition 2.8, en utilisant la formule de Bayes, est l'estimation consistante de la survie conditionnelle aux covariables $S(t|Z \in H)$.

Corollaire 2.9. *Soit $0 = t_0 < t_1 < \cdots < t_k$ les temps de décès distincts. $S(t|Z \in H)$ est estimée de manière consistante par*

$$\tilde{S}(t|Z \in H) = \frac{\int_t^\infty \tilde{P}(Z \in H|u)d\tilde{F}(u)}{\int_0^\infty \tilde{P}(Z \in H|u)d\tilde{F}(u)} = \frac{\sum_{t_i>t} \tilde{P}(Z \in H|t_i)W_{par}(t_i)}{\sum_{j=1}^k \tilde{P}(Z \in H|t_j)W_{par}(t_j)}.$$

On trouvera une illustration de cette estimation dans la Section 2.5.

2.3 Simulations

Dans cette section, nous présentons des simulations effectuées pour étudier $\tilde{\beta}$ et le comparer avec l'estimateur de la vraisemblance partielle $\hat{\beta}_{PL}$ et l'estimateur de Kaplan-Meier $\hat{\beta}_{KM}$. Tout d'abord, nous étudions des cas de modèles à hasards proportionnels *i.e.* des modèles avec un paramètre de régression $\beta_0(t) = \beta_0$ pour tout $t \in [0, \tau]$; ensuite nous passons à quelques cas de modèles à hasards non-proportionnels. Précisément, nous considérons pour la fonction de régression des modèles constants par morceaux avec un seul point de rupture : $\beta_0(t) = \beta_1 \mathbb{1}_{t<t_0} + \beta_2 \mathbb{1}_{t\geq t_0}$, pour tout $t \in [0, \tau]$. La question de l'estimation de t_0 n'est pas abordée ici, elle le sera au Chapitre 3. La taille de l'échantillon est $n = 1500$. Le temps de décès T suit une loi exponentielle de paramètre 2. La covariable Z suit une loi uniforme sur $[0, 1]$. Le temps de censure C suit une loi uniforme sur $[0, t_c]$ (Table 2.1 et Table 2.2) ou une loi exponentielle de paramètre t_c (Table 2.3), où t_c est paramétré pour fixer le pourcentage de censure. Nous avons mené 500 simulations pour chaque cas et calculé les espérances et les écart-types empiriques de chaque estimateur.

TABLE 2.1 – Comparaison de $\hat{\beta}_{PL}$, $\hat{\beta}_{KM}$ et $\tilde{\beta}$ sous un modèle à hasards proportionnels. $C \sim \mathcal{U}[0, t_c]$. Ecart-types entre parenthèses.

β_0	% de censure	$\hat{\beta}_{PL}$	$\hat{\beta}_{KM}$	$\tilde{\beta}$	$E[\beta_0(T)]$
1	0%	1.000 (0.117)	1.000 (0.117)	1.000 (0.117)	1
	50%	0.996 (0.115)	1.001 (0.182)	1.000 (0.186)	1
0.5	0%	0.498 (0.103)	0.498 (0.103)	0.498 (0.103)	0.5
	50%	0.503 (0.100)	0.505 (0.184)	0.506 (0.190)	0.5

Les résultats de la Table 2.1 montrent que les trois estimateurs ont de bonnes performances dans le cas de modèles à hasards proportionnels. De plus, $\hat{\beta}_{KM}$ et $\tilde{\beta}$ sont légèrement moins efficaces que $\hat{\beta}_{PL}$ comme attendu. En effet, dans le cas de hasards proportionnels, l'estimateur $\hat{\beta}_{PL}$ est optimal.

Les résultats pour les modèles à hasards non-proportionnels sont donnés en Table 2.2 et Table 2.3. On remarque que, dans le cas de modèles à hasards non-proportionnels, $\hat{\beta}_{PL}$ dépend de la censure. En effet sa valeur varie fortement quand le taux de censure augmente, notamment quand la censure est à support infini (*cf* Table 2.3). On peut aussi noter que les estimateurs $\hat{\beta}_{KM}$ et $\tilde{\beta}$ sont consistants, même sous un modèle à hasards non-proportionnels, quelque soit le pourcentage de censure. Ceci n'est pas une surprise au regard de la Propriété 2.2 et du Théorème 2.4.

TABLE 2.2 – Comparaison de $\hat{\beta}_{PL}$, $\hat{\beta}_{KM}$ et $\tilde{\beta}$ sous un modèle à hasards non-proportionnels. $C \sim \mathcal{U}[0, t_c]$. Ecart-types entre parenthèses.

β_1	β_2	t_0	% de censure	$\hat{\beta}_{PL}$	$\hat{\beta}_{KM}$	$\tilde{\beta}$	$\mathbb{E}[\beta_0(T)]$
1	0	0.2	0%	0.330 (0.096)	0.331 (0.096)	0.330 (0.096)	0.330
			17%	0.373 (0.092)	0.330 (0.101)	0.329 (0.103)	0.330
			32%	0.418 (0.094)	0.351 (0.122)	0.348 (0.126)	0.330
			50%	0.512 (0.094)	0.438 (0.159)	0.437 (0.164)	0.330
3	0	0.2	0%	0.981 (0.110)	0.984 (0.110)	0.981 (0.112)	0.989
			17%	1.123 (0.111)	1.003 (0.119)	0.999 (0.122)	0.989
			32%	1.268 (0.116)	1.073 (0.142)	1.067 (0.149)	0.989
			50%	1.569 (0.123)	1.342 (0.196)	1.337 (0.204)	0.989

Table 2.3 – Comparaison de $\hat{\beta}_{PL}$, $\hat{\beta}_{KM}$ et $\tilde{\beta}$ sous un modèle à hasards non-proportionnels. $C \sim \mathcal{E}(t_c)$. Ecart-types entre parenthèses.

β_1	β_2	t_0	% de censure	$\hat{\beta}_{PL}$	$\hat{\beta}_{KM}$	$\tilde{\beta}$	$\mathbb{E}[\beta_0(T)]$
1	0	0.2	0%	0.331 (0.090)	0.332 (0.090)	0.331 (0.091)	0.330
			17%	0.382 (0.090)	0.332 (0.104)	0.331 (0.105)	0.330
			32%	0.444 (0.090)	0.340 (0.112)	0.337 (0.113)	0.330
			50%	0.549 (0.091)	0.342 (0.161)	0.338 (0.164)	0.330
3	0	0.2	0%	0.983 (0.113)	0.986 (0.113)	0.984 (0.116)	0.989
			17%	1.135 (0.114)	0.993 (0.124)	0.989 (0.127)	0.989
			32%	1.369 (0.121)	1.061 (0.145)	1.054 (0.148)	0.989
			50%	1.769 (0.128)	1.197 (0.197)	1.186 (0.207)	0.989

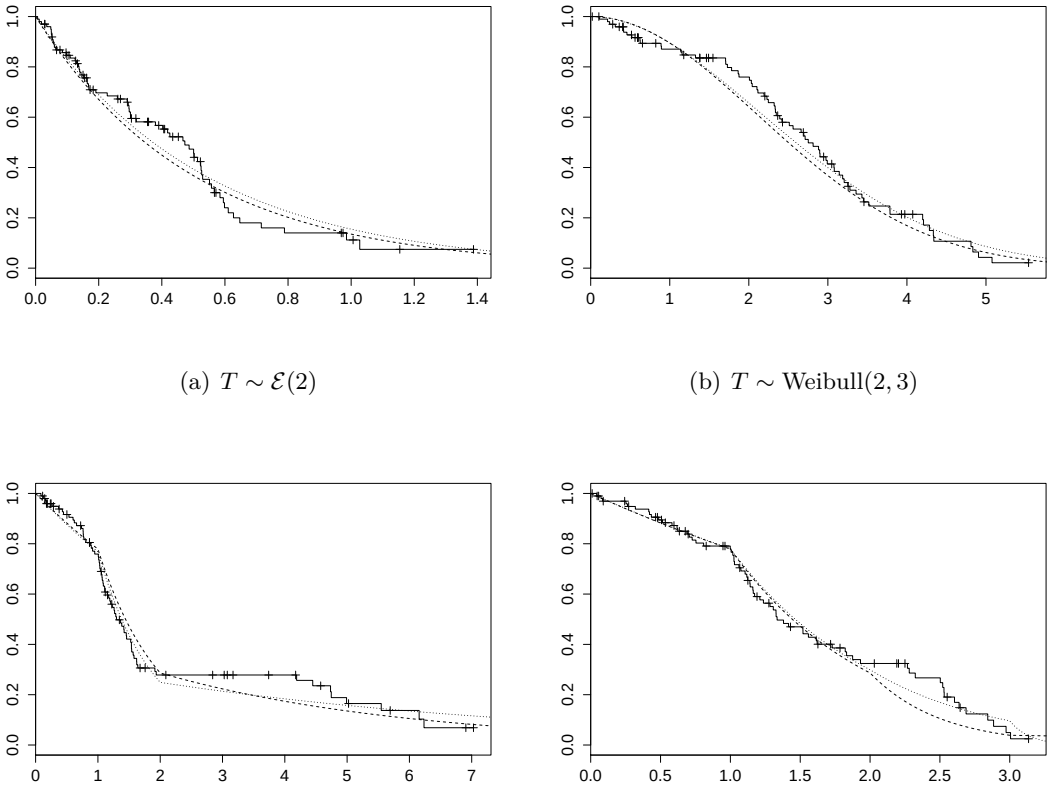
La Figure 2.1 présente la comparaison entre l'estimateur paramétrique de la fonction de survie présenté dans la Section 2.2 et l'estimateur de Kaplan-Meier. La vraie fonction de survie S de T est aussi tracée pour plus de clarté. Le paramètre fonctionnel β_0 choisi pour ces simulations est $\beta_0(t) = \mathbb{1}_{t<0.1}$, pour tout $t \in [0, \tau]$. La taille de l'échantillon est $n = 100$.
La loi de T est une exponentielle de paramètre 2 pour la première figure, et une Weibull de paramètres 2 et 3 pour la seconde. La troisième loi, notée E_3 est une exponentielle par morceaux de paramètres 0.25 sur $[0, 1[$, 1 sur $[1, 2[$ et 0.25 sur $[2, +\infty[$. La dernière loi, notée E_4 est aussi une exponentielle par morceaux de paramètres 0.25 sur $[0, 1[$, 1 sur $[1, 2[$, 2 sur $[2, 3[$ et 0.25 sur $[3, +\infty[$. Dans tous les cas, le pourcentages de censure est de 30%. Les paramètres de ces lois sont estimés par maximisation de la vraisemblance.

Nous pouvons voir que l'estimateur paramétrique est plus lisse que l'estimateur de Kaplan-Meier. Comme attendu, on évite les sauts à la fin du jeu de données de ce dernier.

On peut noter que, dans une étude où on possède de l'information sur les données, on peut choisir une estimation paramétrique pour les temps de décès qui colle mieux aux données que l'estimateur de Kaplan-Meier. Par exemple, si on anticipe que T est sans mémoire, on pourrait estimer T par une loi exponentielle.

Nous avons mené d'autres simulations pour comparer les précisions de l'estimation paramétrique et celle de Kaplan-Meier sous différents β_0, différents pourcentages de censure et différentes tailles d'échantillons. Les paramètres utilisés pour ces simulations sont identiques à ceux de la Table 2.2. Les résultats sont présentés en Table 2.4.

Il semble que les estimateurs $\hat{\beta}_{KM}$ et $\tilde{\beta}$ aient une précision similaire pour un modèle et une taille d'échantillon donnée. C'est pourquoi il peut être intéressant d'utiliser l'un ou l'autre de ces estimateurs suivant l'étude. Par exemple, on peut donner la priorité à $\tilde{\beta}$ dans le cas de données supervisées.



(a) $T \sim \mathcal{E}(2)$

(b) $T \sim \text{Weibull}(2, 3)$

(c) $T \sim E_3$

(d) $T \sim E_4$

Figure 2.1 – Comparaison entre l'estimateur de Kaplan-Meier (ligne pleine) et l'estimateur paramétrique (ligne pointillée) de la fonction de survie. La ligne en tirets représente la vraie courbe de survie.

β_1	β_2	t_0	% de censure	n	$\hat{\beta}_{KM}$	$\hat{\beta}_{par}$	$\mathbb{E}[\beta_0(T)]$
1	0	0.1	0%	50	0.250 (0.521)	0.247 (0.526)	0.181
				100	0.219 (0.365)	0.217 (0.365)	0.181
				500	0.187 (0.163)	0.187 (0.163)	0.181
				1000	0.177 (0.114)	0.176 (0.114)	0.181
			32%	50	0.213 (0.656)	0.204 (0.663)	0.181
				100	0.205 (0.481)	0.202 (0.490)	0.181
				500	0.177 (0.195)	0.175 (0.197)	0.181
				1000	0.187 (0.147)	0.185 (0.151)	0.181
3	0	0.05	0%	50	0.316 (0.543)	0.318 (0.543)	0.285
				100	0.275 (0.369)	0.270 (0.372)	0.285
				500	0.255 (0.170)	0.254 (0.170)	0.285
				1000	0.266 (0.120)	0.265 (0.120)	0.285
			32%	50	0.300 (0.670)	0.286 (0.686)	0.285
				100	0.321 (0.476)	0.311 (0.490)	0.285
				500	0.294 (0.218)	0.294 (0.222)	0.285
				1000	0.280 (0.151)	0.278 (0.156)	0.285

Table 2.4 – Comparaison de la précision de l'estimateur de Kaplan-Meier et de l'estimateur paramétrique.

2.4 Efficacité relative sous des modèles à hasards proportionnels

Nous avons vu dans la Section 2.2 que, sous des modèles à hasards proportionnels, $\tilde{\beta}$ est consistant pour la constante β_0 ; comme $\hat{\beta}_{PL}$. Une question naturelle est alors de trouver lequel de ces deux estimateurs est le plus efficace. Cependant, comme $\tilde{\beta}$ est obtenu en introduisant des poids dans l'équation score, on sait que $\hat{\beta}_{PL}$ est l'estimateur le plus efficace sous le modèle de Cox (Efron, 1977). On cherche donc maintenant à savoir à quel point $\tilde{\beta}$ est moins efficace que $\hat{\beta}_{PL}$.

Les résultats de Lin (1991) donnent l'efficacité relative asymptotique de $\tilde{\beta}$ par rapport à $\hat{\beta}_{PL}$:

$$R_{eff}(\tilde{\beta}, \hat{\beta}_{PL}) = \frac{(\Sigma_1)^2}{\Sigma_0 \Sigma_2},$$

où

$$\Sigma_0 = \int_0^\infty v(\beta_0, t) s^{(0)}(\beta_0, t) dt,\ \Sigma_1 = \int_0^\infty v(\beta_0, t) dF(t),\ \Sigma_2 = \int_0^\infty v(\beta_0, t) \frac{S(t)}{s^{(0)}(0, t)} dF(t).$$

Pour illustrer ceci, on étudie le cas suivant : le taux de hasard de base est constant égal à 1, Z suit une loi de Bernoulli de paramètre p et C suit une loi lognormale de paramètres 0 et t_c. Cette dernière distribution est choisie pour assurer la convergence des intégrales Σ_i. On a alors

$$\Sigma_0 = \int_0^\infty A(\beta_0, t) \mathrm{P}(C \geq t) dt,\ \Sigma_1 = \int_0^\infty A(\beta_0, t) dt,\ \Sigma_2 = \int_0^\infty \frac{A(\beta_0, t)}{\mathrm{P}(C \geq t)} dt,$$

où

$$A(\beta, t) = \frac{(1-p)e^{-t} p e^{\beta} \exp(-te^{\beta})}{(1-p)e^{-t} + pe^{\beta} \exp(-te^{\beta})}.$$

Les résultats pour différentes valeurs de p, β_0 et t_c sont présentés en Table 2.5.

TABLE 2.5 – Efficacité relative asymptotique de $\tilde{\beta}$ par rapport à $\hat{\beta}_{PL}$ sous un modèle à hasards proportionnels. Pourcentages de censure entre parenthèses.

β_0	t_c	$p = 0.25$	$p = 0.5$	$p = 0.75$
0.5	1	0.797 (35%)	0.772 (32%)	0.736 (29%)
1		0.911 (33%)	0.892 (27%)	0.863 (21%)
2		0.990 (30%)	0.986 (21%)	0.979 (12%)
0.5	0.5	0.192 (35%)	0.150 (32%)	0.100 (29%)
1		0.746 (33%)	0.675 (27%)	0.564 (21%)
2		0.996 (30%)	0.993 (21%)	0.988 (12%)

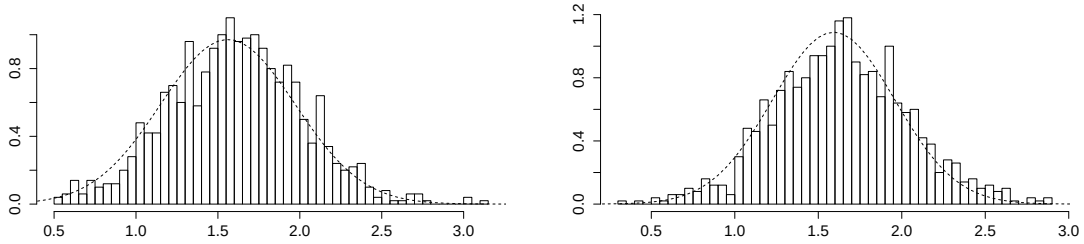
On peut noter que l'efficacité relative asymptotique de $\tilde{\beta}$ par rapport à $\hat{\beta}_{PL}$ est mauvaise pour un fort mécanisme de censure (quatrième ligne de la Table 2.5). De plus, plus la valeur de β_0 est grande, plus l'efficacité relative asymptotique l'est.En effet, quand $|\beta_0| \to \infty$, $A(\beta_0, t) \to 0$ et donc $R_{eff}(\tilde{\beta}, \hat{\beta}_{PL}) \to 1$.

2.5 Une illustration de variance sur le jeu de données Freireich

Nous commençons par une illustration dans un cas de hasards proportionnels. Une question naturelle à ce stade est de déterminer la variance de $\tilde{\beta}$ et la comparer à la variance de $\hat{\beta}_{PL}$ sur un cas pratique. Pour mieux visualiser la comparaison, on se place dans le cas réel des données de leucémie de Freireich et co. (1963). On obtient $\hat{\beta}_{PL} \approx 1,56$ et $\tilde{\beta} \approx 1.59$. On a généré 500 estimateurs ré-échantillonés en introduisant des poids aléatoires dans les équations estimatrices. Pour être plus précis, on a généré à chaque itération un n_t-échantillon de loi exponentielle de paramètre 1, où n_t est le nombre de temps de décès dans le jeu de données. On note l'un d'eux $(e_1, \cdots, e_{n_t})$. On introduit ensuite les poids de la forme $e_i / \sum_{j=1}^{n_t} e_j$ dans les équations estimatrices (2.3) et (2.5). On obtient ainsi deux histogrammes présentés en Figure 2.2. Nous avons aussi tracé les lois asymptotiques théoriques de $\hat{\beta}_{PL}$ et $\tilde{\beta}$: une loi gaussienne d'espérance $\hat{\beta}_{PL}$ et de variance donnée par Andersen and Gill (1982) pour $\hat{\beta}_{PL}$, et une loi gaussienne d'espérance $\tilde{\beta}$ et de variance donnée par Lin (1991) pour $\tilde{\beta}$. Cette dernière variance peut être estimée de manière consistante par $A(\tilde{\beta})^{-1}B(\tilde{\beta})A(\tilde{\beta})^{-1}$, où

$$A(\beta) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i W_{\text{par}}(X_i) V(\beta, X_i) \text{ et } B(\beta) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i W_{\text{par}}^2(X_i) V(\beta, X_i).$$

La première variance peut être estimée de manière consistante par la même formule avec une fonction de pondération W identiquement égale à 1.



(a) Estimateur $\hat{\beta}_{PL}$ (b) Estimateur $\tilde{\beta}$

Figure 2.2 – Comparaison entre l'histogramme de 500 estimateurs ré-échantillonés et de la loi gaussienne théorique (en pointillé).

L'écart-type estimé de $\hat{\beta}_{PL}$ est 0.41 et l'écart-type estimé de $\tilde{\beta}$ est 0.37. On peut voir, dans cette configuration, que $\tilde{\beta}$ et $\hat{\beta}_{PL}$ ont des lois théoriques et empiriques très proches. Les données utilisées ici sont connues pour satisfaire le modèle à hasards proportionnels de Cox et on sait aussi que $\hat{\beta}_{PL}$ est le meilleur estimateur du vrai paramètre de régression β_0 sous un tel modèle. Ce résultat montre l'intérêt d'utiliser $\tilde{\beta}$ sous un modèle à hasards proportionnels et la validité de son utilisation sous un modèle à hasards non-proportionnels a été montrée à la Section 2.2.

Sur ce jeu de données, il est intéressant de mettre en pratique les résultats de la Section 2.2.5. En fait la covariable Z est une covariable binaire valant 0 si le patient a reçu un placebo et 1 s'il a reçu le traitement à l'étude. Pour visualiser l'effet du traitement, nous avons tracé en Figure 2.3 les estimateurs consistants des courbes de survie conditionnelles

à la covariable Z, c'est-à-dire l'estimateur de la courbe de survie des patients ayant reçu un placebo et l'estimateur de la courbe de survie des patients ayant reçu le traitement.

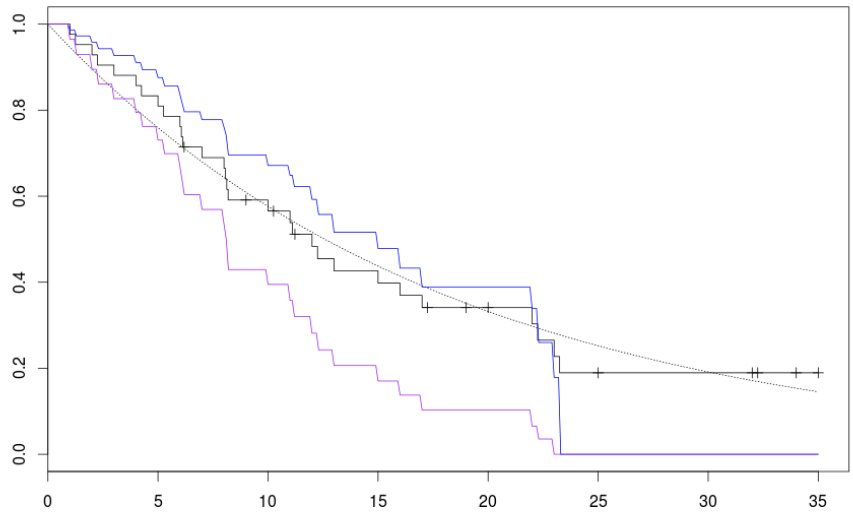


Figure 2.3 – Comparaison de la courbe de survie empirique (trait plein noir) avec la courbe de survie des patients ayant reçu un placebo (violet) et celle des patients ayant reçu le traitement (bleu). L'estimation paramétrique de la courbe de survie est représentée en pointillés pour plus de clarté.

On constate qu'il y a bien un "effet traitement" ce qui correspond bien à la valeur élevée de $\tilde{\beta}$.

2.6 Une première application : données de cancer du sein

Dans cette section, nous illustrons l'utilisation des poids introduits en Section 2.2 sur des données de cancer du sein collectées à l'Institut Curie. Nous supposons que la loi de T est une exponentielle par morceaux, même si cela implique de prendre un grand nombre de points de rupture. On applique le modèle (2.1) à ces données. En choisissant une loi exponentielle par morceaux avec quatre points de rupture pour la loi de T, on obtient la courbe de survie présentée en Figure 2.4.

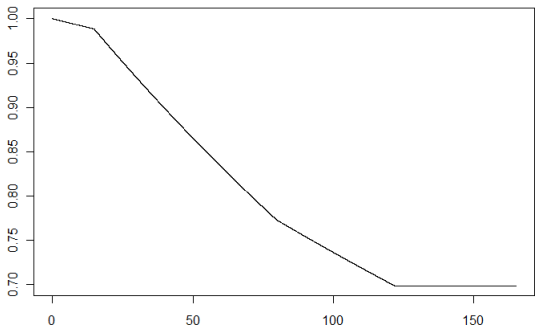


Figure 2.4 – Estimateur de la fonction de survie S obtenue par estimation paramétrique.

On calcule ensuite $\tilde{\beta}$ pour les covariables suivantes : l'âge, le grade histologique, le stade du cancer, le statut du récepteur de progestérone et la taille de la tumeur. On génère 500 échantillons de bootstrap (Efron and Tibshirani, 1994) basés sur les données de l'Institut Curie pour obtenir des estimations des écart-types. Les résultats sont résumés en Table 2.6.

Table 2.6 – Effets moyens estimés pour le jeu de données du cancer du sein. Ecart-types entre parenthèses.

	Age	Grade	Stade	Récepteur	Taille
$\tilde{\beta}$	-0.015 (0.006)	0.417 (0.059)	0.459 (0.055)	-0.580 (0.129)	0.016 (0.002)

Bien sûr, il serait plus précis de supposer que β_0 est constant par morceaux et d'estimer sa valeur sur chaque morceaux. Pour cela, on peut utiliser une recherche de points de rupture, ce qui est l'objet du Chapitre 3. On peut trouver des informations à ce sujet dans Xu and Adak (2002) par exemple. Ce que nous avons obtenu ici sont des effets moyens des différentes covariables étudiées.

2.7 Une deuxième application dans le cadre de la survie relative

Cette section se focalise sur une application en survie relative de la méthode proposée en Section 2.2. Les données utilisées sont des données d'infarctus du myocarde aigu collectées au Centre Clinique de Ljubljana. On peut trouver ces données dans le fichier `rdata` du package R `relsurv` développé par Pohar and Stare (2006). Nous avons utilisé les tables de population slovénienne contenues dans le fichier `slopop` pour estimer la loi de T. La fonction de survie empirique de T est présentée en Figure 2.5. Nous avons choisi de l'estimer par une exponentielle à 2 morceaux. Sa représentation est aussi en Figure 2.5, en bleu.

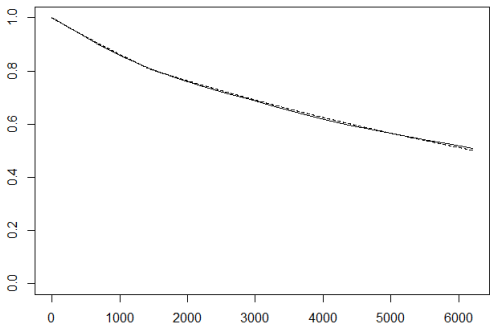


Figure 2.5 – Survie empirique de T (ligne continue) obtenue avec les tables de population slovénienne. Estimation paramétrique de S (ligne en tirets). Le temps est exprimé en jours depuis l'année 1982.

Nous avons calculé $\tilde{\beta}$ pour les covariables suivantes : l'âge, le sexe, l'année de diagnostique de la maladie et la classe d'âge ("sous 54", "54-61", "62-70" et "71-95"). Nous avons généré, comme en Section 2.6, 500 échantillons bootstrap basés sur ces données pour obtenir des estimations des écart-types. Les résultats sont résumés en Table 2.7. L'importance de l'écart-type dans le cas de la covariable "sexe" est peut-être expliqué par un coefficient de régression β qui varie fortement avec le temps.

Table 2.7 – Effets moyens estimés pour les données d'infarctus du myocarde aigu. Ecart-types entre parenthèses.

	Âge	Sexe	Année	Classe d'âge
$\tilde{\beta}$	0.046 (0.012)	-0.069 (0.358)	0.005 (0.001)	0.444 (0.144)
$\hat{\beta}_{PL}$	0.061 (0.004)	0.530 (0.089)	0.000 (0.0001)	0.573 (0.044)

Chapitre 3

Détection de points de rupture

3.1 Introduction

Le modèle à hasards proportionnels de Cox (1972) représente une avancée importante en analyse de survie. Son taux de hasard s'écrit

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta_0^T Z), \tag{3.1}$$

où $Z \in \mathbb{R}^d$ est un vecteur de covariables, λ_0 une fonction inconnue de t et $\beta_0 \in \mathbb{R}^d$ les coefficients de régression associés aux covariables Z. L'utilisation du modèle de Cox s'est rapidement développée dans le cadre de la régression pour des données censurées, notamment grâce à sa facilité d'interprétation. On peut citer par exemple les travaux de Kay (1977); Kalbfleisch and Prentice (1980); Andersen and Gill (1982); Lin (1991). Certains d'entre eux sont consacrés à l'étude des propriétés asymptotiques de l'estimateur de la vraisemblance partielle, sous le modèle (3.1), et permette ainsi de faire de l'inférence sur le paramètre d'intérêt β_0. Cependant, le modèle (3.1) n'est pas toujours réaliste. On peut penser, par exemple, à des études sur la mortalité en cancérologie, où l'effet d'un traitement diminue avec le temps à cause de l'accoutumance. Ce cas de figure n'est malheureusement pas pris en compte par le modèle de Cox qui suppose un effet constant dans le temps. De nombreux auteurs se sont consacrés à l'étude d'un modèle où β_0 est maintenant une fonction de régression $\beta_0(.)$: Moreau et al. (1985); O'Quigley and Pessione (1989, 1991); Liang et al. (1990); Zucker and Karr (1990); Murphy and Sen (1991); Gray (1992); Hastie and Tibshirani (1993); Verweij and Houwelingen (1995); Lausen and Schumacher (1996); Marzec and Marzec (1997) pour n'en citer que quelques uns.

Nous nous intéressons ici à une extension particulière du modèle de Cox (3.1) qui est le cas où la fonction β_0 est constante par morceaux. Les points de discontinuité de la fonction β_0 sont alors appelés "points de rupture" ou "changepoints". Anderson and Senthilselvan (1982) se sont penchés sur l'estimation des paramètres d'un tel modèle, c'est-à-dire les coefficients de régression et les changepoints, dans le cas d'un unique changepoint. Nous voulons pousser leur analyse un peu plus loin et proposer une méthode d'inférence sur le changepoint. Nous utiliserons pour cela les travaux de Davies (1977). Nous souhaitons également proposer une méthode d'estimation pour un modèle plus général avec K changepoints, où K est fixé à l'avance.

Nous commençons par introduire les notations nécessaires en Section 3.2 et présentons plus formellement les différents modèles que nous étudions dans ce chapitre. La Section 3.3 se focalise sur le modèle de Anderson and Senthilselvan (1982). Nous y rappelons la méthode d'estimation qu'ils proposent et établissons une région de confiance pour le changepoint. Nous changeons de modèle en Section 3.4 pour nous placer dans un modèle

plus général de K changepoints et proposons une méthodes d'estimation par moindres carrés à l'aide du processus du score standardisé (Chauvel and O'quigley, 2014). Des simulations pour illustrer les résultats des deux sections précédentes sont présentés en Section 3.5. Enfin, en Section 3.6, une application à des données de cancer du sein fournies par l'Institut Curie clôt ce chapitre.

3.2 Notations

Pour tout $i \in \{1, \dots, n\}$, on note (T_i, C_i, Z_i) une de suite de variables aléatoires i.i.d. de même distribution que le triplet (T, C, Z), où T est la variable aléatoire représentant le temps de décès, de fonction de répartition F, $Z \in \mathbb{R}$ est le vecteur des covariables et C est la variable aléatoire de censure, independante de T sachant Z. On suppose qu'il existe $\tau > 0$ tel que $[0, \tau]$ est le support de T et C. On suppose toujours que les variables T et Z suivent le modèle (2.1). On s'intéresse dans ce chapitre à deux modèles en particulier. Le premier est le modèle introduit par Anderson and Senthilselvan (1982) qui suppose le coefficient de régression constant par morceaux avec deux morceaux. On y fera référence sous le nom de "modèle réduit". Il s'écrit

$$\beta_0(t) = \beta_{01} \mathbb{1}_{t \leq \gamma_0} + \beta_{02} \mathbb{1}_{t > \gamma_0}, \ \forall t \in [0, \tau], \tag{3.2}$$

où β_{01} et β_{02} sont des constantes réelles et γ_0 est une constante strictement positive. On nomme ce dernier point de rupture ou changepoint. Cette définition est naturelle puisque γ_0 est le point de discontinuité de la fonction de régression β_0.

Le second modèle sera appelé "modèle général" et s'écrit

$$\beta_0(t) = \beta_{01} \mathbb{1}_{t \leq \gamma_{01}} + \beta_{02} \mathbb{1}_{\gamma_{01} < t \leq \gamma_{02}} \dots + \beta_{0K} \mathbb{1}_{t > \gamma_{0(K-1)}}, \ \forall t \in [0, \tau], \tag{3.3}$$

où $\beta_{01}, \dots, \beta_{0K}$ sont des constantes réelles et $\gamma_{01}, \dots, \gamma_{0(K-1)}$ sont des constantes strictement positives telles que $\gamma_{01} < \gamma_{02} < \dots < \gamma_{0(K-1)}$. Ces derniers sont, pour la même raison que précédemment, nommés des changepoints. Un point important ici est l'entier naturel $K > 0$. En effet, on considère dans ce chapitre que K est fixe et connu. Autrement dit, le nombre de points de discontinuité du coefficient de régression sont connus avant l'étude clinique.

Pour tout $i \in \{1, \dots, n\}$, définissons de plus $X_i = \min(T_i, C_i)$ et $\Delta_i = \mathbb{1}_{T_i \leq C_i}$ tels que X_i est le temps observé associé au patient i et Δ_i est le statut de ce patient : "décédé" $(= 1)$ ou "censuré" $(= 0)$. On note $\mathcal{D}$ l'ensemble des fonctions discontinues de $[0, \tau]$ dans $\mathbb{R}$, et on a donc $\beta_0 \in \mathcal{D}$. Pour tout $i \in \{1, \cdots, n\}$ et tout $t \in [0, \tau]$, on définit $Y_i(t) = \mathbb{1}_{X_i \geq t}$. Le processus $Y_i(t)$ précise si le patient i est encore à risque au temps t $(= 1)$, ou non $(= 0)$.

Au regard des résultats obtenus par O'Quigley and Pessione (1991), la première idée à apparaître quant à l'inférence du paramètre γ_0 dans le cas du modèle réduit est l'utilisation des propriétés énoncées dans Davies (1977). Cependant un problème survient immédiatement. En effet, Davies (1977) étudie la vraisemblance classique. Or, dans notre cas, nous avons accès à la vraisemblance partielle. Sous le modèle (3.2) la log-vraisemblance partielle s'écrit sous la forme $L(\beta_1, \beta_2, \gamma) = L_1(\beta_1, \gamma) + L(\beta_2, \gamma)$, où

$$L_1(\beta_1, \gamma) = \frac{1}{n} \sum_{X_i \leq \gamma} \Delta_i \left[\beta_1 Z_i - \log \left\{ \sum_{j=1}^{n} Y_j(X_i) \exp(\beta_1 Z_j) \right\} \right] \tag{3.4}$$

$$L_2(\beta_2, \gamma) = \frac{1}{n} \sum_{X_i > \gamma} \Delta_i \left[\beta_2 Z_i - \log \left\{ \sum_{j=1}^{n} Y_j(X_i) \exp(\beta_2 Z_j) \right\} \right]. \tag{3.5}$$

Donc, en suivant la démarche de Davies (1977), on introduit la statistique à valeurs dans $\mathbb{R}^2$ suivante

$$S(\beta_1, \beta_2, \gamma) = (S_1(\beta_1, \gamma), S_2(\beta_2, \gamma)),$$

avec

$$S_1(\gamma) = \sqrt{n} V_1(\gamma)^{1/2} \hat{\beta}_1^{(n)}(\gamma)$$
$$S_2(\gamma) = \sqrt{n} V_2(\gamma)^{1/2} \hat{\beta}_2^{(n)}(\gamma),$$

où, pour $i \in \{1, 2\}$, $\hat{\beta}_i^{(n)}(\gamma)$ est la valeur maximisant la vraisemblance partielle $L_i(\beta_i, \gamma) = 0$ à $\gamma \in [0, \tau]$ fixé, et

$$V_1(\gamma) = -\frac{\partial^2 L_1}{\partial \beta_1^2}(0, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \mathbb{1}_{X_i \leq \gamma} V(0, X_i)$$
$$V_2(\gamma) = -\frac{\partial^2 L_2}{\partial \beta_2^2}(0, \gamma) = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \mathbb{1}_{X_i > \gamma} V(0, X_i)$$

avec

$$V(\beta(t), t) = \frac{S^{(2)}(\beta(t), t)}{S^{(0)}(\beta(t), t)} - \left[\frac{S^{(1)}(\beta(t), t)}{S^{(0)}(\beta(t), t)} \right]^2,$$

où les $S^{(r)}(\beta(t), t)$ pour $r \in \{0, 1, 2\}$ sont définis en (2.2).

Ces notations prises, nous passons dans la Section 3.3 à un travail d'inférence sur le changepoint γ_0 en utilisant les résultats de Davies (1977). La Section 3.4 se focalise sur l'estimation dans le cas du modèle général (3.3) à l'aide du processus du score standardisé.

3.3 Inférence dans le cas du modèle réduit

On note, pour $i \in \{1, 2\}$,

$$\text{Corr}\left\{S_i(\gamma_1), S_i(\gamma_2)\right\} = \rho_i(\gamma_1, \gamma_2).$$

Sur le segment $[0, \gamma_0]$ d'une part et sur $[\gamma_0, \tau]$ d'autre part, nous sommes dans le cas d'un modèle à hasards proportionnels. C'est pourquoi, d'après Andersen and Gill (1982), on a la convergence suivante pour $i \in \{1, 2\}$.

$$\frac{1}{\sqrt{n}} S_i(\beta_i, \gamma_0) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_i(\gamma_0),$$

où $\mathcal{N}_i(.)$ est un processus gaussien d'espérance

$$E_{\beta_{0i},\gamma_0}\left\{\mathcal{N}_i(\gamma_1)\right\} = \beta_{0i} V_i(\gamma_0) \rho_i(\gamma_1, \gamma_0),$$

et de fonction de corrélation

$$\text{Corr}\left\{\mathcal{N}_i(\gamma_1), \mathcal{N}_i(\gamma_2)\right\} = \rho_i(\gamma_1, \gamma_2),$$

pour $\gamma_1, \gamma_2 \in [0, \tau]$. On se place maintenant à n suffisamment grand pour que la déviation de $S_i(\gamma)$ par rapport à la variable gaussienne $\mathcal{N}_i(\gamma)$ soit négligée. On suppose les $\rho_i(\gamma_1, \gamma_2)$ sont des fonctions $\mathcal{C}^2$.

3.3.1 Estimation du point de rupture

Anderson and Senthilselvan (1982) ont proposé une méthode d'estimation, dans le cas du modèle réduit (3.2), de β_{01}, β_{02} et γ_0. On rappelle que la log-vraisemblance partielle s'écrit alors $L(\beta_1, \beta_2, \gamma) = L_1(\beta_1, \gamma) + L_2(\beta_2, \gamma)$ où les fonctions L_1 et L_2 sont définies en (3.4) et (3.5). Une maximisation directe de $L(\beta_1, \beta_2, \gamma)$ est complexe, puisque les méthodes d'optimisation convexes nécessite souvent des conditions de régularité sur la fonction L. Par exemple, pour la méthode de Newton-Raphson, L aurait besoin d'être $\mathcal{C}^2$ sur $\mathbb{R}^2 \times [0, \tau]$, ce qui n'est pas le cas à cause de la variable γ.

On peut cependant estimer $\beta 01$ et β_{02} pour chaque valeur possible du changepoint γ, en supposant par exemple que ce dernier ne peut avoir lieu que sur un temps de décès. A γ fixé, en maximisant $L_1(\beta_1, \gamma)$ d'une part et $L_2(\beta_2, \gamma)$ d'autre part, on obtient respectivement un processus $\hat{\beta}_{01}(\gamma)$ d'une part et $\hat{\beta}_{02}(\gamma)$ d'autre part vérifiant

$$\hat{\beta}_{01}(\gamma) = \arg\max_{\beta_1} L_1(\beta_1, \gamma),$$
$$\hat{\beta}_{02}(\gamma) = \arg\max_{\beta_2} L_2(\beta_2, \gamma).$$

Enfin, le triplet $(\hat{\beta}_{01}, \hat{\beta}_{02}, \hat{\gamma}_0)$ choisi est celui défini par la relation

$$(\hat{\beta}_{01}, \hat{\beta}_{02}, \hat{\gamma}_0) = (\hat{\beta}_{01}(\hat{\gamma}_0), \hat{\beta}_{02}(\hat{\gamma}_0), \hat{\gamma}_0) = \arg\max_{\gamma} L(\hat{\beta}_{01}(\gamma), \hat{\beta}_{02}(\gamma), \gamma).$$

Autrement dit, parmi tous les triplets $(\hat{\beta}_{01}, \hat{\beta}_{02}, \gamma)$ où $\gamma \in \{X_i \ / \ i \in \{1, \ldots, n\},\ \Delta_i = 1\}$, on choisit celui maximisant la vraisemblance totale L.

L'étape d'estimation étant maintenant établie, nous nous intéressons dans la partie suivante à l'obtention d'un intervalle de confiance pour le changepoint γ_0.

3.3.2 Une région de confiance

Les travaux de Davies (1977) sont utilisés pour la première fois en analyse de survie par O'Quigley and Pessione (1991). Ces derniers s'intéresse au modèle

$$\beta_0(t) = \beta_0 \mathbb{1}_{t\leq\gamma_0} - \beta_0 \mathbb{1}_{t>\gamma_0}. \tag{3.6}$$

Il cherche à effectuer un test où l'hypothèse nulle est "$\beta_0 = 0$" contre l'alternative "$\beta_0 > 0$". Ils proposent alors un test basé sur la statistique

$$M = \sup\{|S(\gamma)| \ / \ 0 \leq \gamma \leq \tau\},$$

où

$$S(\gamma) = \left\{\frac{\partial L}{\partial \beta}(\beta,\gamma)\right\}_{\beta=0} \left\{-\frac{\partial^2 L}{\partial \beta^2}(\beta,\gamma)\right\}_{\beta=0}^{-1/2},$$

avec $L(\beta,\gamma)$ la log-vraisemblance partielle associée au modèle (3.6). Nous nous inspirons de la forme de ce test pour construire un intervalle de confiance pour γ_0, dans le modèle réduit (3.2), à partir des statistiques

$$M_1 = \sup\{S_1(\gamma) \ / \ 0 \leq \gamma \leq \tau\}, \tag{3.7}$$
$$M_2 = \sup\{S_2(\gamma) \ / \ 0 \leq \gamma \leq \tau\}. \tag{3.8}$$

Soit $z \in \mathbb{R}$. Nous détaillons ici l'obtention d'une région de confiance à partir de la statistique M_1. Le cas de la statistique M_2 est tout à fait similaire. Définissons la quantité $q_\alpha(z,\gamma_0)$ par

$$P_{\beta_{01},\gamma_0}\left(M_1 = \sup_\gamma S_1(\gamma) > z + q_\alpha(z,\gamma_0) \middle| S_1(\gamma_0) = z\right) = \alpha. \tag{3.9}$$

Alors une région de confiance de niveau $1-\alpha$ pour γ_0 est

$$\{\gamma \ / \ S_1(\gamma) > M_1 - q_\alpha(S_1(\gamma_0),\gamma_0)\}. \tag{3.10}$$

Bien sûr, (3.10) n'est pas directement exploitable. C'est pourquoi nous effectuons un travail de réécriture du membre gauche dans (3.9), pour centrer et réduire $S_1(\gamma)$ conditionnellement à $S_1(\gamma_0) = z$. On obtient alors

$$P_{\beta_{01},\gamma_0}\left(\sup_\gamma \left[\frac{S_1(\gamma) - z\rho_1(\gamma,\gamma_0)}{(1-\rho_1(\gamma,\gamma_0)^2)^{1/2}} - \frac{z + q_\alpha(z,\gamma_0) - z\rho_1(\gamma,\gamma_0)}{(1-\rho_1(\gamma,\gamma_0)^2)^{1/2}}\right] > 0 \middle| S_1(\gamma_0) = z\right). \tag{3.11}$$

On peut remarquer que, quand $\rho_1(\gamma,\gamma_0)$ est proche de 1, le second terme dans (3.11) tend vers l'infini. Donc on peut ne s'intéresser qu'à des valeurs de γ pour lesquelles $\rho_1(\gamma,\gamma_0)$ est proche de 0. On peut alors faire l'approximation que le premier terme de (3.11) est indépendant de γ, avec toutes fois un changement de signe en γ_0. On approche donc ce terme par $\mathcal{N}_0 \text{sgn}(\gamma - \gamma_0)$, où $\mathcal{N}_0$ est une variable gaussienne centrée réduite. La quantité (3.11) devient

$$P\left(\mathcal{N}_0^2 > \{z + q_\alpha(z,\gamma_0)\}^2 - z^2\right).$$

On choisit alors $q_\alpha(z,\gamma_0)$ tel que $\{z + q_\alpha(z,\gamma_0)\}^2 - z^2 = \chi^2_{1,\alpha}$, où $\chi^2_{1,\alpha}$ est le quantile de niveau α d'un loi du χ^2 à 1 degré de liberté. On obtient la région de confiance de niveau $1-\alpha$ pour le changepoint γ_0

$$\left\{\gamma \ / \ S_1(\gamma)^2 > M_1^2 - \chi^2_{1,\alpha}\right\}. \tag{3.12}$$

3.4 Etude du modèle général

3.4.1 Processus du score standardisé

On s'intéresse dans cette section au modèle général (3.3). Avant d'aller plus avant dans l'étude de l'estimation des $K-1$ points de rupture γ_{0i} et des K constantes de régression β_{0i}, nous rappelons quelques notations et résultats sur le processus du score standardisé.

On note $N_i(t) = \mathbb{1}_{T_i \leq t, T_i \leq C_i}$ le processus de comptage et on pose $\bar{N}(t) = \sum_{i=1}^n N_i(t)$. Soit $t \in [0, \tau]$, on définit l'espérance et la variance de Z par rapport à la famille de probabilité $(\pi_i(\beta(t), t))_{i \in \{1,\dots,n\}}$, où

$$\pi_i(\beta(t), t) = \frac{Y_i(t) \exp(\beta(t) Z_i(t))}{\sum_{j=1}^n Y_j(t) \exp(\beta(t) Z_j(t))},$$

par

$$\mathcal{E}_{\beta(t)}(Z|t) = \sum_{i=1}^n Z_i(t) \pi_i(\beta(t), t), \;\; \mathcal{V}_{\beta(t)}(Z|t) = \sum_{i=1}^n Z_i^2(t) \pi_i(\beta(t), t) - \mathcal{E}_{\beta(t)}(Z|t)^2.$$

Le processus du score a été introduit par Wei (1984) dans le cas d'une covariable Z binaire. Le processus du score $U(\beta, t)$ évalué au temps $t \in [0, \tau]$ pour la fonction de régression β est défini par

$$U(\beta, t) = \int_0^t \{Z_i(s) - \mathcal{E}_{\beta(s)}(Z|s) d\bar{N}(s).$$

On remarque qu'à chaque temps de décès, le processus est incrémenté de la différence entre la valeur de la covariable de l'individu qui décède et son espérance sous le modèle. Au dernier temps de décès, le processus est égal à la dérivée de la log–vraisemblance. Wei (1984) a démontré la convergence de ce processus vers un pont Brownien lorsque le β est l'estimateur du maximum de vraisemblance partielle. Haara (1987) a étendu ce résultat aux cas plus généraux de covariables non binaires.

Nous souhaitons maintenant introduire le processus du score standardisé proposé dans Chauvel and O'quigley (2014) avec de légères modifications. Pour cela, nous effectuons au préalable un changement d'échelle du temps. Posons $\hat{k}_n = \sharp\{i \;/\; i \in \{1, \dots, n\}, \; \Delta_i = 1\}$, où $\sharp A$ est le cardinal de l'ensemble A, *i.e.*, $\hat{k}_n$ est le nombre de décès de l'étude. D'après la loi forte des grands nombres,

$$\frac{\hat{k}_n}{n} \xrightarrow{p.s.} \alpha_0 = E[\Delta_1] = \mathbb{P}(T \leq C).$$

On suppose que $\alpha_0 > 0$, ce qui est raisonnable puisqu'une étude ne comporte jamais uniquement des patients censurés. De plus, la loi du logarithme itéré nous fournit une vitesse de convergence de $\hat{k}_n/n$ vers α_0. En effet, pour tout $\varepsilon' > 0$, à partir d'un certain rang, presque sûrement,

$$-(1+\varepsilon') \frac{\sqrt{2\alpha_0(1-\alpha_0) \log\log n}}{\sqrt{n}} \leq \frac{\hat{k}_n}{n} - \alpha_0 \leq (1+\varepsilon') \frac{\sqrt{2\alpha_0(1-\alpha_0) \log\log n}}{\sqrt{n}}. \tag{3.13}$$

Grâce aux inéquations (3.13), on peut donc, pour n'importe quel ε_0, déterminer un N tel que, pour tout $n \geq N$,

$$\left| \frac{\hat{k}_n}{n} - \alpha_0 \right| \leq \varepsilon_0. \tag{3.14}$$

$\hat{k}_n$ est, par définition, une variable aléatoire. Pour pouvoir effectuer une étude théorique sans perdre en performance pratique, nous allons travailler avec une suite $(k_n)_{n\in\mathbb{N}^*}$ déterministe ayant un comportement proche de $\hat{k}_n$ mais plus simple à utiliser. Soit $0 < \varepsilon < \alpha_0$ fixé. On pose maintenant $k_n = \lfloor n(\alpha_0 - \varepsilon) \rfloor \in \mathbb{N}^*$, où $\lfloor x \rfloor$ désigne la partie entière du réel x. On a alors la convergence $k_n/n \to \alpha_0 - \varepsilon$ et donc, pour tout ε_1, à partir d'un certain rang, $|k_n/n - (\alpha_0 - \varepsilon)| \leq \varepsilon_1$. En choisissant $\varepsilon_0 = \varepsilon/2$ et $\varepsilon_1 = \varepsilon/2$, on a, à partir d'un certain rang N, d'après (3.14)

$$\frac{\hat{k}_n}{n} \in [\alpha_0 - \frac{\epsilon}{2}, \alpha_0 + \frac{\epsilon}{2}] \text{ p.s., et } \frac{k_n}{n} \in [\alpha_0 - \frac{3\epsilon}{2}, \alpha_0 - \frac{\epsilon}{2}].$$

Ainsi, pour tout $n \geq N$, $k_n \leq \hat{k}_n$ presque sûrement et k_n est à valeurs entières comme $\hat{k}_n$ et possède bien un comportement proche de celle-ci asymptotiquement. Effectuons maintenant le changement d'échelle proposé dans Chauvel and O'quigley (2014) en posant,

$$\phi_n(X_i) = \frac{\bar{N}(X_i)}{k_n}\left[1 + (1-\Delta_i)\frac{\sharp\{j \ / \ j \in \{1,\dots,n\},\ X_j < X_i,\ \bar{N}(X_j) = \bar{N}(X_i)\}}{\sharp\{j \ / \ j \in \{1,\dots,n\},\ \bar{N}(X_j) = \bar{N}(X_i)\}}\right]. \tag{3.15}$$

Grâce à ce changement, les temps $\{0, 1/k_n, 2/k_n \dots, 1\}$ correspondent à des temps de décès et les temps de censure sont répartis uniformément entre les temps de décès (en respectant l'ordre original). Par exemple, si $T_1 < C_2 < C_3 < T_4$, alors $\phi_n(T_1) < \phi_n(C_2) < \phi_n(C_3) < \phi_n(T_4)$, et $\phi_n(C_2)$ et $\phi_n(C_3)$ sont répartis uniformément entre $\phi_n(T_1)$ et $\phi_n(T_4)$. On peut définir toutes les quantités utiles dans cette échelle. Pour ne pas confondre le travail dans l'échelle de départ et celui dans la nouvelle échelle, on notera x^* la quantité x dans la nouvelle échelle (3.15). Donc, pour tout $t \in [0,1]$ et tout $i \in \{1,\dots,n\}$, $Y_i^*(t) = \mathbb{1}_{\phi_n(X_i)\leq t}$, $N_i^*(t) = \mathbb{1}_{\phi_n(X_i)\leq t, \Delta_i = 1}$ et

$$\bar{N}^*(t) = \sum_{i=1}^{n} \mathbb{1}_{\phi_n(X_i)\leq t, \Delta_i = 1}.$$

Nous avons maintenant tous les outils pour définir le processus du score standardisé.

Définition 3.1 (Processus du score standardisé)**.** Le processus du score standardisé $U^*(\beta(t), t)$ évalué au temps $t \in \{0, 1/k_n, 2/k_n, \dots, 1\}$ pour la fonction de régression β est défini par

$$U^*(\beta(t), t) = \frac{1}{\sqrt{k_n}} \int_0^t \mathcal{V}_{\beta(s)}(Z|s)^{-1/2}\{\mathcal{Z}(s) - \mathcal{E}_{\beta(s)}(Z|s)\} d\bar{N}^*(s). \tag{3.16}$$

Il est ensuite défini sur tout le segment $[0,1]$ par interpolation linéaire.

La différence avec celui de Chauvel and O'quigley (2014) se situe dans l'utilisation de k_n à la place de $\hat{k}_n$. L'utilité du caractère déterministe de k_n dans la preuve du théorème suivant est bien expliqué dans Chauvel (2014). C'est sur ce théorème que nous nous appuierons pour la détection de changepoints. Les hypothèses sont détaillées juste après.

Théorème 3.2. *Pour tout $t \in [0,1]$, sous le modèle (2.1) et sous les hypothèses* **H1-5**, *il existe des constantes strictement positives $C_1(\beta_0)$ et C_2 telles que*

$$U^*(0,t) - \sqrt{k_n} C_2 \int_0^t \beta_0(s) ds \xrightarrow[n\to\infty]{P} C_1(\beta_0) W,$$

où W désigne un mouvement brownien standard.

On note $D([0,1],\mathbb{R})$ l'espace des fonctions continues à droite avec limite à gauche et on le munit de la topologie de la convergence uniforme. On définit maintenant, pour tout $r \in \{0,1,2\}$, l'équivalent de $S^{(r)}(\beta(t),t)$, $t \in [0,\tau]$ introduit à l'équation (2.2) dans la nouvelle échelle. Pour tout $t \in [0,1]$,

$$S^{(r)}(\beta(t),t) = \frac{1}{n}\sum_{i=1}^{n} Y_i^*(t) Z_i \left(\phi_n^{-1}(t)\right)^r \exp\left(\beta(t) Z_i \left(\phi_n^{-1}(t)\right)\right).$$

Énonçons maintenant les hypothèses de Chauvel (2014) suffisant à montrer le Théorème 3.2. On rappelle qu'on se place sous le modèle (2.1)

Hypothèses. **(H1)** *(Stabilité asymptotique) Il existe $\delta_1 > 0$, un voisinage de β_0 de rayon δ_1 contenant la fonction nulle, noté $\mathbb{B} = \{\beta,\ \sup_{t\in[0,1]} |\beta(t) - \beta_0(t)| < \delta_1\}$, et des fonctions $s^{(r)}$ définies sur $\mathbb{B}\times[0,1]$, pour $r \in \{0,1,2\}$, telles que*

$$\sqrt{n} \sup_{t\in[0,1],\ \beta\in\mathbb{B}} \left| S^{(r)}(\beta(t),t) - s^{(r)}(\beta(t),t) \right| \xrightarrow[n\to\infty]{P} 0.$$

(H2) *(Régularité asymptotique) Les fonctions déterministes $s^{(r)}$, définies en* **H1** *sont uniformément continues en $t \in [0,1]$ et bornées sur $\mathbb{B}\times[0,1]$. De plus, pour $r \in \{0,1,2\}$, et $t \in [0,1]$, $s^{(r)}(.,t)$ est continue sur $\mathbb{B}$. La fonction $s^{(0)}$ est minorée par une constante strictement positive.*

On définit, pour tout $t \in [0,1]$ et tout $\beta \in \mathbb{B}$, les quantités suivantes.

$$e(\beta(t),t) = \frac{s^{(1)}(\beta(t),t)}{s^{(0)}(\beta(t),t)},$$

et

$$v(\beta(t),t) = \frac{s^{(2)}(\beta(t),t)}{s^{(0)}(\beta(t),t)} - e(\beta(t),t)^2.$$

(H3) *(Homoscédasticité) Pour tout $t \in [0,1]$ et $\beta \in \mathbb{B}$, $\frac{\partial}{\partial t} v(\beta(t),t) = 0$.*

(H4) *(Covariables uniformément bornées) Il existe $L \in \mathbb{R}_+^*$ tel que*

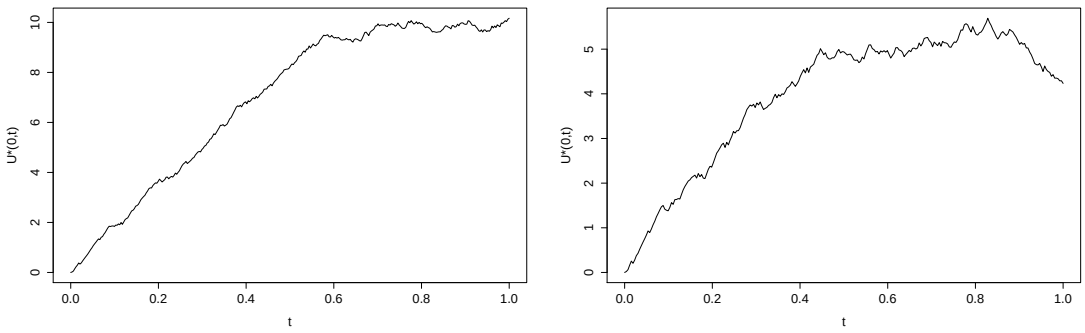
$$\sup_{i\in\{1,\dots,n\}} \sup_{t\in[0,\tau]} |Z_i(t)| \leq L.$$

(H5) *(Non dégénérescence de la variance) Il existe une constante $C_{\mathcal{V}}$ telle que, pour tout $i \in \{1,\dots,n\}$ vérifiant $\Delta_i = 1$, $\mathcal{V}_0(Z|X_i) > C_{\mathcal{V}}$.*

Les hypothèses **H1-2** sont introduites par Andersen and Gill (1982). L'hypothèse **H3** est souvent rencontrée dans l'utilisation de modèles à risques proportionnels de manière implicite, pour l'estimation de la variance du paramètre β_0 ou l'expression de la statistique du log-rank par exemple. On peut remarquer que l'hypothèse **H5** porte uniquement sur les temps de décès. La preuve du Théorème 3.2 nécessite que **H5** soit vérifiée pour k_n temps de décès, ce qui est bien le cas quand **H5** est réalisée puisque, par définition de k_n, $k_n \leq \hat{k}_n$ presque sûrement, à partir d'un certain rang.

3.4.2 Lien avec la détection de points de rupture

Commençons par quelques illustrations du processus (3.16) pour mieux comprendre le Théorème 3.2 et l'utilisation que nous allons en faire pour la détection de points de rupture. La Figure 3.1 présente le processus du score standardisé dans deux cas, tous les deux des cas particuliers du modèle général (3.3). Dans les deux situations, Z suit une loi uniforme sur $[0,1]$, C suit une loi uniforme sur $[0,t_c]$ où t_c est choisi pour obtenir à peu près 30% de censure dans un jeu de données de $n = 500$ observations, et T suit le modèle (2.1) pour $\lambda_0(t) = 1$, pour tout $t \in [0,\tau]$. Pour la première situation, la fonction de régression β_0 vérifie $\beta_0(t) = 3\mathbb{1}_{t\leq 0.1}$ pour tout $t \in [0,\tau]$. Pour le second, $\beta_0(t) = 2\mathbb{1}_{t\leq 0.1} - \mathbb{1}_{t>0.4}$ pour tout $t \in [0,\tau]$.



(a) $\beta_0(t) = 3\mathbb{1}_{t\leq 0.1}$ (b) $\beta_0(t) = 2\mathbb{1}_{t\leq 0.1} - \mathbb{1}_{t>0.4}$

FIGURE 3.1 – Processus du score standardisé

On observe, en Figure 3.1, la dérive annoncé par le Théorème 3.2, qui est l'intégrale de la fonction de régression β_0. On voit bien alors l'intérêt du processus du score standardisé pour la détection de points de rupture. En effet, si la fonction de régression suit le modèle général (3.3), alors le processus du score standardisé (3.16) évalué en la fonction $\beta(t) = 0$ pour tout $t \in [0,\tau]$ est affine par morceaux, d'après le Théorème 3.2. Il suffit alors, pour trouver les points de rupture γ_{0i}, d'effectuer une régression linéaire par morceaux. On obtient ensuite les constantes de régression β_{0i} classiquement avec l'estimateur de la vraisemblance partielle sur chaque morceau. On rappelle que la convergence de l'estimateur de la vraisemblance partielle vers la vraie constante de régression sous un modèle à hasards proportionnels à été montrée par Andersen and Gill (1982)

3.4.3 Modèle linéaire par morceaux

Nous donnons ici quelques références sur l'estimation des points de rupture dans le cas de modèles linéaires par morceaux. Rappelons la forme du modèle de régression linéaire classique

$$y_i = x_i^T \beta + u_i. \tag{3.17}$$

Dans beaucoup d'applications, comme celle que nous étudions ici, il est raisonnable de supposer qu'il y a m points de rupture et donc $(m+1)$ segments sur lesquels les coefficients de régression sont constants. On peut alors réécrire le modèle (3.17) sous la forme suivante

$$y_i = x_i^T \beta_j + u_i \ \ (i \in \{i_{j-1}+1,\dots,i_j\},\ j \in \{1,\dots,m+1\}), \tag{3.18}$$

où j est le numéro du segment. Bai (1994) a donnée les bases de l'estimation des points de rupture dans les séries temporelles. Elles ont été étendues à d'autres types de points de rupture par Bai (1997); Liu et al. (1997); Hawkins (2001); Sullivan (2002) et Bai and Perron (2003) par exemple. Le package R `strucchange` a été proposé par Kleiber et al. (2002). Les idées derrière l'algorithme, `breakpoints`, pour l'estimation de ces points de rupture sont détaillés dans Zeileis et al. (2003) et sont basées sur la minimisation de la somme des carrés des résidus de l'équation (3.18). C'est ce package que nous utilisons dans les simulations présentées en Section 3.5.

On peut noter que, dans les articles cités, on s'intéresse à l'estimation des points de rupture, mais aussi à l'estimation des coefficients de régression. Cette dernière ne nous est cependant d'aucune utilité puisque la dérive annoncée dans le Théorème 3.2 est réalisée à une constante près, notée C_2 dans son énoncé, qu'on ne peut pas déterminer. L'estimation par moindres carrés peut donc gérer les changepoints, mais il faut revenir à la vraisemblance partielle pour l'estimation des coefficients de régression β_{0i} dans le modèle général (3.3).

3.5 Simulations

3.5.1 Modèle réduit

Nous étudions maintenant, dans un premier temps, le comportement du niveau de la région de confiance 3.12 en fonction de la taille de l'échantillon n, des lois de C et Z, et du modèle sur la fonction de régression β_0. Comme énoncé dans la Section 3.3, on peut choisir d'utiliser les statistiques M_1 (3.7) ou M_2 (3.8) pour déterminer une région de confiance pour γ_0. Cependant, ils nous est apparu pendant les simulations que le choix du coefficient de régression le plus élevé fournissait de meilleurs résultats. En pratique, on choisira donc plutôt M_1 si $\hat{\beta}_{01} > \hat{\beta}_{02}$, et M_2 si $\hat{\beta}_{02} > \hat{\beta}_{01}$. C'est ce que nous faisons dans l'ensemble de ces simulations. La taille de l'échantillon est fixée successivement à 500 et 1000 pour que les statistiques utilisées ne dévient pas trop de leurs approximations gaussiennes. La loi de C est une exponentielle de paramètre μ, où μ est calibré pour que le pourcentage de censure soit de 30%, 50% ou 70%. La covariable $Z \in \mathbb{R}$ suit une loi de Bernoulli $\mathcal{B}er$ de paramètre 1/2, une loi uniforme $\mathcal{U}$ sur $[0,1]$, une loi gaussienne $\mathcal{N}$ d'espérance 1/2 et de variance 1/4, ou une loi exponentielle $\mathcal{E}$ de paramètre 1/2. Remarquons que les résultats énoncés en Section 3.3 sont établis pour des variables à support dans un segment $[0,\tau]$. Cependant, il arrive que ce ne soit pas le cas dans les applications pratiques. C'est pourquoi ces scénarii sont également intéressants à étudier dans nos simulations. On considère pour la fonction de régression β_0 les trois modèles suivants : $\beta_0(t) = \mathbb{1}_{t\leq 0.3}$, $\beta_0(t) = \mathbb{1}_{t\leq 0.5}$ et $\beta_0(t) = \mathbb{1}_{t\leq 0.7}$. Pour chaque scénario ainsi créé, 1000 échantillons sont générés pour évaluer le niveau empirique de la région de confiance du changepoint. Ces régions de confiance sont calibrées pour un niveau de 10%. On constate avec les Tables 3.1 et 3.2 que le test se comporte mieux pour des covariables continues. On remarque aussi une légère amélioration du niveau empirique quand la censure diminue.

TABLE 3.1 – Niveaux empiriques des régions de confiance du changepoint (en %) pour une covariable à support fini

		$\beta_0(t) = \mathbb{1}_{t\leq 0.3}$		$\beta_0(t) = \mathbb{1}_{t\leq 0.5}$		$\beta_0(t) = \mathbb{1}_{t\leq 0.7}$	
n	% censure	$Z \sim \mathcal{B}er$	$Z \sim \mathcal{U}$	$Z \sim \mathcal{B}er$	$Z \sim \mathcal{U}$	$Z \sim \mathcal{B}er$	$Z \sim \mathcal{U}$
500	0	10.4	13.3	9.8	13.4	11.5	12.9
500	30	12.6	13.2	11.3	13.7	12.6	13.0
500	50	12.9	13.6	13.1	13.8	12.7	13.5
1000	0	10.3	1.7	9.4	7.5	10.5	2.9
1000	30	11.8	1.4	9.7	7.6	11.8	2.8
1000	50	13.2	1.9	9.9	8.4	13.4	3.4

TABLE 3.2 – Niveaux empiriques des régions de confiance du changepoint (en %) pour une covariable à support infini

		$\beta_0(t) = \mathbb{1}_{t\leq 0.3}$		$\beta_0(t) = \mathbb{1}_{t\leq 0.5}$		$\beta_0(t) = \mathbb{1}_{t\leq 0.7}$	
n	% censure	$Z \sim \mathcal{N}$	$Z \sim \mathcal{E}$	$Z \sim \mathcal{N}$	$Z \sim \mathcal{E}$	$Z \sim \mathcal{N}$	$Z \sim \mathcal{E}$
00	0	0.4	0.6	0.3	10.4	0.8	12.4
500	30	0.5	0.4	0.7	11.8	1.4	13.3
500	50	0.3	0.7	0.9	12.8	1.5	13.1
1000	0	0.2	0.7	0.6	9.9	2.4	8.4
1000	30	0.1	0.6	1.2	10.1	3.7	8.3
1000	50	0.4	0.7	0.8	10.5	4.2	9.3

On peut, dans un second temps, s'intéresser au comportement de la région de confiance (3.12) en fonction de différents paramètres. On commence par regarder son évolution en fonction de la distance entre les coefficients de régression, *i.e.*, en fonction de $|\beta_{01}-\beta_{02}|$. Une illustration des résultats obtenues est présentée en Figure 3.2. Pour obtenir ces graphiques, nous avons simulé des jeux de données de taille $n = 1000$ suivant le modèle réduit (3.2) avec $\beta_{02} = 0$, β_{01} valant successivement 0.5, 1, 1.5 et 2, et γ_0 valant successivement 0.5, 0.4, 0.3 et 0.2. Z suit une loi uniforme sur $[0, 1]$, C une loi uniforme sur $[0, 2]$ et $\lambda_0(t) = 1$ pour tout $t \in [0, \tau]$. On obtient les estimations suivantes :

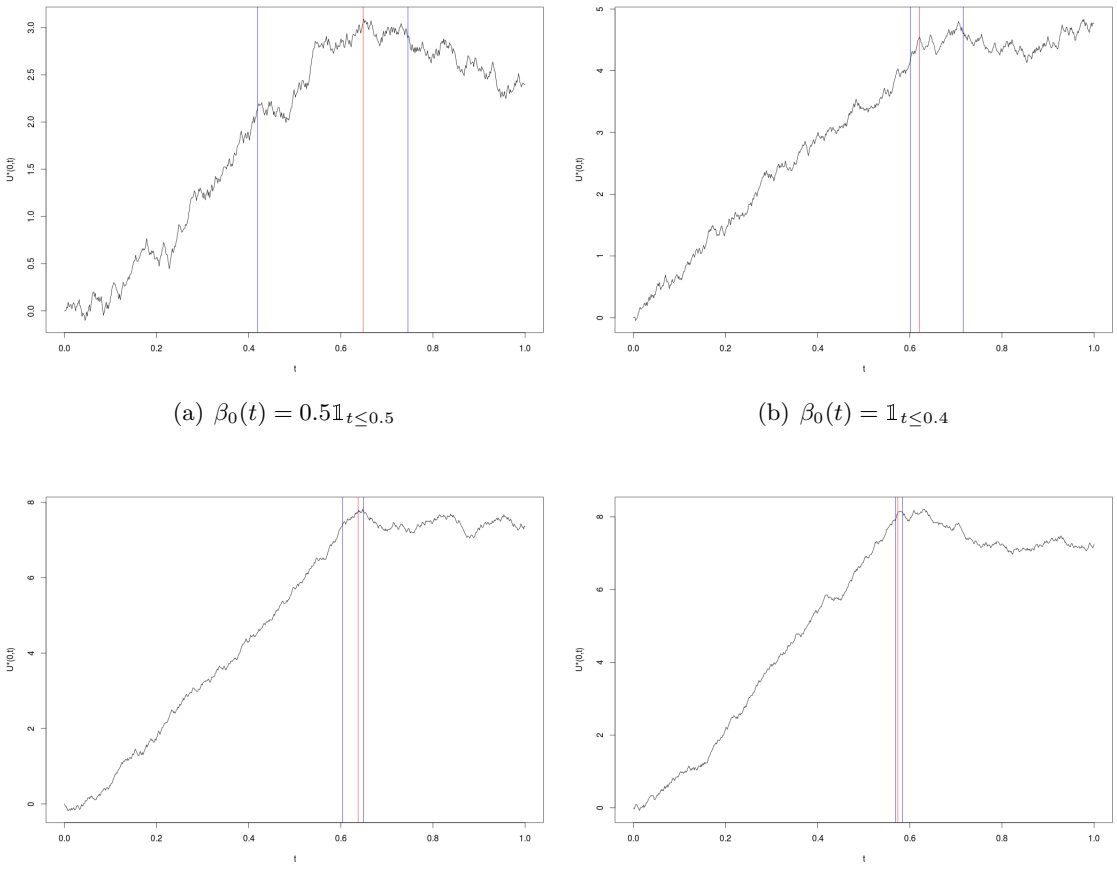
— Pour le modèle $\beta_0(t) = 0.5\mathbb{1}_{t\leq 0.5}$, on trouve $\hat{\gamma}_0 = 0.429$ et $IC_{95\%} = [0.233, 0.563]$.

— Pour le modèle $\beta_0(t) = \mathbb{1}_{t\leq 0.4}$, on trouve $\hat{\gamma}_0 = 0.328$ et $IC_{95\%} = [0.310, 0.461]$.

— Pour le modèle $\beta_0(t) = 1.5\mathbb{1}_{t\leq 0.3}$, on trouve $\hat{\gamma}_0 = 0.290$ et $IC_{95\%} = [0.275, 0.304]$.

— Pour le modèle $\beta_0(t) = 2\mathbb{1}_{t\leq 0.2}$, on trouve $\hat{\gamma}_0 = 0.198$ et $IC_{95\%} = [0.196, 0.206]$.

Avec ces résultats, illustrés en Figure 3.2, on constate sans surprise que la longueur de l'intervalle de confiance au niveau 95% de γ_0 diminue quand la distance entre β_{01} et β_{02} augmente. En effet, il est raisonnable de penser que, plus cette distance est grande, plus il est facile de trouver le changepoint du modèle.



(a) $\beta_0(t) = 0.5\mathbb{1}_{t\leq 0.5}$ (b) $\beta_0(t) = \mathbb{1}_{t\leq 0.4}$

(c) $\beta_0(t) = 1.5\mathbb{1}_{t\leq 0.3}$ (d) $\beta_0(t) = 2\mathbb{1}_{t\leq 0.2}$

Figure 3.2 – Evolution de l'intervalle de confiance à 95% sur γ_0 en fonction de la distance $|\beta_{01}-\beta_{02}|$. Tracé du processus du score standardisé (en noir), de l'estimateur $\hat{\gamma}_0$ (en rouge) et des bornes de l'intervalle de confiance à 95% (en bleu)

3.5.2 Modèle général

Nous nous intéressons maintenant à l'estimation de plusieurs changepoints, dans le cas du modèle général (3.3). Nous commençons par l'étude de quelques modèles réduits (3.2) et comparons les précisions d'estimation par méthode de Anderson and Senthilselvan (1982), *i.e.* par maximisation de la vraisemblance partielle, et méthode de moindres carrés, introduites en Section 3.4. Nous étudions en suite la précision de l'estimation par moindres carrés dans le cas de modèles à plusieurs changepoints. Nous rappelons que l'estimation par moindres carrés s'effectue à l'aide du package `strucchange` (Kleiber et al., 2002) du logiciel R.

Pour l'étude des modèles réduits, nous choisissons une loi exponentielle pour la censure C, de paramètre calibré pour obtenir un pourcentage de censure de 30% ou 50%. Le hasard de base λ_0 est identiquement égal à 1. La covariable Z suit une loi uniforme sur $[0, 1]$. Enfin, les modèles sur β_0 sont $\beta_0(t) = 0.5\mathbb{1}_{t\leq 0.5}$, $\beta_0(t) = \mathbb{1}_{t\leq 0.4}$ et $\beta_0(t) = 2\mathbb{1}_{t\leq 0.3}$. On obtient les résultats présentés en Table 3.3. 100 échantillons de taille $n = 1000$ nous permettent d'obtenir un estimateur moyen et un écart-type pour chaque cas. Dans la Table 3.3, "PL" signifie "Vraisemblance partielle" (partial likelihood) et "LS" signifie "Moindres carrés" (least squares).

TABLE 3.3 – Comparaison des méthodes du maximum de vraisemblance partielle et des moindres carrés pour l'estimation d'un changepoint. Ecarts-type entre parenthèses.

Modèle	% censure	PL	LS
$\beta_0(t) = 0.5\mathbb{1}_{t\leq 0.5}$	0	0.648 (0.379)	0.576 (0.256)
$\beta_0(t) = 0.5\mathbb{1}_{t\leq 0.5}$	30	0.689 (0.440)	0.468 (0.218)
$\beta_0(t) = 0.5\mathbb{1}_{t\leq 0.5}$	50	0.670 (0.497)	0.407 (0.163)
$\beta_0(t) = \mathbb{1}_{t\leq 0.4}$	0	0.506 (0.280)	0.432 (0.136)
$\beta_0(t) = \mathbb{1}_{t\leq 0.4}$	30	0.438 (0.234)	0.386 (0.092)
$\beta_0(t) = \mathbb{1}_{t\leq 0.4}$	50	0.494 (0.341)	0.329 (0.130)
$\beta_0(t) = 2\mathbb{1}_{t\leq 0.3}$	0	0.307 (0.025)	0.313 (0.046)
$\beta_0(t) = 2\mathbb{1}_{t\leq 0.3}$	30	0.313 (0.043)	0.304 (0.044)
$\beta_0(t) = 2\mathbb{1}_{t\leq 0.3}$	50	0.324 (0.138)	0.295 (0.091)

On constate, avec la Table 3.3, que la méthode des moindres carrés semble avoir de meilleures performances que celle par vraisemblance partielle. En effet, la méthode par moindres carrés perd très peu d'efficacité quand le pourcentage de censure augmente et semble mieux estimer le changepoint en moyenne. Comme on pouvait s'y attendre pour les deux méthodes, plus la distance entre les valeurs des coefficients de régression de chaque côté du changepoint est grande, plus l'efficacité s'améliore, c'est-à-dire plus les méthodes estiment précisément le changepoint.

On s'intéresse maintenant à l'évolution de la précision de l'estimation des changepoints par moindres carrés si on augmente le nombre de changepoints dans le modèle (3.3). Pour cela nous faisons varier le taux de censure de 0 à 50% comme précédemment. La loi de Z et le hasard de base restent les mêmes. La taille de l'échantillon n prend les valeurs 200, 500 et 1000. On génère 100 échantillons qui nous permettent d'obtenir une moyenne et un écart-type pour les estimateurs des changepoints. Les modèles étudiés sont listés ci-dessous. Pour une meilleure visualisation de ces derniers, nous avons tracé, en Figure 3.3, le processus du score standardisé pour ces trois modèles, avec une censure de 30% et un échantillon de taille 1000, en insistant sur les points de rupture.

— **Scénario 1** $\beta_0(t) = \mathbb{1}_{t\leq 0.2} - \mathbb{1}_{t>0.6}$.

— **Scénario 2** $\beta_0(t) = -\mathbb{1}_{t\leq 0.5} + 0.5\mathbb{1}_{1.1<t\leq 2.4} + \mathbb{1}_{t>2.4}$.

— **Scénario 3** $\beta_0(t) = 2\mathbb{1}_{t\leq 0.1} - \mathbb{1}_{0.2<t\leq 0.3} + 1.5\mathbb{1}_{t>0.6}$.

Les résultats sont présentés en Table 3.4 pour le **Scénario 1**, Table 3.5 pour le **Scénario 2** et Table 3.6 pour le **Scénario 3**. On constate encore une fois que la précision de l'estimation de dépend pas ou peu de la censure. On remarque aussi qu'elle augmente entre une taille de l'échantillon de 200 à 500, mais que la différence de précision est minime quand on passe de 500 à 1000 observations. Comme on peut le voir sur la Figure 3.3 (b), le dernier changepoint se situe à la limite du jeu de données. C'est pourquoi, dans la Table 3.5 on constate une mauvaise estimation du dernier changepoint γ_3.

Table 3.4 – Evolution de la précision de l'estimation des changepoints en fonction de leur nombre dans le **Scénario 1**. Ecarts-type entre parenthèses.

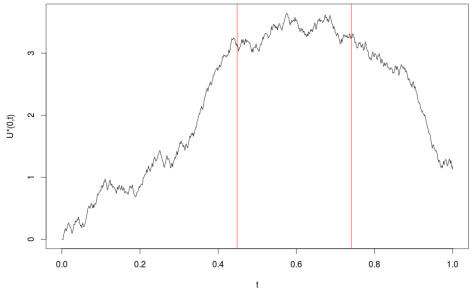
		Scénario 1	
n	% censure	γ_1	γ_2
200	0	0.266 (0.168)	1.179 (0.633)
200	30	0.171 (0.094)	0.694 (0.370)
200	50	0.128 (0.075)	0.417 (0.211)
500	0	0.243 (0.138)	0.672 (0.255)
500	30	0.183 (0.104)	0.665 (0.132)
500	50	0.135 (0.071)	0.434 (0.188)
1000	0	0.218 (0.066)	0.751 (0.276)
1000	30	0.200 (0.066)	0.681 (0.241)
1000	50	0.143 (0.061)	0.643 (0.169)

Table 3.5 – Evolution de la précision de l'estimation des changepoints en fonction de leur nombre dans le **Scénario 2**. Ecarts-type entre parenthèses.

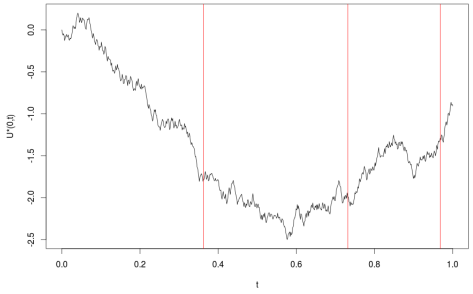
		Scénario 2		
n	% censure	γ_1	γ_2	γ_3
200	0	0.407 (0.198)	0.901 (0.280)	1.566 (0.429)
200	30	0.299 (0.160)	0.703 (0.198)	1.269 (0.327)
200	50	0.185 (0.107)	0.452 (0.166)	0.857 (0.222)
500	0	0.428 (0.141)	0.880 (0.228)	1.604 (0.433)
500	30	0.309 (0.144)	0.679 (0.211)	1.193 (0.272)
500	50	0.248 (0.121)	0.559 (0.155)	1.026 (0.262)
1000	0	0.423 (0.138)	0.864 (0.257)	1.589 (0.454)
1000	30	0.305 (0.149)	0.641 (0.171)	1.174 (0.268)
1000	50	0.202 (0.114)	0.483 (0.135)	0.875 (0.225)

TABLE 3.6 – Evolution de la précision de l'estimation des changepoints en fonction de leur nombre dans le **Scénario 3**. Ecarts-type entre parenthèses.

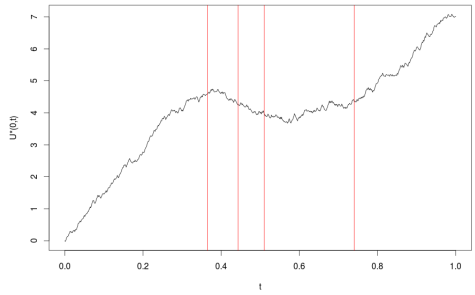
		Scénario 3			
n	% censure	γ_1	γ_2	γ_3	γ_4
200	0	0.086 (0.065)	0.307 (0.195)	0.662 (0.185)	1.121 (0.264)
200	30	0.054 (0.029)	0.168 (0.117)	0.451 (0.179)	0.793 (0.183)
200	50	0.033 (0.017)	0.093 (0.055)	0.237 (0.147)	0.555 (0.172)
500	0	0.077 (0.040)	0.291 (0.191)	0.637 (0.162)	1.095 (0.304)
500	30	0.059 (0.033)	0.175 (0.114)	0.473 (0.175)	0.798 (0.176)
500	50	0.040 (0.024)	0.103 (0.053)	0.246 (0.135)	0.587 (0.134)
1000	0	0.085 (0.031)	0.271 (0.157)	0.577 (0.138)	1.016 (0.282)
1000	30	0.059 (0.033)	0.165 (0.105)	0.433 (0.170)	0.755 (0.176)
1000	50	0.042 (0.022)	0.101 (0.042)	0.267 (0.131)	0.590 (0.111)



(a) $\beta_0(t) = \mathbb{1}_{t\leq 0.2} - \mathbb{1}_{t>0.6}$



(b) $\beta_0(t) = -\mathbb{1}_{t\leq 0.5} + 0.5\mathbb{1}_{1.1<t\leq 2.4} + \mathbb{1}_{t>2.4}$



(c) $\beta_0(t) = 2\mathbb{1}_{t\leq 0.1} - \mathbb{1}_{0.2<t\leq 0.3} + 1.5\mathbb{1}_{t>0.6}$

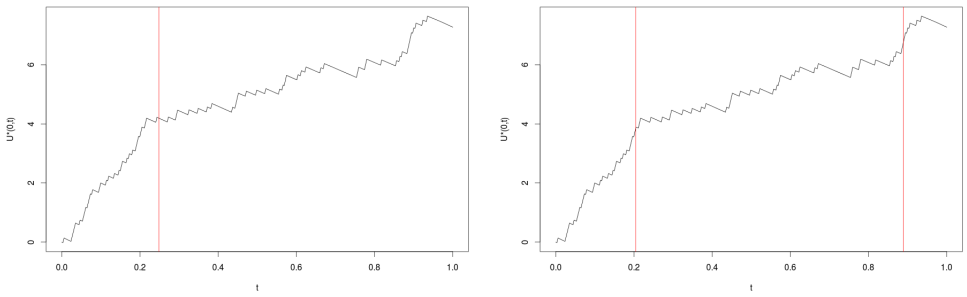
FIGURE 3.3 – Illustration des **Scénarii 1-3** avec le processus du score standardisé (en noir) et la localisation de leurs changepoints respectifs (en rouge)

3.6 Application

Nous illustrons maintenant les résultats présentés en Section 3.4 et 3.5 sur les données de cancer du sein collectées à l'Institut Curie introduites en Section 2.6. Nous nous focalisons ici sur la variable "taille de la tumeur". Nous séparons d'un côté les patients dont la taille de la tumeur est inférieure à 60mm, et ceux pour lesquels elle est strictement supérieure à 60mm. On crée ainsi une covariable binaire. La Figure 3.6 présente le processus du score standardisé pour la taille de la tumeur. On constate bien une dérive et donc d'un effet dépendant du temps. On voit même plus précisément que la pente du processus diminue, c'est-à-dire que l'effet de la taille de la tumeur diminue avec le temps. Au regard de la Figure 3.6, on peut se demander si, à la fin de l'étude, l'effet de la taille de la tumeur n'augmente pas. On hésite donc entre deux modèles réduits (3.2) : un modèle avec un changepoint et un modèle avec deux changepoints. On utilise les moindres carrés pour leurs estimations, puis on estime les coefficients de part et d'autre de ces changepoints par vraisemblance partielle. On obtient les modèles suivants

$$\textbf{M1} \qquad \hat{\beta}(t) = 1.68\mathbb{1}_{t\leq 28.06} + 0.58\mathbb{1}_{t>28.06}$$
$$\textbf{M2} \qquad \hat{\beta}(t) = 1.82\mathbb{1}_{t\leq 26} + 0.64\mathbb{1}_{26<t\leq 73.00} + 1.03\mathbb{1}_{t>73.00}.$$

L'estimation des points de rupture est aussi représentée en Figure 3.6.



(a) **M1** (b) **M2**

Figure 3.4 – Tracé du processus du score standardisé pour la covariable "taille de la tumeur" (en noir) et du ou des changepoints estimés (en rouge)

Chapitre 4

Une approche non paramétrique : les forêts de survie aléatoires

Ce chapitre est le résultat d'un travail réalisé en collaboration avec M. Erwan Scornet.

4.1 Introduction

Les forêts aléatoires sont une classe d'algorithmes d'apprentissage utilisés pour résoudre des problèmes de reconnaissance de formes. Le principe est de faire grandir plusieurs arbres qui permettent d'apprendre sur le jeu de données, puis ils sont agrégés pour former une prédiction. Pour avoir plusieurs arbres différents à partir d'un seul jeu de données, il faut introduire de l'aléa dans le processus de construction de l'arbre, en échantillonnant le jeu de données par exemple. En conséquence, il existe une grande variété de forêts aléatoires, suivant la manière dont les arbres sont construits et l'aléa introduit dans ce processus de construction.

L'une des forêts aléatoires les plus populaires est celle de Breiman (2001) pour laquelle la croissance de l'arbre est basée sur la procédure CART (Classification and Regression Trees, Breiman et al., 1984) et la randomisation s'effectue à la fois sur le jeu d'apprentissage et les directions de coupure. Les forêts de Breiman (2001) sont intensivement étudiées depuis la dernière décennie, principalement grâce à leurs bonnes performances en pratique et leur faculté à gérer des jeux de données de grande dimension. De plus, leur utilisation est simple puisqu'elles ne dépendent que de quelques paramètres qui sont facilement réglables (Liaw and Wiener, 2002; Genuer et al., 2008). Ces méthodes font déjà partie de l'état de l'art dans des domaines comme la génomique (Qi, 2012) et la reconnaissance de forme (Rogez et al., 2008), pour ne citer qu'eux. Elles progressent par ailleurs dans d'autres domaines tels que les données de survie (Ishwaran et al., 2008). C'est celui qui nous intéresse ici.

Cependant, même si les forêts aléatoires sont connues pour leurs performances dans plusieurs contextes, on sait peu de choses sur leurs propriétés mathématiques. Les premières tentatives pour comprendre les principes mathématiques en jeu dans les forêts aléatoires se sont concentrées sur un modèle de forêt simplifié (voir par exemple, Biau et al., 2008; Ishwaran and Kogalur, 2010; Denil et al., 2013) dont la construction ne dépend pas du jeu de données. D'un autre côté, des études récentes ont tenté d'analyser l'algorithme original de Breiman pour prouver sa normalité asymptotique (Mentch and Hooker, 2014; Wager, 2014) ou sa consistance (Scornet et al., 2014). Mais l'apparente simplicité de ces résultats contraste avec la complexité des preuves correspondantes. C'est pourquoi, plutôt que de s'intéresser directement aux forêts de survie aléatoires qui sont utilisées dans le cas de

données censurées, nous nous plaçons d'abord dans le cadre plus simple de régression sans censure et nous nous focalisons sur l'étude des forêts de Breiman.

Les forêts médianes sont un bon compromis entre la simplicité de forêts construites indépendamment du jeu de données et les forêts de Breiman qui dépendent à la fois des positions et des labels du jeu de données. La construction des forêts médianes dépend seulement de la position du jeu de données et peut être réglée pour faire en sorte que chaque feuille de chaque arbre contienne exactement un point. De cette manière, les forêts médianes sont plus proches de celles de Breiman que les forêts simplifiées dont les feuilles ne peuvent pas contenir un nombre prédéfini de points, car leur construction est indépendante des observations.

L'objectif de ce chapitre est d'étudier l'influence du taux de sous-échantillonnage sur la performance des forêts de Breiman. Pour cela nous nous basons sur l'influence de ce paramètre sur la performance des forêts médianes. Nous présentons une majoration de la vitesse de convergence des forêts médianes et montrons que la performance de ces forêts ne dépend pas du taux de sous-échantillonnage, pourvu qu'il soit assez grand, mais plutôt du niveau de chaque arbre dans la forêt. Nous démontrons que les bénéfices d'une forêt médiane totalement développée sur des forêts médianes moins développées ne se trouvent pas dans la précision de la prédiction mais plutôt dans la diminution du temps de calcul.

La Section 4.2 est consacrée à l'introduction des notations et fournit l'algorithme des forêts de Breiman. Les principaux résultats théoriques sur les forêts médianes sont rassemblés en Section 4.3. Ces résultats s'adaptent aux forêts de Breiman, comme présenté en Section 4.4. Nous revenons ensuite au problème de données de survie en Section 4.5 à travers une application à des données de cancer du sein de l'Institut Curie. Les preuves sont détaillées en Annexe A.

4.2 Notations

Dans ce chapitre, on note $\mathcal{D}_n = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ un échantillon d'apprentissage à valeurs dans $[0,1]^d \times \mathbb{R}$ indépendantes et identiquement distribuées selon la loi d'un couple de variables aléatoires $(\mathbf{X}, Y)$, où $\mathbb{E}[Y^2] < \infty$. La variable $\mathbf{X}$ représente la variable prédicteur et Y la variable réponse. On souhaite estimer la fonction de régression non paramétrique $m(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. Dans ce contexte, on utilise les forêts aléatoires pour construire un estimateur $m_n : [0,1]^d \to \mathbb{R}$ de m, basé sur le jeu de données $\mathcal{D}_n$.

Les forêts aléatoires sont des méthodes de classification et de régression basées sur une collection de plusieurs, disons M, arbres aléatoires. On note $m_n(\mathbf{x}, \Theta_j, \mathcal{D}_n)$ la valeur prédite du j-ème arbre, au point $\mathbf{x}$, où $\Theta_1, \ldots, \Theta_M$ sont des variables aléatoires indépendantes, distribuées selon une variable aléatoire générique Θ, indépendante de l'échantillon $\mathcal{D}_n$. En pratique, la variable Θ peut être utilisée pour échantillonner le jeu d'apprentissage ou pour sélectionner les directions ou positions candidates à une coupure. Les prédictions de M arbres aléatoires sont moyennées pour donner la prédiction finale, *i.e.*,

$$m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M, \mathcal{D}_n) = \frac{1}{M} \sum_{m=1}^{M} m_n(\mathbf{x}, \Theta_m, \mathcal{D}_n). \tag{4.1}$$

Pour plus de simplicité, on note $m_n(\mathbf{x})$ la quantité $m_{M,n}(\mathbf{x}, \Theta_1, \ldots, \Theta_M, \mathcal{D}_n)$. D'après la loi forte des grands nombres, pour un $\mathbf{x}$ fixé, conditionnellement à $\mathcal{D}_n$, l'estimateur de la forêt finie tend vers l'estimateur de la forêt infinie

$$m_{\infty,n}(\mathbf{x}) = \mathbb{E}_\Theta [m_n(\mathbf{x})],$$

qu'on peut écrire

$$m_{\infty,n}(\mathbf{X}) = \sum_{i=1}^{n} W_{ni}^{\infty}(\mathbf{X}) Y_i,$$

où

$$W_{ni}^{\infty}(\mathbf{X}) = \mathbb{E}_\Theta \left[\frac{\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i}}{N_n(\mathbf{X}, \Theta)} \right]$$

Comme nous nous plaçons dans le cadre de l'estimation par régression $\mathbb{L}^2$, on définit le risque de $m_{\infty,n}$ par

$$\mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2,$$

le risque de $m_{M,n}$ par

$$\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2,$$

et on dit que $m_{\infty,n}$ (resp. m_n) est $\mathbb{L}^2$ consistante si $\mathbb{E}[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})]^2$ (resp. $\mathbb{E}[m_n(\mathbf{X}) - m(\mathbf{X})]^2$) tend vers zéro quand $n \to \infty$.

Parmi la grande variété de forêts, on s'intéresse en particulier aux forêts de Breiman (2001) originales. Dans cette procédure, à chaque nœud de chaque arbre, on coupe sur un sous-ensemble aléatoire de directions et en minimisant la variance à l'intérieur de chaque cellule fille. La manière de construire les arbres est détaillée dans l'**Algorithme 1**, où m_{try} est le nombre de directions pré-sélectionnées pour couper, a_n est le nombre de points sous-échantillonnés dans chaque arbre et t_n est le nombre de feuilles dans chaque arbre.

Algorithm 1: Valeur prédite d'une forêt aléatoire de Breiman au point $\mathbf{x}$.

Input: Échantillon d'apprentissage $\mathcal{D}_n$, nombre d'arbres $M > 0$, $m_{\text{try}} \in \{1, \dots, p\}$, $a_n \in \{1, \dots, n\}$, $t_n \in \{1, \dots, a_n\}$ et $\mathbf{x} \in [0,1]^p$.

Output: Prédiction de la forêt aléatoire en $\mathbf{x}$.

for $j = 1, \dots, M$ **do**
- Sélectionner a_n points, sans remise, uniformément dans $\mathcal{D}_n$.
- Initialiser $\mathcal{P}_0 = \{[0,1]^p\}$ la partition associée à la racine de l'arbre.
- Pour tout $1 \leq \ell \leq a_n$, on pose $\mathcal{P}_\ell = \emptyset$.
- On pose $n_{nodes} = 1$ et $level = 0$.
- **while** $n_{nodes} < t_n$ **do**
  - **if** $\mathcal{P}_{level} = \emptyset$ **then**
    - $level = level + 1$
  - **else**
    - Soit A le premier élément de $\mathcal{P}_{level}$.
    - **if** *A contient exactement un point* **then**
      - $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$
      - $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A\}$
    - **else**
      - Sélectionner uniformément, sans remise, un sous-ensemble $\mathcal{M}_{try} \subset \{1, \dots, p\}$ de cardinal m_{try}.
      - Sélectionner la meilleure coupure dans A en optimisant le critère de coupure CART (Breiman et al., 1984) sur chaque coordonnée de $\mathcal{M}_{try}$.
      - Couper la cellule A selon la meilleure coupure. Nommer A_L et A_R les deux cellules filles.
      - $\mathcal{P}_{level} \leftarrow \mathcal{P}_{level} \backslash \{A\}$
      - $\mathcal{P}_{level+1} \leftarrow \mathcal{P}_{level+1} \cup \{A_L\} \cup \{A_R\}$
      - $n_{nodes} = n_{nodes} + 1$
    - **end**
  - **end**
- **end**
- Calculer la valeur prédite $m_n(\mathbf{x}; \Theta_j, \mathcal{D}_n)$ en $\mathbf{x}$ en moyennant les Y_i des points tombant dans la même cellule que $\mathbf{x}$ dans la partition $\mathcal{P}_{level} \cup \mathcal{P}_{level+1}$.

end

Calculer l'estimateur de la forêt aléatoire $m_{M,n}(\mathbf{x}; \Theta_1, \dots, \Theta_M, \mathcal{D}_n)$ au point $\mathbf{x}$ selon la formule (4.1).

Parce qu'elle minimise la variance à l'intérieur des cellules, la procédure de construction d'arbre de Breiman (2001) dépend de tout l'échantillon $\mathcal{D}_n$. Cette caractéristique spécifique permet à ces forêts d'avoir, en pratique, de bonnes performances, mais nous avons une compréhension moindre de leur comportement théorique. Au contraire, pour des forêts totalement indépendantes des observations (Biau, 2012), on peut prouver leur consistance et trouver des vitesses de convergence. Cependant, à cause de leur construction simplifiée, elles sont loin de modéliser précisément une forêt de Breiman (2001). Pour faire un pas de plus vers la compréhension du comportement des forêts de Breiman (2001), on étudie la forêt aléatoire médiane, qui satisfait la X-propriété. En effet, leur construction dépend seulement des X_i, ce qui est un bon compromis entre la complexité des forêts de Breiman (2001) et la simplicité des forêts totalement non adaptatives. Cela permet également d'essayer de comprendre pourquoi les forêts aléatoires sont toujours consistantes même quand

il reste exactement un point dans chaque feuille. Dans l'esprit de l'algorithme de Breiman (2001), avant la construction de chaque arbre, les données sont sous-échantillonnées, c'est-à-dire a_n points ($a_n < n$) sont sélectionnés sans remise. Ensuite, chaque coupure est effectuée sur la médiane empirique d'une coordonnée choisie uniformément aléatoirement parmi les d coordonnées. Rappelons que la médiane de $\mathbf{X}_1, \ldots, \mathbf{X}_n$ est définie comme l'unique $\mathbf{X}_{(\ell)}$ satisfaisant $F_n(\mathbf{X}_{(\ell-1)}) \leq 1/2 < F_n(\mathbf{X}_{(\ell)})$, où les $\mathbf{X}_{(i)}$ sont ordonnées de manière croissante. Pour que la loi des observations soit uniforme sur les cellules filles, les points sur lesquels sont effectuées les coupures ne sont pas reportés dans les cellules filles. En effet, sinon il y aurait au moins un point sur le bord de la cellule fille, et donc la loi des observations ne serait plus uniforme dans cette cellule. Finalement, l'algorithme s'arrête quand chaque cellule contient exactement un point. On note k_n la profondeur d'un arbre et on considère que k_n vérifie

$$\frac{a_n}{2^{k_n}} \geq 4.$$

Tout au long de ce chapitre, on utilise les forêts de Breiman (2001) et les forêts médianes pour illustrer nos résultats.

4.3 Résultats théoriques

Les forêts médianes totalement développées sont connues pour être consistantes sous les hypothèses de régularité suivantes sur le modèle de régression (Devroye et al., 1996).

Hypothèses 4.3.1. (H) *On a*

$$Y = m(\mathbf{X}) + \varepsilon,$$

où ε est un bruit centré tel que $\mathbb{V}[\varepsilon|\mathbf{X} = \mathbf{x}] \leq \sigma^2$, où $\sigma^2 < \infty$ est une constante. De plus, $\mathbf{X}$ est uniformément distribuée sur $[0,1]^d$ et m est L-Lipschitz.

Le Théorème 4.1 présente une majoration du risque $\mathbb{L}^2$ de m_n.

Théorème 4.1. *Supposons que* **(H)** *est satisfaite. Alors, pour tout n,*

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq 2\sigma^2 \frac{2^k}{n} + dL^2 C_1 \left(1 - \frac{3}{4d}\right)^k. \tag{4.2}$$

Le membre de droite est minimal pour

$$k_n = \frac{1}{\ln 2 - \ln \beta} \left[\ln(n) + C_3 \right], \tag{4.3}$$

sous la condition $a_n \geq C_4 n^{\frac{\ln 2}{\ln 2 - \ln \beta}}$. Pour ce choix de k_n et a_n, on a

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq C n^{\frac{\ln\left(1-\frac{3}{4d}\right)}{\ln 2 - \ln\left(1-\frac{3}{4d}\right)}}. \tag{4.4}$$

L'équation (4.2) provient d'une décomposition standard erreur d'estimation / erreur d'approximation des forêts médianes.

Commençons par regarder l'erreur d'estimation, on peut noter que l'erreur d'estimation d'un seul arbre est de l'ordre de $2^{k_n}/a_n$. Donc, à cause du sous-échantillonnage (*i.e.*, puisque $a_n < n$), l'erreur d'estimation des forêts médianes est plus petite que celle d'un seul arbre. Ceci souligne un premier avantage des forêts sur les arbres seuls.

Le second terme dans l'inéquation (4.2) est prévisible. En effet, dans les niveaux proches de la racine, une coupure se situe très près du centre du côté d'une cellule (puisque $\mathbf{X}$ est uniformément distribuée sur $[0,1]^d$). Donc, pour tout k suffisamment petit, l'erreur d'approximation des forêts médianes devrait être proche de celle des forêts centrées étudiées par Biau (2012). Une étude détaillée de la Proposition 2.2 dans Biau (2012) montre que cette dernière peut être facilement modifiée pour concorder avec notre erreur d'approximation.

Plus précisément, la vitesse de convergence des forêts médianes est plus rapide que celle des forêts centrées qui est égale à

$$\mathbb{E}\big[m_n^{cc}(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq C n^{\frac{-3}{4d \ln 2 + 3}}, \tag{4.5}$$

où m_n^{cc} est l'estimateur des forêts centrées. Notons que le majorant (4.5) est plus maniable que (4.4). Un regard approfondi sur la preuve montre que la simplicité de la borne (4.5) provient d'une approximation imprécise. La preuve de la Proposition 2.2 peut être facilement adaptée pour obtenir la même majoration que (4.4).

On peut noter que le fait que la majoration (4.4) soit plus précise que (4.5) semble important dans le cas $d = 1$. En effet, dans ce cas, d'après le Théorème 4.1,

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq Cn^{-2/3},$$

qui est la borne minimax sur la classe des fonctions Lipschitz (voir, par exemple, Stone, 1980, 1982). Malheureusement, dans ce cas, le majorant (4.5) n'est pas précis puisqu'il mène à la majoration

$$\mathbb{E}\big[m_n^{cc}(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq Cn^{\frac{-3}{4\ln 2+3}}.$$

Le Théorème 4.1 nous permet de déduire des vitesses de convergence pour deux forêts particulières. La première est la forêt médiane semi-développée, où on n'effectue aucun sous-échantillonnage avant de construire chaque arbre. La seconde est la forêt médiane totalement développée, où chaque feuille contient un petit nombre de points. Le Corollaire 4.2 traite des forêts semi-développées.

Corollaire 4.2 (Forêts médianes semi-développées)**.** *Supposons que* **(H)** *est satisfaite. Considérons une forêt médiane sans sous-échantillonnage, i.e.* $a_n = n$*, et telle que le paramètre* k_n *satisfait (4.3). Alors,*

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq Cn^{\frac{\ln\left(1-\frac{3}{4d}\right)}{\ln 2-\ln\left(1-\frac{3}{4d}\right)}}.$$

A une approximation près, le Corollaire 4.2 est la contrepartie du Théorème 2.2 dans Biau (2012) mais adaptée aux forêts médianes. En effet, à une petite modification de la preuve du Théorème 2.2 près, la vitesse de convergence fournie par le Théorème 2.2 pour les forêts centrées et celle du Corollaire 4.2 pour les forêts médianes sont identiques. Notons que, pour les deux forêts, la profondeur optimale k_n de chaque arbre est la même.

Le Corollaire 4.3 traite du cas des forêts médianes complètement développées.

Corollaire 4.3 (Forêts médianes complètement développées)**.** *Supposons que* **(H)** *est satisfaite. Considérons une forêt médiane complètement développée pour laquelle les paramètres* k_n *et* a_n *satisfont* $k_n = \log_2(a_n) - 2$*. Le choix optimal de* a_n*, qui minimise l'erreur* $\mathbb{L}^2$ *dans (4.2)) est donné par (4.3), c'est-à-dire*

$$a_n = C_4 n^{\frac{\ln 2}{\ln 2-\ln \beta}}.$$

Dans ce cas,

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq Cn^{\frac{\ln\left(1-\frac{3}{4d}\right)}{\ln 2-\ln\left(1-\frac{3}{4d}\right)}}.$$

Comme $k_n = \log_2(a_n) - 2$, le nombre d'observations dans chaque feuille est compris entre 4 et 8. C'est pourquoi cette forêt est appelée une forêt médiane *complètement développée*.

Alors que chaque arbre pris individuellement dans la forêt médiane complètement développée n'est pas consistant, puisque chaque feuille contient un petit nombre de points, la forêt est consistante et sa vitesse de convergence est fournie par le Corollaire 4.3. De plus, le Corollaire 4.3 nous permet de déduire le taux de sous-écantillonage optimal pour les forêts médianes totalement développées.

4.4 Simulations

En pratique, les forêts aléatoires de Breiman sont parmi les algorithmes de forêts les plus utilisés. C'est pourquoi, à la lumière des résultats de la Section 4.3, nous avons effectué quelques simulations afin de trouver une vitesse de convergence similaire pour les forêts aléatoires de Breiman.

Nous commençons, dans cette Section 4.4, par valider la théorie suggérée en Section 4.3 pour différents modèles décrits ci-après. Nous étudions ensuite l'effet de la profondeur des arbres puis du sous-échantillonnage sur la performance des forêts aléatoires de Breiman en terme de risque $\mathbb{L}^2$.

Les comparaisons que nous effectuerons dans la suite de cette section utiliseront les modèles suivants. Pour n'importe lequel de ceux-ci, on considère toujours que les covariables $\mathbf{X} = (X_1, \ldots, X_d)$ sont uniformément distribuées sur $[0,1]^d$. On pose également $\tilde{X}_i = 2(X_i - 0.5)$ pour tout $1 \leq i \leq d$.

- **Modèle 1** : $n = 800, d = 50, Y = \tilde{X}_1^2 + \exp(-\tilde{X}_2^2)$
- **Modèle 2** : $n = 600, d = 100, Y = \tilde{X}_1\tilde{X}_2 + \tilde{X}_3^2 - \tilde{X}_4\tilde{X}_7 + \tilde{X}_8\tilde{X}_{10} - \tilde{X}_6^2 + \mathcal{N}(0, 0.5)$
- **Modèle 3** : $n = 600, d = 100, Y = -\sin(2\tilde{X}_1) + \tilde{X}_2^2 + \tilde{X}_3 - \exp(-\tilde{X}_4) + \mathcal{N}(0, 0.5)$
- **Modèle 4** : $n = 600, d = 100, Y = \tilde{X}_1 + (2\tilde{X}_2 - 1)^2 + \sin(2\pi\tilde{X}_3)/(2 - \sin(2\pi\tilde{X}_3)) + \sin(2\pi\tilde{X}_4) + 2\cos(2\pi\tilde{X}_4) + 3\sin^2(2\pi\tilde{X}_4) + 4\cos^2(2\pi\tilde{X}_4) + \mathcal{N}(0, 0.5)$
- **Modèle 5** : $n = 700, d = 20, Y = \mathbb{1}_{\tilde{X}_1 > 0} + \tilde{X}_2^3 + \mathbb{1}_{\tilde{X}_4 + \tilde{X}_6 - \tilde{X}_8 - \tilde{X}_9 > 1 + \tilde{X}_{10}} + \exp(-\tilde{X}_2^2) + \mathcal{N}(0, 0.5)$
- **Modèle 6** : $n = 500, d = 30, Y = \sum_{k=1}^{10} \mathbb{1}_{\tilde{X}_k^3 < 0} - \mathbb{1}_{\mathcal{N}(0,1) > 1.25}$
- **Modèle 7** : $n = 600, d = 300, Y = \tilde{X}_1^2 + \tilde{X}_2^2\tilde{X}_3 \exp(-|\tilde{X}_4|) + \tilde{X}_6 - \tilde{X}_8 + \mathcal{N}(0, 0.5)$
- **Modèle 8** : $n = 500, d = 1000, Y = \tilde{X}_1 + 3\tilde{X}_3^2 - 2\exp(-\tilde{X}_5) + \tilde{X}_6$

Certains de ces modèles sont des modèles jouets (**Modèles 1, 5-8**). Le **Modèle 2** peut être trouvé dans van der Laan et al. (2007) et les **Modèles 3-4** sont présentés dans Meier et al. (2009).

Toutes les applications numériques ont été réalisées en usilisant le logiciel libre R. Pour chaque expérience, le jeu de données est divisé en deux parties : un échantillon d'apprentissage (80% du jeu de données) et un échantillon test (les 20% restants). Le risque empirique (l'erreur $\mathbb{L}^2$) est ensuite évalué sur l'ensemble test.

Complexité algorithmique La complexité de l'algorithme de la forêt aléatoire médiane réside principalement dans la complexité à trouver la médiane d'un échantillon. Trouver la médiane d'un échantillon de taille n est d'une complexité en $O(n)$, la complexité informatique des forêts médianes est, à un facteur constant près,

$$\sum_{j=0}^{k-1} 2^j \left(\frac{a_n}{2^j}\right) = k a_n.$$

Notons que la valeur optimale de k_n, c'est-à-dire celle qui minimise l'erreur empirique, est la même, peu importe la valeur de a_n. C'est pourquoi minimiser la complexité algorithmique

revient à minimiser a_n, ce qui est le cas pour une forêt médiane *complètement développée* (Corollary 4.3).

Alors qu'il n'y a pas d'intérêt statistique à choisir la forêt médiane *complètement développée* par rapport à celle *semi-développée*, il y a à l'évidence un gain en terme de temps de calcul.

4.4.1 Profondeur des arbres

On commence ici par comparer les forêts de Breiman originales aux forêts semi-développées. La Figure 4.1 présente, pour les **Modèles 1-8** introduits précédemment, les risques $\mathbb{L}^2$ des forêts de Breiman classiques et les comparent à ceux des forêts de Breiman semi-développées. Plus précisément, sur chaque sous-figure sont représentées les erreurs $\mathbb{L}^2$ de forêts de Breiman semi-développées pour des profondeurs d'arbres différentes. A cause du compromis erreur d'approximation/estimation, on devrait naturellement constater que, pour chaque modèle, l'erreur $\mathbb{L}^2$ décroît puis croît quand le nombre de feuilles augmente, c'est-à-dire plus la profondeur des arbres est grande. Il semblerait ici que dans la majorité des modèles, l'erreur d'estimation soit trop faible pour être visible, ce qui explique la décroissance des erreurs présentées en Figure 4.1. Toutes les sous-figures de la Figure 4.1 proviennent de forêts à 500 arbres. Les erreurs présentées sont obtenues en moyennant les erreurs des forêts de 50 échantillons.

Remarquons que pour chacun des huit modèles, les performances des forêts semi-développées sont comparables à celles des forêts de Breiman lorsque le paramètre de profondeur est bien choisi. Dans le cas du **Modèle 1**, une forêt de Breiman semi-développée laissant environ 110 feuilles dans chaque arbre donne la même erreur $\mathbb{L}^2$ qu'une forêt de Breiman originale. Dans la procédure originale des forêts de Breiman, un échantillon bootstrappé est utilisé dans la construction de chacun des arbres. Dans les forêts semi-développées considérées, la totalité du jeu de données est utilisée pour construire chacun d'entre eux, l'aléatoire ne provenant alors que de la présélection des directions de coupure. Les simulations montrent que les performances de ces deux types de forêts sont similaires sous réserve que le paramètre de profondeur soit bien choisi. Les performances obtenues par bootstrap et avec des arbres de petite profondeur étant similaires, il n'y a donc pas un "effet bootstrap" qui permettrait de rendre la forêt aléatoire compétitive par rapport aux autres algorithmes de régression. Comme le montre le Corollaire 4.2, et comme tendent à le prouver les simulations, la profondeur et l'échantillonnage du jeu de données (ici le bootstrap) sont équivalents.

Afin d'étudier la valeur optimale de la profondeur (paramètre `maxnodes` dans l'algorithme R), nous avons tracé les courbes précédentes pour différentes tailles de l'échantillon d'apprentissage (100, 200, 300 et 400). Nous avons également reporté dans un graphique les valeurs optimales de la profondeur pour chacune de ces quatre valeurs. Les résultats sont présentés dans la Figure 4.2. La valeur optimale a été choisie comme étant la plus petite valeur de `maxnodes`, notée `m`, vérifiant

$$|\ell_{\mathtt{m}} - \min_i \ell_i| < 0.05 \times (\max_i \ell_i - \min_i \ell_i),$$

où ℓ_i est l'erreur de la forêt pour le paramètre $\mathtt{maxnodes} = i$.

D'après la dernière sous-figure de la Figure 4.2, la valeur optimale de la profondeur des arbres semble être linéaire en fonction de la taille de l'échantillon. Pour le **Modèle 1**, la valeur optimale `m` semble satisfaire $0.25n < \mathtt{m} < 0.33n$. Les autres modèles donnent également des valeurs optimales linéaires en n. Les résultats sont présentés en Figure 4.3.

Nous avons tracé les erreurs des forêts de Breiman semi-développées pour différentes valeurs du paramètre de la profondeur (10%, 30 %, 63%, 80% et 100%) en fixant la taille du jeu de données afin de respecter les **Modèles 1-8** définis plus haut. Les résultats sont présentés sous forme de boîtes à moustaches en Figure 4.4.

On remarque ainsi que la forêt semi-développée à 30% (*i.e.*, telle que `maxnodes` $= 0.3n$) a des performances comparables (**Modèle 5**) voire meilleures (**Modèle 6**) que celles des forêts de Breiman.

4.4.2 Sous échantillonnage

On cherche maintenant à comparer les forêts de Breiman classiques aux forêts de Breiman sous-échantilonnées. De la même manière que pour la profondeur, la Figure 4.5 présente, pour les **Modèles 1-8**, les risques $\mathbb{L}^2$ des forêts de Breiman classiques et les comparent à ceux des forêts de Breiman sous-échantillonnées. Plus précisément, sur chaque sous-figure sont représentées les erreurs $\mathbb{L}^2$ de forêts de Breiman sous-échantillonnées pour des valeurs différentes de sous-échantillonnage. Comme pour la profondeur, l'erreur $\mathbb{L}^2$ devrait être décroissante puis croissante quand le nombre de points sous-échantillonnés augmente (compromis approximation / estimation). Il semblerait encore une fois que dans la majorité des modèles, l'erreur d'estimation soit trop faible pour être visible, ce qui explique la décroissance des erreurs présentées en Figure 4.5. Toutes les sous-figures de la Figure 4.5 proviennent de forêts à 500 arbres. Les erreurs présentées sont obtenues en moyennant les erreurs des forêts de 50 échantillons.

De la même manière que pour la profondeur, les performances des forêts sous-échantillonnées sont comparables à celles des forêts de Breiman lorsque le taux de sous-échantillonnage est bien choisi. Par exemple, les forêts sous-échantillonnées avec un tirage sans remise de 200 observations parmi les 400 totales ont la même performance que les forêts de Breiman classiques, dans le cas du **Modèle 2**. Dans les forêts sous-échantillonnées que nous étudions ici, le tirage s'effectue sans remise et seules a_n observations sont sélectionnées parmi les n, contrairement aux forêts de Breiman où on effectue un tirage avec remise de n point parmi n avant la construction de chaque arbre de la forêt. Les performances obtenues par bootstrap et par sous-échantillonnage étant similaires, il n'y a donc pas, comme pour la profondeur, un "effet bootstrap" qui permettrait de rendre la forêt aléatoire compétitive par rapport aux autres algorithmes de régression. Comme le montre le Corollaire 4.3, et comme tendent à le prouver les simulations, le sous-échantillonnage et le bootstrap sont équivalents.

On souhaite une nouvelle fois étudier la valeur optimale du sous-échantillonnage (paramètre `sampsize` dans l'algorithme R). Nous avons donc tracé les courbes précédentes pour différentes tailles de l'échantillon d'apprentissage (100, 200, 300 et 400) et reporté dans un graphique les valeurs optimales du sous-échantillonnage pour chacune de ces quatre valeurs. Les résultats sont présentés en Figure 4.6. La valeur optimale a été choisie comme étant la plus petite valeur de `sampsize`, notée `a`, vérifiant

$$|\ell_{\mathtt{a}} - \min_i \ell_i| < 0.05 \times (\max_i \ell_i - \min \ell_i),$$

où ℓ_i est l'erreur de la forêt pour le paramètre `sampsize` $= i$.

Comme pour la partie sur la profondeur, la valeur optimale de sous-échantillonnage semble de nouveau être linéaire en fonction de la taille de l'échantillon. Pour le **Modèle 1**, la valeur optimale `a` semble être proche de $0.8n$. Les autres modèles donnent également des valeurs optimales linéaires en n. Les résultats sont présentés en Figure 4.7.

Nous avons donc tracé les erreurs des forêts sous-échantillonnées pour différentes valeurs du paramètre de sous-échantillonage ($0.4n$, $0.5n$, $0.63n$ et $0.9n$). Les résultats sont présentés sous forme de boîtes à moustaches en Figure 4.8.

La forêt sous-échantillonnée à 63% possède des performances similaires aux forêts de Breiman, ce qui n'est pas surprenant. En effet, il y a en moyenne 63% de points différents dans un échantillon bootstrap. D'autre part, on voit que les taux de sous-échantillonnage élevés ($0.9n$ par exemple) ont de bonnes performances. Cela est probablement dû au rapport signal/bruit qui semble élevé ici. Si on augmente le bruit des modèles, comme illustré en Figure 4.9, on obtient alors des résultats plus nuancés, corroborant ainsi l'utilisation du sous-échantillonnage comme paramètre d'optimisation de la performance des forêts.

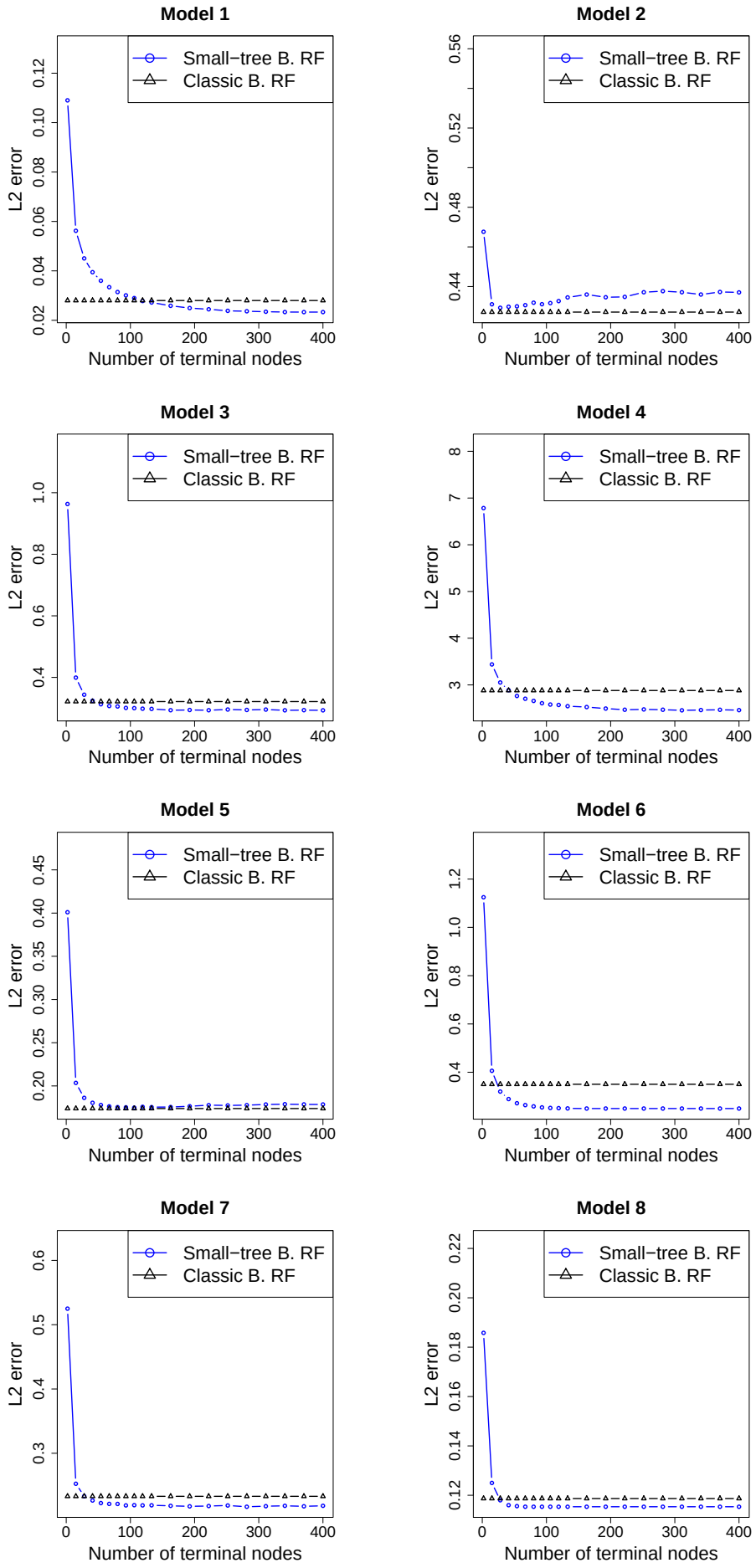


FIGURE 4.1 – Comparaisons de l'erreur $\mathbb{L}^2$ des forêts de Breiman originales et celle des forêts de Breiman semi-développées.

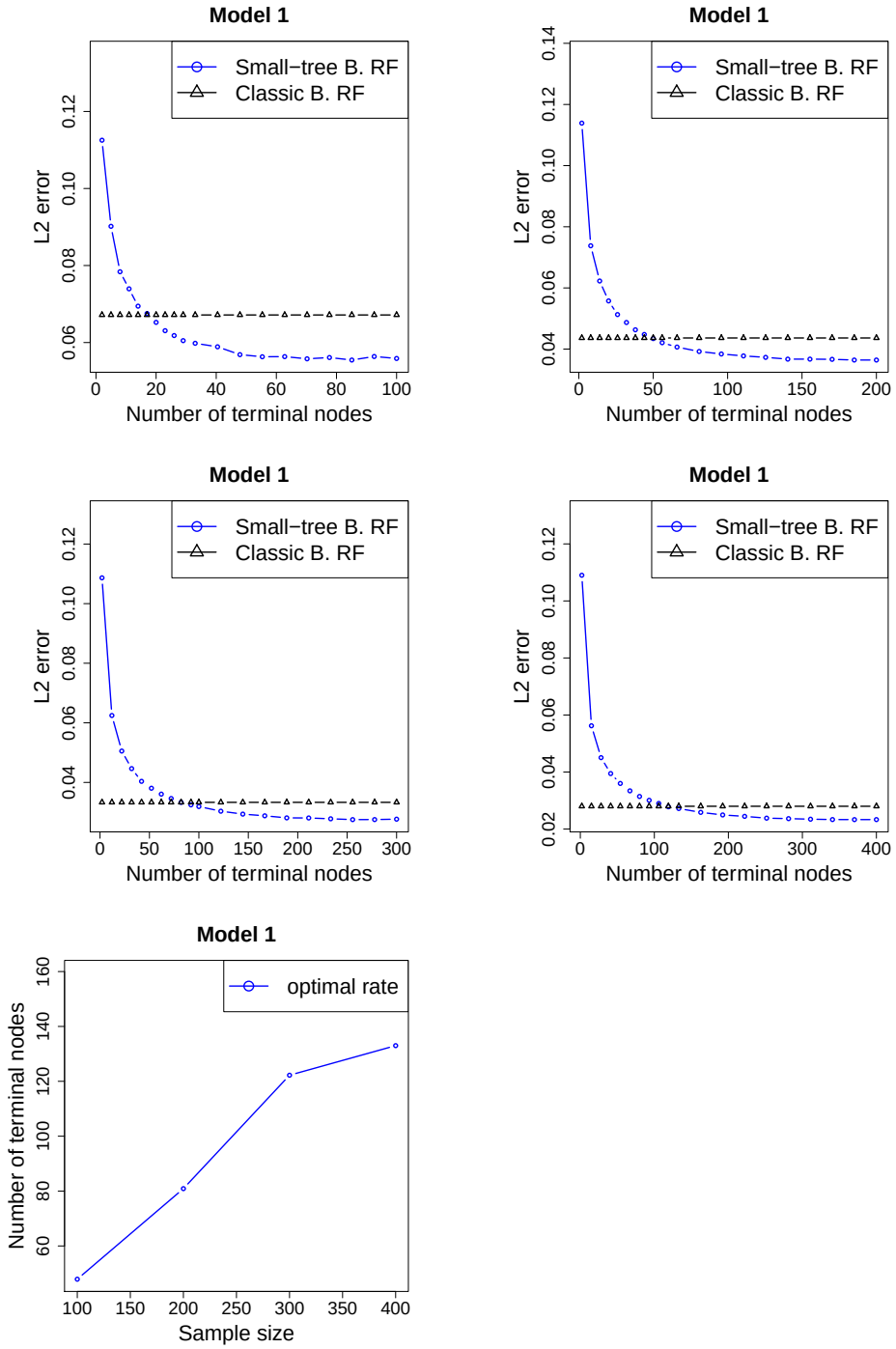


Figure 4.2 – Quatre premiers graphes : erreur $\mathbb{L}^2$ des forêts semi-développées et standard de Breiman pour le **Modèle 1** pour différentes tailles de l'échantillon d'apprentissage (de 100 à 400) ; dernier graphe : valeurs optimales du nombre de noeuds terminaux pour le **Modèle 1**.

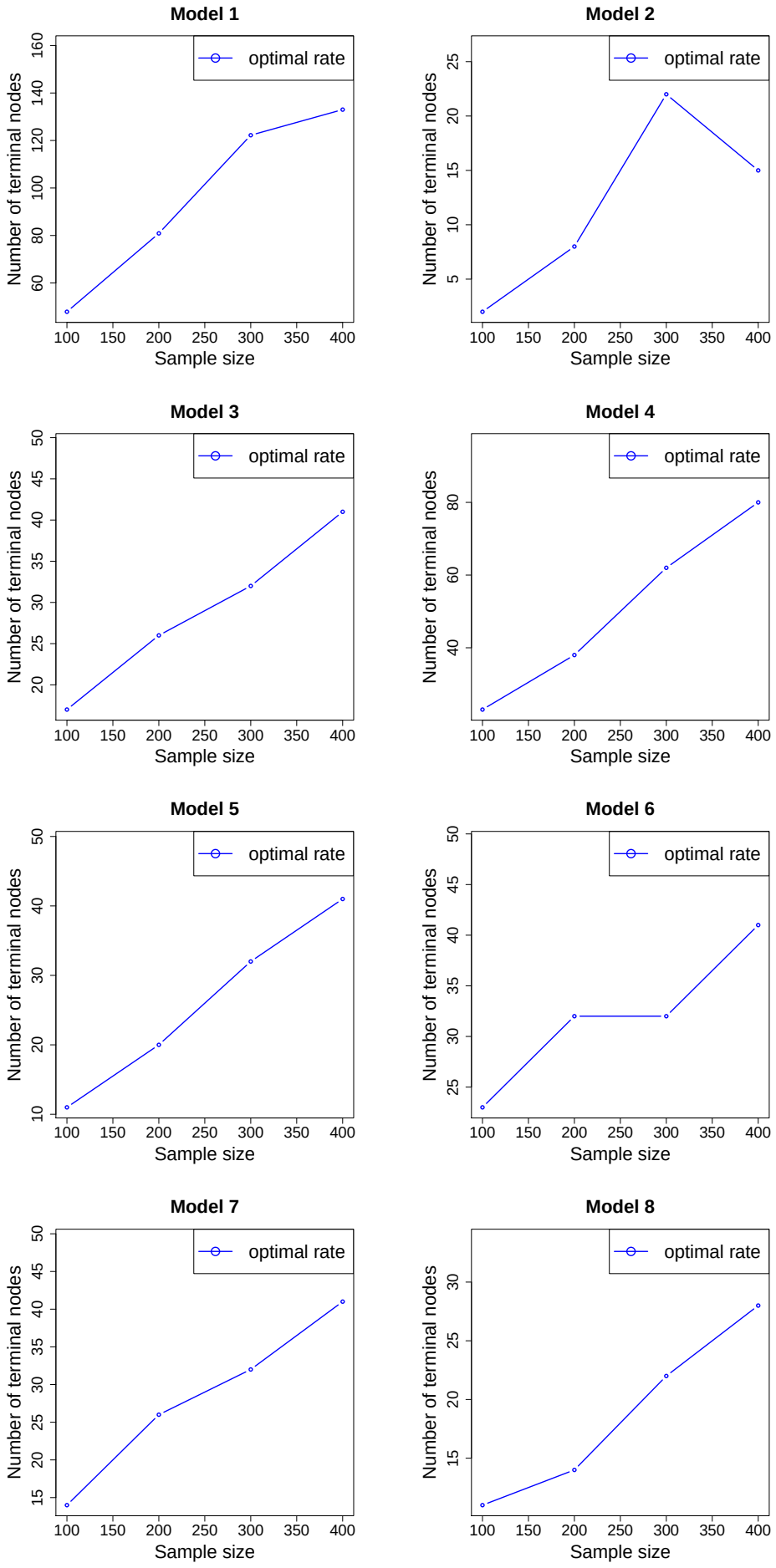


FIGURE 4.3 – Valeurs optimales du paramètre de profondeur des arbres pour les **Modèles 1-8**

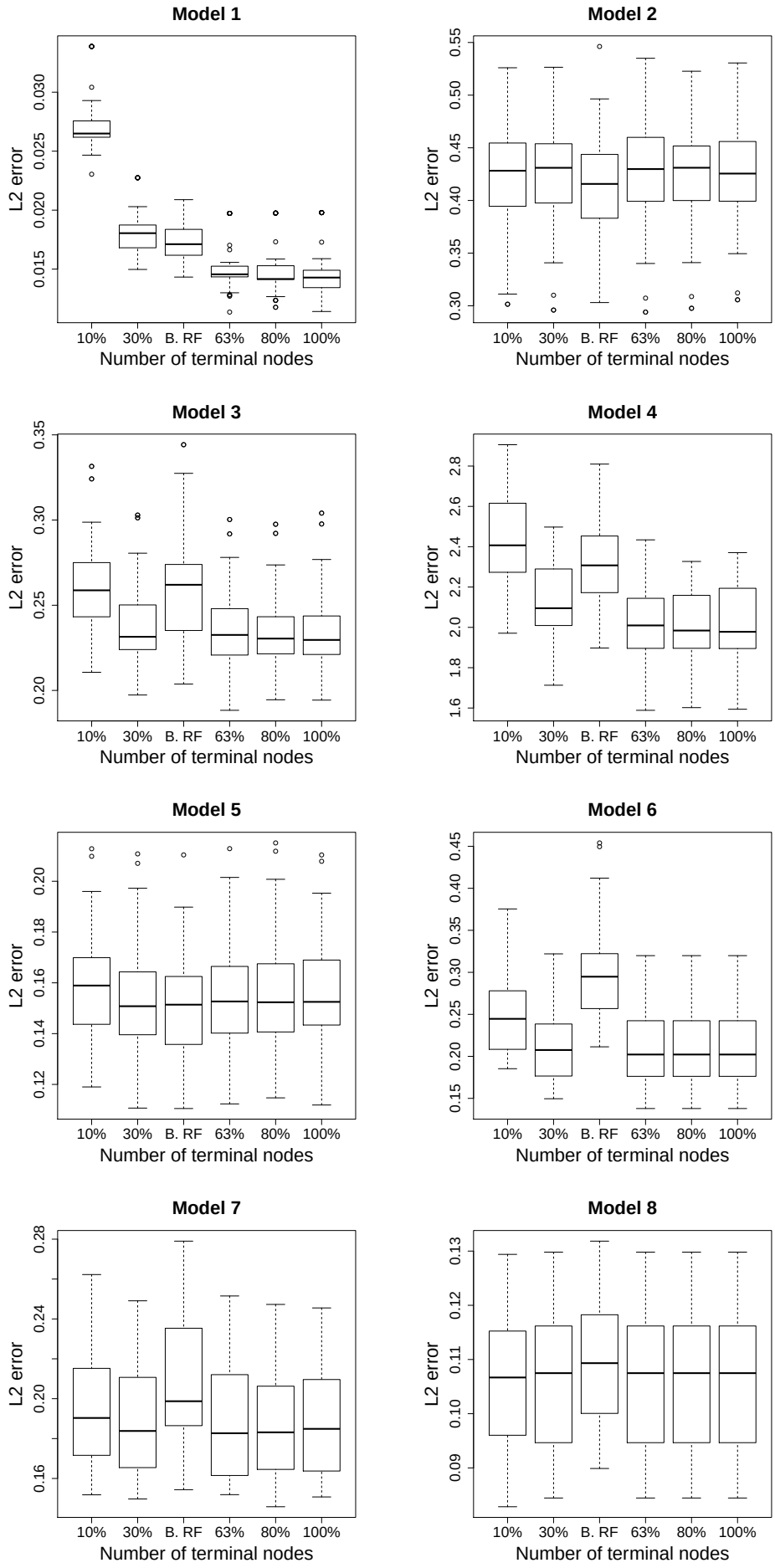


Figure 4.4 – Comparaison d'erreurs $\mathbb{L}^2$ de forêts de Breiman standards et de plusieurs forêts de Breiman semi-développées.

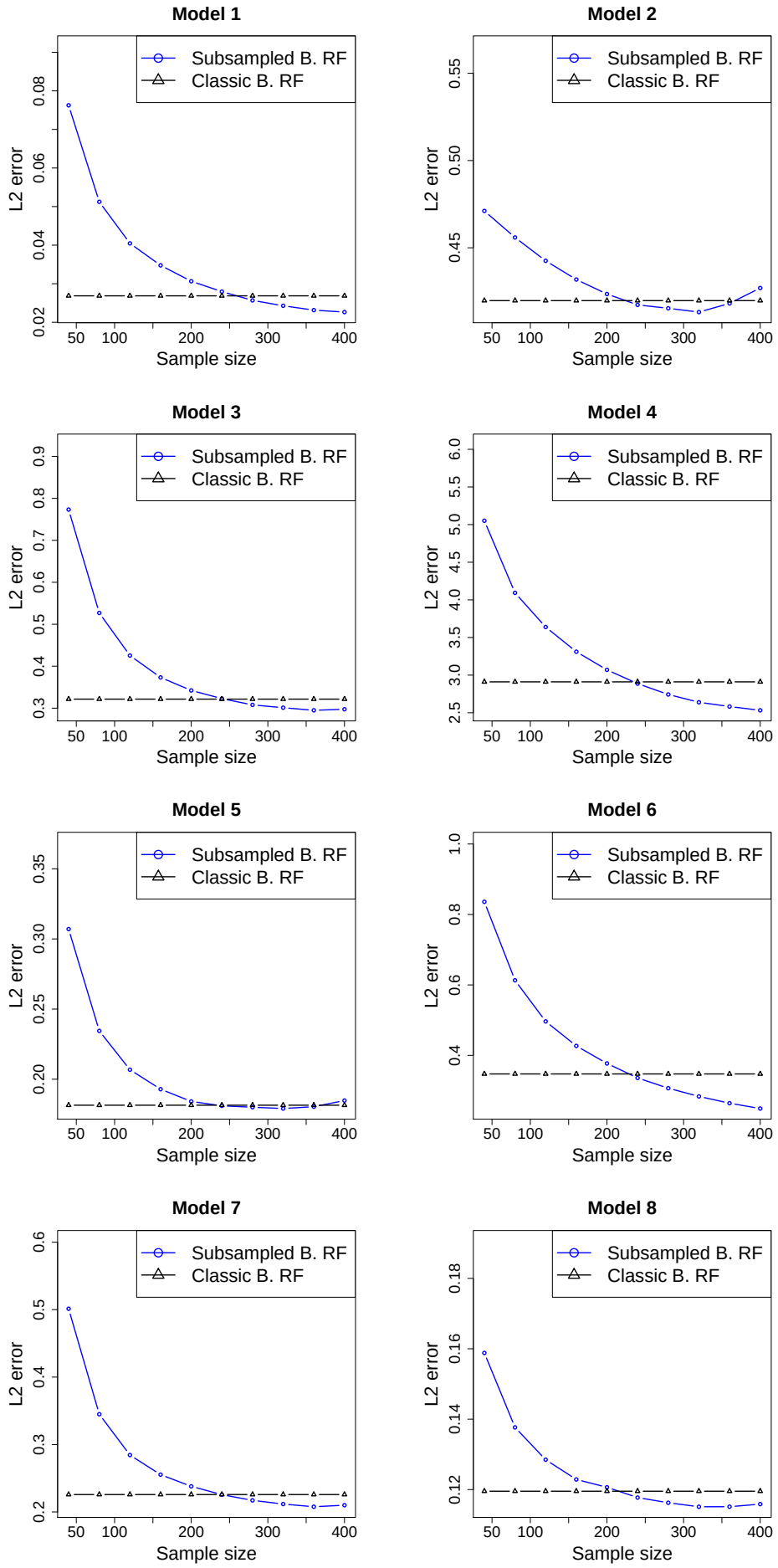


FIGURE 4.5 – Comparaisons de l'erreur $\mathbb{L}^2$ des forêts de Breiman originales et celle des forêts de Breiman sous-échantillonnées

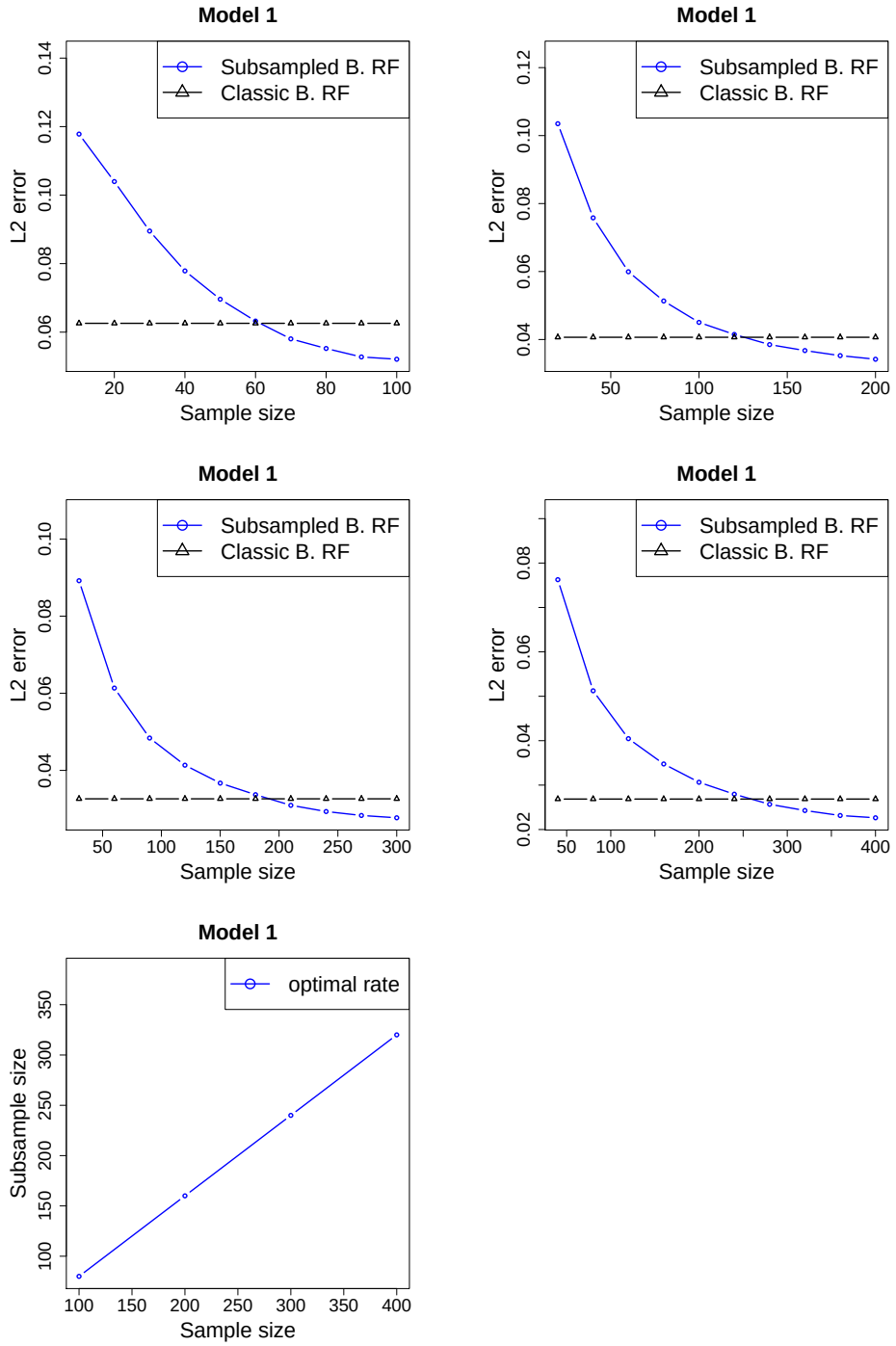


Figure 4.6 – Quatre premiers graphes : erreur $\mathbb{L}^2$ des forêts sous-échantillonnées et standard de Breiman pour le **Modèle 1** pour différentes tailles de l'échantillon d'apprentissage (de 100 à 400) ; dernier graphe : valeurs optimales de la taille du sous-échantillonnage pour le **Modèle 1**.

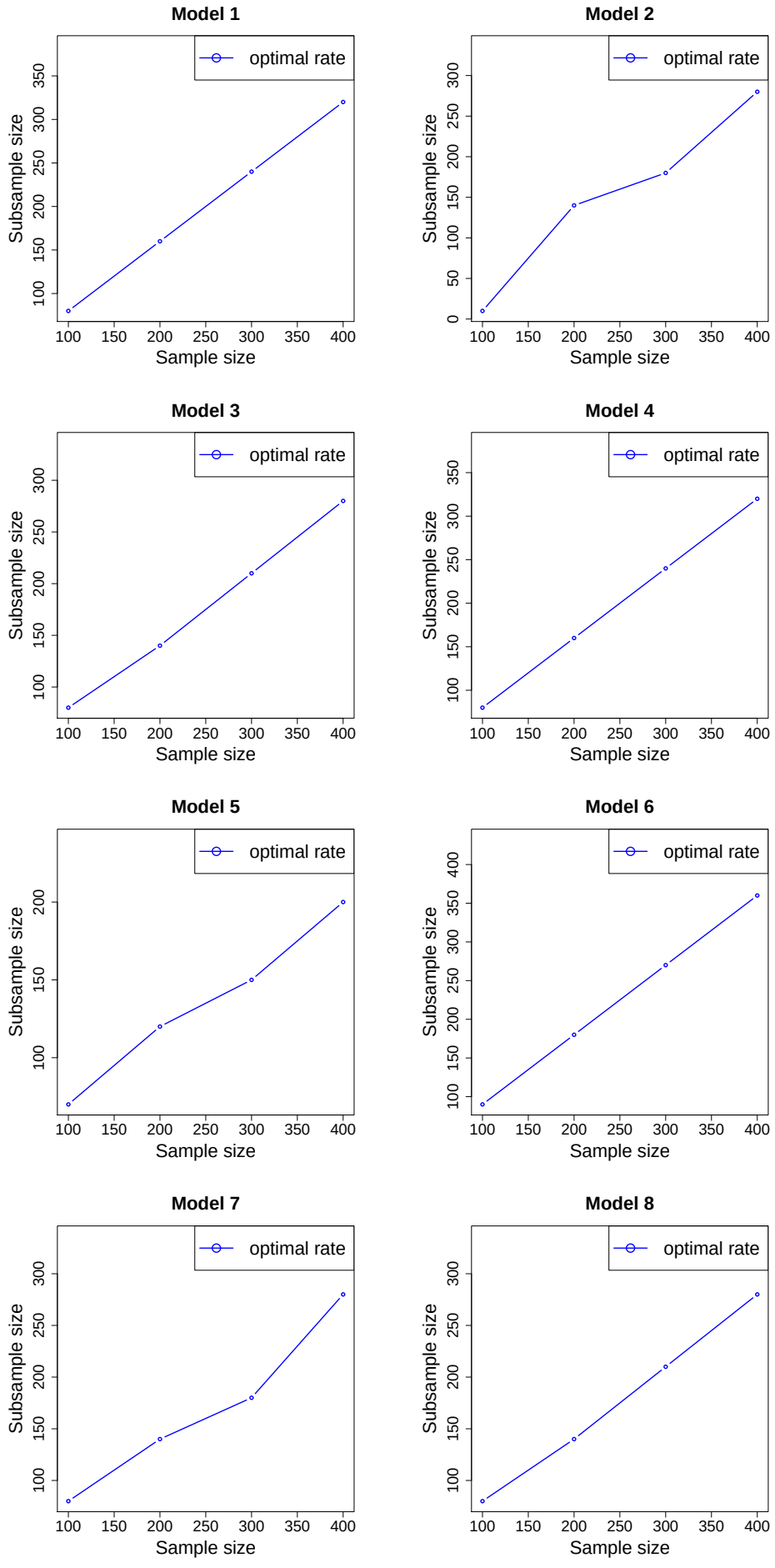


FIGURE 4.7 – Valeurs optimales du paramètre de sous-échantillonnage

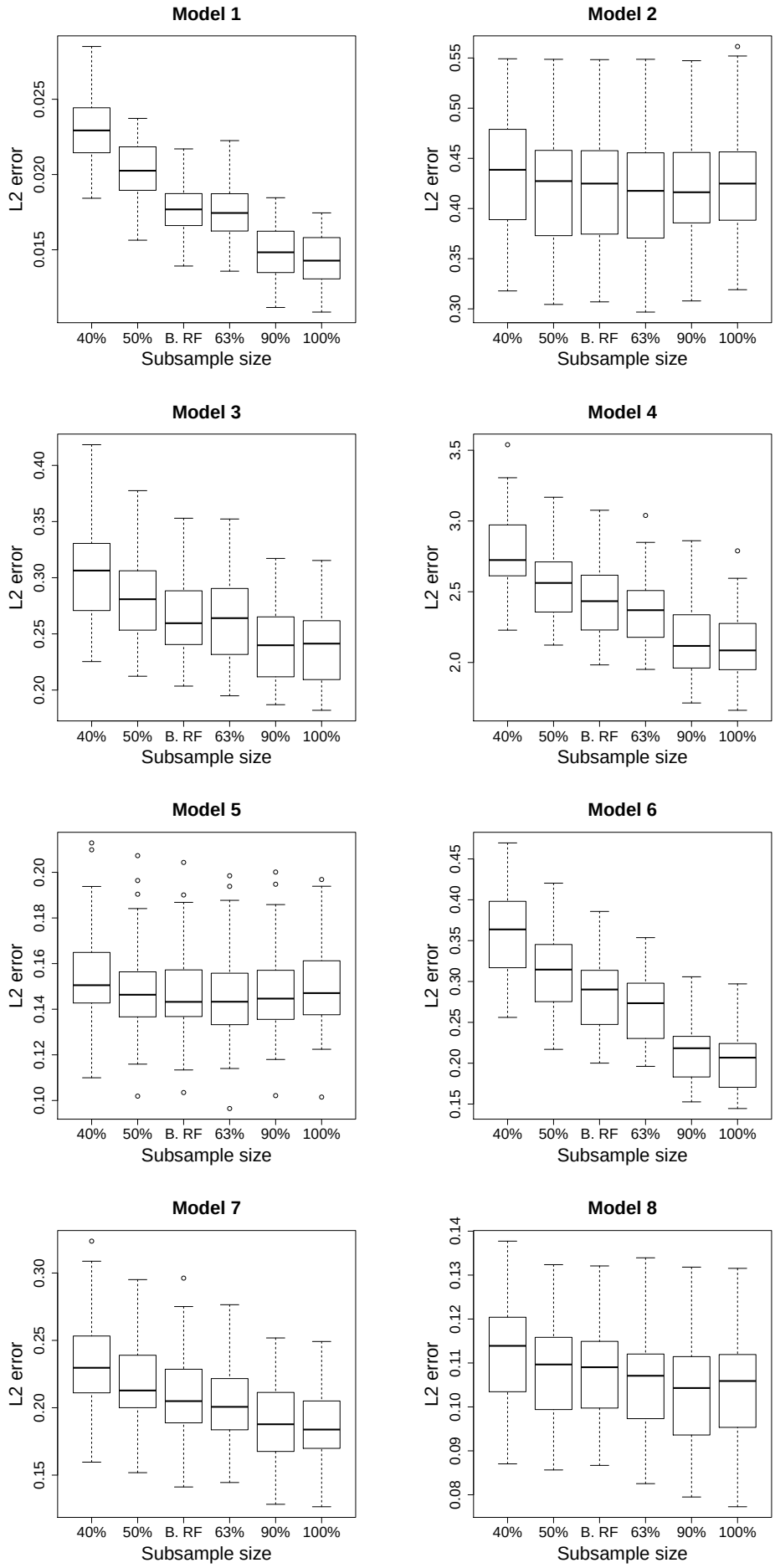


Figure 4.8 – Comparaison de l'erreur $\mathbb{L}^2$ de forêts de Breiman standards avec celles de différentes forêts de Breiman sous-échantillonnées

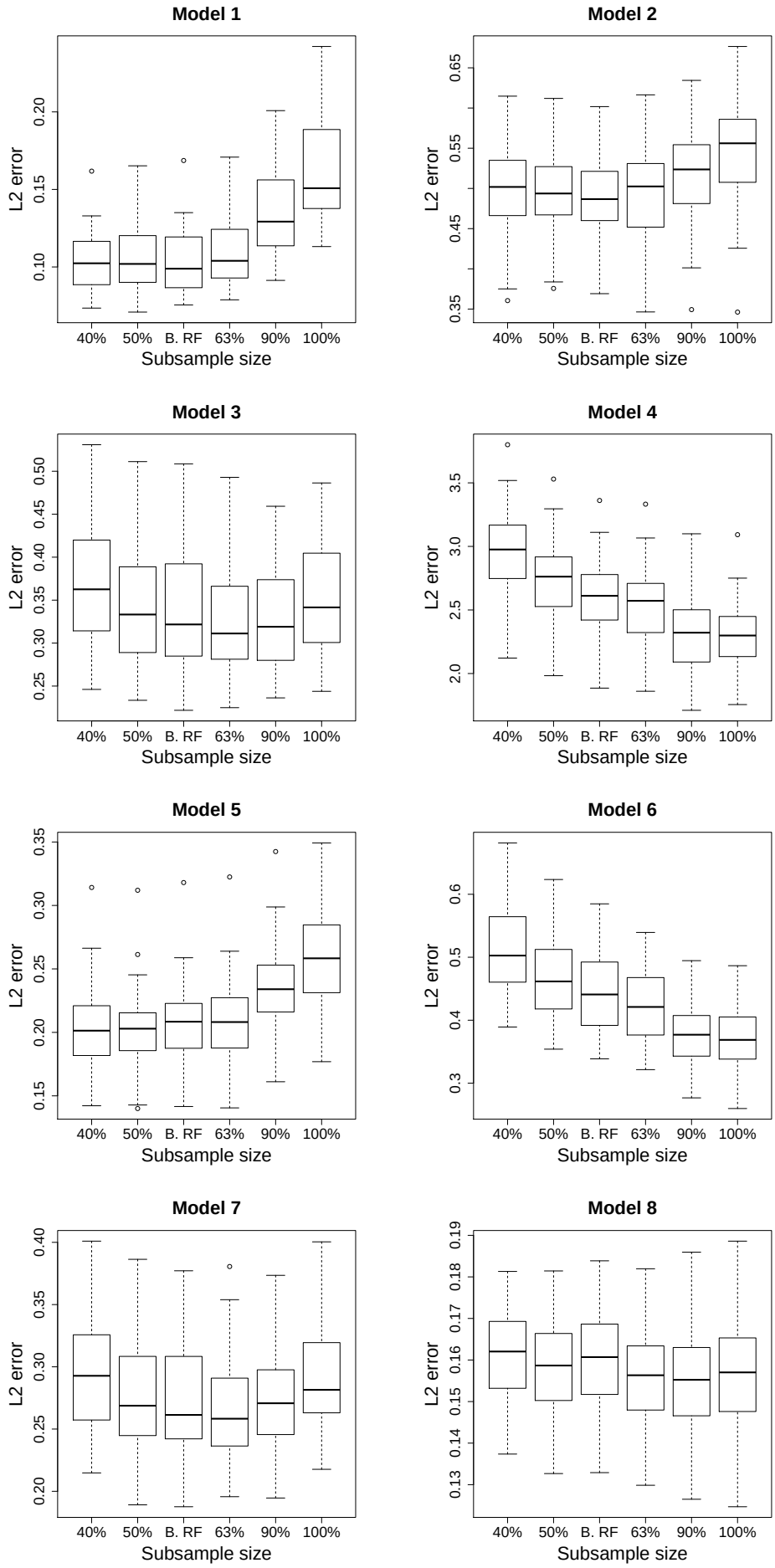


Figure 4.9 – Comparaison de l'erreur $\mathbb{L}^2$ de forêts de Breiman standards avec celles de différentes forêts de Breiman sous-échantillonnées pour des modèles bruités

4.5 Application à la survie

Nous avons étudié précédemment les forêts aléatoires de Breiman et médiane dans le cadre de la régression pour des données non censurées, ceci afin de faciliter l'émergence de leurs propriétés. Revenons maintenant aux forêts de survie aléatoires (RSF). On rappelle que, dans cette procédure, à chaque nœud de chaque arbre, on coupe sur un sous-ensemble aléatoire de directions et en maximisant la différence de survie entre chaque cellule fille. Comme pour les forêts de Breiman, un échantillon bootstrappé est utilisé dans la construction de chacun des arbres. On peut retrouver la construction de ces arbres en détails dans Ishwaran et al. (2008).

Nous illustrons maintenant les résultats présentés en Section 4.3 et 4.4 sur les forêts de survie aléatoires avec les données de cancer du sein collectées à l'Institut Curie introduites en Section 2.6. Nous commençons par comparer les ereurs $\mathbb{L}^2$ d'une forêt de survie classique (Ishwaran et al., 2008) et d'une forêt de survie semi-développée. En effet, nous avons vu en Section 4.4 que ces deux forêts semblaient posséder des risques quadratiques similaires quand le paramètre de profondeur des arbres imposait une profondeur de 30%. La Figure 4.5 présente ces erreurs en fonction du nombre d'arbres qui la composent (de 1 à 500). Les covariables considérées ici pour la régression sont l'âge, le grade histologique, le stade du cancer, le statut du récepteur de progestérone et la taille de la tumeur.

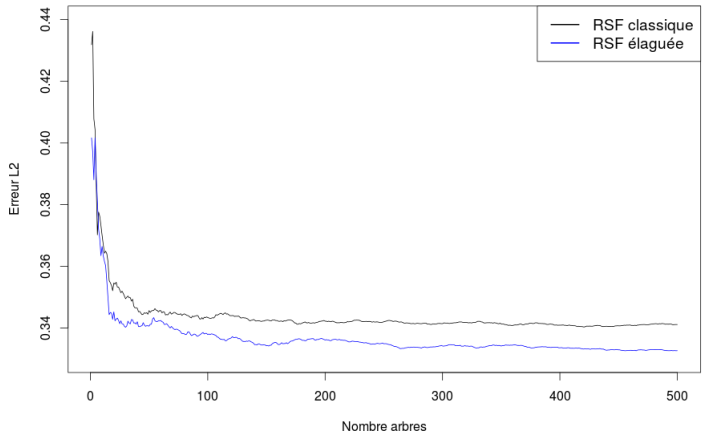
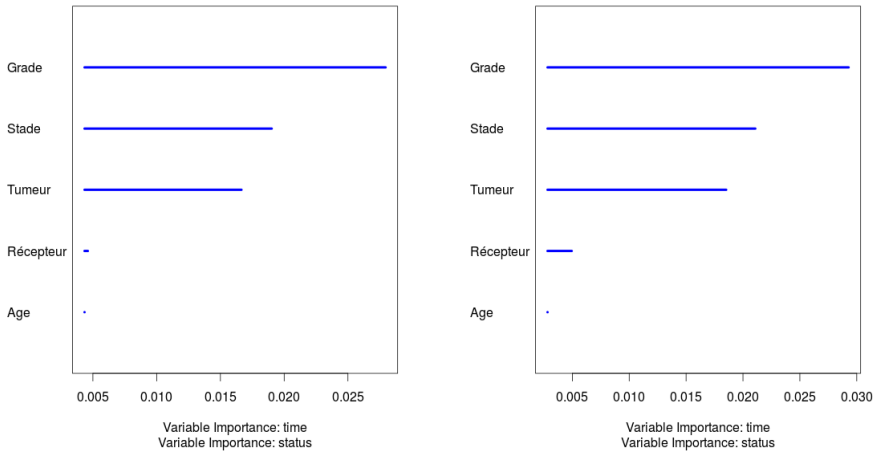


Figure 4.10 – Comparaisons de l'erreur $\mathbb{L}^2$ de la forêt de survie classique et celle de la forêt de survie semi-développée à 30%

Comme pour les forêts de Breiman et médianes, on constate que les performances de la forêt de survie semi-développée sont similaires à celles de la forêt de survie classique, voire un peu meilleur ici. Pour ajouter un point de comparaison entre ces deux forêts, nous les avons étudiées du point de vue de la classification. La Figure 4.11 présente donc l'importance de chaque covariable pour les deux forêts. Ici, l'importance d'une variable $\mathbf{x}$ est la différence entre l'erreur de prédiction quand $\mathbf{x}$ est bruité par permutation aléatoire de sa valeur et l'erreur de prédiction sous le prédicteur original (Breiman, 2001; Liaw and Wiener, 2002; Ishwaran, 2007). On constate encore une fois, grâce à la Figure 4.5, que la forêt de survie classique et la forêt de survie semi-développée donnent à peu près la même importance aux mêmes variables. On remarque que le grade histologique et le stade du cancer semblent être les covariables qui expliquent le mieux le temps de décès. Ceci n'est pas surprenant au regard des résultats obtenus en Section 2.6.



(a) Forêt de survie classique (b) Forêt de survie semi-développée à 30%

FIGURE 4.11 – Importance des différentes covariables pour deux forêts

Chapitre 5

Méthode semi-paramétrique en essais cliniques de phase I dans les cas d'ordre partiel

Ce chapitre est issu d'un article en cours de rédaction réalisé en collaboration avec M. Mathieu Clertant, doctorant du LSTA, et peut également être trouvé dans sa thèse (Clertant, 2015).

5.1 Introduction

This study puts itself in the context of dual-agent Phase I clinical trials. The place of these trials in drugs development is more and more important because of the increasing use of drugs combination in therapies. The practical benefits of drugs combination are numerous : several modes of action can be combined or side effects of a drug can be lightened by another one for example. The aim of Phase I trials, typically in oncology, is to find the one or several drug combinations having the toxicity probability the closest to a threshold α, fixed in advance by clinicians. In general, this threshold is around 25%. We name these combinations the maximum tolerated dose (MTD).

There exist some publications about algorithmic designs used to identify the MTD within a discrete set of doses for Phase I trials Storer (1989); Skolnik et al. (2008) for single-agent trials, Huang et al. (2007) for dual-agent trials. These designs have no modeling, and the escalation, de-escalation rules are determined as a function of some set of the most recent observations, that is why they are used so often. Furthermore they have a Markov property, sometimes referred to in this context as a lack-of-memory property. However such models are flawed and so not fit for purpose, such as almost sure convergence. Multiple model-based designs have been suggested, but they are hardly used in practice. This is even more the case with dual-agent trials because of the complexity and lack of interpretability of the proposed models, but also because of the difficulty to implement these designs with software.

Existing model-based designs can be sorted in two categories. The first one is the class of parametric models. For dual-agent trials, we can cite Wages et al. (2011) with the partial ordering continual reassessment method (CRM), Wang and Ivanova (2005) and Braun and Jia (2013) with extensions of the CRM, Yin and Yuan (2009a) and Yin and Yuan (2009b) with copula models, Thall et al. (2003) with a six-parameter logistic-type model and Braun and Wang (2010) with the use of beta distributions and log-linear

models. The second one is the class of non-parametric models. For dual-agent trials, we can cite Mander and Sweeting (2015) with the product of independent beta probabilities escalation design (PIPE).

In Phase I trials, we are interested in several criteria. Here we consider three of them. Two of them are well explained by Azriel et al. (2011). The first one is called the treatment principle : we want to treat patients at the MTD as often as possible. The second one is called the experimentation principle : we want to obtain a good estimate for the MTD at the end of the study. We know (Azriel et al., 2011) that if the first principle is verified, then the second cannot be, in the sense that we do not have almost sure convergence to the MTD. So there is a trade-off to make, but ideally, we want a design who outperforms the others for these two criteria. The last one is called the coherence, was introduced by Cheung (2005) and can be explain that way : if we observe a dose-limiting toxicity (DLT) for the n-th patient, we do not want the design to recommend a higher dose to the $(n+1)$-th patient ; in the same way, if we do not observe a DLT for the n-th patient, we do not want the design to recommend a lower dose to the $(n+1)$-th patient. This last criterion is essential to be confident in the design.

In this chapter, we introduce a new model-based design for dual-agent trials. It is based on the semi-parametric method (SPM) introduced by Clertant (2015) for single-agent trials. In Section 5.2, we recall some useful notations and results presented in Clertant (2015). In Section 5.3, we present our design and some theoretical results, when the target of the trial is the MTD. In Section 5.4, we recall the definition of the Maximum Tolerated Contour (MTC) introduced by Mander and Sweeting (2015) and extend the SPM in order to target the MTC. Section 5.5 is dedicated to illustrations of our results for combinations of two drugs and comparisons with other dual-agent trials designs (Wages et al., 2011; Mander and Sweeting, 2015). The proofs are available in Appendix B.

5.2 Context and Notations

In this chapter, we are interested in estimating the root of a dose-toxicity regression function as observations are accumulated sequentially. Let (i, j) represents the combination of the i-th dose of a drug A1 and the j-th dose of a drug A2, where $i \in \{1, \cdots, I\}$ and $j \in \{1, \cdots, J\}$. Let D be the set of all the combinations of doses : $D = \{1, \cdots, I\} \times \{1, \cdots, J\}$. We denote the observations by the sequence $(X_n, Y_n)_{n\in\mathbb{N}}$. At step n, corresponding to the n-th patient enrolled in the trial, the variable X_n is the dose selected among the IJ available doses ; and the variable Y_n is the observed binary response at this dose taking values $\{0, 1\}$ (1 for a DLT and 0 otherwise). The conditional distribution of Y_n given $X_n = (i, j)$ follows a Bernoulli distribution with parameter $\beta_{(i,j)}$. This can be written as follow.

Hypothèses 5.2.1. $\forall n \in \mathbb{N}, \quad \mathbb{P}(Y_n = 1 | X_n = (i, j)) = \beta_{(i,j)}$.

For a threshold α, fixed by clinicians, we want to find the dose with a probability of toxicity the closest to α. This is achieved by estimating the root of the regression function. This dose is called the maximum tolerated dose (MTD) and we denote it d^*. It is assumed that the ranges $\{1, \cdots, I\}$ and $\{1, \cdots, J\}$ are ordered in terms of the probability of toxic response. This leads naturally to an order on D. Sign $<$ or $\leq$ will be used for the total ordering on $\mathbb{R}$ and the partial ordering on the set of doses D : $(i, j) \leq (r, s)$ if and only if $i \leq r$ and $j \leq s$.

Hypothèses 5.2.2. $(i, j) < (r, s) \Rightarrow \beta_{(i,j)} < \beta_{(r,s)}$.

For all $X \in D$, this partial ordering partitions D. We denote $A_X = \{(k,l) \in D : (k,l) > X\}$ the set of doses associated with toxicity higher than β_X and $B_X = \{(k,l) \in D : (k,l) < X\}$ the set of doses associated with toxicity lower than β_X. The set $C_X = D \setminus (A_X \cup B_X)$ contains the doses which are not ranked with X. The above sets are illustrated in Figure 5.1. We suppose that all the sample $(X_1^n, Y_1^n) = ((X_1, Y_1), \cdots, (X_n, Y_n))$ can be used for

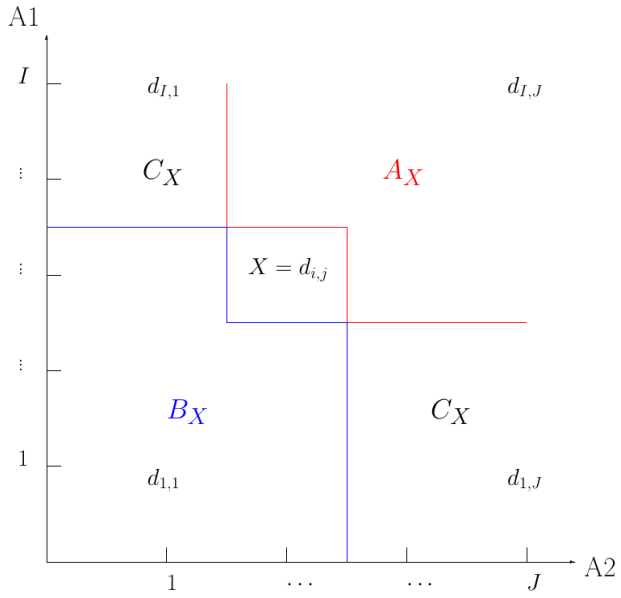


Figure 5.1 – Illustration of the sets A_X, B_X and C_X.

the recommendation of the dose X_{n+1}. More precisely, for all the methods exposed in this chapter, X_{n+1} is completely determined by the history (X_1^n, Y_1^n).

Hypothèses 5.2.3. *The current estimator of the method $\mathcal{M}$ satisfies $\mathcal{M}(X_1^n, Y_1^n) \in \sigma(X_1^n, Y_1^n)$, where $\sigma(X_1^n, Y_1^n)$ is the sigma-algebra generated by the sample.*

We now introduce a generalization, for the case of partial ordering, of the coherence principle introduced by Cheung (2005) for the case of total ordering.

Définition 5.1 (Partial ordering coherence)**.** A method $\mathcal{M}$ is coherent if its current estimator satisfies :

$$(X_n, Y_n) = (d, 0) \Rightarrow \mathcal{M}(X_1^n, Y_1^n) \in D \backslash B_d \text{ and } (X_n, Y_n) = (d, 1) \Rightarrow \mathcal{M}(X_1^n, Y_1^n) \in D \backslash A_d.$$

First condition is the coherence in de-escalation and second the coherence in escalation.

This latter definition matches the one of Cheung (2005) in case of a single-agent trial, and so is a reasonable extension of this criteria in case of partial ordering.

5.3 Modelling from an MTD point of view

5.3.1 General setting

The Semi-Parametric Method in case of Partial Ordering (poSPM) is the generalisation of SPM introduced by Clertant (2015). Let $F = [0,1]^D$ be the set of curves (or functions)

which go from D to $[0,1]$. An element of F is denoted by $q = (q_{(i,j)})$ where $i \in \{1, \cdots, I\}$ and $j \in \{1, \cdots, J\}$. Then, $q_{(i,j)}$ is a probability of toxicity for the dose (i,j). In particular, $\beta \in F$. In the context of Bayesian hierarchical framework, let θ be the random variable of interest taking values in D. We now partition F according to θ :

$$F = \bigcup_{\theta \in D} F_\theta, \text{ where } F_\theta = \left\{ q \in F \ / \ \forall (i,j) \in D, \ |q_\theta - \alpha| \leq |q_{(i,j)} - \alpha| \right\}.$$

We can see that F is partitioned in IJ classes F_θ, where every curve of a same class have the same MTD, by definition. For a fixed $q \in F$, we can make explicit the probability of the history (X_1^n, Y_1^n) :

$$\begin{aligned} \mathbb{P}_q(X_1^n, Y_1^n) &= \mathbb{P}_q(Y_1|X_1) \times \prod_{p=1}^{n} \left[\mathbb{P}_q\left(Y_p|X_p\right) \mathbb{P}_q\left(X_p|X_1^{p-1}, Y_1^{p-1}\right) \right] \\ &= \prod_{p=1}^{n} \left(q_{X_p}\right)^{Y_p} \left(1 - q_{X_p}\right)^{1-Y_p} = \prod_{(i,j) \in D} q_{(i,j)}^{n^1_{(i,j)}} \left(1 - q_{(i,j)}\right)^{n^0_{(i,j)}}, \end{aligned}$$

where $n^1_{(i,j)}$ represents the number of DLTs occurred at dose (i,j), and $n_{(i,j)} = n^1_{(i,j)} + n^0_{(i,j)}$ the number of patients treated at dose (i,j). We endow F with a probability measure $\Lambda \otimes \Pi$, where Π is a measure on D, and the topological support of the measure $\Lambda(.|\theta) = \Lambda_\theta(.)$ is included in F_θ. We can now express the posterior distribution of θ given the history (X_1^n, Y_1^n) :

$$\Pi_n(\theta) = \Pi(\theta|X_1^n, Y_1^n) = \frac{\int \{(i,j) \in D\} \prod q_{(i,j)}^{n^1_{(i,j)}} \left(1 - q_{(i,j)}\right)^{n^0_{(i,j)}} \Lambda_\theta(dq)\, \Pi(\theta)}{\sum_{\theta \in D} \int \prod_{(i,j) \in D} q_{(i,j)}^{n^1_{(i,j)}} \left(1 - q_{(i,j)}\right)^{n^0_{(i,j)}} \Lambda_\theta(dq)\, \Pi(\theta)}. \tag{5.1}$$

Let us assume that Λ_θ is absolutely continuous with respect to a measure ν with density λ_θ. Then the posterior distribution of F_θ given (X_1^n, Y_1^n) is

$$\lambda_{\theta,n}(q) = \frac{\prod_{(i,j) \in D} q_{(i,j)}^{n^1_{(i,j)}} \left(1 - q_{(i,j)}\right)^{n^0_{(i,j)}} \lambda_\theta(q)}{\int \prod_{(i,j) \in D} q_{(i,j)}^{n^1_{(i,j)}} \left(1 - q_{(i,j)}\right)^{n^0_{(i,j)}} \lambda_\theta(q) \nu(dq)}. \tag{5.2}$$

Using Equation (5.1) and Equation (5.2) together, we obtain :

$$\Pi_n(\theta) = \frac{\left[\int q_{X_n}^{Y_n} \left(1 - q_{X_n}\right)^{1-Y_n} \lambda_{\theta,n-1}(q) \nu(dq) \right] \Pi_{n-1}(\theta)}{\sum_{\theta \in D} \left[\int q_{X_n}^{Y_n} \left(1 - q_{X_n}\right)^{1-Y_n} \lambda_{\theta,n-1}(q) \nu(dq) \right] \Pi_{n-1}(\theta)}.$$

The distribution Π is updated by weighting according to the expected likelihood. At step n, we integrate the likelihood according to the posterior $\Lambda_{\theta,n-1}$. Thus, the amount of information of a fixed observation (X_n, Y_n) varies according to the whole history (X_1^{n-1}, Y_1^{n-1}). The family of distributions $(\Lambda_\theta)_{\theta \in D}$ plays a predictive and adaptive model-like role and will be called prior model. We then choose the estimator of the MTD (or current estimator) such as

$$\hat{\theta}_n = \underset{\theta \in D}{\operatorname{argmax}}\ \Pi_n(\theta). \tag{5.3}$$

Note that we can also obtain estimators for every probability $\beta_{(i,j)}$:

$$\hat{\beta}^{(n)}_{(i,j)} = \mathbb{E}_{(\Lambda\otimes\Pi)_n}\left[q_{(i,j)}\right] = \sum_{q\in D}\left[\int q_{(i,j)}\Lambda_{\theta,n}(dq_{(i,j)})\right]\Pi_n(\theta).$$

Now let S_θ and S^d_θ be the topological support of Λ_θ and its marginal Λ^d_θ, with $d \in D$. The interval of toxicity $[0,1]$ is partitioned into three sets : $I = [\alpha-\varepsilon_1, \alpha-\varepsilon_2]$, $B = [0, \alpha-\varepsilon_1]$ and $A = [\alpha+\varepsilon_2, 1]$, with $\varepsilon_1, \varepsilon_2 \in [0, \max(\alpha, 1-\alpha)]$. For all doses θ and d, the marginal support S^d_θ satisfies :

$$\begin{cases} d \in A_\theta & \Rightarrow & S^d_\theta \subset A \\ d \in B_\theta & \Rightarrow & S^d_\theta \subset B \\ d \in C_\theta & \Rightarrow & S^d_\theta \subset [0,1] \\ d = \theta & \Rightarrow & S^d_\theta \subset I \end{cases}$$

5.3.2 A simple and coherent prior model

As explained in Clertant (2015), we are now interested in the prior distribution inside the class (Λ) and call the family $(\Lambda_\theta)_{\theta\in D}$ the prior-model because we use it to make inference. In a single class F_θ, the Bernoulli's parameters at each dose $q_{(i,j)}$ are considered independent, which is summarised by the following assumption.

Hypothèses 5.3.1. *Λ_θ is a product of unidimensionnal distribution, i.e.*

$$\Lambda_\theta(dq) = \prod_{i=1}^{I}\prod_{j=1}^{J}\Lambda^{(i,j)}_\theta(dq_{(i,j)}).$$

When there is no confusion, we write $\Lambda_\theta(dq_{(i,j)})$ for $\Lambda^{(i,j)}_\theta(dq_{(i,j)})$. We then state a stochastic partial ordering assumption on the prior-model.

Hypothèses 5.3.2. *Let d and d' be two doses such that $d < d'$. For all marginal $(i,j) \in D$, the posterior $\Lambda^{(i,j)}_{d,n}$ is stochastically greater than $\Lambda^{(i,j)}_{d',n}$:*

$$\Lambda^{(i,j)}_{d,n}([0,x]) \leq \Lambda^{(i,j)}_{d',n}([0,x])\ \forall x \in [0,1].$$

These assumptions enable us to state the following theorem.

Théorème 5.2. *Under Assumptions 5.3.1 and 5.3.2, if $(\Lambda_\theta)_{\theta\in D}$ is the single prior model used, then the poSPM is coherent in the sense of Definition 5.1.*

Proof. We only need to deal with the following cases :

— Progression in drug $A1$: Let $s \in \{1,\cdots,J\}$ and $(r,t) \in \{1,\cdots,I\}^2$ such as $r < t$,

— Progression in drug $A2$: Let $s \in \{1,\cdots,I\}$ and $(r,t) \in \{1,\cdots,J\}^2$ such as $r < t$.

Suppose that $Y_{n+1} = 1$. The case $Y_{n+1} = 0$ can be solved identically. By construction, we have

$$\Pi_{n+1}(\theta) \propto \left[\int q\Lambda_{\theta,n}(dq)\right]\Pi_n(\theta).$$

Furthermore,

$$\begin{aligned}\int q_{(r,s)}\Lambda_{\theta,n}(dq) &= \int \left[\int \mathbb{1}_{\{0\leq x\leq q_{(r,s)}\}}\mu(dx)\right]\Lambda_{\theta,n}(dq_{(r,s)})\\ &= \int \Lambda_{\theta,n}^{(r,s)}(]x,1])\mu(dx).\end{aligned}$$

If $\hat{\theta}_n = (r,s)$, then for all $\theta \in D$, $\Pi_n(\theta) \leq \Pi_n(r,s)$. Let $t > r$. According to Assumption 5.3.2, we know that $\Lambda_{(r,s),n}^{(r,s)}(.)$ is stochastically greater than $\Lambda_{(t,s),n}^{(r,s)}(.)$, *i.e.*

$$\int \Lambda_{(r,s),n}^{(r,s)}(]x,1])\mu(dx) \geq \int \Lambda_{(t,s),n}^{(r,s)}(]x,1])\mu(dx),$$

that is

$$\int q_{(r,s)}\Lambda_{(r,s),n}(dq) \geq \int q_{(r,s)}\Lambda_{(t,s),n}(dq).$$

Finally $\Pi_{n+1}(r,s) \geq \Pi_{n+1}(t,s)$. With the same arguments, we can show that $\Pi_{n+1}(s,r) \geq \Pi_{n+1}(t,r)$ for $t > s$. So the poSPM is coherent by definition of $\hat{\theta}_n$ (Equation (5.3)). □

In the following example, the prior model satisfies the assumptions 5.3.1 and 5.3.2 and the conjugacy for the likelihood.

Example 5.3.1. *The prior model is defined by a triplet* $\left((S_\theta)_{\theta\in D}, (q^\theta)_{\theta\in D}, c\right)$. *The sets* S_θ *are the topological supports of the distributions* Λ_θ *and for all* $(i,j) \in D$, *the marginal supports,* $S_\theta^{(i,j)}$, *fulfill the following assumption.*

Hypothèses 5.3.3. *The topological supports satisfy :*

$$\begin{cases} (i,j) \in A_\theta & \Rightarrow \quad S_\theta^{(i,j)} = A \\ (i,j) \in B_\theta & \Rightarrow \quad S_\theta^{(i,j)} = B \\ (i,j) \in C_\theta & \Rightarrow \quad S_\theta^{(i,j)} = [0,1] \\ (i,j) = \theta & \Rightarrow \quad S_\theta^{(i,j)} = I, \end{cases}$$

with θ *and* (i,j) *in* D.

The vectors $q^\theta \in [0,1]^m$ *are the modes of the distributions* Λ_θ. *Thus* $q^\theta_{(i,j)}$ *takes the maximum value of the density function* $\lambda_\theta^{(i,j)}$ *:*

$$q^\theta = \arg\max_{q\in F_\theta} \Lambda_\theta(q)\ .$$

These modes are ranked according to the product order : if $(i,j) < (r,s)$ *then* $q^\theta_{(i,j)} < q^\theta_{(r,s)}$. *The positive vector* $c = (c_1, c_2)$ *is a dispersion parameter of the distributions ;* c_1 *corresponds to a number of pseudo-patients observed at the doses ranked with* θ, *i.e. the doses in the set* $D \setminus C_\theta$ *; and* c_2 *corresponds to a number of pseudo-patients observed at the doses non-ranked with* θ, *i.e. the doses in the set* C_θ. *Uniform priors on the topological supports are updated. We denote* l *the real function such as for all* $x, y, z \in [0,1]$, $l(x,y,z) = x^{yz}(1-x)^{y(1-z)}$. *Then, for all* $\theta \in D$ *and all* $q \in F_\theta$, *we have the following result.*

$$\lambda_\theta^{(i,j)}(q_{(i,j)}) \propto \begin{cases} l(q_{(i,j)}, c_1, q^\theta_{(i,j)})\mathbb{1}_{\{q_{(i,j)}\in A\}} & \text{if} \quad (i,j) \in A_\theta, \\ l(q_{(i,j)}, c_1, q^\theta_{(i,j)})\mathbb{1}_{\{q_{(i,j)}\in B\}} & \text{if} \quad (i,j) \in B_\theta, \\ l(q_{(i,j)}, c_2, q^\theta_{(i,j)})\mathbb{1}_{\{q_{(i,j)}\in [0,1]\}} & \text{if} \quad (i,j) \in C_\theta, \\ l(q_{(i,j)}, c_1, q^\theta_{(i,j)})\mathbb{1}_{\{q_{(i,j)}\in I\}} & \text{if} \quad (i,j) = \theta. \end{cases}$$

Thus, Λ_θ *is a product of beta priors on* $[0,1]$ *and incomplete beta priors on* A, B *and* I.

In that case, the marginals of the prior model are absolutely continuous according to the Lebesgue measure. This last assumption is based on the regularity of the prior model.

Hypothèses 5.3.4. *The following conditions are valid except when Λ_θ^θ is a Dirac measure.*

(a) For all $(i,j) \in D$, the marginal distribution $\Lambda_\theta^{(i,j)}$ is absolutely continuous with respect to the Lebesgue measure and $\lambda_\theta^{(i,j)}$ denotes its density function.

(b) There exist two numbers s and S in $\mathbb{R}_+^*$, such that, for all θ and (i,j) in D, we have :

$$\forall\, q_{(i,j)} \in S_\theta^j\,,\ s < \lambda_\theta^{(i,j)}(q_{(i,j)}) < S\,.$$

The second point is only useful for the sake of the demonstration when some toxicities $\beta_{(i,j)}$ are equal to 0 or 1.

5.3.3 Asymptotical results and perspective

In the Bayesian paradigm, the choice of a topological support for the prior model determines consistency. It is therefore a central step for poSPM and we will design it according to the goal of our study. In the previous example, the support of marginal Λ_θ^θ is $I = [\alpha - \varepsilon_1, \alpha + \varepsilon_2]$. In that case, observing at dose θ leads eventually to recommend θ if and only if the toxicity β_θ is included in I. Such an assumption on the scenario is necessary for the almost sure convergence to the MTD (Azriel et al., 2011, Theorem 1). In case of a range of doses totally ordered, this leads to consider the ε-sensitivity behaviour of a method (Cheung, 2011). The definition does not change in case of partial ordering.

Définition 5.3. Let $\varepsilon \geq 0$ and $I = [\alpha - \varepsilon; \alpha + \varepsilon]$. We consider the set $\mathcal{E}(I,\beta)$ of the collection of doses associated with a toxicity belonging to I_ε, *i.e.* $\mathcal{E}(I,\beta) = \{j \in D : \beta_j \in I\}$. A method, $\mathcal{M}$, is called ε-sensitive, if for all β such that $\mathcal{E}(I,\beta) \neq \emptyset$, we have :

$$\mathbb{P}_\beta\left[\exists N\,,\ \forall n > N : \mathcal{M}(X_1^n, Y_1^n) \in \mathcal{E}(I,\beta)\right] = 1 \quad .$$

The ε-sensitivity corresponds to a strong consistency to one dose associated with toxicity close to the threshold α. The almost sure convergence to the MTD is obtained if the MTD is the single dose in $\mathcal{E}(I,\beta)$. This involves to choose a little interval I which could contain no toxicity associated with a dose in our range. For this reason, the complementary behaviour of a method, called ε-balanced, is introduced by Clertant (2015) in case of total ordering. We extend this definition to the case of partial ordering. Let $\delta(.,.)$ be the euclidean distance. A sequence $(X_n)_{n\in\mathbb{N}}$ converges to a set B, denoted by $X_n \xrightarrow{S} B$, if

$$\sup_{x\in B}\left(\liminf_{n\to\infty} \delta(X_n, x)\right) = 0\,.$$

The key notion is the minimal set of doses on which we need to have observations in order to determine almost surely the MTD. Let L and U be the partition of D into the sets of doses associated with toxicities respectively lower and upper than α, *i.e.* $L = \{d \in D : \beta_d \leq \alpha\}$ and $U = \{d \in D : \beta_d \geq \alpha\}$. The minimal set M_D satisfies $M_D = M_L \cup M_U$, where $M_L = \{d \in L : A_d \cap L = \emptyset\}$ and $M_U = \{d \in L : B_d \cap U = \emptyset\}$. In case of total ordering, M_D is equal to $\{a, b\}$, where b (below) and a (above) are the two consecutive doses associated to toxicities either side of the target α. This allow us to introduce the following general definition.

Définition 5.4. Let D be a range of doses. A method, $\mathcal{M}$, is called ε-balanced, if for all β such that $\mathcal{E}(I,\beta) = \emptyset$, we have :

$$\mathcal{M}(X_1^n, Y_1^n) \xrightarrow{S} M_D\,,\ a.s.$$

Note that if $I = \{\alpha\}$, that is if $\varepsilon = 0$, this behaviour is referred as balanced.

This means that if no doses are associated with a toxicity in I, the doses recommended infinitely often by the current estimator are all the doses in M_D and only these doses. In order to state asymptotical properties of the poSPM, we introduce the set $\tilde{D}$ of doses infinitely observed :

$$j \in \tilde{D} \Leftrightarrow n_j \underset{n\to\infty}{\longrightarrow} \infty.$$

According to the prior model parametrization, the poSPM has the following asymptotic properties.

Théorème 5.5. *Under assumptions 5.3.1, 5.3.3 and 5.3.4 the poSPM is ε-sensitive and ε-balanced.*

The ε-balanced property allows the poSPM to concentrate the observations on the minimal set M_D which, in a asymptotical point of view, leads to determine almost surely the MTD. Indeed, the MTD is included in the minimal set for which each dose is infinitely observed. However, this behaviour is only asymptotical and, at a finished rank, the simulations seem often to account a convergence to a single dose identified as the MTD. In the usual case of one dimensional space of drugs, some questions are recently raised about the pragmatical aim of an early phase dose finding study in relation with the expansion cohort. The opportunity to maintain more than one dose in the study could lead to better results (Clertant, 2015). From this point of view and under certain conditions of proximity between the toxicities and the threshold, the minimal set $\{a, b\}$ has the property to be the best candidate. This new paradigm is particularly relevant in the two dimensional context in which the goal of finding the MTD is less obvious. Indeed, the dose/toxicity and dose/efficiency relations can both verify the partial ordering assumption. In this case, there exist some configurations where a dose less toxic than the MTD is more efficient. Such a scenario is impossible in case of total ordering for the two relation considered previously. Moreover, in the partial ordering case, and from the point of view of extrapolation, the observations are usually less informative than in the total order case. Thus, in our situation, the MTD which is not necessarily the most adapted response for the clinicians is even more difficult to find. That is why it seems important to have the capacity to recommend to the phase II a sensitive set of doses which contains the MTD with high probability and other doses close enough to it.

5.4 Modelling from an MTC point of view

In the partial ordering case, the minimal set is less intuitive than in the case of total ordering. However, it is possible to consider the minimal set on each line $l_j = \{(i, j)\,,\ i \in \{1, \cdots, I\}\}$ and column $c_i = \{(i, j)\,,\ j \in \{1, \cdots, J\}\}$. This leads to the notion of Maximum Tolerated Contour introduced by Mander and Sweeting (2015) and for which we give the following formal definition :

$$MTC = \bigcup_{(i,j)\in\{1,\ldots,I\}\times\{1,\ldots,J\}} (M_{c_i} \cup M_{l_j})\,. \tag{5.4}$$

From a treatment efficiency perspective, clinicians could prefer to find the MTC than the MTD. Note that M_D is also the minimal set of doses on which we need to have observations in order to determine with certainty the MTC. This provides a particular interest to the balanced property as it unifies different points of view.

The class of Partial Ordering and Semi-Parametric Methods on the Contour (poSPMc) are inspired by the PIPE model introduced by Mander and Sweeting (2015) and is an extension of unidimensional semi-parametric methods on cut (Clertant, 2015) to the partial ordering case. The modelization includes the PIPE model as a particular case. Now let $\mathcal{K}$ be the set of the available contours which are not in contradiction with the Assumption 5.2.2. The letter $\mathcal{C}$ denotes an element of $\mathcal{K}$ and $\mathcal{C}^*$ is the true MTC. We first define the MTC, using the general notion of contour, as the set of couple of doses which are the minimal set on each lines and columns of D, see Equation (5.4). Each contour can also

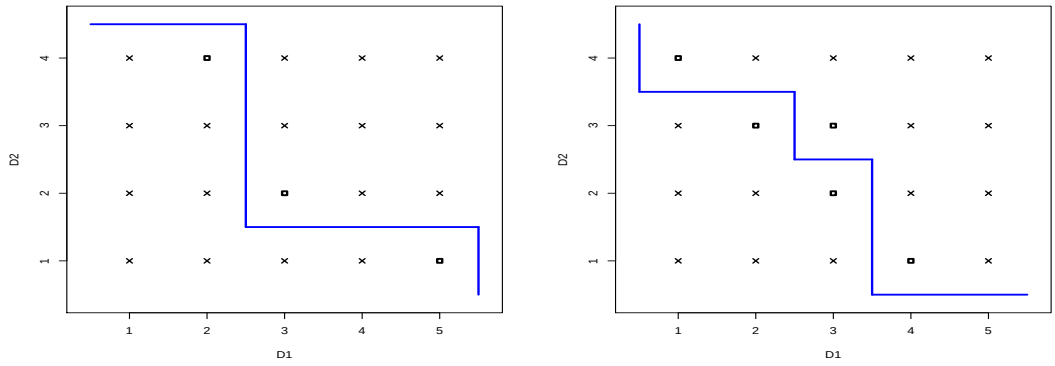


Figure 5.2 – **Two contours**, the symbol □ represents the doses of minimal set and × the other doses.

be seen as a polygonal chain going from $(0.5, J+0.5)$ to $(I+0.5, 0.5)$ by step of size 1 in abscissa and -1 in ordinate. Examples of two different contours can be found in Figure 5.2. This second definition allows us to find easily the total number of contours : $\#\mathcal{K} = \binom{I+J}{I}$. For each contour $\mathcal{C}$, we introduce three sets of dose. $L(\mathcal{C}) = \{d \in D : \exists d' \in \mathcal{C},\ d' > d\}$ and $U(\mathcal{C}) = \{d \in D : \exists d' \in \mathcal{C},\ d' < d\}$ are respectively the set of doses lower and upper than $\mathcal{C}$, and $M(\mathcal{C}) = \{d \in L(\mathcal{C}) : A_d \cap L(\mathcal{C}) = \emptyset\} \cup \{d \in U(\mathcal{C}) : B_d \cap U(\mathcal{C}) = \emptyset\}$ denotes the minimal set corresponding to the contour $\mathcal{C}$. In particular, $M(\mathcal{C}^*) = M_D$. Note that finding the contour is equivalent to find the minimal set. We endow the set of all scenario, F, with a probability measure $\tilde{\Lambda} \otimes \tilde{\Pi}$, where $\tilde{\Pi}$ is a measure on $\mathcal{K}$. The distribution $\tilde{\Lambda}_{\mathcal{C}} = \tilde{\Lambda}(.|\mathcal{C})$ and its support $S_{\mathcal{C}}$ satisfy the following hypotheses.

Hypothèses 5.4.1. *(a) $\tilde{\Lambda}_{\mathcal{C}}$ is a product of unidimensionnal distribution :*

$$\tilde{\Lambda}_{\mathcal{C}}(dq) = \prod_{i=1}^{I} \prod_{j=1}^{J} \tilde{\Lambda}_{\mathcal{C}}^{(i,j)}(dq_{(i,j)}).$$

Let $S_{\mathcal{C}}^d$ be the marginal support for dose d. We have

$$\begin{cases} d \in L(\mathcal{C}) & \Rightarrow \quad S_{\mathcal{C}}^d = [0, \alpha] \\ d \in U(\mathcal{C}) & \Rightarrow \quad S_{\mathcal{C}}^d = [\alpha, 1]. \end{cases}$$

(b) For all $d \in D$, the marginal distribution $\tilde{\Lambda}_{\mathcal{C}}^d$ can have an atom in α and is absolutely continuous with respect to the Lebesgue measure on the rest of its support. $\tilde{\lambda}_{\mathcal{C}}^d$ denotes its density function. There exist two numbers s and S in $\mathbb{R}_+^*$, such that, for all $\mathcal{C}$ and d,

$$\forall q_d \in S_{\mathcal{C}}^d,\ \ s < \tilde{\lambda}_{\mathcal{C}}^d(q_d) < S.$$

We obtain the posterior distribution $\tilde{\Pi}_n$ of $\mathcal{C}$ in the same way as Π (Equation (5.1)) and a natural estimator of the MTC is $\tilde{\mathcal{C}}_n = \arg\max_{\mathcal{C}\in\mathcal{K}} \tilde{\Pi}_n(\mathcal{C})$. From an asymptotical point of view, Assumption 5.4.1 (b) may be deleted when the marginals of $\tilde{\Lambda}$ satisfy

$$\forall \mathcal{C}, \mathcal{C}' \in \mathcal{K},\ \forall d \in (U(\mathcal{C}) \cap U(\mathcal{C}')) \cup (L(\mathcal{C}) \cap L(\mathcal{C}')),\ \tilde{\Lambda}^d_{\mathcal{C}} = \tilde{\Lambda}^d_{\mathcal{C}'}. \tag{5.5}$$

This assumption is fulfilled by the PIPE method. It is not sure that such a restriction on the prior model and the localization of the weights leads to weaker performances. Note that, in the modelization of PIPE, the prior model $\tilde{\Lambda}$ and the distribution $\tilde{\Pi}$ are never explicitly mentioned and are, in some way, indifferentiated. The whole parameter of the poSPMc is summarized by the family $(a_d, b_d)_{d\in D}$, where the couple (a_d, b_d) is a fictive amount of observations on the dose d. We integrate this piece of information into a uniform prior on the parametric set of toxicity at dose d. When we include the PIPE method in the poSPMc settings, this design fulfils Assumptions 5.4.1 (a) and Equation (5.5), and its probability distribution $\tilde{\Lambda} \otimes \tilde{\Pi}$ satisfies, for all $\mathcal{C}$ and d, $\tilde{\Lambda}^d_{\mathcal{C}} \propto B(\alpha; a_d, b_d)\mathbb{1}_{d\in L(\mathcal{C})} \times (1 - B(\alpha; a_d, b_d))\mathbb{1}_{d\in U(\mathcal{C})}$ and

$$\tilde{\Pi}(\mathcal{C}) \propto \prod_{d\in L(\mathcal{C})} B(\alpha; a_d, b_d) \times \prod_{d\in U(\mathcal{C})} (1 - B(\alpha; a_d, b_d)),$$

where $B(\alpha; a_d, b_d)$ is the incomplete beta function. There are some practical problems with the indifferentiation of $\tilde{\Pi}$ and $\tilde{\Lambda}$. The modelization of PIPE is forced to use a very weak prior-model because the weight of information contained in $(a_d, b_d)_{d\in D}$ has a strong impact on the distribution $\tilde{\Pi}$. When this model is parametrized for a threshold at 0.25 with $\sum_{d\in D} a_d + b_d = 1$, as recommended in the original article, and the prior strength is constant across the doses (with $\#D = 12$), $B(\alpha, a_d, b_d)$ varies in the interval $]0.2227, 0.2677[$. The distribution Π is then almost proportional to a product of interval's length ($\simeq 0.25$ and $\simeq 0.75$). In this configuration, when the aim of the trial is to find a lower threshold as 0.1, the strength of the prior $\tilde{\Pi}$ is then increased. From our point of view, it is more convenient to choose directly the prior on $\mathcal{K}$ and the prior model without hidden effect due to the threshold considered or the local strength of prior model. This is the main theoretical benefit of the poSPMc which corresponds to a broader class of methods more easily calibrated through the distinction of $\tilde{\Pi}$ and $\tilde{\Lambda}$.

The poSPMc is a class of sequential methods for which the law $\tilde{\Pi}_n$ does not provide an obvious estimator of the next dose. The estimation of the next dose is made in two steps : 1) update the estimator of the contour $\tilde{C}_n$, and 2) according to the result of step 1, use a specific selection strategy to choose the next dose. One strategy is introduced in Mander and Sweeting (2015, 2.1). It satisfies the following assumption.

Hypothèses 5.4.2. *(a) The dose X_{n+1} is selected in $M(\tilde{\mathcal{C}}_n)$.*

(a) If the contour $\mathcal{C}$ is infinitely recommended then all the doses in $M(\mathcal{C})$ are infinitely selected. In other terms,

$$\left\{ n_{\mathcal{C}} = \sum_{i=1}^{n} \mathbb{1}_{\{\tilde{\mathcal{C}}_n = \mathcal{C}\}} \underset{n\to\infty}{\longrightarrow} \infty \right\} \Longrightarrow \forall d \in M(\mathcal{C}),\ \left\{ n_d \underset{n\to\infty}{\longrightarrow} \infty \right\},\ a.s.$$

Assumption 5.4.2 (*a*) is natural while Assumption 5.4.2 (*b*) is often satisfied and allows us to lay out an asymptotical property. This result is also valid for the PIPE method as particular case of the poSPMc.

Théorème 5.6. *Under Assumptions 5.4.1 and 5.4.2, the poSPMc leads to a strongly consistent estimator of the MTC and its minimal set, i.e. $\tilde{\mathcal{C}}_n \underset{n\to\infty}{\longrightarrow} \mathcal{C}^*$, a.s.*

Proof. We note, as in the proof of ε-balanced behaviour, that the regularity assumption 5.4.1 (or assumption alternative (5.5)) involves

$$\delta(\beta_k, S_r^k) = \delta(\beta_k, S_t^k) \Longrightarrow 0 < \liminf_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} \leq \limsup_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} < \infty, \ a.s.$$

We will show the following assertion

$$\mathcal{C} \in \mathcal{K} \setminus \mathcal{C}^* \Longrightarrow \mathbb{P}(n_{\mathcal{C}} \to \infty) = 0 \tag{5.6}$$

As $\mathcal{C} \neq \mathcal{C}^*$, we have $M(\mathcal{C}) \neq M(\mathcal{C}^*)$ and there exists $k \in M(\mathcal{C})$ such that β_k is included in $S_{\mathcal{C}^*}^k$ and not in $S_{\mathcal{C}}^k$. By using Proposition B.1, we have

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{M_{n,\mathcal{C}}^k}{M_{n,\mathcal{C}^*}^k} = 0 \middle| n_{\mathcal{C}} \to \infty\right) = \mathbb{P}\left(\lim_{n\to\infty} \frac{M_{n,\mathcal{C}}^k}{M_{n,\mathcal{C}^*}^k} = 0 \middle| n_k \to \infty\right) = 1,$$

where the first equality arises from Assumption 5.4.2(b). For all doses $d \in D$, the distribution $\Lambda_{\mathcal{C}^*}$ models properly the toxicity, in other words $\beta_d \in S_{\mathcal{C}^*}$. We then have

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{I_{n,\mathcal{C}}}{I_{n,\mathcal{C}^*}} = 0 \middle| n_{\mathcal{C}} \to \infty\right) = 1 \text{ and } \mathbb{P}\left(\left\{\lim_{n\to\infty} \frac{I_{n,\mathcal{C}}}{I_{n,\mathcal{C}^*}} = 0\right\} \cap \{n_{\mathcal{C}} \to \infty\}\right) = 0,$$

which proves implication (5.6). □

The convergence to $\mathcal{C}^*$ is provided by the convergence to the minimal set in the same way as the balanced behaviour of the poSPM. Indeed, there exist some relations between the modelling from the MTC and the MTD point of view. For illustrate these links, we introduce some modifications of the poSPM. The range of doses on which is build the prior Π is extended to the set $\overline{D} = D \cup \{u, l\}$. The doses u and l are fictive. They can be seen as an additional parametrization of doses $(1,1)$ and (I,J), respectively. From the point of view of l and u, all the toxicities are respectively lower and upper than the threshold α. The topological support of Λ satisfies now the following assumption.

Hypothèses 5.4.3. *All the distributions Λ_θ are product of unidimensional distributions. Their topological supports satisfy Assumption 5.3.3(ii) with $\varepsilon = 0$, and we have $S_u = A^m$ and $S_l = B^m$.*

Let $\overline{\Pi}_n$ be the distribution provided by Π on the range of doses D such as $\overline{\Pi}_n(1,1) = \Pi_n(u) + \Pi_n(1,1)$, $\overline{\Pi}_n(I,J) = \Pi_n(I,J) + \Pi_n(l)$ and $\overline{\Pi}_n(d) = \Pi_n(d)$ otherwise. The natural estimator of the MTD is then $\hat{\theta}_n = \arg\max_{\theta\in D} \overline{\Pi}_n(\theta)$. Finally, for modification, we do not take into account the update of Π for the non-ordered dose. Such an assumption is reasonable having regard to the willingness to explore the whole contour. The posterior of Π according to the observations is then

$$\Pi_n(\theta) \propto \int \prod_{(i,j)\in D\setminus C_\theta} q_{(i,j)}^{n_{(i,j)}^1} \left(1 - q_{(i,j)}\right)^{n_{(i,j)}^0} \Lambda_\theta(dq)\, \Pi(\theta). \tag{5.7}$$

This update fosters doses poorly observed, that is doses d with only a few observations on d, A_d and B_d. This latest modification ensures the almost sure convergence of the poSPM to the minimal set. The modelling from the MTD provides naturally a modelling from the MTC :

$$\forall \mathcal{C} \in \mathcal{K},\ \tilde{\Pi}(\mathcal{C}) \propto \sum_{\theta\in M(C)} \Pi(\theta) \text{ and } \tilde{\Lambda}_{\mathcal{C}} = \sum_{\theta\in M(\mathcal{C})} r_{\mathcal{C}}^\theta \times \Lambda_\theta, \tag{5.8}$$

where $r_{\mathcal{C}}^{\theta} = \Pi(\theta)/\sum_{\theta\in M(C)}\Pi(\theta)$. Note that in Equation (5.8), the minimal set $M(\mathcal{C})$ might be replaced by the whole contour. When the poSPM works under Relation (5.7) and Assumption 5.4.3, the poSPMc defined by Equation (5.8) satisfies Assumption 5.4.1 (a). Furthermore, the relation between the distribution Π and $\tilde{\Pi}$ carries on.

Propriété 5.7. *If $\tilde{\Lambda}\otimes\tilde{\Pi}$ satisfies Equation (5.8), we have $\tilde{\Pi}_n(\mathcal{C}) \propto \sum_{\theta\in M(\mathcal{C})}\Pi_n(\theta)$.*

Proof. We have the following equalities.

$$\begin{aligned}\tilde{\Pi}_n(\mathcal{C}) &\propto \left[\int L_n(q)\tilde{\Lambda}_{\mathcal{C}}(dq)\right]\times\tilde{\Pi}(\mathcal{C})\\ &= \left[\int L_n(q)\sum_{\theta\in M(\mathcal{C})} r_{\mathcal{C}}^{\theta}\times\Lambda_{\theta}(dq)\right]\times\left(\sum_{\theta\in M(C)}\Pi(\theta)\right)\\ &= F(\theta),\end{aligned}$$

where

$$\begin{aligned}F(\theta) &= \sum_{\theta\in M(C)}\left[\int L_n(q)\Lambda_{\theta}(dq)\right]\times r_{\mathcal{C}}^{\theta}\Pi(\theta)\\ &= \sum_{\theta\in M(C)}\Pi_n(\theta).\end{aligned}$$

□

We can notice that this result is still valid when we replace the minimal set $M(\mathcal{C})$ by the whole contour in Equation (5.8). Conversely, modelling from the MTC provides a distribution on D. The probability of toxicity of a dose θ is then proportional to the expected probability of all the contour $\mathcal{C}$ for which θ lies in the minimal set $M(\mathcal{C})$. Let $N(\theta)$ be the set of all contour $\mathcal{C}$ such that $\theta\in M(\mathcal{C})$. We set

$$\forall\theta\in D,\ \Pi(\theta)\propto\sum_{\mathcal{C}\in N(\theta)}\tilde{\Pi}(\mathcal{C}) \text{ and } \Lambda_{\theta} = \sum_{\mathcal{C}\in N(\theta)}\tilde{r}_{\mathcal{C}}^{\theta}\times\tilde{\Lambda}_{\mathcal{C}},$$

with $\tilde{r}_{\mathcal{C}}^{\theta} = \tilde{\Pi}(\theta)/\sum_{\mathcal{C}\in N(\theta)}\tilde{\Pi}(\theta)$. This definition of Π on D, coming from a distribution on the contour, leads to a natural strategy selection : maximising Π on the minimal set of the selected contour. When we use such a strategy selection, the methods stemmed from a MTD or MTC modelling are exactly the same. However, every MTC model is not issued of an MTD modelling. The class of MTC methods is all the greater given that a wide variety of strategy selections can be used.

5.5 Experiments

We begin the experiments by studying the poSPM, modelling from an MTD point of view. In practice, this parametrization is relevant when the true MTD is sufficiently far from the other doses. Otherwise the difference of toxicity between the MTD and its nearest doses may be too small to be of interest, leading the MTC to be a more interesting model. Thus, in a first part, we focus on realities where the toxicity of the MTD is clearly distinct from the other doses. In a second part, we are interested in illustrating the more general point of view of the MTC, and the convergence to the minimal set. In both cases, we compare the experimentation and recommendation percentages between different designs.

Table 5.1 – True toxicity probabilities for **Scenario 1**. Maximum tolerated dose in bold.

	Drug A2					
Drug A1	0.20	0.29	0.31	0.43	0.47	0.50
	0.18	0.19	0.29	0.34	0.41	0.48
	0.16	0.20	0.21	0.32	0.36	0.42
	0.10	0.15	0.20	**0.25**	0.30	0.37
	0.03	0.09	0.16	0.19	0.21	0.32
	0.02	0.05	0.10	0.17	0.21	0.30

Table 5.2 – True toxicity probabilities for **Scenario 2**. Maximum tolerated dose in bold.

	Drug A2					
Drug A1	0.37	0.45	0.51	0.54	0.55	0.58
	0.30	0.38	0.41	0.43	0.46	0.47
	0.23	0.25	0.35	0.36	0.40	0.42
	0.19	0.20	0.26	0.33	0.35	0.39
	0.15	0.17	0.21	0.24	0.31	0.33
	0.05	**0.10**	0.16	0.20	0.25	0.28

5.5.1 MTD modelling

In the light of Section 5.3, we know, with Theorem 5.5, that the poSPM is ε-sensitive. We present two cases to exemplify this convergence, and compare it with the behaviour of the poCRM (Wages et al., 2011) and the PIPE method (Mander and Sweeting, 2015). Then, we carry out some further simulations to investigate the speed of convergence of the poSPM. The two studying scenarios are defined as follow.

— **Scenario 1** : $I = 6$, $J = 6$, $\alpha = 0.25$, $n = 40$. The true toxicity probabilities are presented in Table 5.1.

— **Scenario 2** : $I = 6$, $J = 6$, $\alpha = 0.1$, $n = 40$. The true toxicity probabilities are presented in Table 5.2.

For both scenarios, we allow the poSPM to move only on the closest doses of the dose selected at the previous step, and forbid it to move on the diagonals. We choose $\varepsilon = 0$ to get an almost sure convergence, and the a priori is uniform on the $D = 36$ combinations of doses in order to be non informative. The results of 10000 simulated trials with 40 patients are averaged to obtain the percentages of experimentation and recommendation of the doses. We use the same parameters for the PIPE design. For the poCRM algorithm, we choose six orders verifying the partial ordering. Otherwise, the parameters are the same as the other designs. Condensed results can be found in Table 5.3 for the **Scenario 1** and Table 5.4 for the **Scenario 2**, where we provide the percentages of experimentation and recommendation for groups of doses.

For **Scenario 1**, where $\alpha = 0.25$, we see in Table 5.3 that the poSPM gives better experimentation and recommendation percentages for the interval of toxicity containing the threshold α. The percentages of the PIPE design and the poCRM are alike. Furthermore, the poSPM is rather conservative, *i.e.*, it experiments and recommends below the threshold with a higher percentage than above the threshold. We can deduce a well-known behaviour : the poCRM is also a conservative design, approximately as conservative as the poSPM. This consideration is of importance because treatment or recommendation of a dose too toxic can be damaging. For **Scenario 2**, where $\alpha = 0.1$, we notice in Table 5.4 that the conservative behaviour of the poCRM enables it to have better performances

Table 5.3 – Experimentation and recommendation percentages for the poSPM, the PIPE design and the poCRM in **Scenario 1**. Column containing the MTD in bold.

	Experimentation				
	< 0.1	$0.1 \leq . < 0.2$	$0.2 \leq . < 0.3$	$0.3 \leq . < 0.4$	> 0.4
poSPM	12.1	26.6	**32.7**	17.8	10.4
PIPE	7.8	17.7	**29.9**	31.9	12.8
poCRM	11.6	30.3	**28**	20.6	10.2
	Recommendation				
	< 0.1	$0.1 \leq . < 0.2$	$0.2 \leq . < 0.3$	$0.3 \leq . < 0.4$	> 0.4
poSPM	0.4	21.9	**40.7**	27.5	9.6
PIPE	0.2	10.6	**37.7**	41.8	9.6
poCRM	1.2	21.6	**37.5**	30.4	9.2

Table 5.4 – Experimentation and recommendation percentages for the poSPM, the PIPE design and the poCRM in **Scenario 2**. Column containing the MTD in bold

	Experimentation			
	≤ 0.05	$0.05 < . \leq 0.15$	$0.15 < . \leq 0.25$	> 0.25
poSPM	20.6	**25.9**	44.2	9.1
PIPE	5.9	**16.1**	55.1	22.7
poCRM	29.5	**31.3**	32.9	5.9
	Recommendation			
	≤ 0.05	$0.05 < . \leq 0.15$	$0.15 < . \leq 0.25$	> 0.25
poSPM	23.4	**36**	37.9	2.6
PIPE	2.5	**11.9**	61.6	23.6
poCRM	27.2	**40.6**	29	3

for experimentation and recommendation than the poSPM. This could be expected because the targeted threshold α is low. Nevertheless, the flexibility of the poSPM ensures significantly better results than the PIPE design.

We established, in Section 5.3, the almost sure convergence of the poSPM to the MTD. In order to conjecture its speed of convergence, we study the error $|\beta_{\hat{\theta}_n} - \alpha|$. For this, we averaged the errors of 100 simulated trials. For each trial, we calculate the error of the poSPM, with the same parameters used before, for 8 possible values of n : 10, 50, 100, 150, 200, 300, 400 and 500 patients. In Figure 5.3, we present the evolution, with respect to $\log(n)$, of the logarithm of this error in the case of **Scenario 1**. We can see that the rate of convergence of the poSPM seems to be $O\left(n^{-1/4}\right)$. This is not very surprising, because we can expect that such an algorithm converges in $O\left(n^{-\frac{1}{2r}}\right)$, where r is the number of drugs or, in other words, the dimension of the problem. Obviously, this statement is simply a conjecture and may deserve further theoretical investigations.

5.5.2 MTC modelling

We carry on our experiments by studying the poSPM, modelling from an MTC point of view. As specified at the beginning of this section, we are now interested in cases where several doses may have a toxicity very close from the targeted probability of toxicity α. We recall, from Theorem 5.6 in Section 5.4, that the poSPM converges almost surely to the minimal set. To illustrate this convergence and compare the behaviour of the poSPM

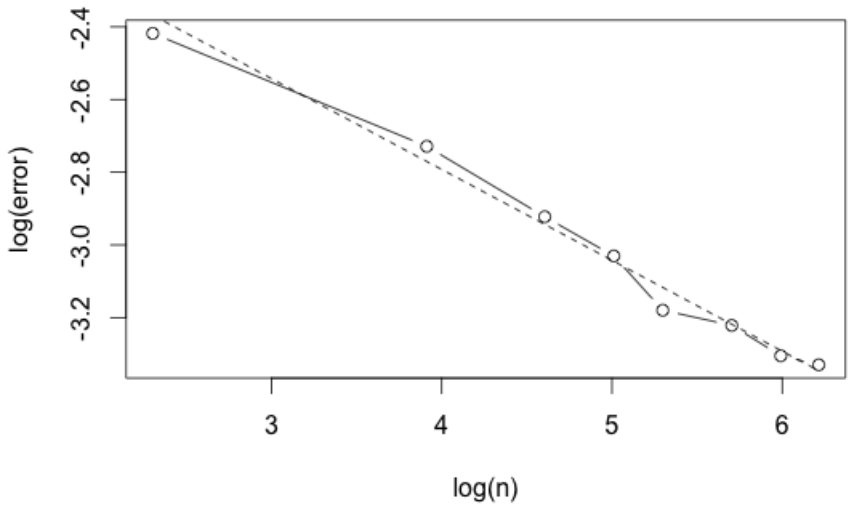


Figure 5.3 – Plots of the logarithm of the error of the poSPM with respect to $\log(n)$ (solid line) and of a straight line with slope $-1/4$ (dashed line) in case of **Scenario 1**.

Table 5.5 – True toxicity probabilities for **Scenario 3** with MTC. Minimal set in bold.

	Drug A2					
	0.27	0.29	0.31	0.43	0.47	0.50
	0.14	0.19	**0.21**	0.39	0.41	0.48
Drug	0.13	0.15	0.18	0.32	0.40	0.42
A1	0.12	0.13	0.14	0.30	0.34	0.37
	0.08	0.10	0.11	**0.28**	0.30	0.32
	0.02	0.05	0.10	0.13	0.17	**0.22**

with the PIPE one, which is a design taking into account a MTC parametrization, and the poCRM, which is not specifically designed for the MTC, we study again two scenarios defined as follow.

— **Scenario 3** : $I = 6$, $J = 6$, $\alpha = 0.25$, $n = 40$. The true toxicity probabilities are presented in Table 5.5.

— **Scenario 4** : $I = 6$, $J = 6$, $\alpha = 0.1$, $n = 40$. The true toxicity probabilities are presented in Table 5.6.

We compare, for each design and each scenario, the experimentation and recommendation percentages for doses in the minimal set M_D, outside the minimal set and with toxicity below α, and outside the minimal set and with toxicity above α. We use for each design the parameters detailed in the previous subsection, and again we average the reuslts of 10000 trials of 40 patients. A summary of these results can be found in Table 5.7 for **Scenario 3** and Table 5.8 for **Scenario 4**.

For **Scenario 3**, we can note that the PIPE design has better performances for both experimentation and recommendation than the poSPM and the poCRM. The better results of the PIPE design and the poSPM against the poCRM are not surprising because these first two methods are designed for the MTC and the minimal set M_D. Besides, we notice

Table 5.6 – True toxicity probabilities for **Scenario 4** with MTC. Minimal set in bold.

	Drug A2					
	0.28	0.31	0.34	0.36	0.42	0.47
	0.27	0.29	0.30	0.32	0.33	0.41
Drug	0.15	**0.22**	**0.29**	0.31	0.32	0.37
A1	0.11	0.12	0.17	**0.21**	0.28	0.33
	0.08	0.09	0.15	0.20	**0.27**	0.30
	0.03	0.08	0.11	0.18	**0.23**	**0.28**

Table 5.7 – Experimentation and recommendation percentages for the poSPM, the PIPE design and the poCRM in **Scenario 3**. Column of the minimal set in bold

	Experimentation		
	$\notin M_D$ and $< \alpha$	M_D	$\notin M_D$ and $> \alpha$
poSPM	48.5	**14.3**	37.2
PIPE	31.5	**18.5**	50
poCRM	48.3	**10.9**	40.8
	Recommendation		
	$\notin M_D$ and $< \alpha$	M_D	$\notin M_D$ and $> \alpha$
poSPM	28.2	**23.1**	48.7
PIPE	16.5	**26.7**	56.8
poCRM	34.4	**15.8**	49.8

again the conservative behaviour of the poSPM and the poCRM. In both **Scenario 3-4**, Table 5.7 and Table 5.8 show high percentages of experimentation and recommendation above the threshold. We can recall that the simulated trials contain only 40 patients each, and so we are far from an asymptotic behaviour. **Scenario 4** is a little more complex than the previous ones. Indeed, some doses out of the minimal set have toxicities comparable to the ones in it. For example, the doses $(3, 5)$ and $(6, 1)$ are outside the minimal set and have a toxicity probability of 0.28, whereas the dose $(4, 3)$ is in the minimal set and has a toxicity probability of 0.29. This explains the particularly high percentage of recommendation shown in the last column of Table 5.8 for the PIPE design. In this scenario, the poSPM experiments and recommends with a higher percentage in the minimal set than the two other designs. Even if the PIPE design has approximately the same results for the minimal set, the poSPM should be preferred because of its conservative behaviour.

Table 5.8 – Experimentation and recommendation percentages for the poSPM, the PIPE design and the poCRM in **Scenario 4**. Column of the minimal set in bold

	Experimentation		
	$\notin M_D$ and $< \alpha$	M_D	$\notin M_D$ and $> \alpha$
poSPM	42.6	**25.8**	31.6
PIPE	24.1	**22.5**	53.4
poCRM	42.7	**20.8**	36.5
	Recommendation		
	$\notin M_D$ and $< \alpha$	M_D	$\notin M_D$ and $> \alpha$
poSPM	23.4	**32**	44.6
PIPE	8.4	**25.1**	66.5
poCRM	27.2	**27.7**	45.1

Conclusion et perspectives

Les travaux de cette thèse portent sur l'étude de l'inférence sous certains modèles statistiques mal spécifiés. En pratique, nous ne connaissons jamais, ou presque, le modèle sous-jacent aux données qui nous intéressent. C'est pourquoi l'étude de modèles mal spécifiés est indispensable. Pour illustrer l'application des différents résultats de cette thèse, nous étudions l'influence des facteurs pronostiques sur le cancer du sein, grâce à des données fournies par l'Institut Curie.

Dans le Chapitre 2, nous avons étudié ce que pouvait nous apporter la connaissance de la survie marginale de T sur l'estimation de la fonction de régression sur les covariables β_0. Nous introduisons un estimateur la prenant en compte. Nous constatons que ce surplus d'informations n'améliore pas l'efficacité de cette estimation dans le cas d'un modèle à hasards proportionnels. Par contre, sous un modèle à hasards non proportionnels, cet estimateur converge en probabilité vers une quantité que l'on peut interpréter comme un effet moyen des covariables sur le temps de décès. Sous certaines hypothèses, Xu and O'Quigley (2000) obtiennent un résultat similaire en utilisant l'estimateur de Kaplan-Meier de la survie. Une conséquence intéressante est que les résultats d'estimation de Xu and O'Quigley (2000) s'étendent aux cas où la survie marginale peut être modélisée de façon paramétrique. Notons qu'il est beaucoup moins fort de supposer un modèle paramétrique pour l'ensemble des lois contionnelles aux covariables. En effet, même si le modèle de la survie marginale est mal spécifié, on obtient toujours un estimateur consistant de la fonction de régression β_0, sous un modèle à hasards proportionnels de Cox. Nous n'avons pas étudié les cas de modèles mal spécifiés pour la survie marginale en conjonction avec des modèles à hasards non proportionnels. Ceci pourrait être une piste intéressante de recherche.

Dans le Chapitre 3, nous nous intéressons à une extension du modèle de Cox qui est le cas où la fonction β_0 est constante par morceaux. Nous poussons l'analyse de Anderson and Senthilselvan (1982) un peu plus loin et proposons une méthode d'inférence sur le changepoint d'un modèle à un unique changepoint. Nous obtenons une région de confiance à partir des travaux de Davies (1977). Nous proposons une méthode d'estimation pour un modèle plus général avec K changepoints, où K est fixé à l'avance. Celle-ci est basée sur le processus du score standardisé (Chauvel and O'quigley, 2014). Les simulations montrent de bonnes performances de cette dernière méthode d'estimation, même en augmentant la censure et le nombre de changepoints. Deux pistes de recherche apparaissent naturellement. Tout d'abord, il serait intéressant d'effectuer de l'inférence à partir du processus du score standardisé. On pourrait par exemple développer un test d'hypothèse nulle "le modèle comporte K changepoints" et d'alternative "le modèle comporte $K+1$ changepoints". La seconde piste de recherche est naturellement d'étudier la question de l'estimation des changepoints d'un modèle à K changepoints, mais où K n'est pas connu à l'avance, ce qui est une situation bien plus réaliste. Une idée serait peut-être d'introduire une pénalisation du type "nombre de changepoints".

Dans le Chapitre 4, nous étudions l'influence du taux de sous-échantillonnage et de la profondeur des arbres sur la performance des forêts de Breiman. Pour cela nous nous basons sur son influence sur la performance des forêts médianes, une version des forêts de Breiman simplifiée mais proche. Nous présentons une majoration de la vitesse de convergence des forêts médianes et montrons que la performance des forêts médianes sous-échantillonnées complètement développées et celle des forêts médianes semi-développées sans sous-échantillonnage sont similaires, pour un bon calibrage des paramètres d'intérêt (respectivement taille du sous-échantillonnage et profondeur des arbres). Les simulations montrent des résultats similaires : les forêts de Breiman peuvent être surpassées par des forêts de Breiman sous-échantillonnées ou semi-développées en calibrant correctement les paramètres. On peut noter que calibrer la profondeur des arbres peut être effectué presque sans aucun coût computationnel supplémentaire pendant que l'algorithme des forêts de Breiman tourne (grâce à la nature récursive des forêts). Cependant, si l'objectif est d'avoir une procédure plus rapide, les forêts de Breiman sous-échantillonnées sont plus intéressantes que les forêts semi-développées. Par ailleurs, notre analyse montre que le bootstrap des données n'a pas d'intérêt particulier comparé au sous-échantillonnage : dans nos simulations, le bootstrap donne des résultats comparables ou moins bons qu'un sous-échantillonnage bien calibré. Une piste de recherche serait d'étendre aux forêts aléatoires de survie les résultats théoriques obtenus sur les forêts médianes. On pourrait ainsi étudier l'effet du sous-échantillonnage sur les forêts de survie.

Dans le Chapitre 5, nous introduisons une nouvelle méthode de recherche de doses, la poSPM, qui trouve son application dans les essais cliniques de phase I comportant plusieurs agents cytotoxiques. Cette méthode repose sur la méthode semi-paramétrique introduite par Clertant (2015) qui s'applique, de son côté, à des essais cliniques comportant un unique agent cytotoxique. Nous montrons que la poSPM vérifie des propriétés de cohérence et de convergence telles que la ε-sensibilité et l'ε-équilibre. Des conséquences de ces propriétés sont la consistance et la cohérence de la méthode PIPE en tant que cas particulier de la poSPM. Les simulations présentées pour les deux paramétrisations, selon la MTD ou le MTC, montrent de bonnes performances de la poSPM par rapport à la méthode PIPE et à la poCRM, en termes d'expérimentation et de recommandation, même avec les paramètres par défaut de l'algorithme. Dans le cas de l'ordre partiel, il semble raisonnable, comme le soulignent Mander and Sweeting (2015) d'utiliser la paramétrisation par MTC. Ce point de vue permet l'utilisation de nombreuses stratégies de sélection et donne ainsi aux cliniciens une flexibilité plus importante. La poSPM est aussi rapide, facile à interpréter et son code est disponible en `R` sur demande. Nous avons aussi conjecturé la vitesse de convergence de la poSPM : $O\left(n^{-1/4}\right)$, et peut-être $O\left(n^{-\frac{1}{2r}}\right)$, où r est le nombre de médicaments que comporte le traitement à tester. Il serait intéressant d'approfondir les recherches théoriques à ce sujet, et en particulier la dépendance par rapport au nombre de médicaments composant le traitement. En effet, la théorie derrière la poSPM peut être généraliser à des dimensions plus grandes grâce à la théorie du Bayésien hiérarchique. Un autre point de recherche intéressant est l'adaptation de la poSPM dans le cas de causes de toxicité connues. Par exemple, si le traitement 1 occasionne des réactions cutanées et le traitement 2 des migraines, on peut déterminer quel traitement cause une toxicité. On pourrait utiliser cette information pour l'estimation, par la poSPM, de la toxicité du couple formé par le traitement 1 et le traitement 2. On pourrait par exemple utiliser des tableaux de contingence 2×2, comme suggéré dans Yin and Yuan (2009a).

Annexe A

Preuves du Chapitre 4

A.1 Un lemme préliminaire

Lemme A.1. *Pour tout ℓ et k, on a*

$$\mathbb{E}\Big[V_l(\mathbf{X},\Theta)^2\Big] \leq C\Big(1-\frac{3}{4d}\Big)^k,$$

avec $C = \exp(12/(4d-3))$, où $V_l(\mathbf{X},\Theta)$ est la longueur du ℓ-ème côté de la cellule contenant $\mathbf{X}$.

Preuve du Lemme A.1. Fixons $\mathbf{X} \in [0,1]^d$ et notons $n_0, n_1, \ldots, n_k$ le nombre de points dans les cellules successives contenant $\mathbf{X}$. Par exemple, n_0 est le nombre de points dans la racine de l'arbre, c'est-à-dire $n_0 = a_n$. Comme $V_\ell(\mathbf{X},\Theta)$ est la longueur du ℓ-ème côté de la cellule contenant $\mathbf{X}$, on a

$$V_\ell(\mathbf{X},\Theta) \stackrel{\mathcal{D}}{=} \prod_{j=1}^{k} \big[B(n_j+1, n_{j-1}-n_j)\big]^{\delta_{\ell,j}(\mathbf{X},\Theta)},$$

où $B(\alpha,\beta)$ est la loi Beta de paramètres α et β, et $\delta_{\ell,j}(\mathbf{X},\Theta)$ vaut 1 si la j-ème coupure de la cellule contenant $\mathbf{X}$ est effectuée sur la ℓ-ème dimension, et 0 sinon. D'où,

$$\begin{aligned}
\mathbb{E}\big[V_\ell(\mathbf{X},\Theta)^2\big] &= \prod_{j=1}^{k} \mathbb{E}\Big[\big[B(n_j+1, n_{j-1}-n_j)\big]^{2\delta_{\ell,j}(\mathbf{X},\Theta)}\Big] \\
&= \prod_{j=1}^{k} \mathbb{E}\Big[\mathbb{E}\Big[\big[B(n_j+1, n_{j-1}-n_j)\big]^{2\delta_{\ell,j}(\mathbf{X},\Theta)} \big| \delta_{\ell,j}(\mathbf{X},\Theta)\Big]\Big] \\
&= \prod_{j=1}^{k} \mathbb{E}\Big[\mathbb{1}_{\delta_{\ell,j}(\mathbf{X},\Theta)=0} + \mathbb{E}\big[B(n_j+1, n_{j-1}-n_j)\big]^2 \mathbb{1}_{\delta_{\ell,j}(\mathbf{X},\Theta)=1}\Big] \\
&= \prod_{j=1}^{k} \Big(\frac{d-1}{d} + \frac{1}{d}\mathbb{E}\big[B(n_j+1, n_{j-1}-n_j)\big]^2\Big) \\
&= \prod_{j=1}^{k} \Big(\frac{d-1}{d} + \frac{1}{d}\frac{(n_j+1)(n_j+2)}{(n_{j-1}+1)(n_{j-1}+2)}\Big)
\end{aligned}$$

$$\leq \prod_{j=1}^{k} \left(\frac{d-1}{d} + \frac{1}{4d} \frac{(n_{j-1}+2)(n_{j-1}+4)}{(n_{j-1}+1)(n_{j-1}+2)} \right)$$
$$\leq \prod_{j=1}^{k} \left(1 - \frac{1}{d} + \frac{1}{4d} \frac{n_{j-1}+4}{n_{j-1}+1} \right). \tag{A.1}$$

On a de plus les inégalités suivantes.

$$\frac{n_{j-1}+4}{n_{j-1}+1} \leq \frac{a_n + 2^{j+1}}{a_n + 2^{j-1}} = \frac{a_n + 2^{j+1}}{a_n(1 + \frac{2^{j-1}}{a_n})}$$
$$\leq \frac{a_n + 2^{j+1}}{a_n} \left(1 + \frac{2^{j-1}}{a_n} \frac{1}{1 - \frac{2^{j-1}}{a_n}} \right)$$
$$\leq \left(1 + \frac{2^{j+1}}{a_n} \right)^2,$$

puisque

$$\frac{2^{j-1}}{a_n} \leq \frac{2^{k-1}}{a_n} \leq \frac{1}{2}.$$

Revenons à l'inéquation (A.1), on trouve

$$\mathbb{E}\Big[V_l(\mathbf{X}, \Theta)^2 \Big] \leq \prod_{j=1}^{k} \left[1 - \frac{1}{d} + \frac{1}{4d} \left(1 + \frac{2^{j+1}}{a_n} \right)^2 \right]$$
$$\leq \prod_{j=1}^{k} \left[1 - \frac{1}{d} + \frac{1}{4d} \left(1 + 2\frac{2^{j+1}}{a_n} + \frac{2^{2j+2}}{a_n^2} \right) \right]$$
$$\leq \prod_{j=1}^{k} \left[1 - \frac{1}{d} + \frac{1}{4d} \left(1 + 3\frac{2^{j+1}}{a_n} \right) \right]$$
$$\leq \prod_{j=1}^{k} \left[1 - \frac{3}{4d} + \frac{3}{d} \frac{2^{j-1}}{a_n} \right]$$
$$\leq \prod_{j=1}^{k} \left[1 - \frac{3}{4d} + \frac{3}{d} \frac{2^k}{a_n} 2^{j-k} \right]$$
$$\leq \prod_{j=0}^{k-1} \left[1 - \frac{3}{4d} + \frac{3}{d} \frac{2^k}{a_n} 2^{-j} \right]$$
$$\leq \prod_{j=0}^{k-1} \left[1 - \frac{3}{4d} + \frac{3}{d} 2^{-j} \right].$$

On peut par ailleurs remarquer que

$$\ln \left(\prod_{j=0}^{k-1} \left[1 - \frac{3}{4d} + \frac{3}{d} 2^{-j} \right] \right) = k \ln \left(1 - \frac{3}{4d} \right) + \sum_{k=0}^{j-1} \ln \left(1 + 6 \frac{2^{-j}}{4d-3} \right)$$
$$\leq k \ln \left(1 - \frac{3}{4d} \right) + \frac{12}{4d-3}.$$

Ceci implique que

$$\mathbb{E}\Big[V_l(\mathbf{X},\Theta)^2\Big] \leq C\Big(1-\frac{3}{4d}\Big)^k,$$

avec $C = \exp(12/(4d-3))$. □

A.2 Preuve du Théorème 4.1

Rappelons que l'estimateur m_n d'une forêt aléatoire peut s'écrire comme un estimateur moyen local

$$m_n(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) Y_i,$$

où l'erreur $\mathbb{L}^2$ de l'estimateur de la forêt aléatoire est de la forme

$$\begin{aligned}\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 &\leq 2\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X})(Y_i - m(\mathbf{X}_i))\Big]^2 \\ &\quad + 2\mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X})(m(\mathbf{X}_i) - m(\mathbf{X})\Big]^2 \\ &= 2I_n + 2J_n.\end{aligned}$$

On étudie ensuite séparément les termes I_n et J_n.

Erreur d'approximation. Commençons par l'étude de J_n. En utilisant l'inégalité de Cauchy-Schwarz, on obtient

$$\begin{aligned}J_n &\leq \mathbb{E}\Big[\sum_{i=1}^{n} \sqrt{W_{ni}(\mathbf{X})}\sqrt{W_{ni}(\mathbf{X})}|m(\mathbf{X}_i) - m(\mathbf{X})|\Big]^2 \\ &\leq \mathbb{E}\Big[\sum_{i=1}^{n} W_{ni}(\mathbf{X})(m(\mathbf{X}_i) - m(\mathbf{X}))^2\Big] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_i \overset{\Theta}{\leftrightarrow} \mathbf{X}}}{N_n(\mathbf{X},\Theta)} \sup_{\mathbf{x},\mathbf{z},|\mathbf{x}-\mathbf{z}|\leq \mathrm{diam}(A_n(\mathbf{X}))} |m(\mathbf{x}) - m(\mathbf{z})|^2\right] \\ &\leq L^2\mathbb{E}\left[\frac{1}{N_n(\mathbf{X},\Theta)}\sum_{i=1}^{n} \mathbb{1}_{\mathbf{X}_i \overset{\Theta}{\leftrightarrow} \mathbf{X}} \left(\mathrm{diam}(A_n(\mathbf{X}))\right)^2\right] \\ &\leq L^2\mathbb{E}\Big[\left(\mathrm{diam}(A_n(\mathbf{X}))\right)^2\Big].\end{aligned}$$

On rappelle que $V_\ell(\mathbf{X},\Theta)$ est la longueur du ℓ-ème côté de la cellule contenant $\mathbf{X}$. Alors,

$$\begin{aligned}J_n &\leq L^2\mathbb{E}\Big[\max_{1\leq l\leq d} V_l(\mathbf{X},\Theta)^2\Big] \\ &\leq dL^2\mathbb{E}\Big[V_l(\mathbf{X},\Theta)^2\Big],\end{aligned}$$

puisque la fonction de régression m est L-Lipschitz. D'après le Lemme A.1, on obtient

$$\mathbb{E}\Big[V_l(\mathbf{X},\Theta)^2\Big] \leq C\Big(1-\frac{3}{4d}\Big)^k,$$

avec $C = \exp(12/(4d-3))$.

Donc, pour tout k suffisamment grand,

$$J_n \leq dL^2 C\left(1 - \frac{3}{4d}\right)^k.$$

Erreur d'estimation Intéressons nous ensuite à I_n. On a

$$\begin{aligned}
I_n &= \mathbb{E}\Big[\sum_{i=1}^n W_{ni}(\mathbf{X})(Y_i - m(\mathbf{X}_i))\Big]^2 \\
&= \sum_{i=1}^n \sum_{i=1}^n \mathbb{E}\Big[W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})(Y_i - m(\mathbf{X}_i))(Y_j - m(\mathbf{X}_j))\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(\mathbf{X})(Y_i - m(\mathbf{X}_i))^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(\mathbf{X})\sigma_i^2\Big] \\
&\leq \sigma^2 \mathbb{E}\Big[\sum_{i=1}^n W_{ni}^2(\mathbf{X})\Big] \\
&\leq \sigma^2 \mathbb{E}\Big[\max_{1\leq i\leq n} W_{ni}(\mathbf{X})\Big].
\end{aligned}$$

Or,

$$\begin{aligned}
\mathbb{E}\Big[\max_{1\leq i\leq n} W_{ni}(\mathbf{X})\Big] &= \mathbb{E}\Big[\max_{1\leq i\leq n} \mathbb{E}_\Theta\Big[\frac{\mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}}{N_n(\mathbf{X},\Theta)}\Big]\Big] \\
&\leq \frac{1}{\frac{a_n}{2^k} - 2}\mathbb{E}\Big[\max_{1\leq i\leq n} \mathbb{P}_\Theta\Big[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\Big]\Big].
\end{aligned}$$

Notons de plus que, dans l'étape de sous-échantillonnage, il y a exactement $\binom{a_n-1}{n-1}$ possibilités de choisir une observation donnée $\mathbf{X}_i$. Comme $\mathbf{x}$ et $\mathbf{X}_i$ appartiennent à la même cellule seulement si $\mathbf{X}_i$ est sélectionnée dans l'étape de sous-échantillonnage, on peut voir que

$$\mathbb{P}_\Theta\Big[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\Big] \leq \frac{\binom{a_n-1}{n-1}}{\binom{a_n}{n}} = \frac{a_n}{n}.$$

Donc,

$$I_n \leq \sigma^2 \frac{1}{\frac{a_n}{2^k} - 2}\frac{a_n}{n} \leq \sigma^2 \frac{2^k}{a_n - 2^{k+1}}\frac{a_n}{n} \leq 2\sigma^2 \frac{2^k}{n},$$

puisque $a_n/2^k \geq 4$. Finalement, on obtient

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq I_n + J_n \leq 2\sigma^2 \frac{2^k}{n} + dL^2 C\left(1 - \frac{3}{4d}\right)^k.$$

A.3 Preuve du Corollaires 4.2

On note $C_1 = \frac{2\sigma^2}{n}$, $C_2 = dL^2C$ et $\beta = \left(1 - \frac{3}{4d}\right)$. Alors,

$$\mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 \leq C_1 2^k + C_2 \beta^k.$$

Soit $f : x \mapsto C_1 e^{x \ln 2} + C_2 e^{x \ln(\beta)}$. Alors,

$$\begin{aligned} f'(x) &= C_1 \ln 2 e^{x \ln 2} + C_2 \ln(\beta) e^{x \ln(\beta)} \\ &= C_1 \ln 2 e^{x \ln 2} \left(1 + \frac{C_2 \ln(\beta)}{C_1 \ln 2} e^{x(\ln(\beta) - \ln 2)}\right) \\ &= C_1 \ln 2 e^{x \ln 2} \left(1 + \frac{C_2 \ln(\beta)}{C_1 \ln 2} e^{x(\ln(\beta) - \ln 2)}\right). \end{aligned}$$

Comme $\beta \leq 1$, $f'(x) \leq 0$ pour tout $x \leq x^\star$ et $f'(x) \geq 0$ pour tout $x \geq x^\star$, où $x^\star$ satisfait

$$\begin{aligned} & f'(x^\star) = 0 \\ \Leftrightarrow & 1 + \frac{C_2 \ln(\beta)}{C_1 \ln 2} e^{x^\star(\ln(\beta) - \ln 2)} = 0 \\ \Leftrightarrow & e^{x^\star(\ln(\beta) - \ln 2)} = -\frac{C_1 \ln 2}{C_2 \ln(\beta)} \\ \Leftrightarrow & x^\star = \frac{1}{\ln(\beta) - \ln 2} \ln\left(-\frac{C_1 \ln 2}{C_2 \ln(\beta)}\right) \\ \Leftrightarrow & x^\star = \frac{1}{\ln 2 - \ln(\beta)} \ln\left(-\frac{C_2 \ln(\beta)}{C_1 \ln 2}\right) \\ \Leftrightarrow & x^\star = \frac{1}{\ln 2 - \ln(\beta)} \left[\ln\left(\frac{1}{C_1}\right) + \ln\left(-\frac{C_2 \ln(\beta)}{\ln 2}\right)\right] \\ \Leftrightarrow & x^\star = \frac{1}{\ln 2 - \ln\left(1 - \frac{3}{4d}\right)} \left[\ln\left(n\right) + \ln\left(-\frac{dL^2C \ln\left(1 - \frac{3}{4d}\right)}{2\sigma^2 \ln 2}\right)\right] \\ \Leftrightarrow & x^\star = \frac{1}{\ln 2 - \ln \beta} \Big[\ln(n) + C_3\Big], \end{aligned}$$

avec $C_3 = \ln\left(-\frac{dL^2C \ln\left(1 - \frac{3}{4d}\right)}{2\sigma^2 \ln 2}\right)$.

Finalement,

$$\begin{aligned} \mathbb{E}\big[m_n(\mathbf{X}) - m(\mathbf{X})\big]^2 &\leq C_1 \exp(k \ln 2) + C_2 \exp(k \ln \beta) \\ &\leq C_1 \exp\left(\frac{1}{\ln 2 - \ln \beta} \Big[\ln(n) + C_3\Big] \ln 2\right) \\ &\quad + C_2 \exp\left(\frac{1}{\ln 2 - \ln \beta} \Big[\ln(n) + C_3\Big] \ln \beta\right) \\ &\leq C_1 \exp\left(\frac{C_3 \ln 2}{\ln 2 - \ln \beta}\right) \exp\left(\frac{\ln 2}{\ln 2 - \ln \beta} \ln(n)\right) \\ &\quad + C_2 \exp\left(\frac{C_3 \ln \beta}{\ln 2 - \ln \beta}\right) \exp\left(\frac{\ln \beta}{\ln 2 - \ln \beta} \ln(n)\right) \end{aligned}$$

$$\begin{aligned}
&\leq C_5 n^{\frac{\ln 2}{\ln 2 - \ln \beta} - 1} + C_6 n^{\frac{\ln \beta}{\ln 2 - \ln \beta}} \\
&\leq \left(C_5 + C_6\right) n^{\frac{\ln\left(1 - \frac{3}{4d}\right)}{\ln 2 - \ln\left(1 - \frac{3}{4d}\right)}},
\end{aligned}$$

où $C_5 = 2\sigma^2 \exp\left(\frac{C_3 \ln 2}{\ln 2 - \ln \beta}\right)$ et $C_6 = C_2 \exp\left(\frac{C_3 \ln \beta}{\ln 2 - \ln \beta}\right)$.

A.4 Preuve du Corollaires 4.3

On détermine le taux d'échantillonnage a_n optimal avec le calcul suivant.

$$\begin{aligned}
k &= \log_2(a_n) - 2 \\
\Leftrightarrow a_n &= 2^{k+2} \\
\Leftrightarrow a_n &= 2^{\frac{1}{\ln 2 - \ln \beta}\left[\ln(n) + C_3\right] + 2} \\
\Leftrightarrow a_n &= 4.2^{\frac{C_3}{\ln 2 - \ln \beta}}.2^{\frac{\ln(n)}{\ln 2 - \ln \beta}} \\
\Leftrightarrow a_n &= 4.2^{\frac{C_3}{\ln 2 - \ln \beta}}.2^{\log_2(n)\frac{\ln 2}{\ln 2 - \ln \beta}} \\
\Leftrightarrow a_n &= 4.2^{\frac{C_3}{\ln 2 - \ln \beta}}.n^{\frac{\ln 2}{\ln 2 - \ln \beta}}.
\end{aligned}$$

Annexe B

Preuves du Chapitre 5

B.1 A general bayesian property

The couple $(\Omega, \mathcal{A})$ denotes an abstract space endowed with its σ-field. We denote by I a finite set and by $(X_k)_{k\in\mathbb{N}}$ a sequence of independent random variables taking their values in I. Let F be the set of functions from I to the segment $[0,1]$. For any element $q \in F$, q_i denotes its value at $i \in I$. Let $S = \{q \in F : \sum_{i\in I} q_i = 1\}$ be the probability space on which we want to work. We say that a random variable X follows the distribution q if $\mathbb{P}\{X = i\} = q_i$, for all $i \in I$. Let Λ_1 and Λ_2 be two probabilities on the Borel σ-field $\mathcal{B}$ of S. Let S_1 and S_2 be the topological supports of Λ_1 and Λ_2 respectively. We would like to know the asymptotic behaviour of the ratio of the expected likelihood under Λ_1 on the one under Λ_2. We define the operator r as follows

$$r(\Lambda_1, \Lambda_2, n) = \frac{\displaystyle\int \prod_{k=1}^{n} q_{X_k}\, \Lambda_1(dq)}{\displaystyle\int \prod_{k=1}^{n} q_{X_k}\, \Lambda_2(dq)} = \frac{\displaystyle\int \prod_{i\in I} q_i^{n_i}\, \Lambda_1(dq)}{\displaystyle\int \prod_{i\in I} q_i^{n_i}\, \Lambda_2(dq)}\,,$$

where $n_i = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=i\}}$, for $i \in I$. We assume that, under the true probability β, the random variables X_k, $k \in \mathbb{N}$ are identically distributed. The convergence of $r(\Lambda_1, \Lambda_2, n)$ depends mainly on the localization of β compared to the supports S_1 and S_2. To deal with this problem, we make use of the usual concept of entropy. The entropy of q relative to p is $H(q|p) = -\sum_{i\in I} p_i \log q_i$, with the conventions $\log 0 = -\infty$ and $0 \times -\infty = 0$. We suppose that β is closer to S_1 than S_2 in terms of entropy.

Hypothèses B.1.1. *Let V be a subspace of S_2 satisfying $\Lambda_2(V) > 0$. There exists $\delta > 0$ such that*

$$\inf_{q\in S_1} H(q|\beta) - \sup_{q\in V} H(q|\beta) > 4\delta.$$

This leads to a simple characterisation of the behavior of $r(\Lambda_1, \Lambda_2, n)$.

Proposition B.1. *Under Assumption B.1.1, we have $r(\Lambda_1, \Lambda_2, n) \underset{n\to\infty}{\longrightarrow} 0$.*

Proof. Let the empirical probabilities of toxicity be $\hat{\beta}_n = (\hat{\beta}_{n,i})_{i\in I}$ where $\hat{\beta}_{n,i} = n_i/n$. By discarding a $\mathbb{P}_\beta$-null set, the law of large number implies that $\lim_{n\to+\infty} \hat{\beta}_{n,i}(\omega) = \beta_i$, for all $i \in I$ and for all $\omega \in \Omega$. The main fact to establish is

$$\inf_{q\in S_1} H(q|\hat{\beta}_n) - \sup_{q\in V} H(q|\hat{\beta}_n) > \delta\,, \tag{B.1}$$

for n large enough. Let $I_+ = \{i \in I : \beta_i > 0\}$, and $S_+ = \{(q_i)_{i\in I_+} : \sum_{i\in I_+} q_i \geq 1\}$. S_+ is the projection of S on the vector space indexed by I_+. In the same way, we defined $S_{1,+}$ and V_+ such that

$$S_{1,+} = \{(q'_i)_{i\in I_+} : \exists q \in S_1,\, \forall i \in I_+,\, q'_i = q_i\},$$

and

$$V_+ = \{(q'_i)_{i\in I_+} : \exists q \in V,\ \forall i \in I_+,\ q'_i = q_i\}.$$

We begin to prove the following inequality for n large enough

$$\left|\sup_{q\in V} H(q|\beta) - \sup_{q\in V} H(q|\hat{\beta}_n)\right| < \delta. \tag{B.2}$$

Assumption B.1.1 implies that $\sup_{q\in V} H(q|\beta) < \infty$. Hence there exists $\kappa > 0$ such that $V_+ \subset T = [\kappa, 1]^{\#I_+}$, where $\overline{V_+}$ denotes the closure of V_+. Let $H_{I_+}(q|p) = -\sum_{i\in I_+} p_i \log q_i$. For all $p \in S_+$, the function $H_{I_+}(.|p)$ defined on T is K_p-Lipschitz continuous for the constant $K_p = (\sup_{i\in I_+} p_i)/\log(\kappa)$. As the sequence $\hat{\beta}_n$ converges to β, there exists a Lipschitz constant K_0 valid for the whole sequence $(H_{I_+}(.|\hat{\beta}_n))_{n\in\mathbb{N}}$ and its limit $H_{I_+}(.|\beta)$ on $\overline{V_+}$. As $\overline{V_+}$ is a compact set, for all $\varepsilon > 0$, there exists a finite family $(p_j)_{1\leq j\leq J}$ satisfying the following property : $\forall q \in V_+$, $\exists j \in \{1, \ldots, J\}$ such that $d(q, p_j) < \varepsilon$. For all $q \in S_+$, we can find $k \in \{1, \ldots, J\}$, such that

$$\begin{aligned}\left|H_{I_+}(q|\beta) - H_{I_+}(q|\hat{\beta}_n)\right| \leq{}& \left|H_{I_+}(q|\beta) - H_{I_+}(p_j|\beta)\right| + \left|H_{I_+}(p_j|\beta) - H_{I_+}(p_j|\hat{\beta}_n)\right| \\ &+ \left|H_{I_+}(p_j|\hat{\beta}_n) - H_{I_+}(q|\hat{\beta}_n)\right|,\end{aligned}$$

where the second term is bounded by $(2K+1)\varepsilon$, for n large enough. Thus, $(H(.|\hat{\beta}_n))_{n\in\mathbb{N}}$ converges uniformly to $H(.|\beta)$ and, for n large enough, we obtain the inequality (B.2). We will continue by establishing the inequality

$$\left|\inf_{q\in S_1} H(q|\beta) - \inf_{q\in S_1} H(q|\hat{\beta}_n)\right| < \delta\,, \tag{B.3}$$

for n large enough and when $(\inf_{q\in S_1} H(q|\hat{\beta}_n))_{n\in\mathbb{N}}$ is bounded. Otherwise the inequality (B.1) is simply obtained. As $(\hat{\beta}_n)_{n\in\mathbb{N}}$ converges to β, there exists κ' such that $S_{1,+} \subset [\kappa', 1]^{\#I_+}$. Thus, we prove (B.3) using the same argument as in Equation (B.2). Assumption B.1.1 combined with Equation (B.2) and Equation (B.3) leads to inequality (B.1). From Equation (B.1) follows

$$\int_V \prod_{i\in I} q_i^{n_i}\, \Lambda_2(dq)/\Lambda_2(V) > \int_V \prod_{i\in I} q_i^{n_i}\, \Lambda_1(dq) \times \exp(n\delta)\,,$$

and

$$r(\Lambda_1, \Lambda_2, n) \leq \frac{\displaystyle\int \prod_{i\in I} q_i^{n_i}\, \Lambda_1(dq)}{\displaystyle\int_V \prod_{i\in I} q_i^{n_i}\, \Lambda_2(dq)} < \exp(-n\delta)\Lambda_2(V)\,,$$

which complete the proof. □

B.2 Proof of asymptotical results for po-SPM

B.2.1 A lemma

Lemme B.2. *Let r be a dose in $\mathcal{E}(I,\beta)\cap\tilde{D}$. Then, there exists no dose d ordered with r in $\mathcal{E}(I,\beta)\cap\tilde{D}$.*

Proof. The proof is based on reductio ad absurdum. Let $r\in\mathcal{E}(I,\beta)\cap\tilde{D}$. Assume that there exists $d\in\mathcal{E}(I,\beta)\cap\tilde{D}$ ordered with r. As $d\in\tilde{D}$, the dose sequence $(x_n)_{n\in\mathbb{N}}$ has the following form

$$(x_n)=(\cdots,\underbrace{d,d,\cdots,d}_{s_1},x_{j_0},\cdots,x_{j_1},\underbrace{d,\cdots,d}_{s_2},x_{j_2},\cdots,x_{j_3},\underbrace{d,\cdots,d}_{s_3},x_{j_4},\cdots),$$

where $j_0<j_1<j_2<\cdots$ and, if we denote S the number of sequences s_i, for all $i\in S$, s_i is a sequence including only the dose d. In words, the dose sequence $(x_n)_{n\in\mathbb{N}}$ switches between sequences including only the dose d and sequences which never include it. s_i is then the i-th sequence of dose d. We denote by A_i the random variable defined by

$$A_i=\begin{cases}\{\text{size}(s_i)<\infty\}, & \text{if } s_i \text{ exists},\\ \varnothing, & \text{otherwise},\end{cases}$$

where $\text{size}(s_i)$ denotes the number of terms in the sequence s_i. We are now interested in the evaluation of $\mathbb{P}(S=\infty)$. First we note B_i the random variable $\{\tilde{s}_i<\infty\}$, where $\tilde{s}_1,\tilde{s}_2,\cdots$ are independent sequences of dose d, such as $\Pi_{\tilde{n}_i}(d)/\Pi_{\tilde{n}_i}(r)=1$ with $\tilde{n}_i$ the start index of $\tilde{s}_i$. In other terms, we place us in the most negative case, *i.e.* the sequence switches from a dose $x_{\tilde{n}_i-1}\neq d$ to the dose d with a ratio of probabilities equals to 1. Then,

$$\begin{aligned}\mathbb{P}(S=\infty)=\mathbb{P}\left(\bigcap_{i\in\mathbb{N}}A_i\right)&=\mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\cdots\mathbb{P}(A_n|A_{n-1},\cdots,A_1)\cdots\\&\leq\mathbb{P}(B_1)\mathbb{P}(B_2)\cdots\mathbb{P}(B_n)\cdots \qquad (B.4)\\&=\mathbb{P}\left(\bigcap_{i\in\mathbb{N}}B_i\right), \qquad (B.5)\end{aligned}$$

where Inequality (B.4) is due to the definition of the random variables B_i and Equation (B.5) to their independence.

Moreover, we have that for all $i\in\mathbb{N}$, $\mathbb{P}(B_i)<1$. Indeed, let us define the sequence of doses $(x_n)_{n\in\mathbb{N}}$ such that $x_n=d$ for all $n\in\mathbb{N}$. By the law of the iterated logarithm, there exists $N\in\mathbb{N}$ such that $\mathbb{P}(B_0)=1$, where $B_0=\{x_N,x_{N+1},\cdots\}$. We denote by A_0 the set $\{x_1,\cdots,x_{N-1}\}$. As the length of A_0 is finite, we know that $\mathbb{P}(A_0)>0$. With the inequality $\mathbb{P}(\bar{B}_i)\geq\mathbb{P}(A_0\cap B_0)$, due to the definition of B_i, we have that $\mathbb{P}(B_i)<1$, for all $i\in\mathbb{N}$. Thus $\mathbb{P}\left(\bigcap_{i\in\mathbb{N}}B_i\right)<1$. By the Kolmogorov's zero-one law, it implies that $\mathbb{P}\left(\bigcap_{i\in\mathbb{N}}B_i\right)=0$ and then $\mathbb{P}(S=\infty)=0$. In words, it means that there exists a last infinite sequence of dose d. This is in contradiction with the fact that $r\in\tilde{D}$. □

B.2.2 Proof of theorem 5.5

Proof. Let us start with the proof of the ε-sensitivity. In this proof, we are interested in an asymptotic behaviour of the po-SPM, that is why we ignore the doses tested only a finite number of times and we reason as if they were never allocated. In other words, the

doses in this proof are always considered in $\tilde{D}$. Assume now that $\mathcal{E}(I,\beta)$ is not empty. Let $r \in \tilde{D} \setminus \mathcal{E}(I,\beta)$. We can distinguish two cases.

The first one is the existence of a dose $d \in \mathcal{E}(I,\beta)$ such that d is ordered with r. We are then reduced to the SPM in case of total ordering and the proof can be found in Clertant (2015).

The second one is the opposite, *i.e.* there exists no dose in $\mathcal{E}(I,\beta)$ ordered with r. So there exists a dose $d \in \mathcal{E}(I,\beta)$ not ordered with r, because we assume that $\mathcal{E}(I,\beta)$ is not empty. We want now to compare the integrals $I_{n,r}$ and $I_{n,d}$.

$$\frac{I_{n,r}}{I_{n,d}} = \prod_{k\in\tilde{D}} \frac{M^k_{n,r}}{M^k_{n,d}} = \prod_{k\in\tilde{D}} \frac{\int_{S^k_r} g\left(q_k, n^1_k, n^0_k\right) \Lambda_r\left(dq_k\right)}{\int_{S^k_d} g\left(q_k, n^1_k, n^0_k\right) \Lambda_d\left(dq_k\right)} \tag{B.6}$$

$$= \prod_{k\in\tilde{D}} \frac{\int_{S^k_r} g\left(q_k, n^1_k, n^0_k\right) \lambda_r\left(q_k\right) dq_k}{\int_{S^k_d} g\left(q_k, n^1_k, n^0_k\right) \lambda_d\left(q_k\right) dq_k} \tag{B.7}$$

$$\leq \frac{S}{s} \prod_{k\in\tilde{D}} \frac{\int_{S^k_r} g\left(q_k, n^1_k, n^0_k\right) dq_k}{\int_{S^k_d} g\left(q_k, n^1_k, n^0_k\right) dq_k}, \tag{B.8}$$

where Equation (B.6) follows from Assumption 5.3.1, Equation (B.7) from Assumption 5.3.4 *(a)* and Inequation (B.8) from Assumption 5.3.4 *(b)*. We state then the following property. For all function f continuous on $[0,1]$, we have

$$\int f\left(q_k\right) \frac{g(q_k, n^1_k, n^0_k)}{\text{Beta}(n^1_k+1, n^0_k+1)} dq_k \underset{n_k\to\infty}{\longrightarrow} \int f\left(q_k\right) pi_{\{\beta_k\}}\left(q_k\right) \gamma\left(dq_k\right) = f\left(\beta_k\right), \tag{B.9}$$

where Beta(.) denotes the Beta function and γ the counting measure. Let then $k \in \tilde{D}$. We are looking for the behaviour, when n_k goes to infinity, of

$$Q_k = \frac{\int_{S^k_r} g\left(q_k, n^1_k, n^0_k\right) dq_k}{\int_{S^k_d} g\left(q_k, n^1_k, n^0_k\right) dq_k}.$$

The case $S^k_r = S^k_d$ is obvious. For the other cases, we use the convergence expressed in Equation (B.9). As d is not ordered with r, with Assumption 5.3.3, we have six cases left to study.

- $S^k_r = A$ or B or I and $S^k_d = [0,1]$. Then $Q_k \underset{n_k\to\infty}{\longrightarrow} pi_A\left(\beta_k\right)$ or $pi_B\left(\beta_k\right)$ or 0. This last result is due to the fact that $r \notin \mathcal{E}(I,\beta)$.
- $S^k_r = [0,1]$ and $S^k_d = I$. Then $Q_k \underset{n_k\to\infty}{\longrightarrow} 1$, because $d \in \mathcal{E}(I,\beta)$.
- $S^k_r = [0,1]$ and $S^k_d = A$. Then $Q_k \underset{n_k\to\infty}{\longrightarrow} 1/\mathbb{1}_A\left(\beta_k\right) = 1$. Indeed, with Lemma B.2, as $d \in \mathcal{E}(I,\beta) \cap \tilde{D}$, $k \in \tilde{D}$ cannot be in $\mathcal{E}(I,\beta)$ and so $\beta_k \in A$.
- $S^k_r = [0,1]$ and $S^k_d = B$. With the same argument as the previous case, we have $Q_k \underset{n_k\to\infty}{\longrightarrow} 1$.

Finally, going back to Inequality (B.8), we conclude that $I_{n,r}/I_{n,d}$ tends to 0 when n goes to infinity. This leads to a contradiction because, as $r \in \tilde{D}$, this ratio is greater than 1 infinitely often. So $\tilde{D} \subset \mathcal{E}(I,\beta)$, *i.e.* the po-SPM is ε-sensitive.

We prove now that the po-SPM is also ε-balanced. Assumption 5.3.1 allows us to focus on the marginal ratio

$$\frac{M^k_{n,r}}{M^k_{n,t}} = \frac{\int g(q_j, n^1_k, n^0_k)\ \Lambda^k_r(dq_k)}{\int g(q_k, n^1_k, n^0_k)\ \Lambda^k_t(dq_k)} \quad \text{and} \quad \frac{I_{n,r}}{I_{n,t}} = \prod_{k\in D} \frac{M^k_{n,r}}{M_{n,t^k}}.$$

We note, as in the lemma 2 of Clertant (2015), that Assumption 5.3.4 involves :

$$d(\beta_k, S_r^k) = d(\beta_k, S_t^k) \Longrightarrow 0 < \liminf_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} \leq \limsup_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} < +\infty, \ a.s. \tag{B.10}$$

The ε-balanced behaviour corresponds to the case where, forall $k \in D,\ \beta_k \notin I$. We show that

$$r \notin M_D \Longrightarrow \mathbb{P}(\{n_r \to \infty\}) = 0 \tag{B.11}$$

By symmetry we can choose $r \in L$. The set $M_D \cap L \cap A_r$ is not empty. Let t be a dose in $M_D \cap L \cap A_r$. By using Equation (B.10), as $A_t \subset A_r$ and $S_t^k = [0,1]$ when $k \in C_t$, we have

$$\forall (i,j) \in A_t \cup C_t,\ \mathbb{P}\left(\limsup_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} < \infty \ \middle|\ n_k \to \infty \right) = 1.$$

If $k \in (B_r \cup C_r) \cap B_t$ then $\beta_{ji} \in B$ and we obtain the same result

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} < \infty \ \middle|\ n_k \to \infty \right) = 1.$$

We consider now the ratios $M_{n,r}^k / M_{n,t}^k$ when $k \in (B_t \cap A_r) \cup \{t, r\}$. In that case, we have $\beta_k \in B$, $S_t^k = B$ and S_r^k equals to I or A. By using Proposition B.1, we have

$$\forall k \in (B_t \cap A_r) \cup \{t, r\},\ \mathbb{P}\left(\lim_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} = 0 \ \middle|\ n_k \to \infty \} \right) = 1.$$

We then have

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{I_{n,r}}{I_{n,t}} = 0 \ \middle|\ n_r \to \infty \right) = 1 \text{ and } \mathbb{P}\left(\left\{ \lim_{n\to\infty} \frac{I_{n,r}}{I_{n,t}} = 0 \right\} \cap \{n_r \to \infty\} \right) = 0,$$

which proves Equation (B.11). We achieve the proof of the ε-balanced property by showing that

$$t \in M_D \Longrightarrow \mathbb{P}(n_t \to \infty) = 1 \tag{B.12}$$

By using Equation (B.11), we have, for all r and t in M_D

$$\forall (i,j) \in D,\ k \neq r \text{ and } k \neq t,\ \mathbb{P}\left(\limsup_{n\to\infty} \frac{M_{n,r}^k}{M_{n,t}^k} < \infty \right) = 1.$$

Moreover, for all $r \in M_D$, S_t^r is equal to the segment $[0,1]$. As $\beta_r \notin I = S_r^r$, we have

$$\forall r \in M_D,\ \mathbb{P}\left(\lim_{n\to\infty} \frac{M_{n,r}^r}{M_{n,t}^r} = 0 \ \middle|\ n_r \to \infty \right) = 1.$$

Let E_r be the event $\{n_r \to \infty\} \cap \{n_t \to \infty\}^c$. Then,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{I_{n,r}}{I_{n,t}} = 0 \ \middle|\ E_r \right) = 1 \text{ and } \mathbb{P}\left(\left\{ \lim_{n\to\infty} \frac{I_{n,r}}{I_{n,t}} = 0 \right\} \cap E_r \right) = 0,$$

As $\mathbb{P}(\sum_{M_D} n_k \to +\infty) = 1$, we have $\mathbb{P}(\cup_{M_D} E_r) = 1$, which proves Equation (B.12) and ends the demonstration of the po-SPM ε-balanced behaviour. □

Table des figures

Liste des tableaux

Bibliographie

Aalen, O. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer, (1980).

Abrahamowicz, M., T. Mackenzie, and Esdaile, J. M. (1996). Time-dependent hazard ratio : modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, 91(436) :1432–1439.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes : A large sample study. *The Annals of Statistics*, 10(4) :pp. 1100–1120.

Anderson, J. A. and Senthilselvan, A. (1982). A two-step regression model for hazard functions. *Applied Statistics*, pages 44–51.

Antoniadis, A., G. Grégoire, and Nason, G. (1999). Density and hazard rate estimation for right-censored data by using wavelet methods. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(1) :63–84.

Arjas, E. (1988). A graphical method for assessing goodness of fit in cox's proportional hazards model. *Journal of the American Statistical Association*, 83(401) :204–212.

Azriel, D., M. Mandel, and Rinott, Y. (2011). The treatment versus experimentation dilemma in dose finding studies. *Journal of Statistical Planning and Inference*, 141(8) : 2759 – 2768.

Bagdonavičius, V. and Nikulin, M. (1997). Asymptotical analysis of semiparametric models in survival analysis and accelerated life testing. *Staistics : a journal of theoretical and applied staistics*, 29(3) :261–283.

Bagdonavičius, V. and Nikulin, M. (2000). Modèle statistique de dégradation avec des covariables dépendant du temps. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 330(2) :131–134.

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, 15(5) :453–472.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4) :551–563.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1) :1–22.

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

Beran, R. Nonparametric regression with randomly censored survival data. Technical report, Technical Report, Univ. California, Berkeley, (1981).

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13 :1063–1095.

Biau, G., L. Devroye, and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9 :2015–2033.

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

Braun, T. M. and Jia, N. (2013). A generalized continual reassessment method for two-agent phase i trials. *Statistics in Biopharmaceutical Research*, 5(2) :105–115.

Braun, T. M. and Wang, S. (2010). A hierarchical bayesian design for phase i trials of novel combinations of cancer therapeutic agents. *Biometrics*, 66(3) :805–812.

Breiman, L. (2001). Random forests. *Machine Learning*, 45 :5–32.

Breiman, L. *Consistency for a simple model of random forests.* Technical Report 670, UC Berkeley, (2004).

Breiman, L., J. Friedman, R.A. Olshen, and Stone, C.J. *Classification and Regression Trees.* Chapman & Hall, New York, (1984).

Brown, C. C. (1975). On the use of indicator variables for studying the time-dependence of parameters in a response-time model. *Biometrics*, pages 863–872.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1) :93–111.

Castellan, G. and Letué, F. (2000). Estimation of the cox regression function via model selection. *Chapter of th PhD thesis of F. Letué.*

Chauvel, C. *Empirical Processes for Inference in the Non-Proportional Hazards model.* PhD thesis, Université Pierre et Marie Curie - Paris 6, Paris, (2014).

Chauvel, C. and O'quigley, J. (2014). Tests for comparing estimated survival functions. *Biometrika*, page asu015.

Cheung, Y. K. *Dose finding by the continual reassessment method.* CRC Press, (2011).

Cheung, Ying Kuen. (2005). Coherence principles in dose-finding studies. *Biometrika*, 92 (4) :863–873.

Clertant, M. *Semi-parametric bayesian model, applications in dose finding studies.* PhD thesis, Université Pierre et Marie Curie - Paris 6, Paris, (2015).

Comte, F., S. Gaïffas, and Guilloux, A. Adaptive estimation of the conditional intensity of marker-dependent counting processes. In *Annales de l'institut Henri Poincaré (B)*, volume 47, pages 1171–1196, (2011).

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

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics*, 17(3) :1157–1167.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2) :247–254.

Denil, M., D. Matheson, and Freitas, N. de. *Consistency of online random forests.* arXiv :1302.4853, (2013).

Devroye, L., L. Györfi, and Lugosi, G. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, (1996).

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359) :557–565.

Efron, B. and Tibshirani, R.J. *An Introduction to the Bootstrap.* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, (1994). ISBN 9780412042317.

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

Gehan, E. A. (1965). A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2) :203–223.

Genuer, R., J.-M. Poggi, and Tuleau, C. Random forests : some methodological insights. arXiv :0811.3619, (2008).

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3) :515–526.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420) :942–951.

Haara, P. A note on the asymptotic behaviour of the empirical score in cox's regression model for counting processes. In *Proceedings of the 1st World Congress of the Bernoulli Society*, pages 139–142, (1987).

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3) :553–566.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4) :757–796.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, 37(3) :323–341.

Hess, K. R. (1994). Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Statistics in medicine*, 13(10) :1045–1062.

Hothorn, T. and Lausen, B. (2003). On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2) :121–137.

Huang, X., S. Biswas, Y. Oki, J. Issa, and Berry, D. A. (2007). A parallel phase i/ii clinical trial design for combination therapies. *Biometrics*, 63(2) :429–436.

Huber, C. (2000). Censored and truncated lifetime data. *Recent Advances in Reliability Theory*, pages 291–305.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1 :519–537.

Ishwaran, H. and Kogalur, U.B. (2010). Consistency of random survival forests. *Statistics & Probability Letters*, 80 :1056–1064.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, pages 841–860.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60(2) :267–278.

Kalbfleisch, J. D. and Prentice, R. L. *The statistical analysis of failure time data.* Wiley series in probability and mathematical statistics : Applied probability and statistics. Wiley, (1980). ISBN 9780471055198.

Kalbfleisch, J. D. and Prentice, R. L. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, (2011).

Kaplan, E. L. and Meier, Paul. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282) :pp. 457–481.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, pages 227–237.

Kleiber, C., K. Hornik, F. Leisch, and Zeileis, A. (2002). strucchange : An r package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2) :1–38.

Klein, J. P. and Moeschberger, M. L. *Survival analysis : techniques for censored and truncated data.* Springer Science & Business Media, (2003).

Klein, J. P., H. C. Van Houwelingen, J. G. Ibrahim, and Scheike, T. H. *Handbook of survival analysis.* CRC Press, (2013).

Lausen, B. and Schumacher, M. (1996). Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Comput. Stat. Data Anal.*, 21(3) :307–326.

Lawless, J. F. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, (2011).

Leblanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422) :457–467.

LeBlanc, M. and Crowley, J. (1999). Adaptive regression splines in the cox model. *Biometrics*, 55(1) :204–213.

Leurgans, S. (1983). Three classes of censored data rank tests : Strengths and weaknesses under censoring. *Biometrika*, 70(3) :651–658.

Leurgans, S. (1984). Asymptotic behavior of two-sample rank tests in the presence of random censoring. *The Annals of Statistics*, pages 572–589.

Liang, K. Y., S. G. Self, and Liu, X. (1990). The cox proportional hazards model with change point : An epidemiologic application. *Biometrics*, 46(3) :783–793.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2 :18–22.

Lin, D. Y. (1991). Goodness-of-fit analysis for the cox regression model based on a class of parameter estimators. *Journal of the American Statistical Association*, 86(415) :pp. 725–728.

Lin, D.Y. and Ying, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *The annals of Statistics*, pages 1712–1734.

Liu, J., S. Wu, and Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, 7(2) :497–525.

Mander, A. P. and Sweeting, M. J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase i trials. *Statistics in Medicine*, 34(8) :1261–1276. ISSN 1097-0258.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1*, 50(3) :163–170.

Marubini, E. and Valsecchi, M. G. *Analysing survival data from clinical trials and observational studies*, volume 15. John Wiley & Sons, (2004).

Marzec, L. and Marzec, P. (1997). On fitting cox's regression model with time-dependent coefficients. *Biometrika*, 84(4) :901–908.

McKeague, I. W. and Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *The Annals of Statistics*, pages 1172–1187.

Meier, L., S. Van Geerde , and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, 37 :3779–3821.

Mentch, L. and Hooker, G. Ensemble trees and clts : Statistical inference for supervised learning. arXiv :1404.6473, (2014).

Moreau, T., J. O'Quigley, and Mesbah, M. (1985). A global goodness-of-fit statistic for the proportional hazards model. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3) :212–218.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1) :153 – 180.

Naftel, D., E. Blackstone, and Turner, M. (1985). Conservation of events. *Unpublished notes*.

O'Quigley, J. (2003). Khmaladze-type graphical evaluation of the proportional hazards assumption. *Biometrika*, 90(3) :577–584.

O'Quigley, J. *Proportional hazards regression*, volume 542. Springer, (2008).

O'Quigley, J. and Pessione, F. (1989). Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics*, 45(1) :135–144.

O'Quigley, J. and Pessione, F. (1991). The problem of a covariate-time qualitative interaction in a survival study. *Biometrics*, 47(1) :101—115.

O'Quigley, J., M. Pepe, and Fisher, L. (1990). Continual reassessment method : a practical design for Phase I clinical trials in cancer. *Biometrics*, 46(1) :33–48.

O'Sullivan, F. (1993). Nonparametric estimation in the cox model. *The Annals of Statistics*, pages 124–145.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, pages 185–207.

Pohar, M. and Stare, J. (2006). Relative survival analysis in r. *Computer methods and programs in biomedicine*, 81(3) :272–278.

Qi, Y. *Ensemble Machine Learning*, chapter Random forest for bioinformatics, pages 307–323. Springer, (2012).

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

Rogez, G., J. Rihan, S. Ramalingam, C. Orrite, and Torr, P. H. Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, (2008).

Satten, G. A. and Datta, S. (2001). The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3) :207–210.

Scheike, T. H. and Zhang, M.-J. (2002). An additive–multiplicative cox–aalen regression model. *Scandinavian Journal of Statistics*, 29(1) :75–88.

Scornet, E. (2014). On the asymptotics of random forests. *arXiv preprint arXiv :1409.2090*.

Scornet, E., G. Biau, and Vert, J.-P. Consistency of random forests. arXiv :1405.2881, (2014).

Segal, M. R. (1988). Regression trees for censored data. *Biometrics*, pages 35–47.

Shen, L. Z. and O'Quigley, J. (1996). Consistency of continual reassessment method under model misspecification. *Biometrika*, 83(2) :395–405.

Skolnik, J. M., J. S. Barrett, B. Jayaraman, D. Patel, and Adamson, P. C. (2008). Shortening the timeline of pediatric phase i trials : the rolling six design. *Journal of Clinical Oncology*, 26(2) :190–195.

Stablein, D. M., W. H. Carter, and Novak, J. W. (1981). Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, 2(2) :149–159.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8 :1348–1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10 :1040–1053.

Storer, B. E. (1989). Design and analysis of phase i clinical trials. *Biometrics*, pages 925–937.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2) :pp. 363–369.

Stute, W. (1986). Conditional empirical processes. *The Annals of Statistics*, pages 638–647.

Stute, W. (1995). The central limit theorem under random censorship. *The Annals of Statistics*, pages 422–439.

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.*, 21(3) :1591–1607.

Sullivan, J. H. (2002). Estimating the locations of multiple change points in the mean. *Computational Statistics*, 17(2) :289–296.

Svetnik, V., A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and Feuston, B.P. (2003). Random forest : A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43 :1947–1958.

Thall, P. F., R. E. Millikan, P. Mueller, and Lee, S. (2003). Dose-finding with two agents in phase i oncology trials. *Biometrics*, 59(3) :487–496.

Therneau, T. M. and Grambsch, P. M. *Modeling survival data : extending the Cox model.* Springer Science & Business Media, (2000).

Laan, M.van der , E.C. Polley, and Hubbard, A.E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6.

Verweij, P. and Houwelingen, H. Van. (1995). Time-dependent effects of fixed covariates in cox regression. *Biometrics*, 51(4) :1550–1556.

Wager, S. Asymptotic theory for random forests. arXiv :1405.0352, (2014).

Wages, N. A., M. R. Conaway, and O'Quigley, J. (2011). Continual reassessment method for partial ordering. *Biometrics*, 67(4) :1555–1563.

Wahba, G. *Spline models for observational data*, volume 59. Siam, (1990).

Wang, K. and Ivanova, A. (2005). Two-dimensional dose finding in discrete dose space. *Biometrics*, 61(1) :217–222. ISSN 1541-0420.

Wei, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association*, 79(387) :649–652.

Xu, R. and Adak, S. (2002). Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics*, 58(2) :305–315. ISSN 1541-0420.

Xu, R. and Harrington, D. P. (2001). A semiparametric estimate of treatment effects with censored data. *Biometrics*, 57(3) :875–885.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4) :423–439.

Yang, F., J. Wang, and Fan, G. (2010). Kernel induced random survival forests. *arXiv preprint arXiv :1008.3952*.

Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, 92(1) :1–17.

Yin, G. and Yuan, Y. (2009). A latent contingency table approach to dose-finding for combinations of two agents. *Biometrics*, 65(3) :866–875.

Yin, G. and Yuan, Y. (2009). Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 58(2) :211–224. ISSN 1467-9876.

Zeileis, A., C. Kleiber, W. Krämer, and Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1) :109–123.

Zucker, David M. and Karr, Alan F. (1990). Nonparametric survival analysis with time-dependent covariate effects : a penalized partial likelihood approach. *Ann. Statist.*, 18 (1) :329–353.