# UNIVERSITÉ CHEIKH ANTA DIOP DE DAKAR


## ÉCOLE DOCTORALE MATHÉMATIQUES ET INFORMATIQUE

Année : 2019 N d'ordre : 143

# THÈSE DE DOCTORAT UNIQUE

pour obtenir le grade de

**DOCTEUR EN MATHÉMATIQUES**
**DE L'UNIVERSITÉ CHEIKH ANTA DIOP**

**Mention : Mathématiques et Modélisation**

**Spécialité : Analyse, Statistique et Applications**

présentée par

**Jean Claude UTAZIRUBANDA**

Sujet :

# Méthode du group-lasso pour la sélection de variables dans le modèle de Cox pour les données en clusters

Soutenue publiquement le 25 janvier 2020, devant le jury composé de :

| | | | |
|---|---|---|---|
| Pésident du jury | **M. Diaraf SECK** | Professeur | UCAD |
| RAPPORTEURS | **M. Abdou Kâ DIONGUE** | Professeur | UGB |
| | **M. Kossi Essona GNEYOU** | Professeur | Université de Lomé |
| EXAMINATEURS | **M. Gabriel Birame NDIAYE** | Professeur | UCAD |
| | **M. Abdoulaye SENE** | Professeur | UCAD |
| DIRECTEUR DE THÈSE | **M. Papa NGOM** | Professeur | UCAD |

# Dédicaces

*À DIEU, le Tout Puissant*
*Le Tout Miséricordieux, le Très Miséricordieux. Je me prosterne devant ta Grandeur pour te remercier de m'avoir comblé de ta grâce et de m'avoir assisté tout au long de ce voyage dans le jardin du savoir et enfin réaliser mes rêves.*

*À mon fils Shiloh Shami IZIHIRWE.*
*À mon épouse Joselyne UMURUNGI.*
*À ma mère Hélène MUKAMWEZI.*
*À toute ma famille et belle famille.*

# Remerciements

Je tiens en premier lieu à remercier chaleureusement mon directeur de thèse, **M. Papa NGOM**, sans qui, cette thèse n'aurait pu avoir lieu. Son professionnalisme ainsi que son sens de l'écoute et de la compréhension font de lui un homme singulier. La confiance et tout le temps qu'il m'a accordés m'ont encouragé à accomplir cette thèse et ses connaissances et conseils au quotidien m'ont été plus que bénéfiques.

Je remercie les membres du jury qui m'ont fait l'honneur d'évaluer ce travail. Merci à M. Diaraf SECK qui a accepté de présider le jury, merci à M. Abdou Kâ DIONGUE et M. Kossi Essona GNEYOU d'avoir aimablement accepté d'être les rapporteurs de cette thèse. C'est pour moi un grand honneur d'être évalué par d'aussi remarquables chercheurs. Merci enfin à M. Gabriel Birame NDIAYE et M. Abdoulaye SENE d'avoir accepté de composer le jury et d'examiner ce travail.

Je remercie ma chère patrie le Rwanda, pays des mille collines et berceau de mes ancêtres. Un grand merci au gouvernement rwandais d'avoir financé cette thèse. Je remercie également l'État sénégalais, la Teranga, de son accueil chaleureux pour l'émergence des chercheurs africains dans tous les secteurs d'activité. Je remercie toute l'équipe de l'Institut Africain des Sciences Mathématiques au Sénégal (AIMS-Sénégal), plus particulièrement mes collègues tuteurs de leurs encouragements tout au long de la réalisation de cette thèse.

À mon épouse **Joselyne UMURUNGI**, merci d'être toujours à mes côtés, par ta présence, par ton amour dévoué et ta tendresse, pour donner du goût et du sens à notre vie de famille. En témoignage de mon amour, de mon admiration et de ma grande affection, je te prie de trouver dans ce travail l'expression de mon estime et mon sincère attachement.

Je ne saurais terminer sans remercier ma famille et ma belle famille, mes amis, l'Ambassade de la République du Rwanda au Sénégal, l'Association de la Communauté Rwandaise au Sénégal (ACRS) et toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail. J'exprime ma profonde gratitude à la famille Innocent NZEYIMANA, à la famille Patrick KARAMAGA, Maman Carolina Rimoldi et mon collaborateur Dr. Tomas. M. Leon.

# Citations

« *N'essayez pas de devenir un homme qui a du succès.*
*Essayez de devenir un homme qui a de la valeur.* »
*Albert Einstein*

« *Ce n'est pas ce qu'il a, ni même ce qu'il fait,*
*qui exprime directement la valeur d'un homme : c'est ce qu'il est.* »
*Henri Frederic Amiel*

# Abréviations & Notations

## Variables aléatoires et modes de convergence

| | |
|---|---|
| $\mathbb{P}(A)$ | : la probabilité de l'événement $A$. |
| $\mathbb{E}(X)$ | : l'espérance mathématique de la variable aléatoire $X$. |
| $\mathbb{V}ar(X)$ | : la variance de la variable aléatoire $X$. |
| $X_n \xrightarrow{\mathbb{P}} Y$ | : la suite de variables aléatoires $(X_n)_n$ converge en probabilité vers $Y$. |
| $X_n \xrightarrow{\mathcal{L}} Y$ | : la suite de variables aléatoires $(X)_n$ converge en distribution vers Y. |
| *i.i.d* | : indépendantes et identiquement distribuées. |
| $p.s$ | : presque sûrement. |
| $\mathcal{F}_t$ | : filtration à l'instant $t$. |
| $\underset{]s,t]}{\mathcal{P}}$ ou $\overset{t}{\underset{s}{\mathcal{P}}}$ | : produit-intégrale (ou produit infini) sur $]s,t]$. |

## Notations d'ordre général

| | |
|---|---|
| $\binom{p}{n} = \frac{n!}{p!(n-p)!}$ | : coefficient binomial. |
| $\beta^T$ | : transposée de $\beta$. |
| $\|U\|$ | : norme du vecteur $U$. |
| $\otimes$ | : produit direct. |
| $Z^{\otimes 2}$ | : $ZZ^\top$ pour tout vecteur colonne transposé $Z$. |
| $\mathbb{I}_A$ | : fonction indicatrice de $A$; $\mathbb{I}_A(x)$ vaut $1$ si $x \in A$ et $0$ sinon. |
| $p \gg n$ | : $p$ très grand devant $n$ |

## Modèle de survie

| | |
|---|---|
| $X$ | : temps de survenue de l'événement d'intérêt. |
| $C$ | : temps de censure. |
| $T = \min(X, C)$ | : temps réellement observé. |
| $\delta = \mathbb{I}_{\{X \leq C\}}$ | : indicateur d'événement. |
| $\tilde{T}$ | : temps observé non censuré (c'est-à-dire $\delta = 1$). |
| $Y(t) = \mathbb{I}_{\{t \leq T_i\}}$ | : indicateur de risque de subir un événement juste avant $T_i$. |
| $Z_{ij}$ | : un vecteur de covariables associé au temps de survie $X_{ij}$ de l'individu $j$ du groupe $i$. |
| $R(T_i)$ | : ensemble des individus encore à risque juste avant $T_i$. |
| $f_\theta(T_i)$ | : utilisée s'il n'y a pas d'embiguité pour désigner $f_T^\theta(T_i)$ (la densité de probabilité de la variable aléatoire $T$ prise à la valeur $T_i$ et indexée par la valeur $\theta$ des paramètres). |
| $F(t)$ | : fonction de repartition. |
| $S(t)$ | : fonction de survie. |
| $S_\theta(T_i)$ | : fonction de survie prise à la valeur de $T_i$ et indexée par la $\theta$ des paramètres. |
| $h(t)$ | : fonction de risque. |
| $h_0(t)$ | : fonction de risque de base. |
| $H(t)$ | : fonction de risque cumulé. |
| $\widehat{H}(t)$ | : estimateur de la fonction de risque cumulé. |
| $H_0(t)$ | : fonction de risque cumulé de base. |
| $\widehat{H}_0(t)$ | : estimateur de la fonction de risque cumulé de base. |
| $\hat{A}(t)$ | : estimateur de Nelson-Aalen de la fonction de risque cumulé. |
| $\lambda(t)$ | : intensité du processus de comptage. |

## Remarque sur la bibliographie

Dans la bibliographie, les numéros qui apparaissent après une référence désignent les numéros de pages où cette référence a été citée.

# Table des matières

# Table des figures

# Liste des tableaux

# Introduction Générale

## Contexte général

De nos jours, les données volumineuses sont au coeur des problématiques émergentes en recherche, que ce soit dans le domaine de la biologie, de l'épidémiologie, de la physique ou de la finance. Par exemple, pour une maladie complexe, les chercheurs tentent toujours à collecter un grand nombre de données sur chacun des sujets de l'étude afin de mieux saisir les mécanismes biologiques et génétiques entourant ces maladies complexes.

Soit $X$ une matrice de données $n \times p$ où $n$ est la taille de l'échantillon et $p$ le nombre de variables. Les données en grande dimension signifie généralement que $p$ est supérieur à $n$. L'analyse de données en grande dimension pose souvent des problèmes qui nécessitent de nouvelles méthodologies et théories statistiques [31]. Par exemple, l'ajustement par les moindres carrés de modèles linéaires et les méthodes statistiques multivariées classiques ne permettent pas de traiter un $X$ de grande dimension, car ils reposent tous les deux sur l'inverse de $X'X$ qui peut être singulier. Il est à noter que l'augmentation de $n$ et de $p$ a des effets très différents et opposés sur les résultats statistiques. En général, l'analyse multivariée a pour objectif de tirer des conclusions statistiques sur la dépendance entre les variables considerées, de sorte que l'augmentation de $n$ a pour effet d'améliorer la précision et la certitude de l'inférence, tandis que l'augmentation de $p$ a l'effet inverse de réduire la précision et la certitude. Par conséquent, le niveau de détail pouvant être déduit des corrélations entre variables s'accroit avec $n$, mais se détériore avec l'augmentation de $p$.

Dans les modèles de regréssion classique, $p \leq n$. Cependant, dans la majorité des cas en statistique appliquée, les données observées sont de grande taille, c'est-à-dire de dimension $p$ très grand devant $n$ ($p \gg n$). Par exemple en analyse de survie d'une maladie génétique, pour chaque individu, on dispose d'un grand nombre de gènes, qui est parfois nettement plus elevé que le nombre d'individus

dans l'échantillon. Bien évidemment, ce ne sont pas tous les gènes qui sont associés à cette maladie génétique en question. Dans cette analyse, nous nous retrouvons avec un vecteur $Y$ de dimension $n \times 1$ contenant les durées de survie d'une certaine maladie et une matrice $X$ de dimension $n \times p$ contenant les expressions génétiques, où le nombre de colonnes est beaucoup plus grand que le nombre de lignes ($p \gg n$).

La réduction des dimensions et la sélection des variables sont d'une importance capitale pour l'analyse de données de grandes dimensions. En effet, le principe de parcimonie (sparsity), qui suppose que seul un petit nombre de prédicteurs contribuent à la réponse, est donc fondamental et il est fréquemment consideré comme déterminant dans l'analyse et l'interprétation statistique de données en grande dimension. Partant du principe de parcimonie (qui permet d'expliquer un phénomène avec un minimum de variables explicatives), un grand nombre de méthodes d'estimation ont été développées au cours de la dernière décennie. Elles sont adaptées au choix des modèles parcimonieux, à l'estimation des coefficients associés et à la sélection des variables statistiquement significatives et ceci de façon simultanée. Ces méthodes reposent sur l'introduction d'un terme de pénalité, dont l'une des plus populaires est sans aucun doute l'approche connue sous le nom de "méthode Lasso" introduite par [111].

# Problématique

L'accroissement de la masse et de la taille des données a en effet nécessité la proposition de nouvelles approches statistiques adaptées aux caractéristiques des données modernes. En particulier, le gros volume des données (nombre important de variables) pose un ensemble de problèmes à la statistique multivariée classique que l'on résume usuellement par le terme "fléau de la dimension" [13]. Parmi les difficultés que pose l'analyse statistique en grande dimension, on peut citer les problème numériques, les problèmes d'inférence ou les problèmes de biais des estimateurs. Il a donc été nécessaire de développer durant ces dernières années, des méthodes capables de pallier avec succès ces problèmes. On note une évolution diversifiée de la litterature portant sur les méthodes d'estimation pénalisée et la sélection des modèles de durées de survie censurées pour les données en cluster de grande dimension. A notre connaissance, malgré ces outils décrivant les aspects

méthodologiques, algorithmiques, computationnels et théoriques, la méthode du "group-Lasso" dans le modèle de régression semi-paramétrique à risques proportionnels de Cox, avec fragilité pour les données en clusters, n'a pas été abordée et fait l'objet principal de cette thèse.

## Objectifs

L'objectif de notre travail était d'abord de réaliser un état de l'art exhaustif sur les méthodes d'estimation pénalisées, souvent utilisées, pour faire de la sélection de variables en grande dimension dans les modèles de durées de vie censurées. Cet état de l'art concerne les aspects méthodologiques, algorithmiques, computationnels et théoriques des méthodes proposées.

Cela commence par une présentation des thóries et méthodologies existantes sur l'intérêt de la sélection de variables dans les modèles de durées de vie censurées. Nous nous sommes également intéressés à la littérature développée sur les modèles de survie avec fragilités (fragilité partgée, fragilité corrélée ainsi que les fragilités emboitées). Ensuite, nous avons présenté les méthodes d'estimations pénalisées (méthodes de régularisation) Ridge [118], Lasso [111], SCAD [35], Elastic net [137], Adptive-Lasso [136] et Mcp [133] pour la sélection des variables dans modèle de régression semi-paramétrique à risques proportionnels de Cox. Un accent particulier a été mis sur l'aspect computationnel. Cet état de l'art inclut également la compréhension des différents packages R développés. Notamment leurs avantages et inconvénients.

Dans une seconde partie, on a adapté la méthode du "group-Lasso" au modèle de Cox pour des données en clusters ou "groupées" en tenant compte de la fragilité propre à chaque cluster. Les aspects algorithmiques ont été traités et des simulations ont été réalisées pour étudier le comportement des estimateurs obtenus. Ensuite la performance de la méthode group Lasso a été comparée avec celle des autres méthodes comme group-MCP et group-SCAD. La consistance des estimateurs obtenus par cette méthode a été établie théoriquement. Le travail a été ensuite généralisé en implémentant l' "Adaptive group-lasso" dans le modèle de Cox avec fragilité pour des données en clusters . Chacune des méthodologies étudiées a été confrontée aux différents jeux de données réelles, issus en grande partie

du domaine biomédical.

# Etat de l'art

L'identification d'un sous-ensemble de variables explicatives pertinentes est l'un des objectifs majeurs d'une analyse de modèle de régression en situation de "petite dimension"($p \leq n$). Dans ce contexte plus traditionnel, on recense plusieurs méthodes de sélection de modèle qui se proposent de comparer les modèles construits sur des sous-ensembles de variables explicatives selon un critère de qualité d'ajustement, pénalisé par le nombre de variables explicatives. Ainsi, dans le contexte du modèle linéaire généralisé, la minimisation du critère AIC introduit par [3] ou celui de BIC proposé par [102], (qui représentent des versions pénalisées de la déviance du modèle par la norme $\ell_0$ du vecteur $\beta$ des paramètres de régression), préfigurent les méthodes d'estimation par régularisation devenues si populaires pour les données en grande dimension et pour lesquelles la pénalisation est définie par les normes $\ell_2$ [56] ou $\ell_1$ [111] de $\beta$.

En effet, l'optimisation de critères pénalisés par :

$$||\beta||_0 = \#\{j \in [1; p], \beta_j \neq 0\}$$

où $\#A$ désigne le cardinal de l'ensemble A, nécessite l'ajustement de tous les sous-modèles possibles ($2^p$ modèles); ce qui pose des problèmes numériques, et ceci même pour des valeurs modérées de $p$. Certes, la sélection pas à pas constitue une alternative raisonnable d'un point de vue computationnel, mais le parcours par cet algorithme séquentiel d'une part très faible du graphe des sous-modèles, au mieux $p(p+1)/2$ sous-modèles, génère une instabilité de la procédure, d'autant plus grande que $p$ est lui-même grand [16, 35].

L'évolution de la technologie en terme de recueil et de stockage de données, notamment en biologie moléculaire pour l'étude du génome [104] pour les puces à ADN [12] et pour l'étude épigénétique de la méthylation de l'ADN, ou en neurosciences, pour l'analyse de l'activité cérébrale par électro-encéphalographie [52], ou l'imagerie par résonance magnétique [92], a conduit à des développements importants de la méthodologie statistique pour l'adapter ensuite à des situations caractérisées par un grand nombre de variables.

Les méthodes classiques, notamment celles dont l'objectif est l'identification de variables d'intérêt par sélection ou tests multiples, ont des propriétés analytiques éprouvées en situation asymptotique, lorsque le nombre $n$ d'individus tend vers l'infini et que le nombre de variables est fixe. Cependant, ces méthodes se montrent peu performantes en grande dimension. Par exemple, la propriété de consistance d'estimation du support par le critère BIC [105, 126] se perd lorsque le nombre de variables n'est pas fixé [18, 22, 69]. A l'instar de la problématique abordée plus haut pour évoquer la sélection pas à pas, un des problèmes est l'explosion combinatoire des associations possibles de variables sélectionnées, qui nécessite aussi le contrôle par des méthodes adaptées du nombre de sélections erronées. Une autre raison plus spécifique du paradigme $n \leq p$ est liée à l'instabilité voire l'impossibilité numérique de l'ajustement de modèles dont le nombre de paramètres dépasse celui des individus par des méthodes impliquant le plus souvent l'inversion de la matrice de variance-covariance des variables explicatives, par exemple la méthode des moindres carrés en régression. Ainsi, au-delà de la recherche de solutions statistiques performantes, un des défis de l'analyse de données de grande dimension est également la simplicité algorithmique des méthodes, garantissant la possibilité effective de leur mise en oeuvre.

Le modèle de régression semi-paramétrique à risques proportionnels de Cox [26] est l'un des modèles de régression de durées les plus utilisées en statistique médicale. Il permet en particulier d'identifier les facteurs de risque d'une maladie, de comparer des traitements, d'estimer les probabilités de survenue d'un événement (décès, rechute, etc) chez un individu identifié par un vecteur donné de variables explicatives. L'inférence statistique dans ce modèle est maintenant bien maîtrisée et repose généralement sur la méthode du maximum de vraisemblance partielle, qui fournit des estimateurs consistants et asymptotiquement gaussiens des paramètres du modèle [78, 37].

De nombreuses extensions de ce modèle ont été proposées, en particulier pour prendre en compte l'existence de "clusters" ou groupes d'individus au sein desquels les durées sont corrélées. Ces groupes peuvent représenter les individus d'une même famille, les patients traités au sein d'un même hôpital, les organes d'un même patient,...etc. Les clusters peuvent également représenter des durées observées de manière répétée sur le même individu: dates de rechute, dates de réapparution d'un symptôme donné,...etc [78].

Les modèles de régression de durées ont également fait l'objet de la litérature abondante développée autour des problèmes de sélection de variables en grande dimension pour le modèle de Cox [111, 112, 134, 9], pour le modèle de Cox avec fragilité [36], pour le modèle à risques convergents [75] et pour la classe des modèles de transformation linéaire [135]. Les méthodes de sélection de variables proposées ici reposent essentiellement sur des versions pénalisées de la méthode du maximum de vraisemblance partielle. Les pénalités Lasso (basée sur une norme $\ell_1$), ridge (basée sur une norme $\ell_2$), elastic-net sont parmi les plus étudiées [54]. L'estimateur du Lasso dans le contexte de l'analyse de survie en grande dimension a été aussi étudié par [79, 43] pour le modèle d'Aalen et par [113, 62, 73] entre autres pour le modèle de Cox.

Un autre aspect plus crucial dans la modélisation de durée de survie, est l'hétérogénéité non observée au sein d'une population en étude qu'on appelle souvent "fragilité" et qui se définit simplement dans le contexte épidémiologique, qu'un sujet peut être plus fragile qu'un autre et donc avoir un risque de mort ou d'autre événement pathologique plus grand. C'est un facteur de proportionnalité aléatoire non observé qui modifie la fonction de risque d'un individu ou d'un groupe de personnes liées les unes aux autres. Donc l'impact de l'hétérogénéité de la population sur l'interprétation de la forme des fonctions de risques est important [2]. Les modèles de fragilité [32, 24, 123, 51] étant des extensions du modèle classique à risques proportionnels de Cox [26], ils ont été proposés, en particulier, pour prendre en compte l'hétŕogénéité liée à l'existence de groupes ou "clusters" d'individus au sein desquels les durées sont corrélées.

Dans le cas de variables catégorielles, les méthodes d'estimation pénalisée où les variables sont sélectionnées individuellement comme dans le Lasso [112], la procedure LARS-Cox [49], la finesse résiduelle [103], l'algorithme LARS généralisé pour le modèle à risques proportionnels de Cox [88] et l'algorithme Lasso à gradient [108] ne sont plus adaptés. Dans cette situation ces méthodes sélectionnent les indicatrices des modalités et non le groupe d'indicatrices, i.e la variable dans sa totalité. Dans de telles situations, il est plus judicieux d'envisager de sélectionner (ou rejeter) les variables par groupes. Cette structure des variables peut être prise en compte en utilisant l'estimateur Group Lasso introduit par [128] pour le modèle linéaire, par [83] pour le modèle logistique et [68] pour le modèle de Cox. Inspirés

par ces ouvrages autour du problème de la modélisation de durées de survie pour les données en grande dimension, nous sommes motivés pour proposer dans cette thèse, la méthode du group Lasso dans la sélection de variables pour le modèle de Cox avec fragilité pour les données en cluster.

# Plan et contribution de la thèse

Pour pouvoir bien mener notre présent travail de thèse, nous avons organisé notre manuscrit en cinq chapitres sans compter l'introduction générale et la conclusion générale. Le reste de la présente étude est répartie comme suit :

**Chapitre 1 :** Le premier chapitre de la thèse est consacré à une brève présentation de l'historique, des notions et des définitions de concepts relatifs à l'analyse de durées de vie et des principales fonctions assocées aux distributions de survie. Nous avons également présenté de façon détaillée trois approches de la modélisations de durées de vie censurées (approche non paramètrique, approche paramètrique et approche semi-paramètrique).

**Chapitre 2 :** Le deuxième chapitre introduit la notion de la fragilité ainsi que ses différents domaines d'application. Nous avons présenté la méthodologie, les algorithmes et l'illustration de méthodes sur quelques modèles de fragilité, particulièrement les modèles de survie à fragilité Gamma (la fragilité partagée, la fragilitée corrélée ainsi que les fragilités emboitées) sur les différentes bases de données biomédicales à l'aide de différents packages R.

**Chapitre 3 :** Le troisième chapitre présente quelques méthodes de régularisation et de sélection de variables qui ont été proposées ces dernières années pour le modèle de Cox. Il s'agit principalement de la régression Ridge, la régression Lasso (Least Absolute Shrinkage and Selection Operator), la régression adaptive Lasso, la regression Elastic-net, et enfin la régression Scad et Mcp. Ce chapitre aussi, présente l'ensemble des solutions (chemin de régularisation) pour une grille de valeurs des paramètres de lissage pour chaque méthode. Ensuite, nous avons illustré chaque méthode présentée sur une base de données de survie en grande dimension. On a 115 patients avec 549 gènes associés au Cancer du sein au Pays-Bas entre les années 1984 et 1995

**Chapitre 4 :** Le quatrième chapitre présente des résultats sur la méthode du group Lasso dans le modèle de Cox avec fragilité. Dans ce chapitre, nous proposons la méthode group Lasso pour la sélection de variables dans le modèle de Cox avec une fragilité gamma partagée afin d'étendre les connaissances antérieures en tenant compte de l'hétérogénéité entre les groupes d'individus liés, exposés et susceptibles à la survenue d'un événement d'intérêt. La consistance théorique de la méthode proposée est établie. Nous avons illustré la comparaison de méthodes group Lasso, group SCAD et group MCP sur une base de données simulées et de données réelles. Il s'agit des données en grande dimension de l'Institut

National de Cancerologie des Etats Unis sur 470+ patients ayant la lymphome à cellules B diffus (DLBCL) recevant un traitement standard avec rituximab plus cyclophosphamide, doxorubicine, vincristine et prednisolone (R-CHOP). Des résultats obtenus dans ce travail réalisé sous la supérvision de mon directeur de thèse Papa Ngom, ont fait l'objet d'une publication d'un article intitulé : **Variable selection with group LASSO approach: Application to Cox regression with frailty model**. Cet article a été publié en 2019 dans une revue internationale de *Communications in Statistics-Simulation and Computation*[114].

**Chapitre 5 :** Le cinquième chapitre présente des résultats sur la méthode d'Adaptive group Lasso dans le modèle de Cox avec fragilité. C'est une généralisation du chapitre 4. Dans ce chapitre, nous avons tout d'abord présenté quelques méthodes d'estimation pénalisées pour la sélection de variables qui ne sélectionnent que des variables individuelles dans le modèle de Cox. Il s'agit principalement de la méthode Lasso (Least Absolute Shrinkage and Selection Operator), la pénalité Hard-thresholding, la pénalité Smoothly Clipped Absolute Deviation (SCAD) et la pénalité concave minimale (MCP). Ces pénalités nous ont permis d'étudier les mèthodes d'estimation pénalisées prenant en compte la structure des variables groupées, plus particulièrement la méthode d'Adaptive group Lasso qui fournit une sélection parcimonieuse à l'intérieur des groupes. La consistance théorique et la sparsité de la méthode ont été presentés. Nous appliquons l'algorithme " Group Coordinate Descent" pour retrouver l'ensemble des solutions de la méthodes proposée pour une grille de valeurs des paramètres de lissage. Un exemple sur une base de données réelles sur la survie au cancer du sein du Programme de l'Institut national de surveillance, d'épidémiologie et des résultats (SEER) des Etats-Unis a été utilisée pour comparer les performances de la méthode proposée avec des méthodes concurrentes, à savoir la méthode group Lasso, la méthode group SCAD et la méthode group MCP. Des résultats obtenus dans ce travail réalisé sous la supervision de mon directeur de thèse Papa Ngom, ont fait l'objet d'un article intitulé : **Cox survival analysis with Adaptive group Lasso** à paraître dans *Journal des sciences et technologie, 2019*.

**Articles publiés**

1 . Utazirubanda, J. C., M. León, T., & Ngom, P. (2019).
*Variable selection with group LASSO approach: Application to Cox regression with frailty model.*
Communications in Statistics-Simulation and Computation, 1-21.
(indexée et abstractée dans Zentralblatt Math et MathSciNet).

2 . Utazirubanda, J. C., M. León, T., & Ngom, P. (2019).
*Cox survival analysis with Adaptive group Lasso.*
A paraître dans "Journal des sciences et technologie".

CHAPITRE 1

# Analyse de durée de survie: Définitions, Notions et Propriétés


## Résumé

*Ce chapitre a pour but de faire une brève présentation des outils de base sur l'analyse de durée de survie que nous utilisons dans cette thèse. Il explique de façon detaillée l'historique des méthodes d'analyse des données de survie, les principales fonctions associées aux distributions de survie, la notion de censure et les propriétés et enfin les trois approches utilisées pour modéliser les durées de survie en particulier l'approche semi-paramétrique (modèle de Cox).*

# Sommaire

L'analyse des données de survie est l'étude du délai de survenue d'un événement particulier pour un ou plusieurs groupes d'individus. Cet événement est souvent associé à un changement d'état; il peut tout aussi bien représenter la mort d'un individu pour une cause déterminée, que l'appartion d'une maladie, ou bien encore la dispartition des symptômes. L'analyse des données de survie est utilisée dans un contexte d'études longitudinales comme les études de cohorte et les essais thérapeutiques. Une de ses caractéristiques est la difficulté à pouvoir observer complètement tous les temps d'événements. Par exemple, quand l'événement étudié est le décès, la date de cet événement n'est pas observée pour les sujets toujours vivants à la fin de l'étude. Ce type d'observation incomplète se nomme "censure à droite" et sera bien discuté dans cette thèse.

## 1.1 Historique des méthodes d'analyse des données de survie

L'étude des données de survie est restée au XVII ème siècle un problème longtemps étudié par les démographes et les actuaires. Le but des analystes de ce siècle est l'estimation, à partir des carnets de décès, de diverses variétés de la population, son groupe, sa durée, etc... Ces analyses, très générales, ne sont affinées qu' à partir du XIX ème siècle, avec l'apparition de particularité suivant des variables exogènes (sexe, nationalité, catégories socio-professionnelles...). Et c'est durant ce siècle que les premières modélisations concernant la probabilité de mourir à un certain âge, probabilité qui sera par la suite désignée sous le terme de fonction de risque ont fait leur appartion.

L'analyse des données de survie commence à prendre une envergure en s'éloignant petit à petit du cadre strict de la démographie et de l'actuariat pour investir, au XX ème siècle, toutes les disciplines susceptibles d'avoir recours à de tels types de données : l'ingénierie avec l'apparition de la théorie de la fiabilité, la sociologie avec l'analyse de l'histoire d'événements et dans le domaine biomédical où on note deux objectifs principaux de l'analyse de survie:

Lors d'un **essai thérapeutique**, il s'agit de tester l'efficacité d'un nouveau traitement en comparant les durées de survie qu'il permet d'obtenir à celles que donne le traitement habituel (ou un placebo).

Lors d'une **étude épidémiologique**, il s'agit d'évaluer la valeur pronostique d'un ou plusieurs facteurs, soit sur la durée de survie, soit sur le délai de survenue d'une maladie. Peu utilisée par la communauté scientifique en particulier celle des statisticiens jusqu'en $1950$, malgré les conceptions d'un modèle paramétrique dans le domaine de la fiabilité par Weibull en $1951$ et celle d'un modèle non-paramétrique de la fonction de survie en $1958$, où l'on note d'importants résultats de Kaplan-Meier en terme d'estimateur , l'analyse de survie connaît son épilogue en $1972$ grâce aux excellents travaux de Cox publiant un article posant les bases d'un cas particulier de modèle semi-paramétrique faisant intervenir des variables explicatives (exogènes).

## 1.2 Les principales fonctions associées aux distributions de survie

Le terme de durée de survie est employé de manière générale pour désigner le temps qui s'écoule dès le début de l'observation jusqu'à la survenue d'un événement particulier. A l'origine cet événement désignait le " décès " mais d'autres événements peuvent être considérés par exemple la rechute d'une maladie en épidémiologie, la panne d'une machine (durée de fonctionnement d'une machine en fiabilité) ou les durées de contrats en actuariat. Cinq fonctions équivalentes sont associées aux distributions de durée de survie. Supposons que la durée de survie $X$ soit une variable aléatoire positive ou nulle et absolument continue. Alors la loi de probabilité peut être définie par l'une des fonctions suivantes.

1. **La fonction de survie** $S$
   La fonction de survie notée $S$ est définie par :

   $$S(t) = \mathbb{P}\{X \geq t\}, \quad t \geq 0.$$

   Pour $t$ fixé, $S(t)$ est la probabilité de survivre jusqu'a l'instat $t$. C'est donc une fonction continue monotone non croissante telle que

   $$S(0) = 1, \lim_{t\to\infty} S(t) = 0.$$

2. **La fonction de repartition** $F$
   La fonction de repartition notée $F$ est définie par :

   $$F(t) = \mathbb{P}\{X < t\} = 1 - S(t), \quad t \geq 0.$$

   Pour $t$ fixé, c'est la probabilité de mourir avant l'instat $t$.

3. **La densité de probabilité** $f$
   La fonction de densité notée $f$ est positive telle que :

   $$F(t) = \int_0^t f(s)ds.$$

   Si la fonction de repartition a une derivée au point $t$ alors

   $$f(t) = \lim_{dt\to 0} \frac{\mathbb{P}(t \leq X < t + dt)}{dt} = F'(t) = -S'(t).$$

   Pour $t \geq 0$ fixé , la densité de probabilité caractérise la probabilité par unité de temps de mourir dans un petit intervalle de temps après l' instant $t$.

4. Le taux d' incidence ou risque instantané $h$
   Le risque instantané est aussi très souvent appelé " le taux de hasard".

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq X < t + dt \mid X > t)}{dt} = \frac{f(t)}{S(t)} = -\ln\left(S(t)\right)'.$$

   Pour $t$ fixé, $h$ caractérise le risque de mourir dans un petit intervalle de temps après l'instant $t$ conditionnellement au fait d' avoir survécu jusqu'à l'instant $t$. Aussi cela signifie -t-il le risque de mort instantané par ceux qui ont survécu.

5. Le risque cumulé $H$

$$H(t) = \int_0^t h(u)du = -\ln\{S(t)\}$$

$$\Longrightarrow \; S(t) = \exp\left(-H(t)\right) = \exp\left(-\int_0^t h(u)du\right).$$

   Le temps moyen de survie $\mathbb{E}(X)$ ainsi que de variance $\mathbb{V}ar(X)$ sont donnés par :

$$\mathbb{E}(X) = \int_0^\infty S(t)dt, \quad \mathbb{V}ar(X) = 2\int_0^\infty tS(t)dt - \left(\mathbb{E}(X)\right)^2.$$

## 1.3 Notions de Censure, Processus ponctuel et Propriétés

Les méthodes standards d'analyse statistique sont inappropriées pour les données de survie. En effet, ces données présentent plusieurs particularités. Une des particularités de données de survie est que l'on dispose de " données incomplètes " dans le sens où l'événement d'intérêt comme le décès n'est pas forcément observé sur quelques individus durant le temps de suivi. Pour ces individus, le temps de survie est dit "censuré". Plusieurs types de censures existent. La censure à droite est l'exemple le plus connu d'observation incomplète en analyse de survie. La censure à gauche est moins souvent rencontrée dans les études épidémiologiques. La censure par intervalles est peu abordée car même lorsqu'elle est présente dans la réalité, elle est rarement prise en compte, le plus souvent dans un souci de simplicité. Nous nous limiterons, dans le cadre de cette thèse, à la censure dite *censure aléatoire à droite* qui indique que le délai exact de décès de sujets (non observés)

est supérieur ou égal (à droite) à son délai de survie. En fait la durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. Il est par conséquent à risque de subir encore l'événement à la fin du suivi. La censure aléatoire à droite est la plus courante par exemple : la perte de vue d'un patient qui quitte l'étude en cours et on ne le voit plus à cause de diverses raisons par exemple :

- les patients qui sont exclus de l'étude du fait qu' il y ait les effets secondaires ou inefficacité du traitement pouvant entrainer un changement ou un arrêt du traitement.
- les patients dits exclus-vivants de l'étude à cause de la fin de l'étude au moment où ils sont encore vivants (ils n'ont pas subi l'événement).

Dans les modèles classiques d'analyse de survie, on fait l'hypothèse de la *censure indépendante*. La censure est dite indépendante si, sachant qu'une personne est vivante en $t$ et connaissant ses caractéristiques individuelles, le fait de savoir que cette personne n'est pas censurée ne change pas son risque instantané [4].

### 1.3.1 Processus de la censure

**Définition 1.3.1.** *La* **variable de censure** $C$ *est définie par la possible non-observation de l'événement. Si l'on observe* $C$ *et non la durée de survie* $X$*, et que l'on sait que* $X > C$ *(respectivement* $X < C, C_1 < X < C_2$*), on dit qu'il y a* **censure à droite** *( respectivement* **censure à gauche, censure par intervalle**).

Si l'événement se produit, $X$ est "réalisée". S'il ne se produit pas (l'individu étant perdu de vue, ou bien exclu vivant), c'est $C$ qui est " réalisée". $X$ peut être considérée comme la durée séparant un événement initial $A$ d'un événement terminal $B$, ou comme la durée pendant laquelle un sujet reste dans un état donné (auquel cas $A$ désigne **l'entrée** dans cet état et $B$ **la sortie** de cet état). Nous allons donner les caractéristiques de la censure.

1. **- Censure de type I** ( Censure non-aléatoire de type I)
   La censure est dite de type I si étant donné un nombre positif $C$ et un $n$-échantillon $X_1, ...., X_n$, les observations consistent en $(T_i, \delta_i)$, où

$$\begin{cases} T_i = & X_i \wedge C \\ \delta_i = & \mathbb{I}_{\{X_i \leq C\}} \end{cases}$$

C'est le cas, par exemple, si les seuls sujets censurés sont des sujets exclus vivants à une date de point ou à la fin de l'étude, fixée à l'avance. La vraisemblance du modèle associée aux observations $(T_1, \delta_1), ..., (T_n, \delta_n)$ s'écrit :

$$L(\theta) = \prod_{i=1}^{n} f_\theta\big(T_i\big)^{\delta_i} S_\theta\big(C\big)^{1-\delta_i}. \tag{1.1}$$

En d'autres termes lorsqu'on observe la sortie avant la censure ($\delta_i = 1$), c'est $f_\theta\left(T_i\right)$ (utilisée s'il n'y a pas d'embiguité pour désigner la densité de probabilité de la variable aléatoire $T$ prise à la valeur $T_i$ et indexée par la valeur $\theta$ des paramètres) qui intervient dans la vraisemblance et, dans le cas contraire ($\delta = 0$), on retrouve $S_\theta\left(C\right)$ (fonction de survie à la date de censure et indexée par $\theta$).
Et pour démontrer la formule (1.1), il suffit de calculer $\mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = d_i\big)$. Puisque $\delta_i$ ne peut prendre que les valeurs 0 et 1, on calcule sur $[0, C]$ :

$$\begin{aligned}
\mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = 1\big) &= \mathbb{P}\big(X_i \wedge C \in \big[t_i, t_i + dt_i, \big], X_i \leq C\big) \\
&= \mathbb{P}\big(X_i \in [t_i, t_i + dt_i]\big) \\
&= f_\theta\big(t_i\big) dt_i. \qquad (1.2)
\end{aligned}$$

On peut toujours supposer $dt_i$ suffisamment petit pour que $t_i + dt_i \leq C$ et

$$\begin{aligned}
\mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = 0\big) &= \mathbb{P}\big(X_i \wedge C \in \big[t_i, t_i + dt_i, \big], X_i \geq C\big) \\
&= \mathbb{P}\big(X_i \geq C\big) \\
&= S_\theta(C). \qquad (1.3)
\end{aligned}$$

Ces deux cas peuvent se résumer en :

$$\mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = d_i\big) = f_\theta\big(t_i\big)^{\delta_i} S_\theta\big(C\big)^{1-\delta_i}. \tag{1.4}$$

Cette expression peut être également obtenue en observant que :

$$\mathbb{P}\big(T_i > t_i, \delta_i = 1\big) = \mathbb{P}\big(X_i > t_i, X_i \leq C\big) = \int_{t_i}^{C} f_\theta(u) du$$

et dans le cas où $\delta_i = 0$ comme alors $T_i = C$, il n'y a pas de densité, mais simplement la probabilité de cet événement est égale à $S_\theta(C)$. Comme pour une observation censurée, par définition $T_i = C$ et l'expression ci-dessus peut se réécrire :

$$L(\theta) = \prod_{i=1}^{n} S_\theta\big(T_i\big) h_\theta\big(T_i\big)^{\delta_i}. \tag{1.5}$$

Cette expression est donc simplement le produit des valeurs de la fonction de survie (qui traduit le fait que les individus sont observés au moins jusqu'en $T_i$), pondérée pour les sorties non censurées par la valeur de la fonction de hasard (qui traduit le fait que pour ces observations la sortie a effectivement lieu à l'instant $T_i$). On utilise en général la $\log$-vraisemblance égale à une constante additive près à :

$$\ln L\big(\theta\big) \;=\; \sum_{i=1}^{n} \big[\delta_i \ln h_\theta\big(T_i\big) + \ln S_\theta\big(T_i\big)\big]\,. \tag{1.6}$$

Ce mécanisme de censure est fréquemment utilisé dans les applications industrielles. Par exemple on peut tester la durée de vie de $n$ ampoules identiques sur un intervalle d'observation fixé $[0, t]$.

2. **- Censure de type II**
   La censure est dite de type II si étant donné un nombre positif fixé $r$ et un $n$-échantillon $X_1, ...., X_n$, les observations consistent en $(T_i, \delta_i)$, où

$$\begin{cases} T_i = & X_i \wedge X_{(r)} \\ \delta_i = & \mathbb{I}_{\{X_i \leq X_{(r)}\}} \end{cases}$$

   $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ sont les statistiques d'ordre. En biologie par exemple on peut tester l'efficacité d'une molécule sur un lot de souris, la durée de l'étude correspondant au temps que mettent $r$ souris à mourir. Un autre exemple concerne l'observation de la durée de fonctionnement de $n$ machines tant que $r$ d'entre elles ne tombent pas en panne. Autrement dit, la censure est présente quand on décide d'observer les durées de vie de $n$ patients jusqu'à ce que $r$ d'entre eux soient décédés et d'arrêter l'étude à cet instant là.

   Soient $X_{(i)}$ et $T_{(i)}$ les statistiques d'ordre des variables $X_i$ et $T_i$. La date de censure est donc $X_{(r)}$.

$$\begin{aligned} T_{(1)} &= X_{(1)} \\ \vdots &= \vdots \\ T_{(r)} &= X_{(r)} \\ T_{(r+1)} &= X_{(r)} \\ \vdots &= \vdots \\ T_{(n)} &= X_{(r)}. \end{aligned}$$

La vraisemblance a une forme proche du cas de la censure de type I. On remarque que, dans la fonction de survie à la date de censure, il convient de choisir les instants des $r$ sorties parmi les $n$ observations. Cela nous permet d'écrire :

$$\begin{aligned} L(\theta) &= \frac{n\,!}{(n-r)\,!}\left[\prod_{i=1}^{r} f_\theta\left(X_{(i)}\right)\right] S_\theta\big(X_{(r)}\big)^{n-r} \\ &= \frac{n\,!}{(n-r)\,!}\prod_{i=1}^{n} f_\theta\big(T_{(i)}\big)^{\delta_i} S_\theta\big(T_{(i)}\big)^{1-\delta_i} \\ &= \frac{n\,!}{(n-r)\,!}\prod_{i=1}^{n} h_\theta\big(T_{(i)}\big)^{\delta_i} S_\theta\big(T_{(i)}\big). \end{aligned} \tag{1.7}$$

Si la loi de référence est la loi exponentielle, on trouve ainsi que :

$$L(\theta) = \frac{n\,!}{(n-r)\,!}\theta^r \exp\big(-\theta T\big),$$

avec $T = \sum_{i=1}^{r} T_{(i)} + \big(n-r\big)T_{(r)}$; la statistique est donc exhaustive pour le modèle.

3. **- Censure de type III** ( Censure aléatoire de type I)
Une censure est dite **aléatoire de type I** si étant donné une $n$-observation $X_1, ..., X_n$, il existe une v.a $n$-dimensionnelle $(C_1, ..., C_n)$ de $(\mathbb{R}^+)^n$ telle que les observations consistent en $(T_i, \delta_i)$, où

$$\begin{cases} T_i = & X_i \wedge C_i \\ \delta_i = & \mathbb{I}_{\{X_i \leq C_i\}} \end{cases}$$

La vraisemblance de l'échantillon $(T_1, \delta_1), ..., (T_n, \delta_n)$ s'écrit avec des notations évidentes:

$$L(\theta) = \prod_{i=1}^{n}\Big[f_X(T_i,\theta)S_C(T_i,\theta)\Big]^{\delta_i}\Big[f_C(T_i,\theta)S_X(T_i,\theta)\Big]^{1-\delta_i}.$$

La forme de la vraisemblance ci-dessus se déduit par exemple du fait que $T_1, ..., T_n$ est un échantillon de la loi $S_T(\theta, .)$ avec :

$$S_T(\theta,t) = \mathbb{P}_\theta(T_i > t) = \mathbb{P}_\theta\big(X_i \wedge C_i > t\big) = \mathbb{P}_\theta(X_i > t)\mathbb{P}_\theta(C_i > t) = S_X(t,\theta)S_C(t,\theta).$$

On écrit comme dans l'équation 1.2 et l'équation 1.3 que :

$$\mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = 1\big) \;=\; \mathbb{P}\big(X_i \wedge C_i \in \big[t_i, t_i + dt_i, \big], X_i \leq C_i\big)$$

$$\begin{aligned} &= \mathbb{P}\big(X_i \in [t_i, t_i + dt_i], t_i < C_i\big) \\ &= S_X(\theta, t_i) f_C(\theta, t_i) dt_i. \end{aligned}$$

et

$$\begin{aligned} \mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = 0\big) &= \mathbb{P}\big(X_i \wedge C_i \in \big[t_i, t_i + dt_i, \big], X_i \geq C_i\big) \\ &= \mathbb{P}\big(C_i \in [t_i, t_i + dt_i], X_i > t_i\big) \\ &= f_X(\theta, t_i) S_C(\theta, t_i) dt_i. \end{aligned}$$

On peut obtenir ces expressions directement à partir de celles obtenues dans la sous-section 1.3.1 en conditionnant par rapport à la censure, puis en intégrant par rapport à la loi de celle-ci. Plus précisement on écrit :

$$\begin{aligned} \mathbb{P}\big(T_i > t_i, \delta_i = 1\big) &= \mathbb{P}\big(X_i \wedge C_i > t_i, X_i \leq C_i\big) \\ &= \mathbb{P}\big(t_i < X_i \leq C_i\big) \\ &= \int_{t_i}^{+\infty} \mathbb{P}\big(t_i < X_i \leq c\big) f_C(\theta, c) dc \\ &= \int_{t_i}^{+\infty} \Big( \int_{t_i}^{c} f_X(\theta, x) dx \Big) f_C(\theta, c) dc, \end{aligned}$$

puis par Fubini on inverse les intégrales pour obtenir :

$$\begin{aligned} \mathbb{P}\big(T_i > t_i, \delta_i = 1\big) &= \int_{t_i}^{+\infty} f_X(\theta, x) \Big( \int_{x}^{+\infty} f_C(\theta, c) dc \Big) dx \\ &= \int_{t_i}^{+\infty} f_X(\theta, x) S_C\big(\theta, x\big) dx \end{aligned}$$

et finalement

$$\mathbb{P}\big(T_i \in [t_i, t_i + dt_i], \delta_i = 1\big) = -\frac{d}{dt_i} \mathbb{P}\big(T_i > t_i, \delta_i = 1\big) = f_X\big(\theta, t_i\big) S_C\big(\theta, t_i\big) dt_i.$$

On fait alors l'hypothèse que la censure est non-informative, c'est à dire que la loi de censure est indépendante du paramètre $\theta$. La vraisemblance se met dans ce cas sous la forme

$$L(\theta) = \prod_{i=1}^{n} f_\theta\big(T_i\big)^{\delta_i} S_\theta\big(T_i\big)^{1-\delta_i}.$$

Cette dernière expression peut s'écrire comme en (1.1) ci-dessus :

$$L(\theta) = \prod_{i=1}^{n} S_\theta(T_i) h_\theta(T_i)^{\delta_i}. \tag{1.8}$$

Enfin le mécanisme de censure est habituellement supposé être indépendant de l'événement étudié : on parle de censure non-informative. On constate ici que la censure fixe est un cas particulier de censure aléatoire non-informative dans laquelle la loi de la censure est une loi de Dirac au point $C$. L'expression établie dans le cas de la censure fixe se généralise donc facilement. Si la censure est informative, alors l'expression classique de la vraisemblance ne correspond plus à une vraisemblance complète, mais à une vraisemblance partielle qui peut être utilisée pour des inférences, bien qu'il y ait une perte d'efficacité des estimateurs produits (car toute l'information n'est pas utilisée). Ainsi, la censure informative est à l'origine d'un biais lors de l'analyse standard basée sur la vraisemblance [101, 66].

### 1.3.2 Processus ponctuel $N_i(t)$

L'étude des durées de survie peut être abordée d'une autre façon. Au lieu de considérer $T$ la durée étudiée, qui est une variable aléatoire réelle positive, généralement continue, de densité $f$, de fonction de répartition $F$ et de fonction de survie $S = 1 - F$, on représente l'expérience par le processus ponctuel associé $N_i(t)$. Notons que certaines notions utilisées dans cette section sont définies en annexe A de cette thèse.

**Définition 1.3.2.** - *Une observation consiste en* $(T_i, \delta_i)$*, où*

$$\begin{cases} T_i = & X_i \wedge C_i \\ \delta_i = & \mathbb{I}_{\{X_i \leq C_i\}} \end{cases}$$

Un processus ponctuel est défini par :

$$N_i(t) = \mathbb{I}_{\{T_i \leq t,\ \delta_i = 1\}},\ t \geq 0.$$

$N_i(t)$ est un processus non décroissant, tel que $N_i(0) = 0$ et qui fait des sauts de 1. Si l'individu $i$ subit l'événement avant $t$, alors $N_i(t) = 1$; sinon, $N_i(t) = 0$. Nous définissons également la fonction

$$Y_i(t) = \mathbb{I}_{\{T_i \geq t\}}$$

qui est l'indicateur de risque pour le sujet $i$ ( n'a pas encore subi l'événement ni censuré) juste avant l'instant $t$.

**Définition 1.3.3.** *Les processus*

$$\lambda_i(t) = h_i(t)Y_i(t)$$

*et*

$$\Lambda_i(t) = \int_0^t h_i(s)Y_i(s)ds$$

*sont appelés respectivement* **processus d'intensité** *et* **processus d'intensité cumulée** *de* $N_i(t)$. $(cf.\ annexe\ A.2)$. $\Lambda_i(t)$ *est également appelé* **compensateur** *du processus* $N_i(t)$. *Notons que* $\lambda_i(t)$ *est une variable aléatoire, contrairement à* $h_i(t)$ *qui est fixe.*

**Proposition 1.3.4.** *Le processus stochastique défini par :*

$$M_i(t) = N_i(t) - \int_0^t h_i(s)Y_i(s)$$

*est une martingale.*

***Preuve.***

$$\begin{aligned} \mathbb{E}\left[dM_i(t)|\mathcal{F}_{t-}\right)] &= \mathbb{E}\left[dN_i(t) - h_i(t)Y_i(t)dt|\mathcal{F}_{t-}\right] \\ &= \mathbb{E}\left[dN_i(t)|\mathcal{F}_{t-}\right] - h_i(t)Y_i(t)dt \\ &= \lambda_i(t)dt - \lambda_i(t)dt \\ &= 0 \quad (cf.\ annexe\ A.2). \end{aligned} \tag{1.9}$$

□

**Modélisation**

Concernant les modèles statistiques proprement dits, trois approches se retrouvent dans la description des données de survie : l'approche non-paramétrique, paramétrique et semi-paramétrique.

## 1.4 L'approche non-paramétrique

L'approche non-paramétrique est utilisée lorsqu'aucune hypothèse n'est faite sur la distribution des temps de survie. Il s'agit dès lors d'un problème d'estimation fonctionnelle, avec les équivoques que cela implique par exemple, la fonction de survie sera estimée par une fonction discontinue, sachant qu'elle est continue.

L'inconvénient d'une telle approche est la nécessité de disposer d'un nombre important d'observations, le problème de l'estimation d'un paramètre fonctionnel étant délicat puisqu'il est non décrit par un nombre fini de paramètres, c'est à dire il appartient à un espace de dimension infinie.

## 1.4.1 Estimation du risque cumulé

L'une des fonctions qui caractérisent la distribution des temps d'événements est la fonction de risque cumulé. Nous traiterons donc de l'estimation de la fonction de risque cumulé, avec l'estimateur de Nelson-Aalen.

### 1.4.1.1 Estimateur de Nelson-Aalen $\widehat{A}(t)$

Lorsque $X$ admet une densité $f$, on a défini la fonction du risque cumulé noté dans cette section par

$$A(t) = \int_0^t h(u)du = \int_0^t \frac{f(u)}{S(u)}du.$$

Dans le cas où la distribution de $X$ n'admet pas de dérivée en tout point de $\mathbb{R}^+$, on peut toujours écrire le risque cumulé comme

$$A(t) = -\int_0^t \frac{S(du)}{S(u^-)}.$$

Dans le cadre de données censurées aléatoirement à droite, on remarque que

$$A(t) = -\int_0^t \frac{H_1(du)}{H(u^-)}, \tag{1.10}$$

où $H(t) = \mathbb{P}(T > t)$, $H_1(t) = \mathbb{P}(T > t, \delta = 1)$. On note $G(t)$ la fonction de survie de la variable de censure $C$. D'après l'hypothèse d'indépendance entre $X$ et $C$, on obtient

$$H(t) = \mathbb{P}(T > t) = \mathbb{P}(X > t, C > t) = S(t)G(t)$$

et

$$H_1(t) = \int_t^{+\infty} G(u^-)f(u)du = -\int_t^{+\infty} G(u^-)S(du).$$

Par conséquent,

$$H(u^-) = G(u^-)S(u^-)$$

et

$$H_1(dt) = G(t^-)S(dt).$$

En remplaçant les fonctions $H$ et $H_1$ dans (1.10) par leurs équivalents empiriques définis par

$$\widehat{H}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_{(i)}>u\}} \quad \text{et} \quad \widehat{H}_1(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_{(i)}>u,\delta_i=1\}},$$

on obtient alors, pour tout $t$ tel que $\mathbb{P}(T > t) > 0$, un estimateur "naturel" de la fonction de risque cumulé de Nelson Aalen :

$$\widehat{A}(t) = -\int_0^t \frac{\widehat{H}_1(du)}{\widehat{H}(u^-)} = \sum_{i,T_i\leq t} \frac{\sum_{j=1}^{n} \mathbb{I}_{\{T_j=T_i,\delta_j=1\}}}{\sum_{j=1}^{n} \mathbb{I}_{\{T_j\geq T_i\}}} = \sum_{i,T_i\leq t} \frac{d_i}{Y_i}.$$

— $Y_i$ est le nombre d'individus à risque juste avant $T_i$.

— $d_i$ est le nombre de décès en $T_i$.

L'estimateur de Nelson-Aalen est une fonction en escalier ayant un saut de taille $\frac{d_i}{Y_i}$ à chaque instant de décès.

**Exemple 1.4.1.** *On observe les durées de vie de 10 diodes exprimées en mois* $1, 2, 4^+, 5, 7^+, 8, 9, 10^+, 11, 13^+$ *où* $(+)$ *désigne les données censurées.*
*Le calcul de l'estimateur de Nelson-Aalen donne :*

$$\begin{array}{lllllll}
\forall & 0 \leq t < 1 & \widehat{A}(t) & = & 0 & & \\
\forall & 1 \leq t < 2 & \widehat{A}(t) & = & \frac{1}{10} & +\widehat{A}(0) & = & 0.1 \\
\forall & 2 \leq t < 5 & \widehat{A}(t) & = & \frac{1}{9} & +\widehat{A}(1) & = & 0.21 \\
\forall & 5 \leq t < 8 & \widehat{A}(t) & = & \frac{1}{7} & +\widehat{A}(3) & = & 0.35 \\
\forall & 8 \leq t < 9 & \widehat{A}(t) & = & \frac{1}{5} & +\widehat{A}(5) & = & 0.55 \\
\forall & 9 \leq t < 11 & \widehat{A}(t) & = & \frac{1}{4} & +\widehat{A}(8) & = & 0.80 \\
\forall & t \geq 11 & \widehat{A}(t) & = & \frac{1}{2} & +\widehat{A}(9) & = & 1.3
\end{array}$$

#### 1.4.1.2 Estimation de la variance de $\widehat{A}(t)$

En utilisant le processus de comptage et en faisant une approximation de l'accroissement du processus de comptage par une loi de poisson, [1] a montré que la variance de l'estimateur de Nelson-Aalen est donné par

$$\widehat{\mathbb{V}}ar(\widehat{A}(t)) = \sum_{i,T_i\leq t}^{n} \frac{d_i}{Y_i^2},$$

où $d_i$ et $Y_i$ sont respectivement le nombre de décès et d'individus à risque en $T_i$.

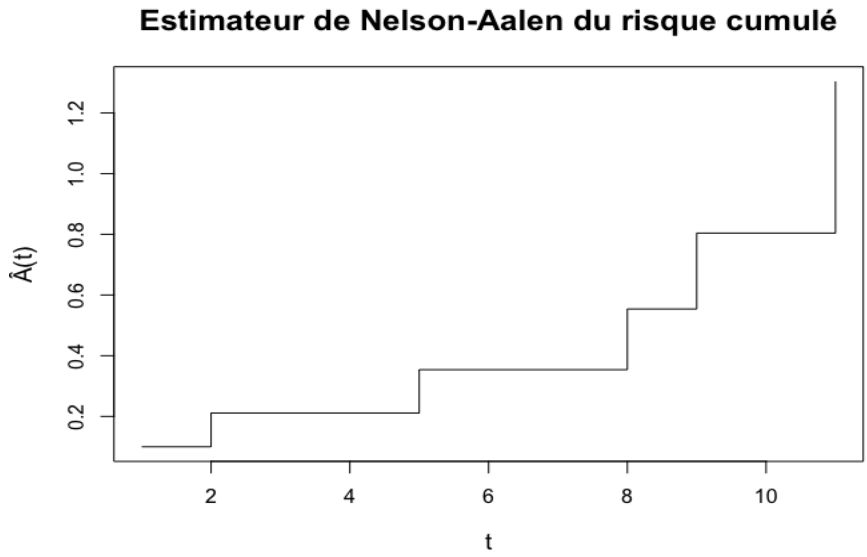



**Figure** 1.1 – Estimateur de Nelson-Aalen sur les durées de vie de 10 diodes exprimées en mois.

#### 1.4.1.3 Loi asymptotique

D'après la décomposition de Doob-Meyer (cf. annexe A.2)

$$M(t) = N(t) - \Lambda(t) = N(t) - \int_0^t h(s)Y(s)ds$$

est une martingale locale de carré intégrable. Heuristiquement, on a

$$dN(t) = dM(t) + h(t)Y(t)dt,$$

où $dM(t)$ peut être considéré comme un "un bruit aléatoire". On déduit que $\frac{dN(t)}{Y(t)} \simeq h(t)dt$ et que

$$\int_0^t \frac{dN(t)}{Y(t)} \simeq \int_0^t h(t)dt.$$

On cherche à estimer le risque cumulé

$$A(t) = \int_0^t h(t)dt.$$

On définit les notations suivantes :

$$N_+(t) = \sum_{i=1}^{n} N_i(t)$$

et

$$Y_+(t) = \sum_{i=1}^{n} Y_i(t).$$

comme étant, respectivement, le nombre total de survenues de l'événement à l'instant $t$ et le nombre total d'individus encore à risque juste avant l'instant $t$.

**Définition 1.4.2.** *L'estimateur de Nelson-Aalen de la fonction de risque cumulé [86, 1] est défini par*

$$\hat{A}(t) = \int_0^t \frac{J(u)}{Y_+(u)} dN_+(u), \tag{1.11}$$

*où* $J(t) = \mathbb{I}_{\{Y_+(t)>0\}}$. *Par convention,* $\dfrac{J(t)}{Y_+(t)} = 0$ *quand* $Y_+(t) = 0$.

**Théorème 1.4.3.** *(cf. section IV.1, pg.193 de l'ouvrage d'[4]).*
$\widehat{A}(t)$ *est un estimateur biaisé de* $A(t)$ *et sous l'hypothèse que sa fonction de répartition* $F(t) < 1$ *(c-à-d. que* $A(t) < \infty$*), nous avons*

$$\sqrt{n}\big(\widehat{A}(t) - A(t)\big) \xrightarrow{\mathcal{L}} U(t) \qquad (n \to +\infty),$$

*avec* $U$ *une martingale gaussienne telle que*

$$\begin{cases} U(0) = & 0 \\ \mathbb{V}ar(U(t)) = & \displaystyle\int_0^t \frac{h(u)}{y(u)} du, \end{cases}$$

*où* $y(s) = \big[1 - F(s)\big]\big[1 - G(s^-)\big]$ *avec* $G$ *la fonction de répartition de censure C.*

***Preuve.*** Pour démontrer le Théorème 1.4.3, on va d'abord calculer le biais et ensuite, on examine le comportement asymptotique de l'estimateur.
Pour le **biais**, calculons $\mathbb{E}\big[\widehat{A}(t)\big]$.

$$\begin{aligned}
\mathbb{E}\big[\widehat{A}(t)\big] &= \mathbb{E}\Big[\int_0^t \frac{J(u)}{Y_+(u)} dN_+(u)\Big] \\
&= \mathbb{E}\Big[\int_0^t \frac{J(u)}{Y_+(u)} dM_+(u) + Y_+(u)h(u)\Big] \\
&= \mathbb{E}\Big[\int_0^t \frac{J(u)}{Y_+(u)} dM_+(u)\Big] + \mathbb{E}\Big[\int_0^t J(u)h(u)du\Big] \\
&= 0 + \int_0^t \mathbb{E}[J(u)]h(u)du \quad (\text{car } M_+(t) \text{ est une martingale}) \\
&= \int_0^t \mathbb{P}\big(Y_+(u) > 0\big)h(u)du
\end{aligned}$$

$$\begin{aligned} &= \int_0^t h(u)du - \int_0^t \mathbb{P}\big(Y_+(u)=0\big)h(u)du \\ &= A(t) - \int_0^t \mathbb{P}(Y_+(u)=0)h(u)du. \end{aligned}$$

**Remarque 1.4.4.** *Le biais de l'estimateur de Nelson-Aalen est extrêmement faible car en pratique la probabilité qu'à un instant $t$ tous les individus aient, soit subi l'événement, soit été censuré est proche de zéro.*

Pour expliciter le caractère **asymptotique** de l'estimateur, nous allons exprimer

$$\begin{aligned} N^{(n)}(t) &= \sum_{i=1}^n N_i(t), \\ Y^{(n)}(t) &= \sum_{i=1}^n Y_i(t), \\ \text{et} \qquad J^{(n)}(t) &= \mathbb{I}_{\{Y^{(n)}(t)>0\}} \end{aligned}$$

Notons aussi $F$ la fonction de répartition des $X_i$ et $G$ celle des $C_i$ $(i \in 1, ..., n)$. Nous obtenons que la fonction de répartition des $T_i$ qui est $[1-(1-F)][1-G]$. D'après le théorème de Glivenko-Cantelli (cf. annexe A.4.4),

$$\sup_{s\in[0,1]} \left| \frac{Y^{(n)}}{n} - [1-F(s)][1-G(s^-)] \right| \xrightarrow{\mathbb{P}} 0 \qquad (n \to +\infty).$$

Par ailleurs,

$$J^{(n)}(t) = \mathbb{I}_{\{Y^{(n)}(t)>0\}}.$$

Nous en déduisons que

$$\begin{aligned} 1 - J^{(n)}(t) &= \mathbb{I}_{\{Y^{(n)}(t)=0\}} \\ &= \mathbb{I}_{\{\mathcal{B}(n,[1-F(t)][1-G(t^-)]=0\}} \xrightarrow{\mathbb{P}} 0 \qquad (n \to +\infty), \end{aligned}$$

et, par suite, que

$$J^{(n)} \xrightarrow{\mathbb{P}} 1 \qquad (n \to +\infty).$$

Par conséquent,

$$\Big\langle \sqrt{n}\big[\widehat{A}^{(n)} - A\big](t) \Big\rangle = \int_0^t n \frac{J^{(n)}(u)}{Y^{(n)}(u)} h(u)du \xrightarrow{\mathbb{P}} \int_0^t \frac{h(s)}{[1-F(s)][1-G(s^-)]}$$

qui est déterministe. Le théorème de Rebolledo donne le résultat. (cf. annexe A.3.1). □

## 1.4.2 Estimation de la fonction de survie.

### 1.4.2.1 Estimateur de Kaplan-Meier $\widehat{S}(t)$

L'estimateur de la fonction de survie le plus utilisé dans l'approche non-paramétrique est celui de Kaplan-Meier. Il découle de l'idée intuitive suivante : survivre après un temps $t$, c'est être en vie juste avant $t$ et ne pas mourir au temps $t$. Il est aussi appelé P.L (Produit-Limite) car il s'obtient comme la limite d'un produit. En considérant les temps d'événements (censure et décès) distincts $T_{(i)}$ $(i = 1, ..., n)$ rangés par ordre croissant, on obtient

$$\mathbb{P}(X > T_{(j)}) = \prod_{k=1}^{j} \mathbb{P}(X > T_{(k)} | X > T_{(k-1)})$$

avec $T_{(0)} = 0$.
Considérons les notations suivantes:
-$Y_i$ le nombre d'individus à risques subissant l'événement juste avant le temps $T_{(i)}$
-$d_i$ le nombre de décès en $T_{(i)}$.
Alors la probabilité $p_i$ de mourir dans l'intervalle $]T_{(i-1)}, T_{(i)}]$ sachant que l'on était vivant en $T_{(i-1)}$ , i.e $p_i = \mathbb{P}(X \leq T_{(i)} | X > T_{(i-1)})$ peut être estimée par

$$\widehat{p_i} = \frac{d_i}{Y_i}.$$

Comme les temps d'événement sont supposés distincts, on a :
$d_i = 0$ en cas de censure en $T_{(i)}$, i.e quand $\delta_i = 0$.
$d_i = 1$ en cas de décès en $T_{(i)}$, i.e quand $\delta_i = 1$.
On obtient alors l'estimateur de Kaplan-Meier :

$$\widehat{S}(t) = \prod_{i=1,..,n;T_i \leq t} \Big(1 - \frac{\delta_i}{Y_i}\Big) = \prod_{i=1,..,n;T_i \leq t}^{n} \Big(1 - \frac{\delta_i}{n-(i-1)}\Big) = \prod_{i=1,..,n;T_i \leq t} \Big(\frac{n-i}{n-i+1}\Big)^{\delta_i}.$$

Par la suite on montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance non paramétrique. $\widehat{S}(t)$ est une fonction en escalier décroissante, constante entre deux temps d'événements consécutifs, continue à droite avec un saut à chaque temps d'événement observé.

**Exemple 1.4.5.** *Freireich, a fait en* $1963$*, un essai thérapeutique pour comparer les durées de rémission des patients atteints de leucémie selon qu'ils ont reçu ou non un médicament appelé*
$6$*-Mercaptopurnie ou un placebo. Les resultats obtenus sont les suivants :*

$6-MP : 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+$

*Les nombres suivis du signe + correspondent à des données censurées.*
*Calculons l'estimateur de Kaplan-Meier sur les données de Freireich en* 1963.

$$\begin{array}{lcll}
\widehat{S}(t) = & 1 & si & 0 \leq t < 6 \\
\widehat{S}(t) = & \big(1 - 3/21\big)\widehat{S}(6-) & = 0.857\ si & 6 \leq t < 7 \\
\widehat{S}(t) = & \big(1 - 1/17\big)\widehat{S}(7-) & = 0.807\ si & 7 \leq t < 10 \\
\widehat{S}(t) = & \big(1 - 1/15\big)\widehat{S}(10-) & = 0.753\ si & 10 \leq t < 13 \\
\widehat{S}(t) = & \big(1 - 1/12\big)\widehat{S}(13-) & = 0.690\ si & 13 \leq t < 16 \\
\widehat{S}(t) = & \big(1 - 1/11\big)\widehat{S}(16-) & = 0.627\ si & 16 \leq t < 22 \\
\widehat{S}(t) = & \big(1 - 1/7\big)\widehat{S}(22-) & = 0.538\ si & 22 \leq t < 23 \\
\widehat{S}(t) = & \big(1 - 1/6\big)\widehat{S}(23-) & = 0.448\ si & 23 \leq t
\end{array}$$

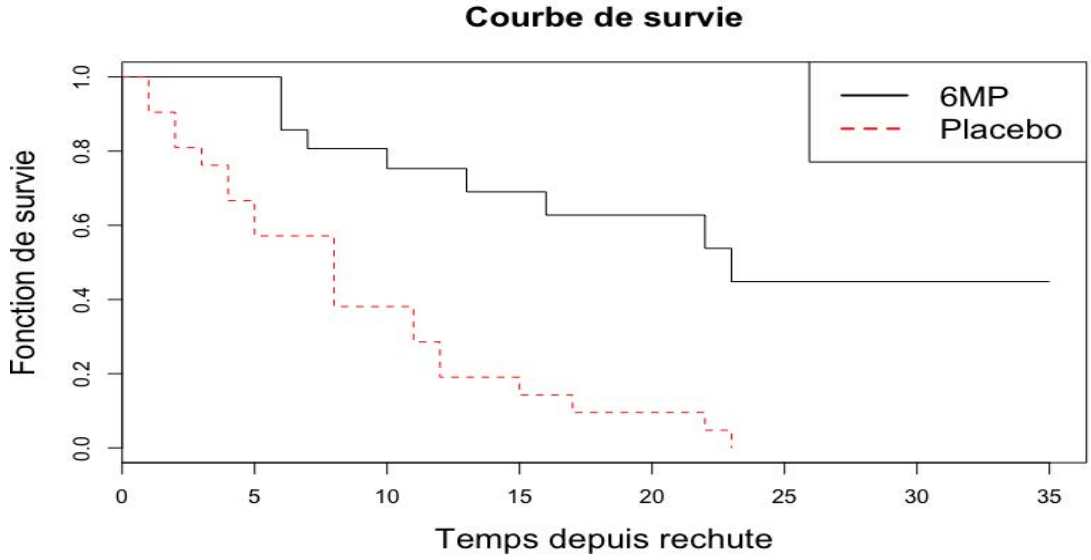



**Figure** 1.2 – Estimateur de Kaplan-Meier sur les données de Freireich en 1963.

L'estimateur de Kaplan-Meier peut être également obtenu dans le cas des données tronquées mais pas dans le cas des données censurées par intervalles car on ignore les temps de décès.

**Remarque 1.4.6.** *Dans le cas où il y a des ex-aequo :*

- *si ce sont des événements de nature différente, on considère que les observations non censurées ont lieu avant les observations censurées.*
- *s'il y a plusieurs décès au même temps* $T_{(i)}$ *alors* $d_i > 1$ *et on a*

$$\widehat{S}(t) = \prod_{i=1,..,n;T_i \leq t} \Big(1 - \widehat{p}_i\Big) = \prod_{i=1,..,n;T_i \leq t} \Big(1 - \frac{d_i}{Y_i}\Big),\ \widehat{p}_i = \frac{d_i}{Y_i}.$$

**Remarque 1.4.7.** *Estimation empirique : Pour un échantillon i.i.d de durées non censurées* $(X_i)_{i=1,\dots,n}$*, un estimateur "naturel" de la survie de la variable* $X$ *est la survie empirique*

$$\hat{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{X_i>t\}}.$$

*Cet estimateur a de bonnes propriétés en terme de convergence : convergence p.s (Glivenko-Cantelli), convergence en loi du processus empirique associé vers un pont brownien. Néamoins, dans le cas des données consurées, la variable d'intérêt n'est plus la variable observée. Ainsi estimer la survie* $S$ *par la survie empirique des données observées* $(T_i)\, i = 1, \dots, n$

$$\hat{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_i>t\}}$$

*fournit un* **estimateur biaisé** *de* $S$ *: il y a une sous estimation de* $S$ *puisque les censures sont considérées comme des décès.*

Notons qu'en l'absence de censure, l'estimateur de Kaplan-Meier se réduit à la fonction de survie empirique.

#### 1.4.2.2 Estimation de la variance de $\widehat{S}(t)$

**Propriété 1.4.8.** *Pour* $t \in [0, \tau[$ *on a asymptotiquement* $\widehat{S}(t) \sim \mathcal{N}\left(S(t), \widehat{\mathbb{V}}ar\left(\widehat{S}(t)\right)\right)$ *où* $\widehat{\mathbb{V}}ar\left(\widehat{S}(t)\right)$ *est la variance de* $\widehat{S}(t)$ *estimée par la formule de [47] :*

$$\widehat{\mathbb{V}}ar\left(\widehat{S}(t)\right) = \widehat{S}(t)^2 \sum_{i:T_{(i)}\leq t} \frac{d_i}{Y_i\left(Y_i - d_i\right)},$$

*où* $Y_i$ *est le nombre d'individus à risques subissant l'événement juste avant le temps* $T_{(i)}$ *et* $d_i$ *est le nombre de décès en* $T_{(i)}$*.*

En conséquence, on peut utiliser la formule suivante pour calculer un intervalle de confiance de $S(t)$ au niveau $1-\alpha$ :

$$IC_{1-\alpha}\left(S(t)\right) = \left[\widehat{S}(t) \pm Z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathbb{V}}ar\big(\widehat{S}(t)\big)}\right].$$

Cet intervalle n'est pas utilisable quand $\widehat{S}(t)$ est proche de $0$ ou de $1$. En effet les bornes peuvent dépasser $0$ ou $1$, l'intervalle étant symétrique autour de $\widehat{S}(t)$.

De ce fait, il est préférable d'utiliser l'intervalle de confiance de Rothman pour contourner cette difficulté [98] :

$$IC_{1-\alpha}\left(S(t)\right) = \frac{K}{K + \left(Z_{\frac{\alpha}{2}}\right)^2}\left[\widehat{S}(t) + \frac{\left(Z_{\frac{\alpha}{2}}\right)^2}{2K} \pm Z_{\frac{\alpha}{2}}\sqrt{\widehat{\mathbb{V}}ar\big(\widehat{S}(t)\big) + \frac{\left(Z_{\frac{\alpha}{2}}\right)^2}{4K^2}}\right]$$

avec $K = \frac{\widehat{S}(t)(1-\widehat{S}(t))}{\widehat{\mathbb{V}}ar\left(\widehat{S}(t)\right)}$.

#### 1.4.2.3 Lois asymptotique

**Théorème 1.4.9.** *Soit $t \in [0, \tau)$, nous suppons qu'il exite une fonction positive $y$ telle que $h/y$ est integrable sur $[0, t]$. Soit*

$$\sigma^2(s) = \int_0^s \frac{h(u)}{y(u)}du.$$

*L'estimateur de Kaplan-Meier $\hat{S}(t)$ est biaisé; de plus, si nous supposons que*

*A- pour tout $s \in [0, t]$,*

$$n\int_0^s \frac{J(u)}{Y(u)}h(u)du \xrightarrow{\mathbb{P}} \sigma^2(s) \qquad (n \to +\infty),$$

*B- pour tout $\varepsilon > 0$,*

$$n\int_0^t \frac{J(s)}{Y(s)}h(s)\mathbb{I}_{\{\sqrt{n}|J(s)/Y(s)|>\varepsilon\}}ds \xrightarrow{\mathbb{P}} 0 \qquad (n \to +\infty),$$

*C-*

$$\sqrt{n}\int_0^t \big(1 - J(u)\big)h(u)du \xrightarrow{\mathbb{P}} 0 \qquad (n \to +\infty),$$

*alors cet estimateur vérifie asymptotiquement*

$$\sqrt{n}\left(\widehat{S}(t) - S(t)\right) \xrightarrow{\mathcal{L}} -U(t)S(t) \qquad (n \to +\infty),$$

*où $U$ est la martingale gaussienne définie dans la sous-section 1.4.1.3.*

***Preuve.*** On montre d'abord que $\hat{S}(t)$ est **biaisé**. La fonction de survie est définie par

$$S(t) = \exp\left(-A(t)\right).$$

Posons

$$S^*(t) = \exp(-A^*(t))$$

où $A^*(t) = \int_0^t J(u)h(u)du$. En utilisant (1.11), on a

$$\hat{A}(t) - A^*(t) = \int_0^t \frac{J(u)}{Y_+(u)} dM_+(u). \tag{1.12}$$

Nous avons, par la formule de Duhamel (cf. annexe A.4.3) et (1.12)

$$\begin{aligned} \frac{\widehat{S}(t)}{S^*(t)} - 1 &= -\int_0^t \frac{\widehat{S}(s^-)}{S^*(t)} d\big(\widehat{A} - A^*\big)(s) \\ &= -\int_0^t \frac{\widehat{S}(s^-)J(s)}{S^*(t)Y_+(t)} dM_+(s) \end{aligned} \tag{1.13}$$

pour $t \in [0, \tau]$ et avec $dM_+(t) = dN_+(t) - Y_+(t)h(t)dt$. Puisque $S^*(t) \geq S(t)$, l'intégrande en (1.13) est borné par $1/S(t)$. Par suite, $\widehat{S}/S^* - 1$ étant une martingale localement de carré intégrable sur $[0, \tau]$ (Théorème $II.3.1$ d'[4]), nous avons

$$\mathbb{E}\Big[\frac{\widehat{S}(t)}{S^*(t)}\Big] = 1$$

pout tout $t \in [0, \tau]$.
Enfin,

$$\begin{aligned} & S(t) && \leq S^*(t) \ (puisque\ A(t) \geq A^*(t)) \\ \Longrightarrow \quad & \tfrac{\widehat{S}(t)}{S^*(t)} && \leq \frac{\widehat{S}(t)}{S(t)} \\ \Longrightarrow \quad & \mathbb{E}\left[\tfrac{\widehat{S}(t)}{S^*(t)}\right] && \leq \mathbb{E}\left[\frac{\widehat{S}(t)}{S(t)}\right] \\ \Longleftrightarrow \quad & 1 && \leq \frac{\mathbb{E}\left(\widehat{S}(t)\right)}{S(t)} \\ \Longleftrightarrow \quad & \mathbb{E}\left(\widehat{S}(t)\right) && \geq S(t). \end{aligned}$$

Pour la **convergence asymptotique** on a, d'après (1.13)

$$\sqrt{n}\left(\frac{\widehat{S}(t)}{S^*(t)} - 1\right) = -\sqrt{n}\int_0^t \frac{\widehat{S}(s^-)}{S^*(t)Y_+(t)} dM(s).$$

D'après les condition $A$ et $B$, et d'après le fait que $\widehat{S}(s)/S^*(s) \leq \frac{1}{S(t)}$ pour tout $s \in [0, t]$, nous déduisons du théorème de Rebolledo (annexe A.3.1) que

$$\sqrt{n}\Big(\frac{\widehat{S}(t)}{S^*(t)} - 1\Big) \xrightarrow{\mathcal{L}} -U(t) \qquad (n \to +\infty).$$

La condition $C$ et le fait que

$$\sqrt{n}\Big|\frac{\widehat{S}(t)}{S^*(t)}-1\Big| = \int_0^s \frac{S(u)}{S^*(u)} d\big(\widehat{A}-A^*\big)(u) \leq \frac{1}{S(t)}\big(1-J(u)\big)h(u)du$$

entrainent que

$$\sup_{s\in[0,t]} \sqrt{n}\Big|\frac{\widehat{S}(t)}{S^*(t)}-1\Big| \xrightarrow{\mathbb{P}} 0 \qquad (n\to+\infty).$$

D'où il s'ensuit que

$$\sqrt{n}\frac{\widehat{S}(s)-S(s)}{S^*(s)} \xrightarrow{\mathcal{L}} -U \qquad (n\to+\infty)$$

et nous obtenons le résultat du théorème. □

Pour les détails de cette démonstration, nous renvoyons le lecteur à l'ouvrage de réference ([4],pg.263).

## 1.4.3 Autres estimateurs

### 1.4.3.1 Estimateur de Breslow du risque cumulé

Les fonctions du risque cumulé $H(t)$ et de survie $S(t)$ sont liées par

$$H(t) = -\log S(t).$$

L'estimateur de Kaplan-Meier $\widehat{S}(t)$ de la fonction de survie induit donc un estimateur non-paramétrique du risque cumulé appelé l'estimateur de [17] :

$$\begin{aligned}\widehat{H}_{Br}(t) &= -\log\Big(\widehat{S}(t)\Big)\\ &= -\sum_{i,T_i\leq t}^{n} \log\left(1-\frac{d_i}{Y_i}\right).\end{aligned}$$

L'estimateur de sa variance est donnée par

$$\widehat{\mathbb{V}}ar(\widehat{H}_{Br}(t)) = \sum_{i,T_i\leq t}^{n} \frac{d_i}{Y_i(Y_i-d_i)}.$$

#### 1.4.3.2 Estimateur de Harrington et Fleming de la survie

Les fonctions de survie $S(t)$ et du risque cumulé $H(t)$ sont liées par

$$S(t) = \exp\left(-H(t)\right).$$

L'estimateur de Nelson-Aalen $\widehat{A}(t)$ du risque cumulé induit donc un estimateur non-paramétrique de la survie appelé l'estimateur de [37] :

$$\begin{aligned}
\widehat{S}_{Hf}(t) &= \exp\left(-\widehat{A}(t)\right) \\
&= \prod_{i,T_i\leq t}^{n} \exp\left(-\frac{d_i}{Y_i}\right) \\
&\approx \prod_{i,T_i\leq t}^{n} \left(1-\frac{d_i}{Y_i}\right) \qquad si \qquad \frac{d_i}{Y_i} \to 0.
\end{aligned}$$

L'estimateur de sa variance est donnée par

$$\widehat{\mathbb{V}}ar(\widehat{S}_{Hf}(t)) = \exp\left(-2\sum_{i,T_i\leq t}^{n}\frac{d_i}{Y_i}\right) \times \left(\sum_{i,T_i\leq t}^{n}\frac{d_i}{Y_i^2}\right).$$

## 1.5 L'approche paramétrique

Une approche alternative pour l'estimation de la fonction de survie est l'approche paramétrique. Elle consiste à faire l'hypothèse que la distribution des temps de survie appartient à une famille de distributions paramétrées par un vecteur de paramètres réels de dimension finie. L'avantage de cette approche est la facilitation attendue de la phase d'estimation des paramètres, ainsi que de l'obtention d'intervalles de confiance et de la construction de tests. L'inconvénient de l'approche paramétrique est la non adéquation pouvant exister entre le phénomène étudié et le modèle retenu. Quand on analyse des durées de survie, quelques formes de risque instantané les plus usuelles sont les suivantes: constante, monotone (croissant ou décroissant), en forme de $\bigcap$.

### 1.5.1 Risque instantané constant

#### 1.5.1.1 Loi exponentielle

La loi exponentielle $\mathcal{E}(\theta)$ est la seule qui admet un risque instantané constant. Cette loi est dite sans mémoire car la probabilité de décès pour un individu dans

un laps de temps est la même. i.e

$$\mathbb{P}(X > s + t|X > t) = \mathbb{P}(X > s).$$

Les quantités associées à cette loi sont :

$$\begin{aligned} f(t|\theta) &= \theta e^{-\theta t} \qquad t \geq 0, \theta > 0 \\ h(t|\theta) &= \theta \\ S(t|\theta) &= e^{-\theta t} \end{aligned}$$

## 1.5.2 Risque instantané monotone

### 1.5.2.1 Loi de Weibull

Considérons une loi de Weibull $W(\theta, \nu)$. Sa densité, fonction du risque et fonction de survie sont données par :

$$\begin{aligned} f(t|\theta,\nu) &= \nu\Big(\frac{1}{\theta}\Big)^{\nu} t^{\nu-1} \exp\bigg(-\Big(\frac{t}{\theta}\Big)^{\nu}\bigg) \qquad t \geq 0, \theta > 0, \nu > 0 \\ h(t|\theta,\nu) &= \nu\Big(\frac{1}{\theta}\Big)^{\nu} t^{\nu-1} \\ S(t|\theta,\nu) &= \exp\bigg(-\Big(\frac{t}{\theta}\Big)^{\nu}\bigg) \end{aligned}$$

**Remarque 1.5.1.** *Pour $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}\Big(\frac{1}{\theta}\Big)$. Si $0 < \nu < 1$, le risque instantané est monotone décroissant; si $\nu > 1$ le risque est monotone croissant.*

### 1.5.2.2 Loi Gamma

Soit une loi Gamma notée $G(\nu, \theta)$, on a alors :

$$\begin{aligned} f(t|\nu,\theta) &= \frac{\theta^{\nu}}{\Gamma(\nu)} t^{\nu-1} e^{-\theta t} \qquad t \geq 0, \theta > 0, \nu > 0 \\ h(t|\nu,\theta) &= \frac{1}{\Gamma(\nu)} \int_0^{\theta t} u^{\nu-1} e^{-u} du \\ S(t|\nu,\theta) &= \frac{f(t|\theta,\nu)}{1 - F(t|\theta,\nu)} \end{aligned}$$

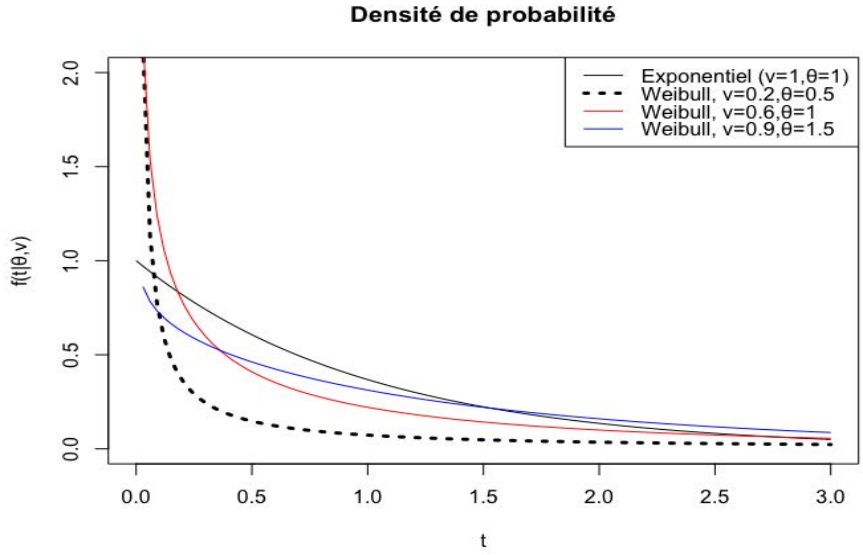



**Figure** 1.3 – Densité de probabilité de la loi de Weibull

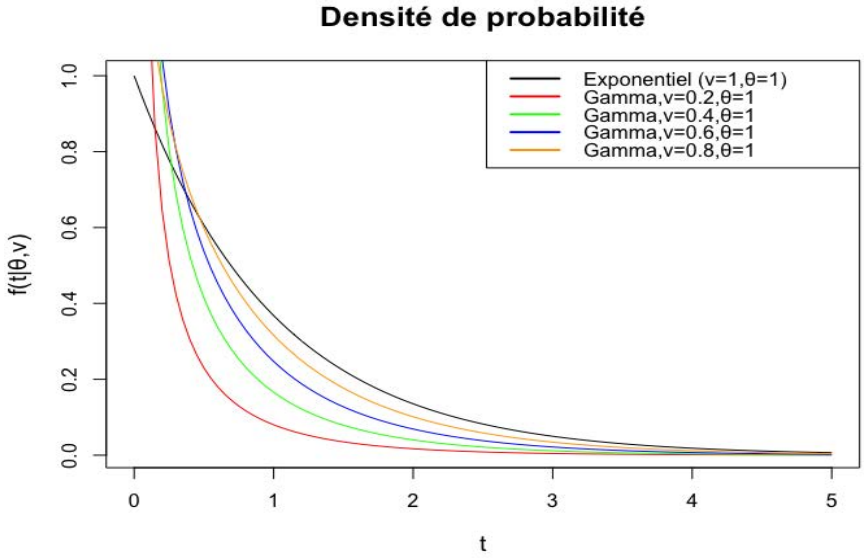



**Figure** 1.4 – Densité de probabilité de la loi Gamma

#### 1.5.2.3 Loi de Weibull généralisée

Cette loi est intéressante pour modéliser des risques monotones. L'évolution de risque est non-monotone au cours du temps en forme de cloche.

$$h(t|\theta,\nu,\gamma) = \left(1+\left(\frac{t}{\theta}\right)^{\nu}\right)^{\frac{1}{\gamma}-1}\frac{\nu}{\gamma\theta^{\nu}}t^{\nu-1} \qquad t\geq 0;\theta,\nu,\gamma>0$$

$$S(t|\theta,\nu,\gamma) = \exp\left[1-\left(1+\left(\frac{t}{\theta}\right)^{\nu}\right)^{\frac{1}{\gamma}}\right]$$

**Remarque 1.5.2.** *Pour $\gamma = 1$, on retrouve la loi de Weibull $W(\theta,\nu)$ et pour $\gamma = 1$ et $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}\left(\frac{1}{\theta}\right)$.*

## 1.5.3 Risque instantané en $\bigcap$

### 1.5.3.1 Loi Log-normale

La distribution des temps d'événements est Log-normale si :

$$\begin{aligned} f(t|\mu,\sigma) &= \frac{1}{t\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left[\frac{\log(t)-\mu}{\sigma}\right]^2\right) \\ h(t|\mu,\sigma) &= \frac{\frac{1}{t\sigma}\Phi\left[\frac{\log(t)-\mu}{\sigma}\right]}{\Phi\left[-\frac{\log(t)-\mu}{\sigma}\right]}, \qquad t>0,\mu>0,\sigma>0. \\ S(t|\mu,\sigma) &= 1-\Phi\left[\frac{\log(t)-\mu}{\sigma}\right] \end{aligned}$$

Cette loi est intéressante pour modéliser des risques monotones. L'évolution de

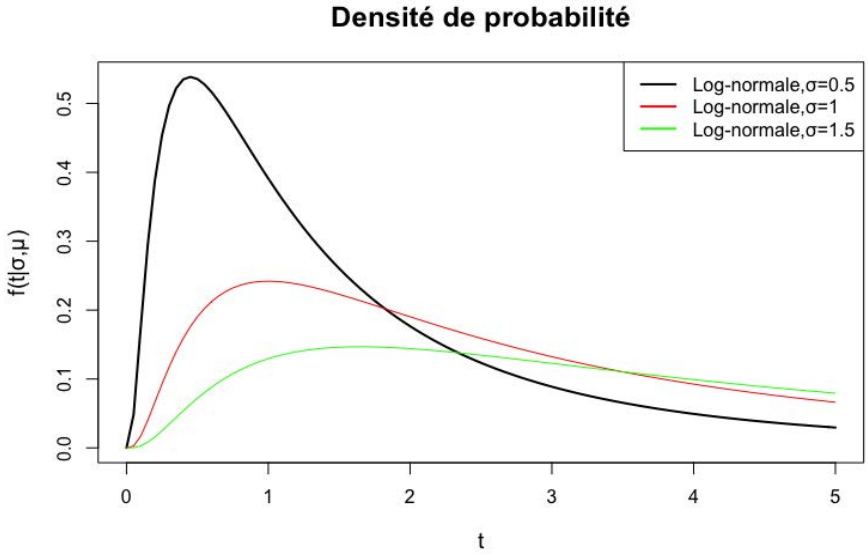



**Figure** 1.5 – Densité de probabilité de la loi Log-normale

risque est non-monotone au cours du temps en forme de cloche. Si $T$ est une variable aléatoire qui suit une loi Log-normale, $Y = \log(T)$ est une distribution normale.

#### 1.5.3.2 Loi Log-logistique

La distribution des temps d'événements est Log-logistique si :

$$\begin{aligned} f(t|\alpha,\gamma) &= \frac{\alpha\gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})^2} \qquad t \geq c; \alpha,\gamma > 0 \\ h(t|\alpha,\gamma) &= \frac{\alpha\gamma t^{\gamma-1}}{1+\alpha t^{\gamma}} \\ S(t|\alpha,\gamma) &= \frac{1}{1+\alpha t^{\gamma}} \end{aligned}$$

On rappelle que si $T$ est une variable aléatoire qui suit une loi Log-logistique, alors $Y = \log T$ suit une loi logistique.

## 1.6 L'approche semi-paramétrique

Elle peut être vue comme une sorte de *médiation* entre l'approche non-paramétrique et l'approche paramétrique. En effet, on utilise cette approche lorsque la famille de lois à laquelle appartient la loi de la variable de durée $X$ n'est pas totalement spécifiée. Apparue au cours des années soixante-dix, cette approche est très répandue en analyse de la survie, notamment à travers le modèle de régression de [26]. Le modèle de régression semi-paramétrique à risques proportionnels de Cox est l'un des modèles de régression de durée les plus utilisés en statistique médicale. Il permet en particulier d'identifier les facteurs de risque d'une maladie, de comparer des traitements, d'estimer des probabilités de survenue d'un événement (décès, rechute) chez un individu identifié par un vecteur donné de variables explicatives.

Le modèle de durées de vie accélérées (Accelerated failure time model ou ATF détaillé dans [27] et le modèle additif de hasard [38] sont présentés comme des alternatives du modèle de Cox. Tous ces modèles ne considèrent que la survenue d'un seul événement tel que le décès, la première récidive, l'apparition de la démence ou la survenue d'un diabète... Le modèle de Cox classiquement utilisé en analyse de survie suppose l'indépendance des temps de survie (au moins conditionnellement à un ensemble de variables explicatives observées). Or cette hypothèse s'avère quelque fois irrecevable au cas de l'existence de $"cluster"$ ou groupes d'individus au sein desquels les durées sont corrélées. Ces groupes peuvent représenter les individus d'une même famille, les patients traités au sein d'un même hôpital,

les organes d'un même patient,... Les clusters peuvent également représenter des durées observées de manière répétée sur le même individu : date de rechute, date de réapparution d'un symptôme donné [78]. L'utilisation sur des données corrélées d'un modèle de Cox conçu pour l'analyse de données indépendantes peut biaiser les paramètres de régression et lorsque la variable explicative est spécifique à chaque groupe, elle conduit à une sous-estimation de la variance de l'estimateur du paramètre. D'où l'introduction du modèle de Cox avec "fragilité".

### 1.6.1 Les modèles à risques proportionnels

Pour utiliser ces modèles exprimant un effet multiplicatif des diverses covariables sur la fonction de hasard (modèle à structure multiplicative), on introduit une fonction de hasard de base (appelée encore fonction de risque de base) et qui est commune à tous les individus. En d'autres termes on se place dans un contexte où l'objectif est le positionnement de différentes populations les unes par rapport aux autres, sans considération du niveau absolu du risque. Cela motive l'intérêt pour une spécification partielle, étudiée ici. Ces modèles se caractérisent par la relation suivante, pour tout $t > 0$

$$h(t|Z) = h_0(t)h(\beta, Z)$$

où $Z$ est un vecteur de covariables, $\beta$ le paramètre d'intérêt et $h$ une fonction positive. La fonction de hasard est le produit d'une fonction qui ne dépend que du temps et d'une fonction qui n'en dépend pas. En général, on suppose que l'effet de covariables se résume en une quantité réelle $\beta^\top Z$. C'est à dire

$$h(t|Z) = h_0(t)h(\beta^\top Z).$$

Ce modèle est dit à risques proportionnels, car quels que soient $i$ et $j$ qui ont pour covariables $Z_i$ et $Z_j$, le rapport des risques instantanés de deux individus ne varie pas au cours du temps :

$$\frac{h(t|Z_i)}{h(t|Z_j)} = \frac{h(\beta^\top Z_i)}{h(\beta^\top Z_j)}.$$

On en déduit que les fonctions de hasard sont donc propotionnelles.
Un cas particulier très important est le modèle de Cox [26]. Ce modèle supposant que la fonction $h$ est positive, impose une transformée logarithmique sur $h$ :

$$h(\beta^\top, Z) = \exp(\beta^\top Z).$$

Soit au total:

$$h(t|Z) = h_0(t)\exp(\beta^\top Z).$$

#### 1.6.1.1 Le modèle de régression à risques proportionnels de Cox

Le modèle de Cox [26] spécifie que le risque instantané ou relatif s'écrit :

$$h(t|Z) = h_0(t)\exp(\beta^\top Z) \tag{1.14}$$

où $Z$ est un vecteur de covariables de dimension $p \times 1$ et $\beta$ un vecteur $(p \times 1)$ de coefficients de régression. Considérons

- $N$ le nombre de décès observés parmi les $n$ sujets à l'étude,
- $T_1 < T_2 < ... < T_N$ les temps d'événements (décès) distincts,
- $(1), (2), ..., (N)$ les indices des individus décédés respectivement en $T_1 < T_2 < ... < T_N$,
- $Z_i$ la valeur des covariables de l'individu $i$,
- $R(T_i)$ l'ensemble des individus encore à risque à $T_i^-$ ( juste avant $T_i$).

#### 1.6.1.2 La vraisemblance partielle de Cox

Le modele de Cox, [26], est un modèle semi-paramétrique à risques proportionnels puisqu'il est paramétré par un vecteur de paramètres réels $\beta$ (la partie paramétrique) et une fonction $h_0(t)$ sur laquelle on ne fait aucune hypothèse (la partie non-paramétrique). C'est le modèle de régression le plus utilisé pour les données de survie, dans le but d'étudier les effets des covariables. Pour l'individu $i$, $i = 1, ..., N$, soit $X_i$ le temps d'événement, $C_i$ le temps de censure (indépendante) et $T_i = \min(X_i, C_i)$. On définit l'indicateur d'événement $\delta_i = \mathbb{I}_{\{X_i<C_i\}}$. L'expression de la vraisemblance pour un échantillon constitué de n individus indépendants est :

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}. \tag{1.15}$$

$f(t_i), S(t_i)$ sont respectivement la dansité de probabilité et la fonction de survie evaluées en $t_i$. Soit $R(t_i)$ , les individus à risque au temps d'événement $t_i$. En supposant pour l'instant qu'il n'y a pas concomitance dans la survenue de l'événement entre plusieurs individus : chaque temps d'événement observé est spécifique à un individu $i$ et un seul, alors il vient :

$$L = \prod_{i=1}^{n} \{h_i(t_i)S_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i} \tag{1.16}$$

$$= \prod_{i=1}^{n} \{h(t_i)\}^{\delta_i} S(t_i) \tag{1.17}$$

$$= \prod_{i=1}^{n} \left\{ \frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)} \right\}^{\delta_i} \left\{ \sum_{j \in R(t_i)} h_j(t_i) \right\}^{\delta_i} S_i(t_i). \tag{1.18}$$

Cox propose de ne considérer que le premier terme de l'expression précédente pour construire la vraisemblance partielle qui se définit donc comme :

$$L_{Cox} = \prod_{i=1}^{n} \left\{ \frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)} \right\}^{\delta_i}. \tag{1.19}$$

Il est très important de comprendre que le terme : $\left\{ \frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)} \right\}$ représente la probabilité qu'un individu $i$ connaisse l'événement du temps $t_i$ sachant q'il s'est produit un événement à cette durée parmi tous les individus à risque $R(t_i)$. On note $Z_i$ le vecteur associé des coéfficients de régression. La fonction de risque pour un individu $i$ au temps $t$ s'écrit :

$$h(t|Z_i) = h_0(t) \exp(\beta^\top Z_i).$$

$h(t|Z_i)$ est le produit de deux fonctions, la première $h_0(t)$ étant dépendante du temps mais pas de caractéristiques individuelles. Elle désigne la fonction de risque de base inconnue et non spécifiée, identique pour tous les individus, alors que la seconde est une fonction exponentielle ne dépendant pas du temps mais uniquement des caractéristiques des individus et plus généralement des explicatives retenues. L'estimation des coefficients de régression est obtenue par maximisation de la vraisemblance obtenue en ne considérant qu'une partie de la vraisemblance totale, d'où le qualificatif de vraisemblance partielle. Les estimateurs obtenus seront naturellement moins efficients que ceux découlant de la maximisation de la vraisemblance complète mais cette perte d'efficience est toutefois contrebalancée par l'énorme avantage de ne pas avoir à spécifier de distribution particulière sur les temps de survie, ce qui doit accroître leur robustesse. Une fois obtenus les estimateurs des coefficients, il est possible de construire un estimateur non paramétrique

du risque de base.
La conjugaison d'une estimation non paramétrique avec une technique de maximisation de vraisemblance explique que ce modèle soit qualifié de semi-paramétrique. Cette solution, proposée par Cox, est l'approche la plus couramment employée pour l'ajustement des modèles de survie [67, 26]. La vraisemblance partielle de Cox, est notée $L_{Cox}$ :

$$L_{Cox}(\beta) = \prod_{i=1}^{N} \left\{ \frac{\exp(\beta^\top Z_i(T_i))}{\sum_{j \in R(T_i)} \exp(\beta^\top Z_j(T_i))} \right\}^{\delta_i},$$

où $R(T_i)$ représente l'ensemble des individus à risque au temps d'événement $T_i$.

## 1.6.2 Estimation

### 1.6.2.1 Estimation des coefficients de régression $\beta$

L'estimation du vecteur des paramètres de dimension $p \times 1$ peut être obtenue à partir de la vraisemblance partielle. Notons

$$\mathcal{L}(\beta) = \log(L_{Cox}(\beta)) = \sum_{i=1}^{N} \left[ \beta^\top Z_i - \log \left( \sum_{j \in R(T_i)} \exp(\beta^\top Z_j) \right) \right]$$

et $U(\beta)$ la fonction score, c'est à dire le vecteur $p \times 1$ des dérivées premières de $\mathcal{L}(\beta)$

$$\begin{aligned} U(\beta) &= \frac{\partial \mathcal{L}(\beta)}{\partial(\beta)} \\ &= \frac{\partial \mathcal{L}(\beta)}{\partial(\beta_1)}, \cdots, \frac{\partial \mathcal{L}(\beta)}{\partial(\beta_p)} \\ &= \left( \sum_{i=1}^{N} \left[ Z_{i,1} - \frac{\sum_{j \in R(T_i)} Z_{j,1} \exp(\beta^\top Z_j)}{\sum_{j \in R(T_i)} \exp(\beta^\top Z_j)} \right], \ldots, \sum_{i=1}^{N} \left[ Z_{i,p} - \frac{\sum_{j \in R(T_i)} Z_{j,p} \exp(\beta^\top Z_j)}{\sum_{j \in R(T_i)} \exp(\beta^\top Z_j)} \right] \right). \end{aligned}$$

L'estimateur de Cox $\widehat{\beta}$ du coefficient de régression est solution de l'équation

$$U(\beta) = 0. \tag{1.20}$$

**Proposition 1.6.1.** *Soit $\widehat{\beta}$ l'estimateur du maximum de vraisemblance de $\beta$ i.e. la quantité vérifiant :*

$$U(\widehat{\beta}) = 0$$

*Alors*

$$\widehat{\beta} \xrightarrow{\mathcal{L}} \mathcal{N}(\beta, I^{-1}(\widehat{\beta}))$$

$I^{-1}$ *est l'inverse de la matrice d'information de* **Fisher** *basée sur la vraisemblance partielle.*

Autrement dit un estimateur consistant de la matrice de variance-covariance de $\hat{\beta}$ peut être calculé à partir de l'inverse de la matrice d'information de Fisher,

$$\widehat{\mathbb{V}}ar(\widehat{\beta}) = \left(I\left(\widehat{\beta}\right)\right)^{-1}$$

où le terme $(i, j)$ de la matrice $I(\hat{\beta})$ est

$$\left[I(\hat{\beta})\right]_{i,j} = \mathbb{E}\left[-\frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta_i \partial\beta_j}|_{\beta=\hat{\beta}}\right].$$

Cela est démontré dans [5] en utilisant la théorie des martingales. Il faut noter qu'il n'y a pas de solution exacte à cette équation (1.20). L'algorithme de Newton-Raphson est souvent utilisé par les logiciels pour obtenir une solution.

### 1.6.3 Résolution numérique

Pour résoudre l'équation du score (1.20), l'algorithme de Newton-Raphson est habituellement employé. Partant d'une solution initiale $\beta_0 = 0$, l'algorithme consiste en la succession d'itérations de la forme

$$\widehat{\beta}^{j+1} = \widehat{\beta}^{j} - \left[\frac{\partial^2 \log \mathcal{L}(\widehat{\beta}^{j})}{\partial\beta^2}\right]^{-1} \frac{\partial \log \mathcal{L}(\widehat{\beta}^{j})}{\partial\beta}.$$

Le second terme du deuxième membre (qui suit le signe moins ) est le pas itératif de l'algorithme de Newton-Raphson. Si la fonction de vraisemblance évalué en $\widehat{\beta}^{j+1}$ est inférieure à celle évaluée en $\widehat{\beta}^{j}$, alors $\widehat{\beta}^{j+1}$ est recalculée en utilisant, cette fois-ci, la moitié du pas itératif. Ces étapes se succèdent jusqu'à ce que la convergence soit obtenue, c'est-à-dire jusqu'à ce que $\widehat{\beta}^{m+1}$ soit suffisamment proche de $\beta$ . L'estimateur du maximum de vraisemblance de $\beta$ est alors $\widehat{\beta}$=$\widehat{\beta}^{m+1}$. Les composantes $\beta_l$ du vecteur de paramètres $\beta$ s'interprètent comme le logarithme du risque relatif associé aux covariables et représentent une augmentation $(\beta_l > 0)$ ou une diminution $(\beta_l < 0)$ du risque d'observer un événement.

### 1.6.4 Evaluation du risque cumulé de base

Après avoir estimé les coefficients de régression, on peut estimer le risque cumulé de base par l'estimateur de Breslow qui est une extension de l'estimateur de Nelson- Aalen,

$$\widehat{H}_0(t) = \sum_{i;T_i \leq t} \frac{d_i}{\sum_{j \in R(T_i)} \exp(\widehat{\beta}^\top Z_j)},$$

où $d_i$ est le nombre de décès en $T_i$.
Si $\widehat{\beta} = 0$, on retrouve l'estimateur de Nelson-Aalen. On peut déduire un estimateur de la fonction de survie pour un vecteur des covariables $Z$, donnée par :

$$S(t|Z) = \exp\left(-\int_0^t h(u|Z)du\right)$$

et l'estimateur correspondant est :

$$\widehat{S}(t|Z) = \exp\left(-\widehat{H}_0(t)\exp(\widehat{\beta}^\top Z)\right).$$

### 1.6.5 Tests

On souhaite souvent vérifier certaines hypothèses sur les coefficients de régression. Trois tests de l'hypothèse nulle $\mathcal{H}_0 : \beta = \beta_0$ peuvent être déduits du resultat concernant la convergence asymptotique de $\widehat{\beta}$.

#### 1.6.5.1 Test du rapport de vraisemblance

Ce test mesure la distance entre $\log L_{Cox}(\widehat{\beta})$ et $\log L_{Cox}(\widehat{\beta}_0)$, très couramment utilisé en statistique, découle d'un développement de Taylor à l'ordre 2 de $\log \mathcal{L}(\beta)$, puis de propriétés de convergence en loi [30]. Il s'énonce comme suit:

$$\chi^2_{LRT} = 2\left[\log L_{Cox}(\widehat{\beta}) - \log L_{Cox}(\widehat{\beta}_0)\right] \overset{\mathcal{H}_0}{\rightsquigarrow} \chi^2(p).$$

#### 1.6.5.2 Test de Wald (ou du maximum de vraisemblance)

Il mesure l'écart entre $\widehat{\beta}$ et $\beta_0$. Si une variable aléatoire $X$ p-dimensionnelle suit une loi normale centrée réduite alors $X^\top X$ suit une loi du Khi-deux à p degrés de liberté. Ainsi

$$\chi^2_w = (\widehat{\beta} - \beta_0)' I(\widehat{\beta})(\widehat{\beta} - \beta_0) \overset{\mathcal{H}_0}{\rightsquigarrow} \chi^2(p).$$

#### 1.6.5.3 Test du Score (ou mesure de la pente en $\beta_0$)

Ce test mesure la pente de la tangente en $\beta_0$. Sous $\mathcal{H}_0$, le maximum de vraisemblance est obtenu pour une valeur $\widehat{\beta}$ proche de $\beta_0$. La pente en $\beta_0$ diffère de peu de $0$, et elle est nulle en moyenne sous $\mathcal{H}_0$.
Il est possible de montrer que

$$\frac{\partial \log L(\beta)}{\partial \beta}\Big|_{\beta=\beta_0} \xrightarrow{\mathcal{L}} \mathcal{N}(0, I)$$

En notant $U(\beta) = \partial \log L(\beta)/\partial\beta$, nous obtenons par conséquent

$$\chi^2_s = (U(\beta_0))'(I(\beta_0))^{-1}U(\beta_0) \overset{\mathcal{H}_0}{\rightsquigarrow} \chi^2(p)$$

Ces trois statistiques suivent, sous $H_0$, une loi de $\chi^2$ à p degrés de liberté (où $\beta$ est un vecteur de dimension $p$). La statistique du rapport de vraisemblance ne nécessite pas de calculer les dérivées secondes de la log-vraisemblance. La statistique du score ne nécessite pas l'estimation de $\widehat{\beta}$.
On peut déduire de ces statistiques des tests partiels permettant de tester des hypothèses concernant certaines coordonnées de $\beta$. En particulier, supposons que l'on souhaite tester l'addition d'une nouvelle variable $Z_p$ dans un modèle avec $(p-1)$ variables. On veut savoir si le modèle contenant la variable $Z_p$ apporte plus d'informations sur la distribution des durées de vie que le modèle sans cette variable. On teste l'hypothèse

$$\mathcal{H}_0 : \beta = \beta_0$$

où $\beta = (\beta_1, ..., \beta_p)$ et $\beta_0 = (\beta_1, ..., \beta_{p-1}, 0)$, c'est à dire on teste

$$\mathcal{H}_0 : \beta_p = 0$$

Dans ce cas on aura:
- La statistique du rapport de vraisemblance:

$$\chi^2_{LRT} = 2\Big[\log L_{Cox}(\widehat{\beta}) - \log L_{Cox}(\widehat{\beta}_0)\Big] \overset{\mathcal{H}_0}{\rightsquigarrow} \chi^2(1).$$

- La statistique de Wald:

$$\chi^2_w = (\widehat{\beta} - \beta_0)' I(\widehat{\beta})(\widehat{\beta} - \beta_0) \overset{\mathcal{H}_0}{\rightsquigarrow} \chi^2(1).$$

- La statistique du Score:

$$\chi^2_s = (U(\widehat{\beta}_0))'(I(\widehat{\beta}_0))^{-1}U(\widehat{\beta}_0) \overset{\mathcal{H}_0}{\rightsquigarrow} \chi^2(1)$$

où $\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$ et $\widehat{\beta}_0 = (\widehat{\beta}_1, ..., \widehat{\beta}_{p-1}, 0)$.

### 1.6.6 Significativité des paramètres

Puisque $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ suit asymptotiquement une loi normale de moyenne $\beta$ et de variance $I(\beta)^{-1}$, il est facile de calculer un intervalle de confiance pour le paramètre $\beta_j (j = 1, \ldots, p)$. L'intervalle de confiance de $\beta_j$ est donné par :

$$IC_{1-\alpha}(\beta_j) = \left[\widehat{\beta}_j - z_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j), \widehat{\beta}_j + z_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)\right],$$

où $z_{\frac{\alpha}{2}}$ est le quantile de la loi normale d'ordre $\frac{\alpha}{2}$ et $\hat{\sigma}(\hat{\beta}_j)$ est l'écart-type de $\hat{\beta}_j$. Rappelons par définition que le rapport des risques à l'instant $t$ pour deux vecteurs de covariables $Z_i$ et $Z_j$, est égal à :

$$RR(t) = \frac{h(t|Z_i)}{h(t|Z_j)}.$$

Dans le modèle de **Cox** le rapport des risques est constant au cours du temps :

$$RR(t, Z_i, Z_j) = RR = \exp\left(\beta^\top (Z_i - Z_j)\right). \tag{1.21}$$

Dans un modèle de **Cox** à $p$ variable explicatives, le modèle s'écrit :

$$h(t|Z, \beta) = h_0(t) \exp) \left(\beta_1 Z_1 + \cdots + \beta_p Z_p\right)$$

et on a $RR_j = \exp(\beta_j)$ qui est le rapport des risques lié à la variable $Z_j$ ajusté sur toutes les autres variables explicatives, c'est à dire le rapport des risques d'un sujet à un autre sujet qui ne diffère que par la valeur de $Z_j$. En général, dans l'interprétation des résultats, ce n'est pas l'intervalle de confiance du paramètre qui est fourni mais plutôt l'intervalle de confiance du rapport des risques. Puisque l'on dispose d'un intervalle de confiance pour le paramètre et que le rapport des risques est égal à l'exponentielle de ce paramètre, l'intervalle de confiance à $1 - \alpha$ pour le rapport des risques correspondant peut être calculé par :

$$IC_{1-\alpha}(RR_j) = \left[\exp\left(\widehat{\beta}_j - z_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)\right), \exp\left(\widehat{\beta}_j + z_{\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)\right)\right].$$

Il faut aussi noter que si les valeurs des statistiques précédentes ( statistique de Wald, statistique du rapport de vraisemblance, statistique du score) sont supérieures à la valeur de la statistique de test du $\chi^2$ à $p$ degrés de libertés au risque $\alpha = 5\%$, alors on rejette l'hypothèse nulle et on conclut à l'effet significatif des covariables sur la survie.

### 1.6.7 Diagnostique du modèle de Cox

Il y a deux hypothèses importantes à vérifier dans l'utilisation du modèle de Cox :

1. La **log-linéarité** pour les variables continues,
2. Les **risques proportionnels**.

**Les résidus**. De nombreuses procédures de vérification des hypothèses du modèle de Cox sont basées sur des quantités que l'on appelle résidus. Il existe plusieurs résidus :

1. Les résidus de **Schoenfeld** servent à tester **l'hypothèse des risques proportionnels**.
   - Si l'hypothèse des risques proportionnels est vérifiée, ces résidus ont en théorie un aspect totalement aléatoire et l'évolution temporelle moyenne est une droite horizontale.
   - Si l'hypothèse des risques proportionnels n'est pas vérifiée, par exemple si le facteur de risque est important au début du suivi du patient mais pas à la fin, alors les résidus seront, sur le schéma, négatifs puis positifs, et l'évolution temporelle moyenne sera une courbe croissante (ou le contraire selon le codage du facteur risque).
2. Les résidus des **martingales** permettent de détecter la **log-linéarité**, c'est-à-dire une forme fonctionnelle mal spécifiée dans la partie paramétrique du modèle. Le tracé des résidus des martingales en fonction des variables explicatives inclues dans le modèle peut être utilisé pour indiquer si certaines variables ont besoin d'être transformées avant d'être incorporées dans le modèle. Pour cela, on ajoute une courbe lisse sur les points obtenus. La forme fonctionnelle est alors suggérée par la forme de la courbe lisse :
   - Une croissance lente suggèrera une transformation logarithmique.
   - À l'inverse, une croissance rapide suggèrera une transformation puissance avec une puissance supérieure à 1.

   Par ailleurs, le tracé des résidus des martingales en fonction des variables explicatives non inclues dans le modèle peut être utilisé pour indiquer si certaines variables devraient être inclues dans le modèle, ce qui est le cas si une dépendance apparaît.
3. Les résidus de **Cox-Snell** servent à valider l'ensemble du modèle. L'idée est que si $T$ suit une loi de fonction de risque cumulé $H(t|Z)$, alors $Y = H(T|Z)$

suit une loi exponentielle de paramètre 1. En effet, on sait que la fonction de survie de la variable $T$ vérifie

$$S(t) = \mathbb{P}(T > t) = \exp(-H(t|Z))$$

où $H(t|Z) = H_0(t)\exp(\beta^\top Z)$. Donc $Y = H(T|Z)$ vérifie

$$\mathbb{P}(Y > y) = \mathbb{P}(T > H^{-1}(y|Z)) = \exp(-y),$$

c'est-à-dire que la variable $Y = H(T|Z)$ suit une loi exponentielle de paramètre $1$. Il y a donc adéquation du modèle si le risque cumulé de la variable $Y$ est proche de la droite $y = x$. On procède de la façon suivante :

(i). On estime $H(.|Z)$ du modèle supposé par l'estimateur semi-paramétrique $\widehat{H}_0(.)\exp(\widehat{\beta}^\top Z)$.

(ii). On calcule les résidus de Cox-Snell pour chaque temps observé $T_i$, $i = 1, \ldots, n$

$$r_i = \widehat{H}(T_i|Z) = \widehat{H}_0(T_i)\exp(\widehat{\beta}^\top Z).$$

(iii). On estime la fonction de risque cumulé des $r_i$ de façon non paramétrique (Nelson-Aalen) par l'estimateur de Nelson-Aalen noté $\widehat{A}$.

(iv). On trace donc les fonctions $y = \widehat{A}(x)$ et $y = x$ sur le même graphique. En effet, si le modèle est correct, le risque cumulé $\widehat{A}(x)$ doit être approximativement égal à la droite $y = x$.

4. Les résidus de **déviance** (resymétrisation des résidus des martingales, pour corriger leur asymétrie) sont compris entre $-1$ et $1$. Ils valent $0$ sous l'hypothèse des risques proportionnels.
5. Les résidus du **score** permettent d'identifier les observations qui contribuent fortement à la détermination des paramètres du modèle.
6. Il y aussi les indices **dfbetas** qui permettent de rechercher les sujets influents (ou marginaux). Cette recherche se limite à la représentation des indices dfbetas de chaque sujet en fonction du temps. Des points extrêmes ou isolés doivent être examinés en détail. Il pourra parfois être intéressant

d'extraire provisoirement ces sujets de l'échantillon et de recalculer le modèle : une variabilité importante fera craindre une instabilité importante et jettera un doute sur les résultats.

**Hypothèse des risques proportionnels**

1. Utiliser l'interaction avec le temps,
2. Méthodes de validation graphique pour les variables qualitatives. Ces méthodes sont peu puissantes mais constituent une première approche intéressante. Le tracé de $\ln(-\ln(S(t)))$ en fonction de $\ln t$ dans chaque groupe, dans le cas de risques proportionnels, devrait se traduire par des courbes translatées. Par exemple, pour une variable à deux modalités, on peut estimer la fonction de survie dans les deux groupes et tracer les courbes

$$\begin{aligned}\ln\big[-\ln(\widehat{S}(t))\big] &= \ln\left[\int_0^t \widehat{h}(u|Z)du\right] \\ &= \ln\big[\widehat{H}_0(t)\exp(\widehat{\beta}Z)\big] \\ &= \ln\Big(\widehat{H}_0(t)\Big) + \widehat{\beta}Z.\end{aligned}$$

3. Test basé sur les résidus de Schoenfeld.

Dans le cas des risques non proportionnels, on peut utiliser une des méthodes suivantes :

(i). Transformer les variables quantitatives en variables qualitatives,

(ii). Pour les variables qualitatives, essayer de changer de classes,

(iii). Si l'hypothèse de proportionalité est vérifiée sur des intervalles de temps courts, faire une modélisation par partie (construire différents modèles sur ces intervalles),

(iv). Stratification.

**Hypothèse 1.6.2.** *Le modèle fait également une hyothèse de* $\log$*-linéarité, c'est à dire que le logarithme du risque est une fonction linéaire de la* $v.a.\ Z$

$$\log h(t|Z) - \log h_0(t) = \beta^\top Z. \tag{1.22}$$

Autrement dit si l'on peut raisonnablement accepter l'hypothèse de $\log$-linéarité, la variable doit rester continue. Cette hypothèse a pour conséquence qu'une augmentation de la variable par $t$ multiplie le risque de survie par $\exp(t\beta)$ que lorsque

l'on passe de $t$ à $t+a$ ou de $2t$ à $2t+a$. *Par exemple si l'événement est le décès et la covariable est l'âge, alors cette hypothèse entraîne qu'une augmentation de l'âge de 5 ans multiplie le risque de décès par* $\exp(5\beta)$ *lorsque l'on passe de* $20$ *à* $25$ *ans ou de* $70$ *à* $75$ *ans*. On peut donner un autre exemple dans lequel, *si l'âge est une variable explicative continue et que l'on étudie une maladie qui touche essentiellement les personnes âgées, le modèle supposera que le risque relatif est le même pour une augmentation de* $1$ *an, que ce soit pour un âge de* $30$ *ans ou pour un âge de* $70$ *ans.* Dans le cas de variables quantitatives, on peut considérer un codage dichotomique (*par exemple , une variable codée* $0, 1, 2$ *peut être recodée en utilisant deux variables binaires:* $(0,0), (0,1), (1,1)$ .

C'est une hypothèse forte qui n'est généralement pas recommandée. Si l'on choisit de transformer une variable quantitative continue en une variable à $n$ classes ordonnées, se pose le problème du nombre de classes et du choix des seuils pour découper cette variable. La transformation peut être simple en classes en utilisant la moyenne ou la médiane de la variable continue ou plus complexe en utilisant les tertiles pour un découpage en 3 classes ou les quartiles pour un découpage en 4 classes. L'information la plus réduite est associée au plus petit nombre de classes. Attention à ce que l'hypothèse de proportionnalité des risques soit vérifiée. On peut également essayer de transformer la variable (logarithme, racine, puissance...) afin qu'elle vérifie l'hypothèse de $\log$-linéarité.

## 1.7 Conclusion

L'analyse de données de survie est l'étude du délai de survenue d'un événement d'intérêt pour un ou plusieurs groupes d'individus. Pour cela on estime la fonction de survie en utilisant les différentes méthodes : l'estimation non-paramétrique, l'estimation paramétrique et l'estimation semi-paramétrique. Ces méthodes classiques supposent l'indépendance des temps de survenue de l'événement d'intérêt. Pour pouvoir bien mener cette étude d'analyse de survie, nous avons collecté les informations nécessaires en se basant en grande partie sur des notions mathématiques pour donner un sens à leur signification. Nous avons présenté une étude ayant pour objectif de modéliser les données de survie et en donner des applications en exposant les différents modèles en particulier le modèle de Cox, leur signification, leur critère d'adéquation ainsi que la nature de la convergence de certains estimateurs. Dans ce chapitre, nous avons aussi présenté l'utilisation des

outils habituels propres aux processus ponctuels qui nous permettent de déterminer les propriétés asymptotiques du meilleur estimateur. Ce qui montre que l'apprentissage de la théorie mathématique des données de survie peut se faire suivant une autre approche à savoir la théorie des processus de comptage qui présente l'avantage d'une grande précision mathématique. Par conséquent la notion de complexité auquel le modèle de Cox fait appel dans l'analyse de survie peut cependant être déjouée : les modèles de Cox, malgré leurs applications épidémiologiques nombreuses sont limités.

CHAPITRE 2

# Modèles de fragilité


## Résumé

*Dans ce chapitre, nous présentons une revue sur des modèles de fragilité. Nous présentons en particulier des modèles à risque multiplicatif avec hypothèse d'une distribution gamma pour la variable de fragilité. Le modèle de base examiné est un modèle avec des fragilités partagées par des sujets du même groupe. Le modèle à fragilité Gamma corrélée ainsi que le modèle à fragilités emboitées ont été discutés. Différentes approches ont été proposées pour s'adapter à ces modèles, notamment l'utilisation de l'algorithme EM et la vraisemblance pénalisée.*

## Sommaire



# 2.1 Introduction

La notion de fragilité offre un moyen pratique d'introduire des effets aléatoires, de dépendance et d'hétérogénéité non observée, dans les modèles de données de survie. Dans sa forme la plus simple, une fragilité est un facteur de proportionnalité aléatoire non observé qui modifie la fonction de risque d'un individu ou d'un groupe de personnes liées les unes aux autres. Fondamentalement, le concept de fragilité remonte aux travaux de [48] sur la "prédisposition à l'accident". Ce n'est qu'en 1979 que le terme de fragilité lui-même a été présenté par [117] pour rendre compte l'hétérogénéité individuelle dans un contexte de mortalité. En général, dans la plupart des applications cliniques, l'analyse de survie suppose implicitement que la population étudiée est homogène. Ceci veut dire que tous les individus

en étude sont soumis, en principe, aux mêmes risques (par exemple, risque de décès, risque de récidive). Or dans de nombreuses applications, cette hypothèse n'est pas réaliste. Par exemple, il peut y avoir une prédisposition génétique à certaines maladies qui est différente d'un individu ou d'un groupe d'individus à l'autre. Ceci veut donc dire que les individus en étude forment un ensemble de personnes avec différents risques. Les modèles de fragilité [32, 24, 123, 51] sont des extensions du modèle classique à risques proportionnels de Cox [26] et ont été proposés en particulier pour prendre en compte l'hétérogénéité liée à l'existence de groupes ou "clusters" d'individus au sein desquels les durées sont corrélées. Les facteurs de risque non observés ou non mesurés et partagés par un groupe d'individus vont créer une dépendance des événements étudiés dans chaque groupe. Dans les analyses de survie, si cette hétérogénéité non observée est ignorée, elle peut créer un biais important dans l'estimation de la variance des paramètres de régression et sur l'estimation de la fonction de risque.

Les données de survie environnementales conduisent très souvent à des analyses statistiques particulières. En effet, ces données sont souvent regroupées en zones géographiques (région, ville, pays) et il est fréquent que les individus d'une même zone partagent des facteurs de risque non identifiés (génétiques, environnementaux).

Les individus appartenant à des groupes, tels que les familles, les zones géographiques, les patients traités au sein d'un même hôpital, les organes d'un même patient, font intervenir la corrélation des temps de survie. Une corrélation peut alternativement provenir des événements récurrents, c'est-à-dire lorsqu'un individu subit un même événement de manière répétée, telles que les réhospitalisations ou la réapparition d'un symptôme donné [78]. Dans ces situations, la façon naturelle de modéliser la durée de survie est l'introduction d'un effet aléatoire spécifique-groupe, la fragilité. Cet effet aléatoire explique la dépendance dans le sens où si la fragilité était connue, alors les événements seraient indépendants. En d'autres termes, les durées de vie sont indépendantes conditionnellement à la fragilité. Cette approche peut être utilisée pour les temps de survie d'unités liées, comme les membres d'une famille ou les observations récurrentes sur une même personne. Ainsi, dans ce chapitre, les modèles de fragilités sont présentés. Un accent particulier est mis sur les modèles à fragilité gamma partagée où l'aspect méthodologique et algorithmique ainsi que l'illustration des propos sur différentes bases de données, en utilisant les différents packages "R" ont été succinctement explorés.

## 2.2 Présentation du modèle de fragilité

Ici, nous introduisons les modèles de fragilité en donnant quelques exemples les plus rencontrés. Nous précisons également les notations et les hypothèses de base.

### 2.2.1 Modèle à fragilité partagé ( Fragilité Gamma)

Le modèle à fragilités partagées $(Shared\ frailty\ model)$ spécifie que la fonction de risque conditionnelle à la variable de fragilité est

$$h_{ij}(t|U_i) = U_i h_0(t) \exp(\beta^\top Z_{ij}) \tag{2.1}$$

où $h_0(t)$ est la fonction de risque de base; $Z_{ij} = (Z_{1ij}, ..., Z_{pij})^\top$ représente le vecteur des variables explicatives pour le $j^{ieme}$ individu du groupe $i$, $\beta$ est le vecteur correspondant des paramètres de régression, et $U_i$ sont des variables aléatoires non observées (ou variables de fragilité) et partagées par tous les individus d'un même groupe. Le plus souvent, on suppose que la variable de fragilité est distribuée selon une loi gamma, pour avoir une forme explicite de la vraisemblance marginale [24]. La distribution gamma peut être paramétrée par un paramètre de forme $\eta$, et un paramètre d'échelle $\nu$. La densité de la distribution est alors :

$$f_U(u) = \frac{u^{\eta-1}\exp(-\nu u)\nu^\eta}{\Gamma(\eta)},$$

avec $\eta > 0$, $\nu > 0$. L'espérance d'une variable $U$ de distribution gamma de paramètres $\eta$ et $\nu$ est :

$$\mathbb{E}(U) = \frac{\eta}{\nu}$$

et sa variance:

$$\mathbb{V}ar(U) = \frac{\eta}{\nu^2}.$$

En considérant par la suite que les effets aléatoires $U_i$ sont indépendants et identiquement distribués d'éspérence 1 et de variance inconnue $\varphi = \frac{1}{\nu}$, on a la densité $f_U(u|\varphi) = \frac{u^{(\frac{1}{\varphi}-1)}\exp(-\frac{u}{\varphi})}{\varphi^{\frac{1}{\varphi}}\Gamma(\frac{1}{\varphi})}$ comme distribution a priori. On peut en déduire la distribution à posteriori des effets aléatoires de la manière suivante :

$$f_{u_i}(U|T_i, h_0(.), \beta, \varphi) = \frac{f(T_i|u_i, h_0(.), \beta, \varphi)\, f_{u_i}(U|\varphi)}{f(T_i, h_0(.), \beta, \varphi)}$$

Le choix de la distribution gamma a été discuté par [58]. D'autres distributions ont été étudiées, telles que la distribution log-normal ou la distribution positive stable, proposée par [57].

Dans un modèle semi-paramétrique où le risque de base n'est pas spécifié, l'estimation directe n'est pas possible. Deux approches sont alors souvent utilisées : l'algorithme EM ou la vraisemblance pénalisée.

#### 2.2.1.1 Ecriture du modèle

Si les effets aléatoires $U_i$ étaient observés, la vraisemblance conditionnelle des observations $(U_i, T_{i1}, ..., T_{iJ_i})$ pour le cluster $i$ dans un modèle à fragilité gamma et des données potentiellement censurées à droite serait :

$$\left[\prod_{j=1}^{J_i} h_{ij}(T_{ij}|U_i)^{\delta_{ij}} \times S_{ij}(T_{ij}|U_i)\right] f(U_i).$$

Donc la vraisemblance conditionnel s'écrit :

$$\begin{aligned} &L_i\left(h_0(.), \beta|U_i\right) = \\ &\left[\prod_{j=1}^{J_i} \left(U_i h_0(T_{ij}) \exp(\beta^\top Z_{ij})\right)^{\delta_{ij}} \exp\left(-H_0(T_{ij}) \exp(\beta^\top Z_{ij}) U_i\right)\right] f(U_i), \end{aligned} \tag{2.2}$$

avec $H_0(t) = \int_0^t h_o(u)\, du$. La log-vraisemblance complète correspondante à la vraisemblance (2.2) peut s'écrire

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i} \left[\delta_{ij}\left(\log h_0(T_{ij}) + \log U_i + \beta^\top Z_{ij}\right) - H_0(T_{ij}) \exp(\beta^\top Z_{ij}) U_i + \log\left(f(U_i)\right)\right]. \tag{2.3}$$

La log-vraisemblance (2.3) peut se décomposer en deux parties, l'une impliquant les paramètres $\beta$ et le risque cumulé de base $H_0(.)$, l'autre impliquant le paramètre $\varphi$ associé à la fragilité :

$$\ell\left(\beta, H_0(.), \varphi\right) = \ell_1\left(\beta, H_0(.)\right) + \ell_2\left(\varphi\right)$$

avec

$$\ell_1\left(\beta, H_0(.)\right) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{J_i}\left[\delta_{ij}\left(\log h_0(T_{ij}) + \beta^\top Z_{ij}\right) - H_0(T_{ij}) \exp\left(\beta^\top Z_{ij}\right) U_i\right]\right\} \tag{2.4}$$

$$\ell_2\left(\varphi\right) = \sum_{i=1}^{n}\left[\left(\frac{1}{\varphi} + d_i - 1\right)\log(U_i) - \frac{U_i}{\varphi}\right] - L\left[\frac{1}{\varphi} + \log\Gamma\left(\frac{1}{\varphi}\right)\right] \tag{2.5}$$

avec $d_i = \sum_{j=1}^{J_i} \delta_{ij}$ le nombre d'événements observés dans le $i^{ieme}$ cluster. Le vecteur de paramètres $\beta$ n'intervient que dans la première partie $\ell_1(\beta, H_0(.))$ de cette log-vraisemblance.
La vraisemblance marginale pour les observations $(U_i, T_{i1}, ..., T_{iJ_i})$ est donc

$$L_{marg,i}(\beta, H_0(.), U_i) = \int_0^\infty \left[ \prod_{j=1}^{J_i} \left( U_i h_0(T_{ij}) \exp(\beta^\top Z_{ij}) \right)^{\delta_{ij}} \exp\left( -H_0(T_{ij}) \exp\left(\beta^\top Z_{ij}\right) U_i \right) \right] f_{U_i}(u) du \tag{2.6}$$

Donc

$$\begin{aligned} L_{marg,i}(\beta, H_0(.), \varphi) &= \int_0^\infty \left[ \prod_{j=1}^{J_i} \left( u h_0(T_{ij}) \exp(\beta^\top Z_{ij}) \right)^{\delta_{ij}} \exp\left( -H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u \right) \right] \\ &\times \frac{u^{(\frac{1}{\varphi}-1)} \exp(-\frac{u}{\varphi})}{\varphi^{\frac{1}{\varphi}} \Gamma\left(\frac{1}{\varphi}\right)} du \end{aligned}$$

Dans cette expression,

$$\begin{aligned} f(Z_i|u_i, h_0(.), \beta, \varphi) f_{U_i}(u|\varphi) &= \prod_{j=1}^{J_i} \left( u h_0(T_{ij}) \exp(\beta^\top Z_{ij}) \right)^{\delta_{ij}} \exp\left( -H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u \right) \\ &= u^{d_i} \prod_{j=1}^{J_i} \left( h_0(T_{ij}) \exp(\beta^\top Z_{ij}) \right)^{\delta_{ij}} \exp\left( -H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u \right) \end{aligned}$$

correspond à la contribution conditionnelle à la vraisemblance du cluster $i$, avec $d_i = \sum_{j=1}^{J_i} \delta_{ij}$ le nombre d'événements observés dans le $i^{ieme}$ cluster. En posant $k = \frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(H_0(T_{ij})\right) \exp(\beta^\top Z_{ij})$, on peut écrire l'expression précedente de la vraisemblance marginale comme

$$\begin{aligned} L_{marg,i}(\beta, H_0(.), \varphi) &= \frac{\prod_{j=1}^{J_i} \left( h_0(T_{ij}) \exp(\beta^\top Z_{ij}) \right)^{\delta_{ij}}}{k^{\frac{1}{\varphi}+d_i} \varphi^{\frac{1}{\varphi}} \Gamma\left(\frac{1}{\varphi}\right)} \int_0^\infty (ku)^{\frac{1}{\varphi}+d_i-1} \exp\left((-ku) d(ku)\right) \\ L_{marg,i}(\beta, H_0(.), \varphi) &= \frac{\Gamma\left(\frac{1}{\varphi}+d_i\right) \prod_{j=1}^{J_i} \left( h_0(T_{ij}) \exp(\beta^\top Z_{ij}) \right)^{\delta_{ij}}}{\left( \frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(H_0(T_{ij})\right) \exp(\beta^\top Z_{ij}) \right)^{\frac{1}{\varphi}+d_i} \varphi^{\frac{1}{\varphi}} \Gamma\left(\frac{1}{\varphi}\right)} \end{aligned} \tag{2.7}$$

On en déduit une expression de la distribution à posteriori des effets aléatoires

$$f_{u_i}(U|Z_i, h_0(.), \beta, \varphi) = \frac{u_i^{\frac{1}{\varphi}+d_i-1} \exp\left[-u_i\left(\frac{1}{\varphi} + \sum_{j=1}^{J_i} H_0(T_{ij})\right)\exp(\beta^\top Z_{ij})\right]}{\left(\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(H_0(T_{ij})\right)\exp(\beta^\top Z_{ij})\right)^{-(\frac{1}{\varphi}+d_i)} \Gamma\left(\frac{1}{\varphi}+d_i\right)}$$

On retrouve ici l'expression d'une distribution gamma de paramètre de forme $(\frac{1}{\varphi}+d_i)$ et de paramètre d'échelle $\left(\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(\hat{H}_0(T_{ij})\right)\exp(\hat{\beta}^\top Z_{ij})\right)$

Le logarithme de la vraisemblance complète correspondant à la vraisemblance (2.7) s'écrit :

$$\begin{aligned}\ell\left(\beta, H_0(.), \varphi\right) = \sum_{i=1}^{n} \Bigg[ & d_i \log\varphi - \log\Gamma\left(\frac{1}{\varphi}\right) + \log\Gamma\left(\frac{1}{\varphi}+d_i\right) \\ & - (\frac{1}{\varphi}+d_i)\log\left(1+\varphi\sum_{j=1}^{J_i}\exp\left(H_0(T_{ij})\right)\exp(\beta^\top Z_{ij})\right) \\ & + \sum_{j=1}^{J_i}\delta_{ij}\left(\beta^\top Z_{ij} + \log h_0(T_{ij})\right)\Bigg]\end{aligned} \tag{2.8}$$

La validité de l'inférence sur $\beta$ et $\varphi$ à partir de cette vraisemblance repose sur l'hypothèse que conditionnellement à la fragilité $U_i$, la censure est indépendante des événements et non informative sur la fragilité.

Dans un cadre semi-paramétrique où le risque de base $h_0(.)$ est considéré comme un paramètre de nuisance, si les fragilités $U_i$ étaient observées, l'estimation de $\beta$ pourrait être obtenue par maximisation de la vraisemblance partielle de Cox :

$$\prod_{i=1}^{n}\prod_{j=1}^{J_i}\left\{\frac{\exp(\beta^\top Z_{ij})U_i}{\sum_{l\in R(T_i)}\exp(\beta^\top Z_l)U_l}\right\}^{\delta_{ij}}.$$

D'où la log-vraisemblance partielle :

$$\ell_{p,1}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\left[\log(U_i) + \beta^\top Z_{ij} - \log\left(\sum_{l\in R(T_i)}\exp(\beta^\top Z_l)U_l\right)\right] \tag{2.9}$$

#### 2.2.1.2 Estimation des paramètres par l'algorithme EM

Plusieurs algorithmes ont été proposés pour estimer les différents paramètres d'intérêts qui sont: les coefficients $\beta$, la fonction de risque de base $h_0(.)$ considérée comme un paramètre de nuisance et la variance $\varphi$ des effets aléatoires,

ou pour prédire les effets aléatoires. Dans le cadre du modèle de fragilité de Cox, l'algorithme le plus utilisé est l'algorithme EM (Estimation-Maximization). L'algorithme EM introduit par [29] permet d'obtenir des estimations par maximisation de la vraisemblance dans le cas de données incomplètes. L'utilisation de l'algorithme EM dans le cadre des modèles à fragilité gamma a été proposée par [45] et développée par [87, 70, 50]. [87] conjecturent que les estimateurs ont les propriétés habituelles des estimateurs obtenus par maximum de vraisemblance. Cela a été démontré par la suite par [85]. Les fragilités individuelles peuvent être vues comme des données manquantes, d'où l'utilité de l'algorithme EM. A l'étape M, les fragilités sont remplacées par les valeurs calculées à l'étape E et la vraisemblance partielle est maximisée comme si les fragilités étaient connues.

Des valeurs initiales $\hat{\beta}^{(0)}$ et $\hat{H}_0^{(0)}(.)$ sont attribuées à $\beta$ et $H_0(.)$ respectivement via maximisation de la log-vrasemblance classique de Cox (2.9) et l'équation (2.12) avec $U_i = 1$.

**1. Etape E.** $U_i$ et $\log(U_i)$ sont remplacés par leurs espérences conditionnelles aux données ($\beta$, $H_0(.)$), pour des valeurs courantes de $\beta$ et de $h_0$. C'est-à-dire,

$$\mathbb{E}\left(U_i\right) = \frac{\frac{1}{\varphi} + d_i}{\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(\hat{H}_0(T_{ij})\right) \exp\left(\hat{\beta}^\top Z_{ij}\right)}, \tag{2.10}$$

d'où à la $k^{ieme}$ iteration

$$\mathbb{E}^{(k)}(U_i) = \frac{\frac{1}{\varphi} + d_i}{\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(\hat{H}_0^{(k-1)}(T_{ij})\right) \exp\left(\hat{\beta}^{\top(k-1)} Z_{ij}\right)}$$

Les fragilités $U_i$ étant distribuées selon une loi gamma conditionnellement aux observations, de paramètre de forme $(\frac{1}{\varphi} + d_i)$ et de paramètre d'échelle $\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(\hat{H}_0(T_{ij})\right) \exp(\hat{\beta}^\top Z_{ij})$, ce qui correspond à l'éspérence en (2.10), en s'appuiyant sur le calcul des moments et cumulants de la statistique suffisante on peut montrer que

$$\mathbb{E}\left(\log(U_i)\right) = \frac{\Gamma'\left(\frac{1}{\varphi} + d_i\right)}{\Gamma\left(\frac{1}{\varphi} + d_i\right)} - \log\left[\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(\hat{H}_0(T_{ij})\right) \exp(\hat{\beta}^\top Z_{ij})\right], \tag{2.11}$$

d'où à la $k^{ieme}$ iteration

$$\mathbb{E}^{(k)}\big(\log(U_i)\big) = \frac{\Gamma'\left(\frac{1}{\varphi}+d_i\right)}{\Gamma\left(\frac{1}{\varphi}+d_i\right)} - \log\left[\frac{1}{\varphi} + \sum_{j=1}^{J_i} \exp\left(\hat{H}_0^{(k-1)}(T_{ij})\right)\exp\left(\hat{\beta}^{\top(k-1)}Z_{ij}\right)\right].$$

Dans cette expression, la variance $\varphi$ est, soit fixée (dans le cas où l'on utilise l'algorithme EM pour une série de valeurs fixes de $\varphi$ avant de maximiser la vraisemblance marginale profilée déterminée graphiquement ou numériquement), soit correspond à la valeur estimée $\hat{\varphi}^{(k)}$ à l'étape M en maximisant (2.5).

**2. Etape M.** Cette étape nécessite respectivement une maximisation de (2.4) et de (2.5) par rapport aux paramètres inconnus $\varphi$ et $\beta$ en remplaçant $U_i$ et $\log(U_i)$ dans (2.4) et dans (2.5) par leurs espérences calculées à l'étape E. Pour estimer $\beta$, on effectue une maximisation de la log-vraisemblance (2.4). Cette log-vraisemblance contient le paramètre de nuisance $H_0(.)$, qui va être remplacé par l'estimateur similaire à l'estimateur de Breslow, soit

$$\hat{H}_0^{(k)}(t) = \int_0^t \frac{dN_.(s)}{\sum_{i,j\in R(t)} \mathbb{E}^{(k)}(U_i)\exp\left(\hat{\beta}^{\top(k)}Z_{ij}\right)} \tag{2.12}$$

où $N_.(t)$ est le processus de comptage qui compte le nombre d'évéments de tous les sujets juste avant le temps $t$: $N_.(t) = \sum_{i,j} N_{i,j}(t)$ et $R(t)$ est l'ensembre des indices des sujets à risque juste avant le temps $t$. Le risque de base estimé $\hat{h}_0^{(k)}(t)$ correspond alors aux sauts du processus $\hat{H}_0^{(k)}(t)$.

Ces étapes sont répétées jusqu'à la convergence de l'algorithme, c'ést-à-dire $\ell\left(\hat{\beta}^{(k)}, \hat{H}_0^{(k)}(.), \hat{\varphi}^{(k)}\right)$ entre deux itérations inférieure à une valeur $\varepsilon$ fixée. Une fois la convergence obtenue, les estimateurs des variances des paramètres $\hat{\beta}$ et $\hat{\varphi}$ sont obtenus par inversion de la matrice d'information de Fisher dont le calcul ne porte plus seulement sur $I\left(\hat{\beta},\hat{\varphi}\right)^{-1}$ mais plutôt sur $I\left(\hat{\beta},\hat{H}_0(.),\hat{\varphi}\right)^{-1}$ de la log-vrasemblance marginale $\ell\left(\hat{\beta},\hat{H}_0(.),\hat{\varphi}\right)$ [6]. L'algorithme pour l'estimation est décrit dans la Figure 2.1.

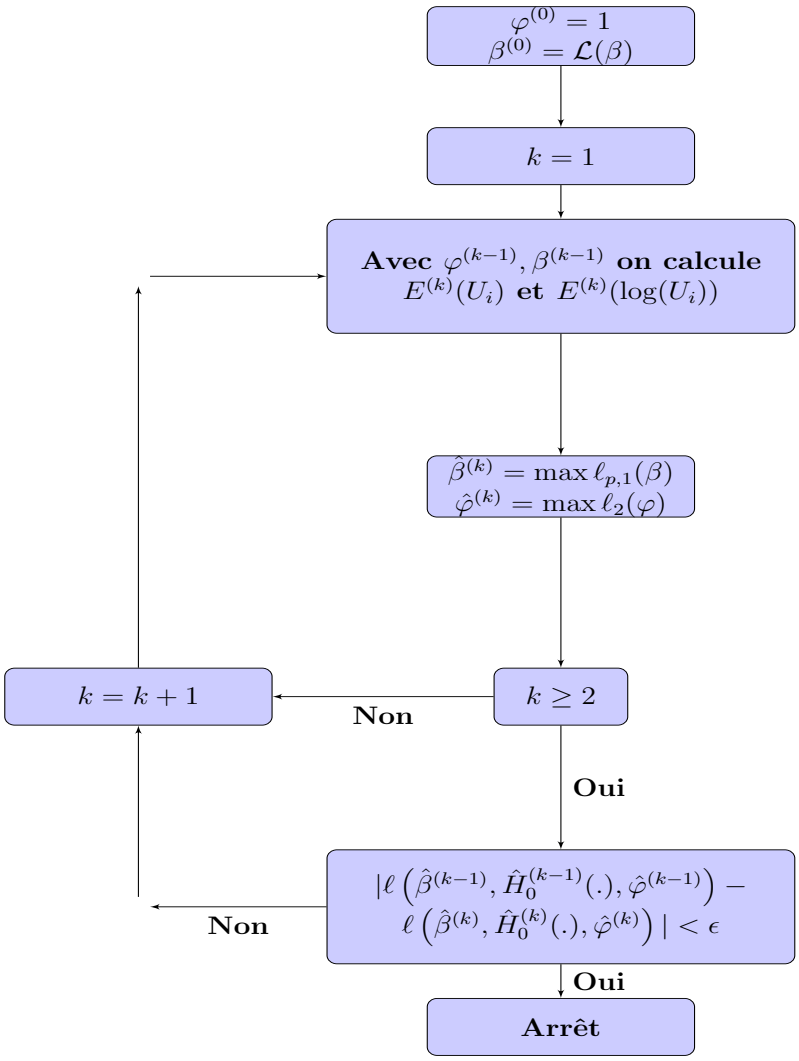



**Figure** 2.1 – Algorithme EM pour estimer des paramètres dans le modèle semi-paramètrique à fragilités partagées

#### 2.2.1.3 Application de l'Algorithme EM avec les packages R *Survival* et R *FrailtyEM*

**Descriptions des données CGD de Fleming et Harrington**

Les données sont issues d'un essai randomisé contre placebo sur le traitement par gamma-interferon ($\gamma-$IFN) dans la granulomatose septique chronique (CGD "Chronic Granulomatous disease"). Des temps d'infection récurrents sur 128 patients de 13 hopitaux différents ont été observés. Ce fichier de données est

disponible dans son intégralité sous le package R $survival$.

### Analyse des données CGD de Fleming et Harrington

Pour analyser ces données du cancer colorectal, on s'est servi du package $frailtyEM$ [11], un package R qui utilise l'algorithme général d' Espèrence-Maximisation (EM) [29] pour ajuster des modèles de fragilité partagés semi-paramétriques.

**Tableau** 2.1 – Résultats de l'estimation des paramètres du modèle semi-paramètrique à fragilités partagées par l'Agorithme EM, pour des données CGD.

| Variables | coef | exp(coef) | se(coef) | adj. se | z | p |
|---|---|---|---|---|---|---|
| sexfemale | -0.23 | 0.80 | 0.40 | 0.40 | -0.57 | 0.57 |
| treatrIFN-g | -1.05 | 0.35 | 0.31 | 0.31 | -3.39 | 0.00 |
| Effets aléatoires | 1.22 | | 0.90 | | | |

Ici, Nous omettons la partie sortie de la fragilité; la variance de fragilité estimée est de $0.820$ avec un intervalle de confiance basé sur la vraisemblance à $95\%$ de $(0.539, 4.326)$ et elle est donc significativement différente de $0$. Ceci veut dire qu'il y a une hétérogénéité entre les patients. Nous avons une valeur $p < 0,05$ pour la variable " treatrIFN-g". Ceci montre que le traitement par $\gamma-$IFN réduit significativement le risque d'infections graves chez les patients atteints de CGD. L'Illustration de fonctions de risques cumulés et fonctions de survie chez les patients sous traitement$\gamma-$IFN et du groupe Placebo peut se voir sur les figures 2.2 et 2.3.

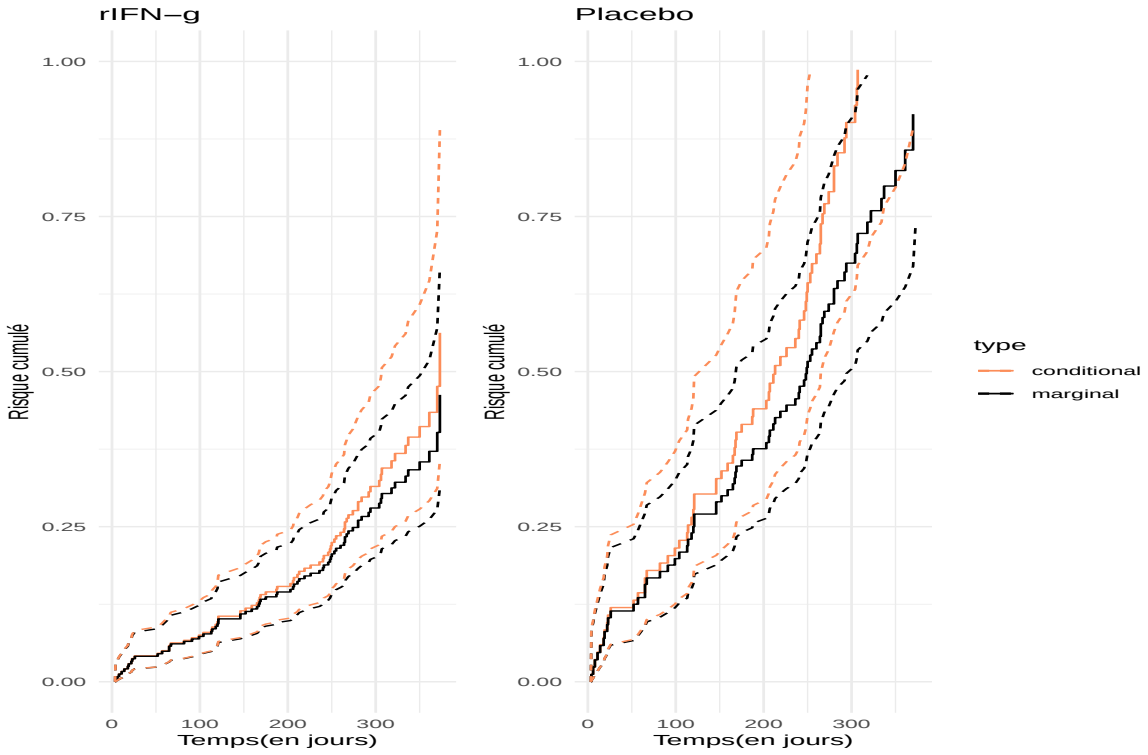



**Figure** 2.2 – Risques cumulés conditionnel (en rouge) et marginal (en noir) avec les intervalles de confiance correspondants en traits chez les hommes, l'un appartenant au groupe de traitement $\gamma$-IFN et l'autre du groupe Placebo. Ici on constate que le risque d'infections graves chez les patients atteints de CGD au cours d'une période donnée est moins élevé pour les pateints atteints de CGD sous traitement $\gamma$-IFN que les patients atteints de CGD du groupe Placebo

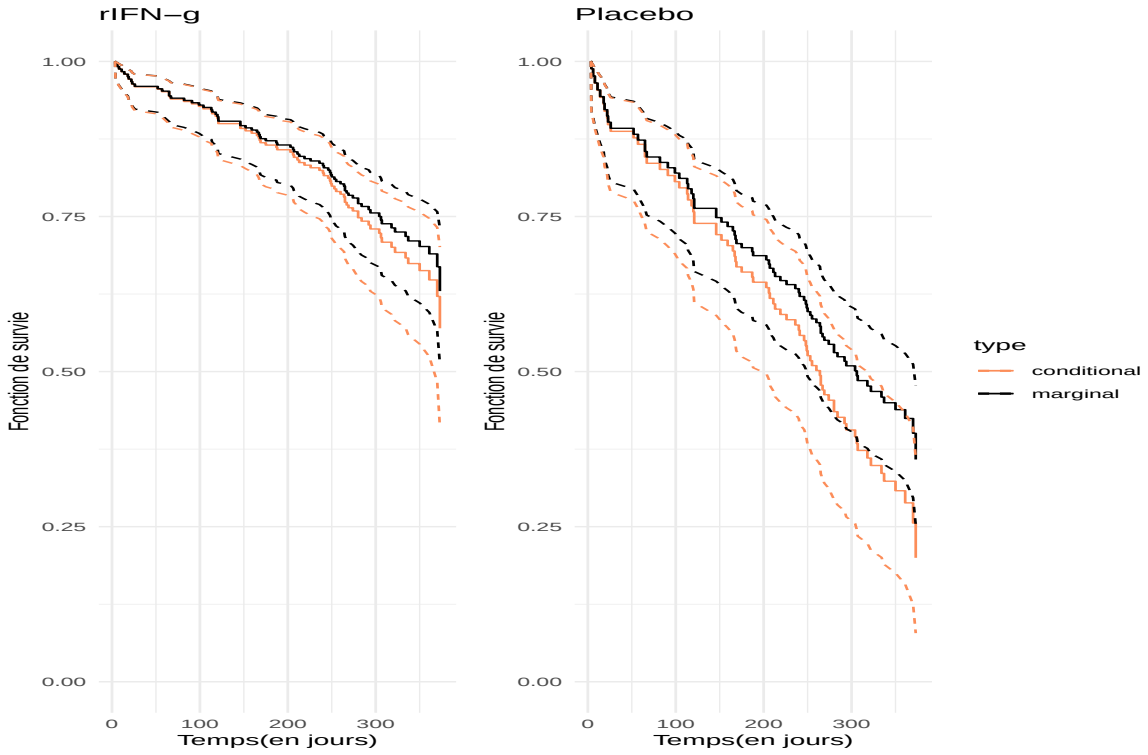



**Figure** 2.3 – Fonctions de survie conditionnelle (en rouge) et marginale (en noir) avec les intervalles de confiance correpondants en traits chez les hommes, l'un appartenant au groupe de traitement $\gamma$-IFN et l'autre du groupe Placebo dans le modèle à fragilités partagées. On constate que la probabilité de survivre jusquà un instant donné chez les patients atteints de CGD sous traitement $\gamma$-IFN est plus élevée que chez les patients atteints de CGD du groupe Placebo.

#### 2.2.1.4 Estimation des paramètres par vraisemblance partielle pénalisée

L'approche par vraisemblance pénalisée a été proposée dans un premier temps pour des fragilités de distribution lognormales par [81, 82, 94] mais est utilisable également pour des fragilités gamma [110, 32, 109]. Pour estimer les paramètres dans un modèle à fragilités gamma partagées en utilisant l'approche par vraisemblance pénalisée, il est préférable de réécrire le modèle (2.1) en remplaçant $v_i = \log(u_i)$.

$$h_{ij}(t|v_i) = h_0(t)\exp(\beta^\top Z_{ij} + v_i) \tag{2.13}$$

Pour les différencier, On appellera fragilité le terme $u_i$ et effet aléatoire le terme $v_i$. De manière similaire à ce qui a été vu dans l'EM, la log-vraisemblance complète

$$\ell\left(\beta, H_0(.), \varphi\right) = \ell_1\left(\beta, H_0(.)\right) + \ell_2\left(\varphi\right)$$

Dans le cadre d'une estimation semi-paramétrique, où $h_0(.)$ est considéré comme un paramètre de nuisance, le premier terme de cette vraisemblance peut être remplacé par une vraisemblance partielle

$$\ell_{p,1}(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \delta_{ij} \left[ v_i + \beta^\top Z_{ij} - \log \left( \sum_{l \in R(T_i)} \exp(\beta^\top Z_l + U_l) \right) \right] \tag{2.14}$$

Le second terme de cette vraisemblance correspond à la distribution des fragilités

$$\ell_2(\varphi) = \sum_{i=1}^{n} f_V(v_i) \tag{2.15}$$

Ce second terme est ici considéré comme un terme de pénalité et par la suite noté $l_{pen}(v, \varphi)$. La log-vraisemblance partielle pénalisée s'écrit alors

$$\ell_{ppl(v,\beta,\varphi)} = \ell_{p,1}(v, \beta) - l_{pen}(v, \varphi). \tag{2.16}$$

Sous l'hypothèse d'une distribution gamma de la fragilité, d'espérance 1 et de variance $\varphi$, la densité de l'effet aléatoire V est :

$$f_V(v) = \frac{(\exp(v))^{\frac{1}{\varphi}} \exp(-\exp(\frac{v}{\varphi}))}{\varphi^{\frac{1}{\varphi}} \Gamma(\frac{1}{\varphi})} \tag{2.17}$$

La procédure consiste à estimer, pour une variance $\varphi$ fixée, les effets aléatoires $v_i$ et le vecteur des paramètres associés aux covariables $\beta$ en maximisant la log-vraisemblance partielle pénalisée (2.16). Puisque le Log de la densité des effets aléatoires en (2.17) est égal $\frac{1}{\varphi}(v - \exp(v))$ plus des termes qui ne dépendent pas de $v_i$, Ils peuvent donc être omis dans le terme de pénalité, d'où dans l'expression (2.16),

$$l_{pen}(v, \varphi) = -\frac{1}{\varphi} \sum_{i=1}^{n} (v_i - \exp(v_i))$$

Le terme $-(v_i - \exp(v_i))$ a son minimum en $v_i = 0$. Dans l'estimation des effets aléatoires $v_i$ et de $\beta$ à partir de (2.16), les valeurs de $v_i$ sont donc d'autant plus pénalisées qu'elles sont éloignées de 0. La variance de l'effet aléatoire correspond au paramètre de "lissage", pré-spécifié ou estimé. C'est-à-dire qu'elle détermine l'importance de la pénalité. Si la variance est faible, le terme de pénalité est important, et donc des valeurs de l'effet aléatoire éloignées de 0 sont très pénalisées. Si la variance est élevée, le terme de pénalité contribue moins à la vraisemblance

complète, et ainsi on autorise l'effet aléatoire à s'éloigner d'avantage de 0. La procédure générale d'estimation des paramètres utilise deux boucles imbriquées l'une dans l'autre.
Pour une valeur fixée de $\varphi$, l'algorithme de Newton-Raphson permet d'estimer $\hat{\beta}(\varphi)$ et $\hat{v}(\varphi)$ en maximisant la log-vraisemblance partielle pénalisée (2.16) et fournir la valeur correspondante de $\ell_{ppl}(\hat{\beta}(\varphi), \hat{v}(\varphi), \varphi)$ .

La seconde étape consiste à estimer la variance $\varphi$ de l'effet aléatoire. Elle est estimée en maximisant la log-vraisemblance marginale profilée, obtenue de manière similaire à celle de l'algorithme EM en remplaçant dans l'expression de la log-vraisemblance marginale (2.8) le vecteur $\beta$ par l' estimation obtenue à l'étape précédente, et le risque de base cumulé de manière similaire qu'en (2.12) en remplaçant l'espérance des fragilités par sa valeur estimée. On itère entre les deux étapes jusqu'à ce que la différence entre deux valeurs successives de $\varphi$ soit inférieure à un certain seuil. [110] ont démontré que la solution obtenue dans un modèle à fragilité gamma partagée par une vraisemblance partielle pénalisée coincide avec la solution obtenue par l'algorithme EM. Une approche alternative par vraisemblance pénalisée a été proposée par [95, 97]. Cette approche repose sur la vraisemblance marginale dans laquelle le risque de base cumulé $H_0(.)$ est estimé de façon non-paramétrique et approché par des splines. L'algorithme pour l'estimation est décrit dans Figure 2.4.

#### 2.2.1.5 Illustration de l'estimation des paramètres par vraisemblance partielle pénalisée sur les données de Fleming et Harrington

Dans cette sous-section, nous illustrons l'estimation des paramètres d'intérêt par approche semi-paramétrique par vraisemblance partielle pénalisée sur les données CGD décrites en section 2.2.1.3.

**Tableau** 2.2 – Résultats de l'estimation des paramètres par vraisemblance partielle pénalisée, pour des données CGD.

| | coef | se(coef) | se2 | Chisq | DF | p |
|---|---|---|---|---|---|---|
| sexfemale | -0.22 | 0.40 | 0.33 | 0.32 | 1.00 | 0.57 |
| treatrIFN-g | -1.05 | 0.31 | 0.26 | 11.61 | 1.00 | 0.00 |
| Effets aléatoires | | 0.91 | | 56.46 | 37.58 | 0.02 |

Ici nous remarquons que cette approche conduit aux mêmes estimateurs qu'en

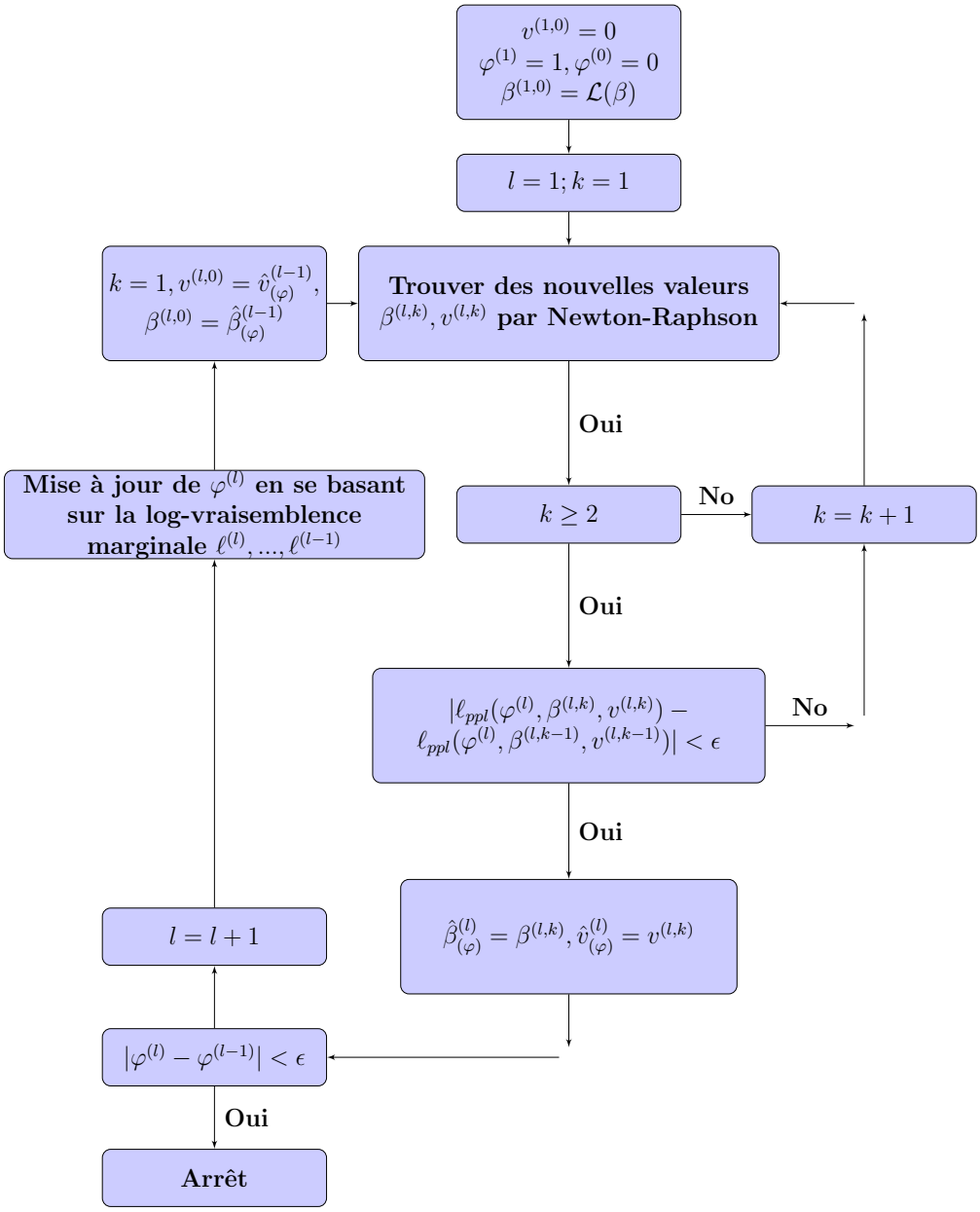



**Figure** 2.4 – Algorithme d'estimation des paramètres dans le modèle semi-paramètrique à fragilités partagées par vraisemblance partielle pénalisé

utilisant l'algorithme EM décrit dans la section 2.2.1.3 dont les résultats sont présentés dans le tableau 2.1. Nous omettons aussi la partie sortie de la fragilité; la variance de fragilité estimée est de $0.828$ et à une valeur $p < 0,05$. Ceci veut dire

qu'il y a une hétérogénéité entre les patients. Nous avons aussi une valeur $p < 0,05$ pour la variable " treatrIFN-g". Ceci montre que le traitement par $\gamma-$IFN réduit significativement le risque d'infections graves chez les patients atteints de CGD. Notons qu'en raison de la difficulté technique, la fonction "predict.coxph" dans le package "survival" n'a pas encore l'option de prédire les modèles contenant un terme de fragilité avec argument "newdata". Les fonctions de risque cumulé et de survie estimées par approche semi-paramétrique par vraisemblance partielle pénalisée sur les données CGD (résultats du tableau 2.2) sont données par la figure 2.5 .

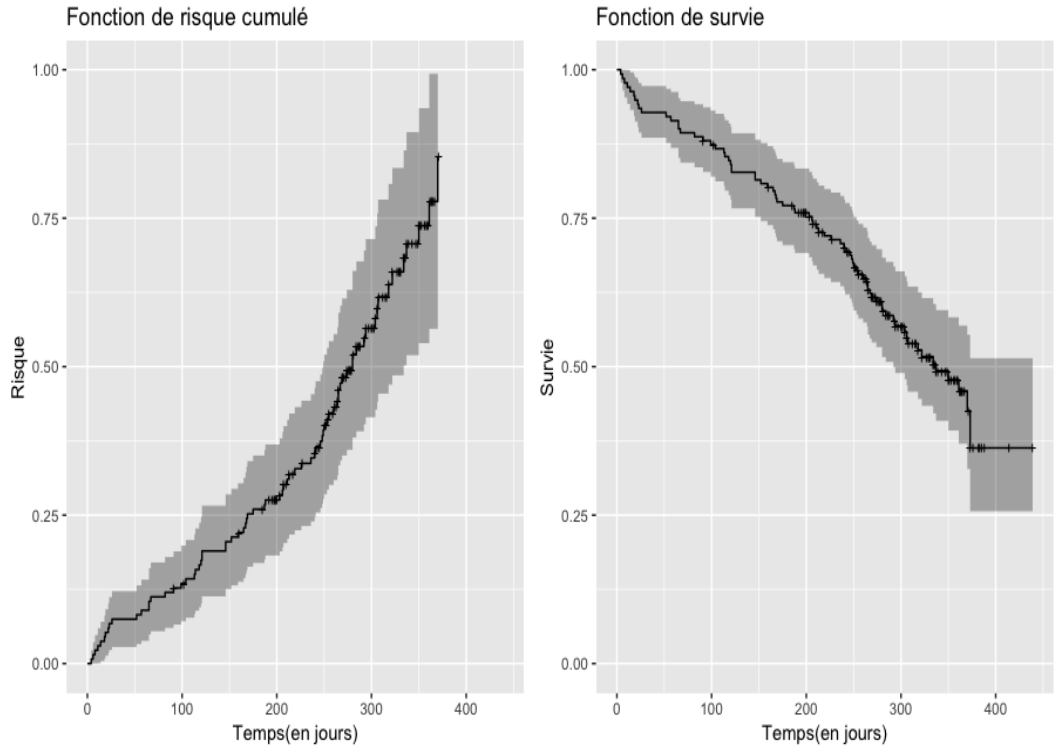



**Figure** 2.5 – Fonction de risque cumulé à gauche et fonction de survie à droite avec les intervalles de confiance estimées par approche semi-paramétrique par vraisemblance partielle pénalisée sur les données CGD dans le modèle à fragilités partagées.

La fragilité partagée explique la corrélation entre les sujets dans les groupes. Cependant, elle comporte certaines limites. Tout d'abord, elle suppose que les facteurs non observés sont les mêmes dans le groupe, ce qui ne reflète pas toujours la réalité. Par exemple, à certains moments, il peut être inapproprié de supposer que tous les partenaires dans un groupe partagent tous leurs facteurs de risque non observés.

Deuxièmement, la dépendance entre les temps de survie au sein du groupe est basée sur les distributions marginales des temps de survie. Toutefois, lorsque les covariables sont présentes dans un modèle à risques proportionnels avec fragilité distribuée selon une loi Gamma, le paramètre de dépendance et l'hétérogénéité de la population sont confondus [23]. Ceci implique que la distribution jointe ne peut être différenciée des distributions marginales [57, 58].

Troisièmement, dans la plupart des cas, une fragilité à une dimension ne peut que provoquer une association positive au sein du groupe. Toutefois, il y a certaines situations où les temps de survie des sujets, dans le même groupe sont associés négativement. Par exemple, dans l'étude de transplantation de Coeur de Stanford, plus une personne doit attendre pour un coeur disponible, moins il ou elle est susceptible de survivre après la transplantation. Par conséquent, les temps d'attente et les temps de survie peuvent être négativement associés. Pour éviter les limitations mentionnées ci-dessus les modèles de fragilité corrélée ont été développés.

## 2.2.2 Modèle à fragilité Gamma corrélée

### 2.2.2.1 Formulation du modèle

Une autre structure de modèle à fragilités a été proposée pour tenir compte à la fois d'une hétérogénéité due à des variables individuelles non observées et d'une corrélation entre certains individus. A l'origine, ce modèle à fragilitées corrélées ou modèle additif à fragilités a été développé pour l'analyse des données de temps d'événements bivariés, dans laquelle deux variables aléatoires associées sont utilisées pour caractériser l'effet de la fragilité pour chaque paire. Par exemple, une variable aléatoire est attribuée pour le partenaire 1 et une pour le partenaire 2 afin qu'ils ne soient plus contraints d'avoir une fragilité commune. Ces deux variables sont associées et ont une distribution conjointe. La connaissance de l'une d'entre elles n'implique pas nécessairement la connaissance de l'autre. Il n'y a plus de restriction sur le type de corrélation. Ces deux variables peuvent également être négativement associées, ce qui induit une association négative entre les temps de survie. Ce modèle a été developpé par plusieurs auteurs [91, 127, 89, 71] pour distinguer l'effet de facteurs de risque environnementaux par rapport aux facteurs génétiques sur la survie des jumeaux. Pour deux personnes dans une paire, les fragilités ne sont pas nécessairement les mêmes, comme dans le modèle à fragilité partagée. Nous supposons que les fragilités agissent de manière multiplicative

sur la fonction de risque de base (modèle à risques proportionnels) et que les observations d'une paire sont conditionnellement indépendantes, compte tenu de la fragilité. La formulation générale du modèle à fragilités corrélées se présente de la manière suivante pour $i = 1, 2, ..., n$ et $j = 1, 2$.

$$h(t|Z_{ij}, U_{ij}) = U_{ij} h_{0j}(t) \exp(\beta^\top Z_{ij}) \tag{2.18}$$

où $h_{0j}(t)$ est la fonction de risque de base en un temps $t$ et $U_{ij}$ est une variable aléatoire de fragilité spécifique au sujet $j$ du groupe $i$. Les modèles de fragilité à corrélation bivariée sont caractérisés par la distribution conjointe d'un vecteur bidimensionnel de fragilités $(U_{i1}, U_{i2})$. Si les deux fragilités sont indépendantes, les durées de vie résultantes le sont et aucun cluster n'est présent dans le modèle. Si les deux fragilités sont égales, le modèle de fragilité partagée est obtenu comme un cas particulier du modèle de fragilité corrélée avec une corrélation 1 entre les fragilités [123]. Afin de dériver une fonction de vraisemblance marginale, l'hypothèse de l'indépendance conditionnelle de la durée de vie, étant donné la fragilité, est utilisée. Soit $\delta_{ij}$ une indicatrice de censure pour l'individu $j(j = 1, 2)$ de la paire $i(i = 1, ..., n)$. L'indicatrice $\delta_{ij}$ est 1 si la personne a vécu l'événement d'intérêt, et 0 sinon. D'après (2.18), la fonction de survie conditionnelle de l'individu $j$ dans la paire $i$ est :

$$S(t|Z_{ij}, U_{ij}) = \exp\left(-U_{ij} H_{0j}(t) \beta^\top Z_{ij}\right) \tag{2.19}$$

avec $H_{0j}$, la fonction de risque de base cumulée. La contribution de l'individu $j(j = 1, 2)$ de la paire $i(i = 1, ..., n)$ à la vraisemblance conditionnelle est donnée par :

$$\left[U_{ij} h_{0j}(t) \exp(\beta^\top Z_{ij})\right]^{\delta_{ij}} \exp\left(-U_{ij} H_{0j}(t) \beta^\top Z_{ij}\right).$$

Le modèle à fragilités corrélées conduit à une expression de la vraisemblance plus compliquée que celle utilisée dans le modèle à fragilités partagées [89]. En supposant que la censure est indépendante et non informative pour $U$, la vraisemblance sur les données non observées est de la forme

$$\prod_{i=1}^{n} \left\{ \prod_{j=1}^{J_i} h_{ij}(t_{ij}|U_{ij})^{\delta_{ij}} \exp\left[-U_{ij} H_{ij}(t_{ij})\right] \right\} f_U(u_{i1}, u_{i2}) \tag{2.20}$$

avec $f_U(.,.)$, une densité de probabilité d'une loi de distribution de la fragilité correspondante.

[127] ont par exemple proposé d'appliquer un modèle à fragilités corrélées sur une cohorte de jumeaux. Pour cela, ils utilisent une décomposition de la variable

de fragilité en une somme de deux variables de fragilité, une partagée par plusieurs individus d'un même groupe, l'autre étant non partagée. La structure de corrélation additive est: $U_{i1} = U_{i0} + U_{i1}$ et $U_{i2} = U_{i0} + U_{i2}$. Ces variables aléatoires dans chaque groupe sont corrélées. Par contre, les variables aléatoires $U_{i0}, U_{i1}$ et $U_{i2}$ sont indépendantes et distribuées selon des lois gamma des paramètres $(k_0, \lambda_0), (k_1, \lambda_1), (k_2, \lambda_2)$ respectivement. Nous ne mentionnons pas toute la théorie découlant de ces hypothèses, le lecteur intéressé par les détails de la théorie peut consulter le travail de [51]. Seulement, il résulte de ce modèle la distribution de survie bivariée de la forme

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-\frac{\sigma_1}{\sigma_2}\rho} S_2(t_2)^{1-\frac{\sigma_2}{\sigma_1}\rho}}{\left(S_1(t_1)^{-\sigma_1^2} + S_2(t_2)^{-\sigma_2^2} - 1\right)^{\frac{\rho}{\sigma_1\sigma_2}}}, \tag{2.21}$$

avec

$$\mathbb{E}\left(U_{i1}\right) = 1, \mathbb{V}ar\left(U_{i1}\right) = k_0 + k_1 = \frac{1}{\lambda_1} = \sigma_1,$$

$$\mathbb{E}\left(U_{i2}\right) = 1, \mathbb{V}ar\left(U_{i2}\right) = k_0 + k_2 = \frac{1}{\lambda_2} = \sigma_2,$$

et

$$\rho = \frac{Cov(U_{i1}, U_{i2})}{\sqrt{\mathbb{V}ar\left(U_{i1}\right)\mathbb{V}ar\left(U_{i2}\right)}} = \frac{k_0}{\sqrt{(k_0 + k_1)(k_0 + k_2)}}$$

représente la corrélation entre le jumeau 1 et le jumeau 2. L'intervalle possible de la corrélation entre les fragilités dépend des valeurs de $\sigma_1$ et $\sigma_2$.

$$0 \leq \rho \leq \min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right). \tag{2.22}$$

Si les deux fragilités sont égales, le modèle de fragilité partagée est obtenu comme un cas particulier où $\sigma_1 = \sigma_2 = \sigma$, $\rho = 1$,

$$S(t_1, t_2) = \left(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1\right)^{\frac{-1}{\sigma^2}} \tag{2.23}$$

D'autres modèles de fragilité Gamma corrélée ont été appliqués au cas des jumeaux [132, 124]. L'algorithme EM peut être utilisé également pour estimer les paramètres d'intérêt selon la même procédure décrite pour les modèles à fragilités partagées.

### 2.2.3 Modèle à fragilités emboitées

L'analyse de données de survie groupées par des modèles à fragilités partagées est très utile, avec des domaines d'application multiples. Cependant, ces modèles ne sont restreints qu'à un seul niveau de regroupement des données. De plus en plus d'études épidémiologiques ont un schéma de regroupement des données plus complexe, avec plusieurs niveaux de regroupement. Les modèles à fragilités emboitées sont particulièrement intéressants dans des cohortes où les données sont naturellement regroupées en clusters avec deux niveaux hiérarchiques. Par exemple le regroupement des familles dans les différentes villes, ou par le schéma d'étude, c'est-à-dire des données collectées selon plusieurs niveaux de regroupement (différents quartiers nichés dans des villes). Cette approche peut être également intéressante dans l'analyse de données de survie récurrentes et groupées.

Dans les modèles à fragilités partagées, en ignorant des composantes aléatoires (ou des niveaux de regroupement des données) on risque d'obtenir des résultats non valides. Des développements ont donc été proposés pour permettre l'analyse de données de survie hiérarchiques à plusieurs niveaux de regroupements, ils concernent les modèles à fragilités emboitées. Les données tronquées à gauche et censurées à droite sont autorisées dans ce modèle.

#### 2.2.3.1 Formulation du modèle

Nous condidérons $n$ groupes indépendants des individus et dans chaque groupe $i = 1, ..., n_i$ il y a $J_i$ sous-groupes des individus. Soient $X_{ijk}$, $C_{ijk}$ et $L_{ijk}$ les temps de survie, de censure à droite et de troncature à gauche respectivement pour le $k^{ime}$ individu ($k = 1, ..., K_{ij}$) du sous-groupe $j(j = 1, ..., J_i)$ et du groupe $i(i = 1, ..., n)$. Les temps observés sont $T_{ijk} = \min(X_{ijk}, C_{ijk})$ et les indicateurs de censure sont $\delta_{ijk}$. Nous supposons que $X_{ijk}, C_{ijk}, L_{ijk}$ soient indépendants. Dans le modèle à fragilités emboitées, la fonction de risque conditionnelle aux deux effets aléatoires $U_i$ et $V_{ij}$ pour l'individu $k(k = 1, ..., K_{ij})$ du sous-groupe $j(j = 1, ..., J_i)$ et du groupe $i(i = 1, ..., n)$ est :

$$h_{ijk}\left(t|Z_{ijk}, U_i, V_{ij}\right) = U_i V_{ij} h_0(t) \exp\left(\beta^\top Z_{ijk}\right) \tag{2.24}$$

avec $h_0(t)$ la fonction de risque de base, $Z_{ijk}$ covariable associée à l'individu $k$ du groupe $i$ et sous-groupe $j$. L'effet aléatoire $U_i$ du groupe $i$ et l'effet aléatoire $V_{ij}$ du sous-groupe $j$ sont tous les deux distribués de façon indépendante et identique avec une loi gamma. C'est-à-dire :

$U_i \overset{i.i.d}{\to} \Gamma\left(\frac{1}{\gamma}, \frac{1}{\gamma}\right)$, avec $\mathbb{E}(U_i) = 1$ et $\mathbb{V}ar(U_i) = \gamma$

$V_{ij} \overset{i.i.d}{\to} \Gamma\left(\frac{1}{\eta}, \frac{1}{\eta}\right)$, avec $\mathbb{E}(V_{ij}) = 1$ et $\mathbb{V}ar(V_{ij}) = \eta$

Si $\eta$ est nul, les observations du même sous-groupe sont indépendantes et si $\gamma$ est nul, les observations du même groupe sont indépendantes. Une variance plus importante implique une plus grande hétérogénéité de la fragilité d'un groupe à l'autre et une plus grande corrélation des temps de survie des individus appartenant au même groupe. La log-vraisemblance complète dans le modèle à fragilité gamma emboitée prend la forme suivante [96]

$$\begin{aligned}\ell\left(h_0(.), \beta, \eta, \gamma\right) = \sum_{i=1}^{n} \Bigg\{ \sum_{j=1}^{J_i} \Bigg\{ \sum_{k=1}^{K_{ij}} \delta_{ijk} \left[\beta^\top Z_{ijk} + \log\left(h_0(T_{ijk})\right)\right] \\ + \mathbb{I}_{\{A_i>1\}} \sum_{k=1}^{A_{ij}} \log\left(1 + \eta(A_{ij} - k)\right) \Bigg\} \\ + \log \int_0^\infty \frac{U_i^{\frac{1}{\gamma}-1+A_i} \exp(-\frac{U_i}{\gamma})}{\prod_{j=1}^{n_i} \left[1 + \eta U_i \sum_{k=1}^{K_{ij}} H_{ijk}(L_{ijk}) + 1\right]^{\frac{1}{\eta}+A_{ij}}} dU_i \\ - \log \int_0^\infty \frac{U_i^{\frac{1}{\gamma}-1} \exp(-\frac{U_i}{\gamma})}{\prod_{j=1}^{n_i} \left[\eta U_i \sum_{k=1}^{K_{ij}} H_{ijk}(L_{ijk}) + 1\right]^{\frac{1}{\eta}}} dU_i \Bigg\}\end{aligned} \tag{2.25}$$

où $H_{ijk}(.)$ est la fonction de risque cumulé de base et $A_i = \sum_{j=1}^{n_i} \sum_{k=1}^{K_{ij}} \mathbb{I}_{\{\delta_{ijk}=1\}}$ est la somme des événements dans le groupe $i$.

Les deux niveaux hiérarchiques de regroupement des données sont pris en compte dans les modèles en incluant deux effets aléatoires emboités. Cependant, les difficultés numériques rencontrées pour étendre les modèles à fragilités partagés ont limité leur développement. Les paramètres peuvent être estimés en utilisant un algorithme EM [100] ou une approche bayésienne [33]. Une méthode d'estimation non paramétrique par vraisemblance pénalisée pour estimer la fonction de risque dans les modèles à fragilités emboitées sur des données de survie censurées à droite et tronquées à gauche a été proposée [96]. Contrairement aux modèles à fragilités partagées, dans les modèles à fragilités emboitées, l'expression de la vraisemblance marginale contient des intégrales non analytiques sur des effets aléatoires du niveau de regroupement le plus haut.

#### 2.2.3.2 Illustration du modèle

Ici, en utilisant le package R $frailtypack$, nous illustrons l'utilisation du modèle à fragilités emboitées sur une base des données [37] décrite dans la section 2.2.1.3. Les données sont issues d'un essai randomisé contre placebo sur le traitement par gamma-interferon ($\gamma-$IFN) dans la granulomatose septique chronique (CGD "Chronic Granulomatous disease"). Des temps d'infection récurrents sur $128$ patients de $13$ hopitaux différents ont été observés. Le nombre de patients par hôpital varie entre 4 et 26. Les résultats exposés dans le tableau 2.3 comparent des modèles à fragilités partagées standards à des modèles à fragilités emboitées qui tiennent compte de deux effets aléatoires. Les résultats sont également exposés pour deux choix de temps de base (en"gap time" ou "calendar time"). En général, toutes les analyses montrent que le traitement par $\gamma-$(IFN) réduit significativement le risque d'infections graves chez les patients atteints de CGD.

**Tableau** 2.3 – Analyse d'événements récurrents avec un modèle à fragilités emboitées pour des données CGD.

| Variables | Modèle à fragilités partagés | | Modèle à fragilités emboitées |
|---|---|---|---|
| | Niveau Hôpital | Niveau patient | |
| | $\hat{\beta}$(S.E) | $\hat{\beta}$(S.E) | $\hat{\beta}$(S.E) |
| $\star$Temps de base"calendrier" | | | |
| Traitement ($\gamma-$IFN) | -1.112(0.262) | -1.059(0.309) | -1.058(0.308) |
| Var$(U_i) = \gamma$ (hôpital) | 0.12(0.13) | - | 0.007(0.003) |
| Var$(V_{ij}) = \eta$ (patient) | - | 0.803(0.391) | 0.802(0.391) |
| Log-vraisemblance pénalisée | -526.63 | -522.58 | -510.99 |
| $\star$Temps de base"gap" | | | |
| Traitement ($\gamma-$IFN) | -1.124(0.268) | -1.166(0.355) | -1.166(0.355) |
| Var$(U_i) = \gamma$ (hôpital) | 0.145(0.141) | - | 0.007(0.003) |
| Var$(V_{ij}) = \eta$ (patient) | - | 1.579(0.707) | 1.577(0.705) |
| Log-vraisemblance pénalisée | -532.38 | -525.54 | -514.09 |

Ici on remarque que, quelque soit l'approche utilisée (Calendrier ou en gap), toutes les analyses montrent que le traitement par $\gamma-$IFN réduit significativement le risque d'infections graves chez les patients atteints de la CGD. Par exemple, dans le modèle à fragilités emboitées, en "gap time", nous obtenons

$RR = 0.311, IC_{95}(0.16 - 0.63)$. Puisqu'une seule partie de la structure de corrélation des données est considérée dans le modèle simple à fragilités partagées, cela fait que la variance des effets aléatoires spécifiques aux hôpitaux ($U_i$) estimée à $0.145$ est supérieure à la variance des effets aléatoires spécifiques des hôpitaux ($U_i$) estimée à $0.007$ dans le modèle à fragilités emboitées où les deux structures de corrélation sont prises en compte.
Les modèles à fragilités emboitées suggèrent que la variance entre patients ($V_{ij} = 1.577$) est largement supérieure à la variance entre hopitaux ($U_i = 0.007$).

## 2.3 Conclusion

Ce chapitre introduit la notion de la fragilité ainsi que ses différents domaines d'application. La fragilité est une variable aléatoire qui permet, entre autres, de prendre en compte l'hétérogénéité des individus issus d'un même groupe par rapport à d'éventuels sous-groupes homogènes. Cette hétérogénéité est considérée comme non observable et elle reflète souvent des facteurs environnementaux ou des facteurs génétiques. Les groupes d'individus peuvent être des familles, les patients d'un même hôpital, les gens de la même ville ou par exemple un événement observé avec répétition pour une même personne. La génétique est l'un des domaines d'application le plus prometteur des modèles à fragilités et l'hétérogénéité inexpliquée d'origine géographique est le plus faible et souvent non detectable [95]. Les modèles à fragilités sont intéressants pour modéliser la durée de survie en tenant compte de l'hétérogénéité de la population qui ne peut pas être expliquée par des facteurs connus et mesurables. L'hétérogénéité peut être présente même si les observations sont indépendantes; cependant, c'est dans le cas où l'on a des données corrélées que l'on peut valablement estimer des paramètres mesurant cette hétérogénéité. Dans le modèle à fragilités, la variance de la variable de fragilité est un indice direct de l'hétérogénéité. Dans ce chapitre, un accent particulier a été mis sur le modèle à fragilité partagé gamma ainsi que certaines extensions de ce modèle : modèle à fragilités corrélées et modèle à fragilités emboitées. En effet, cette distribution possède de bonnes propriétés. Un de ces avantages réside dans ses facilités calculatoires.

Malgré la popularité de modèles à fragilités gamma partagées aujourd'hui, les modèles utilisant d'autres distributions de fragilité peuvent être utilisés pour modéliser cette fragilité : Positive Stable, Log-normale,... Le choix de la distribution

pour l'effet aléatoire est un problème majeur dans les modèles de fragilité. Même si chaque distribution possède ses atouts, dans la plupart des analyses effectuées dans la littérature, il ressort que la distribution gamma est particulièrement satisfaisante [59]. C'est pourquoi nous n'utilisons que celle-ci pour modéliser la fragilité dans les divers modèles de cette thèse. Le modèle à fragilités gamma partagées formulé mathématiquement dans la section (2.2.1) a été illustré sur une base de données CGD de Flemming et Harrington décrite dans la sous-section (2.2.1.3). L'approche de l'algorithme EM et l'approche de la vraisemblance partielle pénalisée ont été utilisées. La formulation mathématique du modèle à fragilité gamma corrélée a été présentée dans la section (2.2.2) et suite à l'unique disponible package "R frailtypack" à notre connaissance pour le moment, qui ne prend que les modèles à fragilités corrélées suivant une loi gamma, on s'est uniquement limité à la présentation d'une formulation mathématique du modèle à fragilité corrélées suivant une loi gamma. Le modèle à fragilités emboitées formulé mathématiquement dans la section (2.2.3) a été illustré sur une base CGD de Flemming et Harrington décrite dans la sous-section (2.2.1.3). Ici, on signale que quelques exemples sur l'estimation de la fonction de risque de base dans les modèles de fragilités ont été graphiquement présentés. Quelle que soit l'approche utilisée dans l'estimation des paramètres d'intérêt, l'introduction de la notion de fragilité dans la modélisation classique de durées de survie produit des résultats qui reflètent la réalité.

CHAPITRE 3

# Estimations pénalisées dans les modèles de durées de vie censurées


## Résumé

*Dans ce chapitre, nous présentons un certain nombre de méthodes de régularisation dans le modèle à risques proportionnels de Cox. Les approches présentées dans ce chapitre sont basées sur l'estimation pénalisée : la régression Ridge, la régression Lasso (Least Absolute Shrinkage and Selection Operator), la régression adaptive Lasso, la régression Elastic-net, la régression Scad (Smoothly Clipped Absolute Deviation) et la régression Mcp (Minimax Concave Penalty). Nous présentons ensuite l'illustration de ces méthodes sur une base de données censurées en utilisant l'algorithme "coordinate descente".*

## Sommaire

## 3.1 Introduction

L'avènement des données génomiques, protéomiques, métaboliques, mais aussi celles issues de l'imagerie médicale ou décrivant l'historique des prescriptions médicamenteuses, ouvre de nouvelles perspectives dans la modélisation statistique. Plusieurs modèles ont ainsi été développés autour de ces nouvelles sources d'information [80]. Cependant, ces données posent également de nouvelles questions d'un point de vue méthodologique. D'une part, du point de vue de la qualité de l'estimation, la plupart des procédures statistiques classiques souffrent du fléau de la dimension [46]. Les modèles de régression paramétriques, par exemple, ont des performances prédictives détériorées lorsqu'ils sont estimés à partir d'un grand nombre de covariables. Or ces performances prédictives sont cruciales, notamment dans le cas des modèles pronostiques. D'autre part, du point de vue de l'interprétation, on cherche à travers ces modèles à déterminer quelles covariables sont effectivement associées à la variable d'intérêt (par exemple pour mieux comprendre les mécanismes biologiques en jeu). L'identification des variables pertinentes est cependant d'autant plus difficile que le nombre de variables "candidates" est grand. Ainsi, les données de grandes dimensions disponibles aujourd'hui posent naturellement la question de la sélection des variables pertinentes, tant pour l'interprétation des modèles obtenus que pour leur garantir de bonnes performances prédictives.

Le problème de la sélection de variables (voire plus généralement de la sélection de modèle) est un des axes de recherche majeur en statistique. Parmi les procédures classiques de sélection de variables figurent celles qui reposent sur la minimisation de critères pénalisés. Un exemple bien connu est le BIC [102] pour lequel la consistance en sélection de variables est garantie sous certaines conditions [69]. Cependant, ce critère reposant sur la "norme" $l_0$ des paramètres n'est pas convexe et sa résolution numérique est dite combinatoire : en général, il n'existe pas d'autres stratégies que celle consistant à calculer le BIC pour l'ensemble des $2^p$ modèles possibles. Dès que $p \geq 30$, il n'est pas raisonnable de construire les $2^p$ modèles et on le combine le plus souvent à des techniques heuristiques qui permettent de ne parcourir qu'un sous-ensemble de ces $2^p$ modèles. Les plus utilisées en épidémiologie et en recherches cliniques sont les approches "gloutonnes" dites pas-à-pas (*stepwise* en anglais), qui peuvent être ascendantes, descendantes, voire hybrides [55].

On parle généralement de grandes dimensions quand le nombre total de va-

riables explicatives est du même ordre de grandeur ou est supérieur au nombre d'individus. En d'autres termes, il y a trop de variables pour pouvoir directement appliquer un modèle de régression de Cox.

Nous partons du postulat qu'il existe un modèle incluant un petit nombre de variables explicatives permettant de bien prédire la variable réponse. Cette hypothèse semble raisonnable, car elle traduit le fait que le nombre important de variables à notre disposition (dans le cas des données socio-épidémiologiques) est dû au fait que les questionnaires sont issus des experts de différents domaines (médecins, sociologues, économistes etc..) qui peuvent poser des questions très proches. Même si cette hypothèse n'est pas verifiable, le fait de décrire les données par un nombre restreint de variables permet d'avoir des modèles interprétables. Plusieurs stratégies de réduction de dimension existent. Nous présentons ici celles basées sur la minimisation d'un contraste pénalisé qui soient simples à résoudre numériquement tout en renvoyant des estimateurs présentant de bonnes propriétés statistiques [111, 35, 21, 19, 46]. Nous nous restreignons ici sur les penalités Ridge, de type Lasso (ou $l_1$), adaptive lasso, elastic net, Group Lasso et de Adaptive Group Lasso, que nous allons maintenant décrire.

## 3.2 Méthodes de régulation dans le modèle de Cox

### 3.2.1 La régression Ridge

La recherche du meilleur estimateur au sens du compromis biais-variance est un problème fondamental en statistique. Les approches pénalisées ont été introduites dans le but d'améliorer la qualité du modèle, en réduisant par exemple le nombre de variables. Dans le cas où chaque variable n'est liée qu'à un coefficient, la réduction du nombre de variables est équivalente à la réduction du nombre de coefficients. Pour obtenir une meilleure prédiction, [56] ont proposé la régression Ridge. Dans le cas du modèle de Cox, c'est la solution du problème pénalisé suivant :

$$\hat{\beta}_{Ridge} = \arg\ \min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \left[ \delta_i \left( \beta^\top Z_i - \log \left( \sum_{j \in R(T_i)} \exp(\beta^\top Z_j) \right) \right) \right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{3.1}$$

où $\lambda > 0$ est un paramètre de régularisation. La régression Ridge rétrécit l'estimateur des moindres carres ordinaires vers 0. Ce faisant, elle donne un esti-

mateur biaisé, mais avec une variance plus petite que celle de l'estimateur des moindres carres ordinaires. La régression Ridge permet de rétrécir les coefficients, mais elle n'en annule aucun, donc elle ne donne pas des modèles interprétables (Figure 3.1b).

### 3.2.2 La régression Lasso (Least Absolute Shrinkage and Selection Operator)

La méthode Lasso, proposée par [111], est certainement la plus populaire des méthodes de régularisation. Cette méthode minimise la log-vraisemblance partielle négative de Cox pénalisée par la norme $l_1$ des coefficients. Elle s'écrit sous la forme suivante :

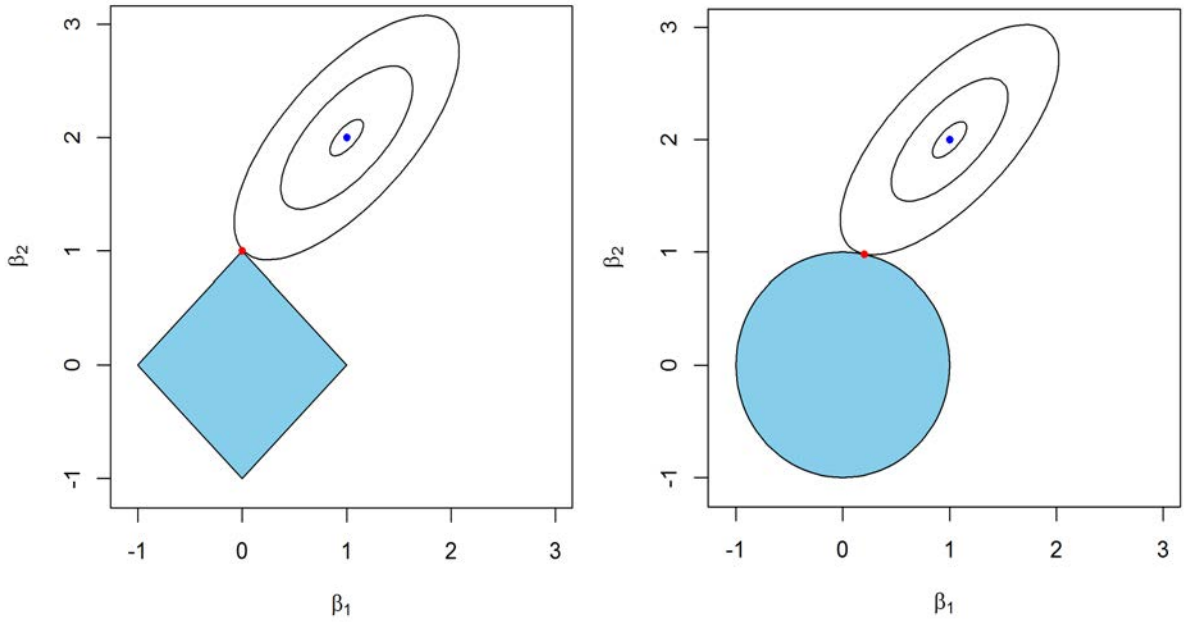
$$\hat{\beta}_{Lasso} = \arg \min_{\beta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \left[ \delta_i \left( \beta^\top Z_i - \log \left( \sum_{j \in R(T_i)} \exp(\beta^\top Z_j) \right) \right) \right] + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.2}$$

où $\lambda > 0$ est un paramètre de régularisation. Ce paramètre de régularisation a le double effet de rétrécir les coefficients de $\beta$, permettant de diminuer le biais à l'instar de la méthode Ridge, mais surtout de faire de la sélection automatique de variables, en annulant certains coefficients de $\beta$, pour des valeurs suffisamment grandes du paramètre $\lambda$ (Figure 3.1a). Même si en général il n'existe pas de forme analytique de la solution pour ce type de problème d'optimisation, des algorithmes existent pour résoudre ce problème, par exemple, l'algorithme couramment utilisé de coordinate descent introduit par [40].

La figure 3.1 permet de visualiser les solutions respectives du Lasso et du Ridge en rouge. Sur cet exemple, on constate que le coefficient $\beta_1$ solution du Lasso est contrainte à zéro. Ce n'est pas le cas de la solution du Ridge pour laquelle les deux coefficients $\beta_1$ et $\beta_2$ sont nuls. Plus généralement, le Lasso introduit de la sparsité, alors que ce n'est pas le cas du Ridge.

### 3.2.3 La régression adaptive Lasso

Dans la méthode Lasso, il est bien connu que, plus le paramètre de régularisation est grand plus le coefficient a de forte chance d'être estimé égal à zéro et inversement, plus le paramètre de régularisation est petit, plus le coefficient a de forte chance d'être estimé différent de zéro. Il est donc judicieux de pénaliser différemment les coefficients du vecteur $\beta$ : affecter aux coefficients non significatifs




(a) Lasso (b) Ridge

**Figure** 3.1 – Les solutions du Lasso et du Ridge

une pénalité considérable (un poids important) et aux coefficients significatifs une petite pénalité (un petit poids). C'est pour cela que [136] a proposé la méthode de l'Adaptive Lasso. C'est une version pondérée de Lasso classique. C'est une solution du problème pénalisé suivant :

$$\hat{\beta}_{AdaLasso} = \arg\min_{\beta} \left\{ -\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\left(\beta^\top Z_i - \log\left(\sum_{j\in R(T_i)}\exp(\beta^\top Z_j)\right)\right)\right] + \lambda\sum_{j=1}^{p}\hat{\omega}_j\,|\beta_j| \right\} \tag{3.3}$$

où $\hat{\omega} = (\hat{\omega}_1, ..., \hat{\omega}_p)^\top$ est un vecteur de poids donné par un estimateur initial et $\lambda > 0$ est un paramètre de régularisation. Elle résout le problème de consistance en sélection du Lasso, en pénalisant différemment les coefficients de $\beta$. Le problème est que l'on ne connait pas à l'avance les paramètres significatifs. En pratique on utilise généralement les poids $\hat{\omega} = \frac{1}{|\hat{\beta}_j|}$, où $\hat{\beta}_j$ est soit l'estimateur du maximum de vraisemblance soit l'estimateur ridge [136].

### 3.2.4 La regression Elastic-net

Malgré la popularité de la méthode Lasso, [137] ont constaté que le Lasso possède certaines limitations.

Le nombre de variables sélectionnées par la méthode Lasso est limité par la taille de l'échantillon, donc ce nombre ne peut pas dépasser $n$. Les résultats théoriques qui garantissent la consistance de l'estimateur Lasso portent en général sur une hypothèse de faible corrélation entre les variables. La méthode Lasso a donc de mauvaises performances en cas de forte multicolinéarité entre les variables explicatives. En effet, lorsque plusieurs variables sont fortement corrélées entre elles, on aimerait que l'ensemble de ces variables soit sélectionné. Cependant, le Lasso a tendance à sélectionner une seule variable parmi un groupe de variables très corrélées. Donc la méthode Lasso présente un problème d'identifiabilité. Toujours dans le cas d'existence d'une forte corrélation entre les variables explicatives, les performances de prédiction de la méthode Lasso ne sont pas aussi bonnes que celles de la régression Ridge.

Pour pallier ces limitations du Lasso, [137] ont proposé l'*Elastic-net*. Cette méthode combine les deux pénalités $l_1$ (norme $L1$) et $l_2$ (norme $L2$). C'est la solution du problème pénalisé :

$$\begin{aligned}\hat{\beta}_{Enet} &= \arg\min_{\beta}\Bigg\{ -\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\left(\beta^\top Z_i - \log\left(\sum_{j\in R(T_i)}\exp(\beta^\top Z_j)\right)\right)\right] \\ &\quad +\lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)\Bigg\} \qquad (3.4)\end{aligned}$$

où, $\lambda \geq 0$ et $\alpha \in [0,1]$ sont deux paramètres de régularisation. Avec un bon équilibre des deux pénalités, l'Elastic-net parvient à combiner les points forts du Lasso et du Ridge tout en minimisant leurs inconvénients. Notons que les régressions Lasso et Ridge sont deux cas particuliers de l'estimateur *Elastic-net* correspondants respectivement à $\alpha = 0$ et $\alpha = 1$.

Cette combinaison de deux pénalités a l'avantage de sélectionner les variables tout en prenant en compte les corrélations entre celles-ci. En effet le premier terme de pénalité ($l_1$) assure la sélection de variables c'est à dire la sparsité de la solution $\hat{\beta}_{Enet}$ et le second terme ($l_2$) permet de prendre en compte la corrélation entre les variables (en encourageant les variables corrélées à être sélectionnées ensemble).

### 3.2.5 La régression Scad et Mcp

Lorsque le nombre de variables $p$ est grand et que le nombre de variables pertinentes est petit, [35] et [133] ont remarqué que :

1. pour écarter les variables parasites, le paramètre de régularisation $\lambda$ du problème Lasso doit être grand, ce qui provoque un biais dans les variables retenues.
2. La diminution de $\lambda$ pour réduire le biais cause l'introduction de beaucoup de variables parasites.

Pour résoudre ces problèmes, [35] ont proposé l'estimateur Scad, tandis que [133] ont proposé l'estimateur Mcp.

#### 3.2.5.1 La regression Scad (Smoothly Clipped Absolute Deviation)

Etant donné $\lambda \geq 0$ et $\gamma > 2$, considérons la fonction définie sur $[0, \infty[$ par :

$$q_{\lambda,\gamma}(t) = \begin{cases} \lambda t, & \text{si } t \leq \lambda \\ \dfrac{\gamma\lambda t - 0.5(t^2+\lambda^2)}{\gamma - 1}, & \text{si } \lambda < t < \lambda\gamma \\ \dfrac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & \text{si } t > \gamma\lambda. \end{cases} \tag{3.5}$$

et le critère pénalisé est donné par :

$$N(\beta,\lambda,\gamma) = -\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\left(\beta^\top Z_i - \log\left(\sum_{j\in R(T_i)}\exp(\beta^\top Z_j)\right)\right)\right] + \sum_{j=1}^{p} q_{\lambda,\gamma}(|\beta_j|) \tag{3.6}$$

L'estimateur Scad est défini par :

$$\hat{\beta}_{Scad} = \arg\min_{\beta} N(\beta,\lambda,\gamma) \tag{3.7}$$

Il consiste à remplacer la pénalité $l_1$ dans le Lasso par la pénalité dans l'équation ((3.5)). Pour avoir une idée du fonctionnement de cette pénalité, il est intéressant de considérer sa dérivée :

$$q'_{\lambda,\gamma}(t) = \begin{cases} \lambda, & \text{si } t \leq \lambda \\ \dfrac{\gamma\lambda - t}{\gamma - 1}, & \text{si } \lambda < t \leq \lambda\gamma \\ 0, & \text{si } t > \gamma\lambda. \end{cases} \tag{3.8}$$

Cette pénalité permet de :

1. pénaliser plus sévèrement que Lasso les petits coefficients, conduisant a des modèles plus parcimonieux.
2. pénaliser moins les grands coefficients, ce qui réduit leur biais.

#### 3.2.5.2 La régression Mcp (Minimax Concave Penalty)

Pour $\lambda \geq 1$ et $\gamma > 1$, considérons la fonction définie sur $[0, \infty)$ par :

$$p_{\lambda,\gamma}(t) = \begin{cases} \lambda t - \dfrac{t^2}{2\gamma}, & \text{si } t \leq \lambda\gamma \\ \dfrac{1}{2}\gamma\lambda^2, & \text{si } t > \lambda\gamma. \end{cases} \tag{3.9}$$

Soit le critère de pénalité :

$$M(\beta, \lambda, \gamma) = -\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\left(\beta^\top Z_i - \log\left(\sum_{j\in R(T_i)}\exp(\beta^\top Z_j)\right)\right)\right] + \sum_{j=1}^{p} p_{\lambda,\gamma}(|\beta_j|) \tag{3.10}$$

L'estimateur Mcp est défini par :

$$\hat{\beta}_{Mcp} = \arg\min_{\beta} M(\beta, \lambda, \gamma) \tag{3.11}$$

Il consiste donc à remplacer la pénalité $l_1$ dans le Lasso par la pénalité (3.9), qui a pour dérivée :

$$p'_{\lambda,\gamma}(t) = \begin{cases} \lambda - \dfrac{t}{\gamma}, & \text{si } t \leq \lambda\gamma \\ 0, & \text{si } t > \lambda\gamma. \end{cases} \tag{3.12}$$

La pénalité Mcp commence en appliquant le même taux de pénalisation que Lasso, mais elle le relaxe d'une manière continue jusqu'à ce que $t > \gamma\lambda$ ou le taux de pénalisation devient nul.

### 3.2.6 Chemin de régularisation

Les statisticiens sont intéressés par l'obtention d'un estimateur $\hat{\beta}$ pour une grille de valeurs $\lambda$ allant d'une valeur maximale $\lambda_{max}$ pour laquelle tous les coefficients pénalisés sont nuls, jusqu'à $\lambda = 0$ ou une valeur minimale $\lambda_{min}$ pour laquelle le modèle devient trop grand ou cesse d'être identifiable(non-sparsité). Nous parlons donc de chemin de régulation : $\hat{\beta}(\lambda), \ \ \lambda_{min} \leq \lambda \leq \lambda_{max}$.

En 2004, l'algorithme LARS [34] a fourni un moyen pour calculer efficacement le chemin de régularisation du Lasso avec un coût de calcul des moindres carrés ordinaires. De 2004 jusqu'à présent, d'autres algorithmes de calcul de chemin de régulation sont apparus pour une variété de problèmes similaires au Lasso : Grouped lasso [128], support-vector machine [53], elastic-net [137], Dantzig selector

[64]. Cependant, beaucoup d'algorithmes parmi eux ne possèdent pas la propriété de la linéarité par morceaux de l'algorithme LARS, ce qui peut poser un certain nombre de problèmes de calculs.

Dernièrement, un grand intérêt a été donné à l'algorithme Coordinate Descent [106, 72, 44, 39, 125, 83]. Cet algorithme consiste à optimiser chaque paramètre séparément tout en fixant les autres et en répétant la procédure jusqu'à convergence.

## 3.3 Illustration des méthodes sur la base de données en grande dimension

Dans cette section, toutes les méthodes suivent l'approche de [40] qui donne les solutions pour une grille de 100 valeurs pour $\lambda$ qui sont equi- espacées sur l'échelle logarithmique. L'application de méthodes se fait sur une base de données de survie pour $115$ patients avec $549$ gènes exposées au cancer du sein au Pays-Bas entre les années $1984$ et $1995$ [116]. Nous sommes donc dans une situation où $p \gg n$. Dans cette partie, nous nous concentrons sur la démonstration de l'utilisation et la performance des méthodes de régularisation présentées dans cette section sur cette base de données en utilisant des librairies R.

### 3.3.1 La régression Ridge dans le modèle de Cox

La Figure 3.2 donne le chemin de régularisation de la régression Ridge, ainsi que l'erreur de validation croisée calculée par des $10 - folds$. Pour la base de données du cancer du sein, la valeur optimale du paramètre de régularisation est $\hat{\lambda} = 0.1105954$, correspondant à une erreur $MSE_{cv} = 10.05765$. L'estimateur $\hat{\beta}_{Ridge}$ est l'intersection du chemin de régularisation (courbe gauche de la figure 3.2) avec la ligne verticale d'abscisse $\hat{\lambda} = 0.1105954$. La régression Ridge permet de rétrécir les coefficients, mais elle n'en annule aucun, donc elle ne donne pas des modèles interprétables.

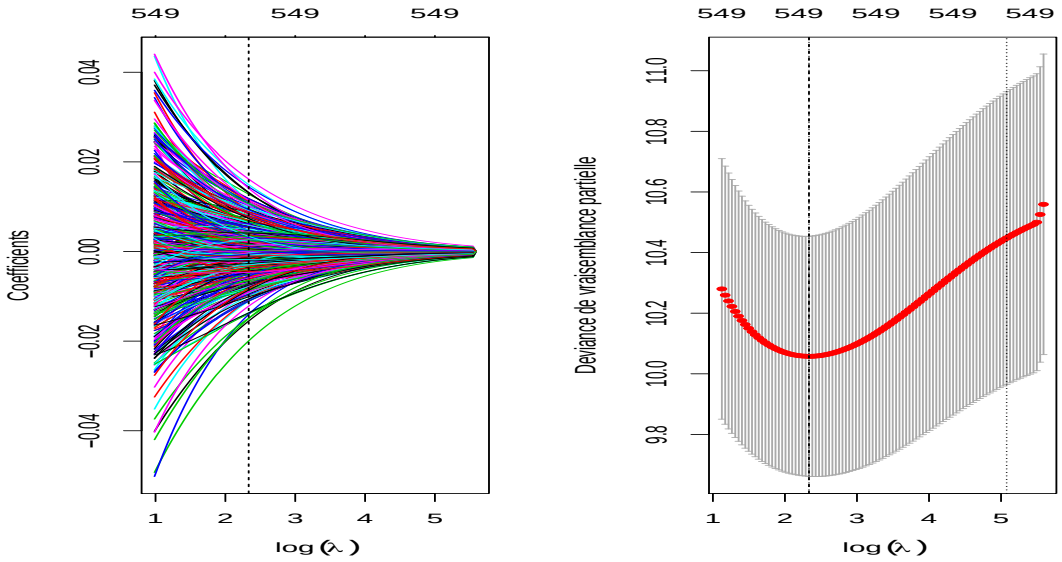



**Figure** 3.2 – La régression Ridge dans le modèle de Cox pour la base de données Breast Cancer. La courbe de gauche représente $\hat{\beta}_{Ridge}$ en fonction de $\log(\lambda)$, tandis que la courbe de droite représente l'erreur moyenne calculée par validation croisée, ainsi qu'un intervalle de confiance de la déviance de la vraisemblance partielle associée à son écart-type. La ligne verticale de gauche correspond à l'erreur minimale $mincv$, tandis que la ligne verticale de droite correspond à la plus grande valeur de $\lambda$ telle que son erreur est inférieure ou égale à $mincv + sdmincv$, où $sdmincv$ est l'écart-type de $mincv$. Le régression Ridge n'effectue pas la séléction de modèles

### 3.3.2 La méthode Lasso dans le modèle de Cox

La Figure 3.3 donne le chemin de régularisation du Lasso, ainsi que l'erreur de validation croisée calculée par des $10 - folds$, c'est-à- dire qu'on a découpé le jeu de données arbitrairement en 10 parties (folds en anglais) à peu près égales et tour à tour, chacune des 10 parties est utilisée comme jeu de test. Le reste (autrement dit, l'union des 9 autres parties) est utilisé pour l'entraînement). Pour la base de données Breast Cancer, la valeur optimale du paramètre de régularisation est $\hat{\lambda} = 0.1105954$, correspondant à une erreur $MSE_{cv} = 10.15128$. L'estimateur $\hat{\beta}_{Lasso}$ est l'intersection du chemin de régularisation (courbe gauche de la Figure 3.3) avec la ligne verticale d'abscisse $\hat{\lambda} = 0.1105954$. Pour cet exemple le Lasso produit 10 variables non nulles.

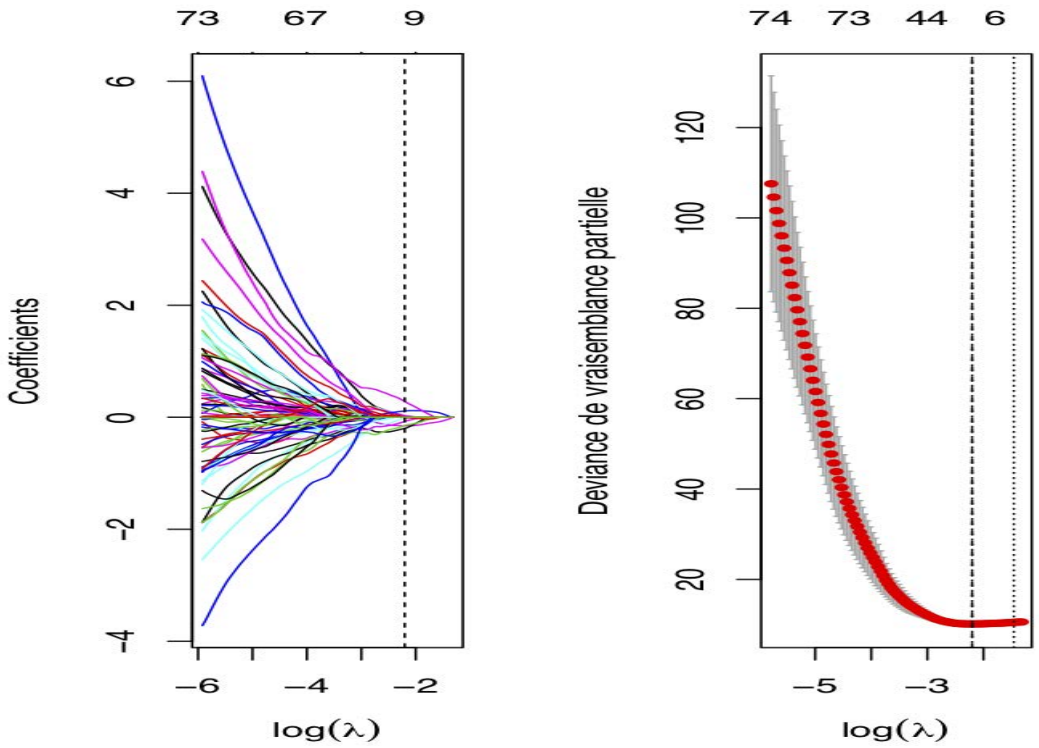



**Figure** 3.3 – La régression Lasso pour la base de données Breast Cancer. La courbe de gauche représente $\hat{\beta}_{Lasso}$ en fonction de $\log(\lambda)$, tandis que la courbe de droite représente l'erreur moyenne calculée par validation croisée, ainsi qu'un intervalle de confiance de la déviance de la vraisemblance partielle associée à son écart-type. La ligne verticale de gauche correspond à l'erreur minimale $mincv$, tandis que la ligne verticale de droite correspond à la plus grande valeur de $\lambda$ telle que son erreur est inférieure ou égale à $mincv + sdmincv$, où $sdmincv$ est l'écart-type de $mincv$. Les nombres en haut des deux courbes représentent la taille des modèles.

### 3.3.3 La méthode Adaptive Lasso dans le modèle de Cox

Pour l'Adaptive-Lasso, un choix usuel pour les poids est le suivant :

$$\hat{\omega}_j = \left( \left| \hat{\beta}_j(Lasso) + \frac{1}{n} \right| \right)^{-\gamma} \tag{3.13}$$

où $\gamma$ est un paramètre de régularisation. Nous faisons une validation croisée bidimensionnelle pour une grille de valeurs du paramètre de régularisation $\theta = (\gamma, \lambda)$. Nous choisissons la grille $\{0.01,\ 0.5, 1, 5\}$ pour le paramètre $\gamma$ et de la même manière que le Lasso, nous faisons tourner l'algorithme pour 100 valeurs de $\lambda$. La Figure 3.4 permet de visualiser les résultats obtenus pour $\gamma = 0.5,\ MSE_{cv} = 9.337551,\ \hat{\theta} = (0.5,\ 0.1147515)$ et 9 variables sont introduites dans le modèle final.

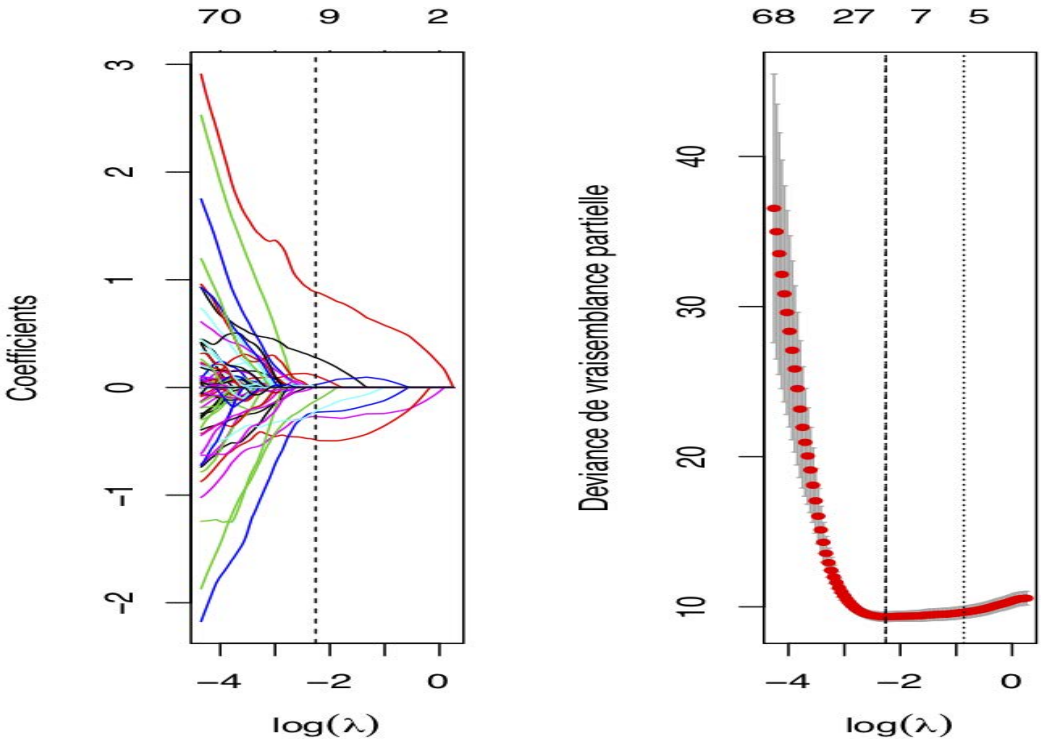



**Figure** 3.4 – La régression Adaptive Lasso dans le modèle de Cox pour la base de données Breast Cancer. La courbe de gauche représente $\hat{\beta}^{(AdaLasso)}$ en fonction de $\log(\lambda)$, tandis que la courbe de droite représente l'erreur moyenne calculée par validation croisée, ainsi qu'un intervalle de confiance de la déviance de la vraisemblance partielle associée à son écart-type. La ligne verticale de gauche correspond à l'erreur minimale $mincv$, tandis que la ligne verticale de droite correspond à la plus grande valeur de $\lambda$ telle que son erreur est inférieure ou égale à $mincv + sdmincv$, où $sdmincv$ est l'écart-type de $mincv$. Les nombres en haut des deux courbes représentent la taille des modèles. Pour cet exemple $\gamma = 0.5$.

### 3.3.4 La méthode Elastic-net dans le modèle de Cox

Pour l'Elastic-net le paramètre de régularisation est $\theta = (\alpha, \lambda)$. En pratique, nous choisissons une petite grille de valeurs de $\alpha$, à savoir $\alpha \in \{0.1, 0.5, 0.9, 1\}$. Ensuite, pour tout $\alpha$ fixé, l'algorithme coordinate descent produit le chemin de régularisation de l'Elastic-net. L'autre paramètre de régularisation $\lambda$ est choisi par validation croisée. La valeur finale de $\alpha$ est celle qui minimise l'erreur de validation croisée. La Figure 3.5 donne les chemins de régularisation de l'Elastic-net ainsi que l'erreur de validation croisée pour les 4 valeurs de $\alpha$. La valeur optimale du paramètre de régularisation est $\hat{\theta} = (\hat{\alpha}, \hat{\lambda}) = (0.5, 0.1752899)$. L'erreur de prédiction est $MSE_{cv} = 9.981516$ et le modèle final contient 28 variables.

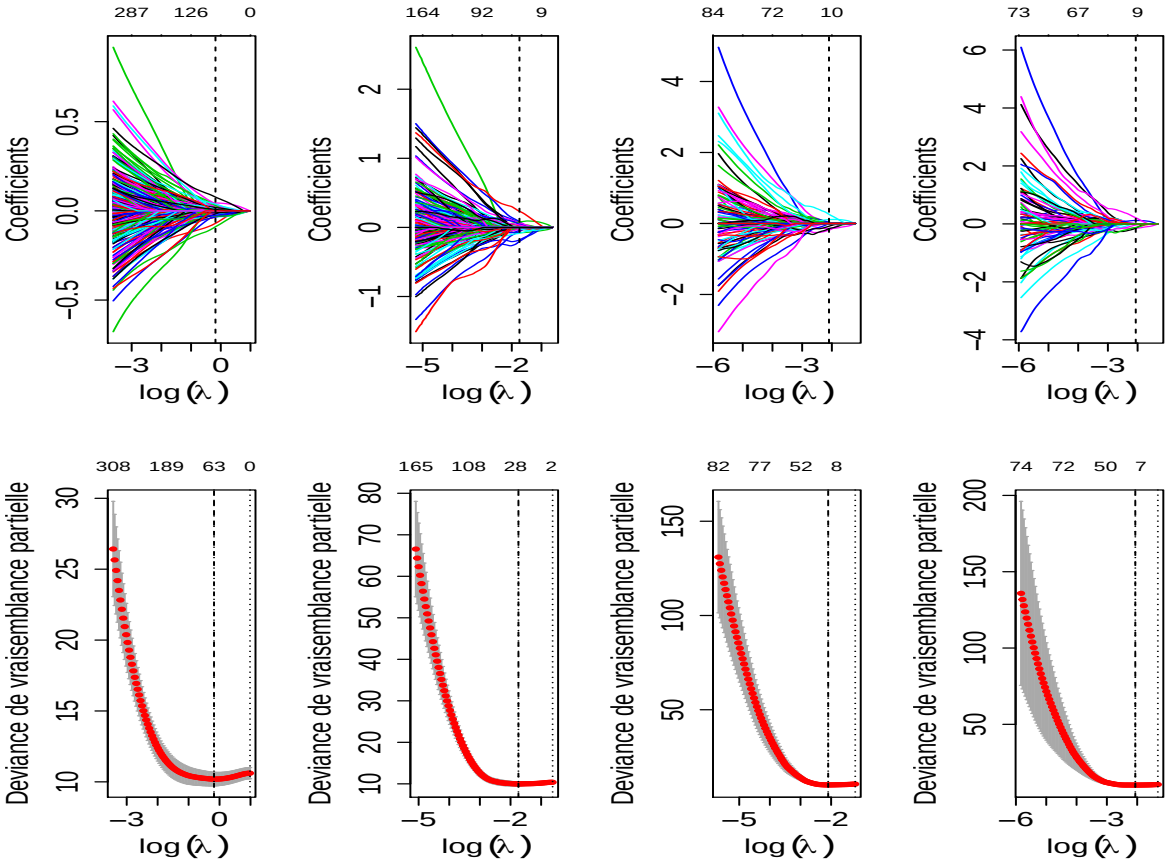



**Figure** 3.5 – La régression Elastic-net pour la base de données Breat Cancer. Les courbes du haut représentent $\hat{\beta}^{Enet}$ en fonction de $\log(\lambda)$ pour $\alpha \in \{0.1, 0.5, 0.9, 1\}$, tandis que les courbes du bas représentent l'erreur moyenne calculée par validation croisée en fonction de $\log(\lambda)$ pour les même valeurs de $\alpha$.

### 3.3.5 La regression Scad (SmoothlyClipped Absolute Deviation) dans le modèle de Cox

Le paramètre de régularisation de la régression Scad est $\theta = (\gamma, \lambda)$. Afin de trouver les meilleures performances, nous devons faire de la validation croisée bidimensionnelle. Nous faisons tourner l'algorithme pour $\gamma \in \{60, 70, 80, 90\}$ et pour 100 valeurs de $\lambda$ choisies par défaut. Les résultats obtenus sont résumés graphiquement sur la Figure 3.6. La meilleure prédiction est $MSE_{cv} = 155.6727$. Elle est atteinte pour $\hat{\theta} = (\hat{\gamma}, \hat{\lambda}) = (80, 0.1050802)$ et nous avons $14$ variables sélectionnées dans le modèle final.

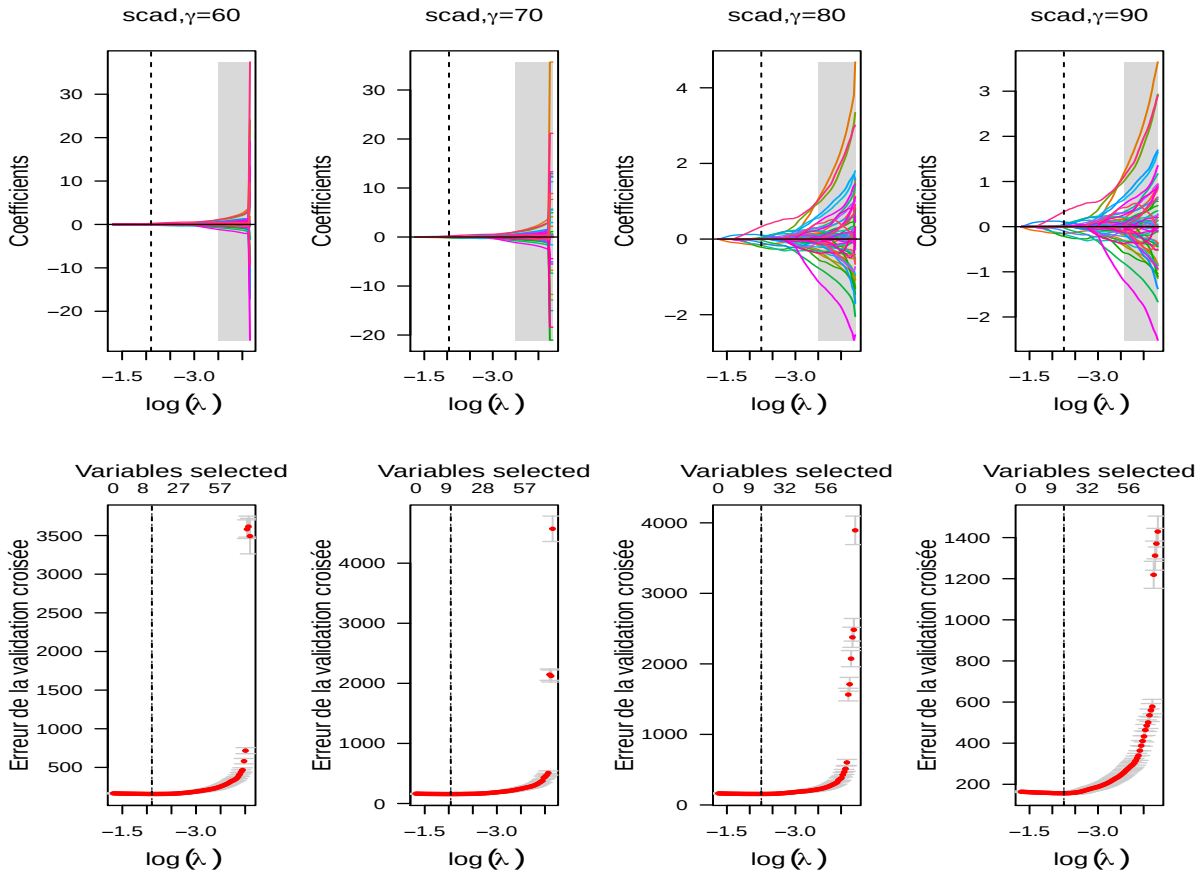



**Figure** 3.6 – La régression Scad pour la base de données Breat Cancer. Les courbes du haut représentent $\hat{\beta}^{Scad}$ en fonction de $\log(\lambda)$ pour $\gamma \in \{60, 70, 80, 90\}$, tandis que les courbes du bas représentent l'erreur moyenne calculée par validation croisée en fonction de $\log(\lambda)$ pour les même valeurs de $\gamma$.

### 3.3.6 La régression Mcp (Minimax Concave Penalty) dans le modèle de Cox

Pour la régression Mcp dans le modèle de Cox, le paramètre de régularisation est $\theta = (\gamma, \lambda)$. Nous devons donc faire de la validation croisée bidimensionnelle pour retrouver l'estimateur optimal. La Figure 3.7 montre le chemin de régularisation et l'erreur de prédiction pour $\gamma \in \{70, 80, 90, 100\}$. Il faut souligner ici que pour certaines valeurs du paramètre $(\gamma, \lambda)$, la fonction objectif peut ne pas être convexe, ce qui peut ralentir ou mme empêcher la convergence de l'algorithme. La Figure 3.7 montre les chemins de régularisation et l'erreur de prédiction pour les 4 valeurs de $\gamma$. Le paramètre de régularisation optimal est $\hat{\theta} = (\hat{\gamma}, \hat{\lambda}) = (90, 0.1050802)$, l'erreur

de prédiction est $MSE_{cv} = 155.7598$ et le nombre de variables non nulles est $14$.

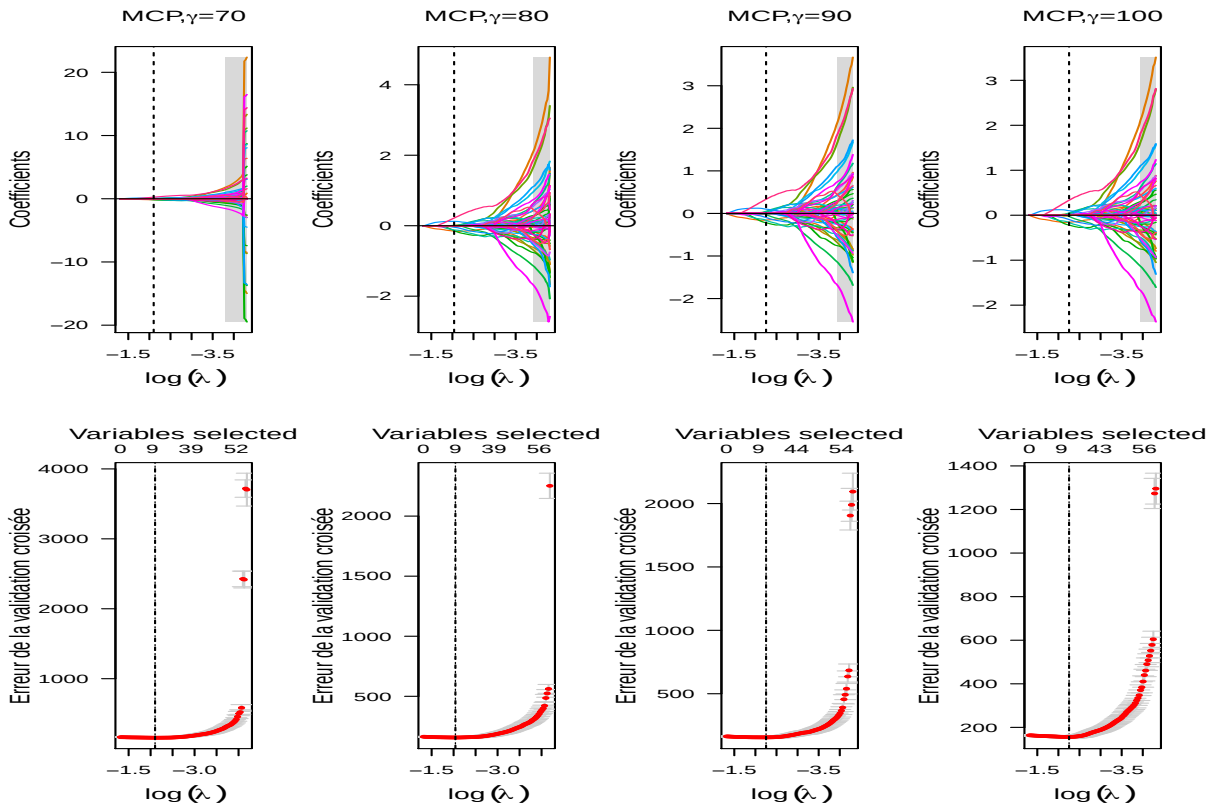



**Figure** 3.7 – La régression Mcp dans le modèle de Cox pour la base de données Breat Cancer. Les courbes du haut représentent $\hat{\beta}^{Mcp}$ en fonction de $\log(\lambda)$ pour $\gamma \in \{70, 80, 90, 100\}$, tandis que les courbes du bas représentent l'erreur moyenne calculée par validation croisée en fonction de $\log(\lambda)$ pour les même valeurs de $\gamma$.

## 3.4 Conclusion

Dans ce chapitre, nous avons vu que le choix d'un modèle et de ses paramètres repose sur la minimisation de la vraisemblance partielle de Cox négative pénalisée. Les approches pénalisées visent à contrôler la complexité d'un modèle en cherchant un compromis entre biais et variance via l'ajout d'une pénalité donnée. Nous avons présenté un certain nombre de méthodes de régularisation dans le modèle de Cox et nous avons illustré la performance de ces méthodes par un exemple de données réelles sur la survie des patients du cancer du sein en utilisant les librairies de R comme $"survival", "glmnet" et "ncvreg"$.

CHAPITRE 4

# Méthode du group Lasso dans le modèle de Cox avec fragilité


## Résumé

*Dans ce chapitre, nous présentons l'intérêt de la méthode group Lasso dans la sélection des variables groupées pour les données en grande dimension. La formulation mathématique de cette méthode se restreint dans le modèle de Cox avec fragilités partagées suivant la loi Gamma. L'algorithme de la résolution du problème d'optimisation sera présenté et la consistance de l'estimateur obtenu est démontrée. Nous discutons les différents domaines d'application de la méthode group Lasso dans le cas des données en grande dimension. Un exemple d'application sur les données simulées et un exemple sur les données réelles seront traités pour comparer les performances de la méthode proposée avec celles des méthodes concurrentes à savoir : la méthode group SCAD et la méthode group MCP.*

## Sommaire

## 4.1 Introduction

Dans le cas de variables catégorielles, les méthodes d'estimation pénalisée où les variables sont sélectionnées individuellement comme dans le Lasso [112], la procedure LARS-Cox [49], la finesse résiduelle [103], l'algorithme LARS généralisé pour le modèle à risques proportionnels de Cox [88] et l'algorithme Lasso à gradient [108] ne sont plus adaptées dans cette situation puisqu'elles sélectionnent les indicatrices des modalités et non le groupe d'indicatrices, c'est-à-dire la variable dans sa totalité. Par exemple, dans le cas de variables corrélées, le Lasso standard a tendance à ne sélectionner qu'une seule variable et ignorer toutes les autres. Dans de telles situations, il est plus judicieux d'envisager de sélectionner (ou rejeter) les variables par groupes. Cette structure des variables peut être prise en compte en utilisant l'estimateur Group Lasso introduit par [128] pour le modèle linéaire, par [83] pour le modèle logistique et [68] pour le modèle de Cox. Le Group Lasso est en effet une méthode de réduction de dimension developpée dans le but de remédier au problème de corrélation dans la régression Lasso.

Nous considérons ici le cas où les variables explicatives ont une structure en groupe qui est connue a priori, structure que l'on souhaite prendre en compte dans la procédure d'estimation. La structure en groupe des variables est présente par exemple en biologie, ou un groupe peut être constitué des variables qui partagent une même propriété biologique ou chimique. Par exemple, si 30 gènes font partie de la même voie biologique, alors ils auront les mêmes coefficients. Ils seront tous différents de 0 ou exactement égaux à 0 selon qu'ils prédisent bien ou non la variable réponse. C'est aussi le cas des variables catégorielles (nombreuses dans les données Actu-Palu), ou chacune d'entre elles est représenté dans la matrice du *design* par un groupe d'indicatrices de modalités.

Cette méthode considère les groupes de variables au lieu de variables individuelles de la façon suivante: notons $(G_l)_{l=1,\cdots,g}$ une partition de $\{1,\cdots,d\}$ en $g$ groupes de $L_j$ coefficients, où chaque groupe est potentiellement constitué d'un nombre différent de coefficients. Pour tout $\beta \in \mathbb{R}^d$, on note $\beta = (\beta_1,\cdots,\beta_d) = (\beta^1,\cdots,\beta^g)$. On note $\beta_{j,l}$ le l-ième coefficient du groupe $j$. L'estimateur group Lasso $\hat{\beta}_{GL}$ est défini par :

$$\hat{\beta}_{GL} \in \text{argmin}_\beta \left\{ \ell_n(\beta) + \lambda \sum_{l=1}^{g} ||\beta^l||_2 \right\} \tag{4.1}$$

où $||\beta^l||_2 = \left(\sum_{l=1}^{L_j} \beta_{j,l}^2\right)^{\frac{1}{2}}$ et $\lambda > 0$, le paramètre de régularisation. Si chaque groupe contient exactement une variable, on retrouve l'estimateur Lasso. Il s' agit donc d'une extension du Lasso. L'estimateur Groupe Lasso defini en (4.1) repose sur l'hypothèse que les groupes forment une partition de $\{1, \cdots, d\}$. Les cas où les groupes ne forment pas une partition sont traités par [63], [65], [61] entre autres.

Un autre défi est qu'en général, dans la plupart des applications cliniques, l'analyse de survie suppose implicitement que la population étudiée est homogène. Ceci veut dire que tous les individus en étude sont soumis, en principe, aux mêmes risques (par exemple, risque de décès, risque de récidive). Or dans de nombreuses applications, cette hypothèse n'est pas réaliste. Par exemple, il peut y avoir une prédisposition génétique à certaines maladies qui est différente d'un individu ou d'un groupe d'individus à l'autre. Ceci veut donc dire que les individus en étude forment un ensemble de personnes avec différents risques. La notion de fragilité offre un moyen pratique d'introduire des effets aléatoires, de dépendance et d'hétérogénéité non observée, dans les modèles de données de survie. C'est pour cela que dans ce chapitre, nous présentons la méthode du group Lasso dans la sélection de variables pour le modèle de Cox avec fragilité pour les données en clusters.

## 4.2 Modélisation

Supposons qu'on a $n$ groupes d'individus et que le $i^{eme}$ groupe est constitué de $J_i$ individus avec une fragilité partagée $u_i$. Soit $Z_{ij}$ un vecteur de covariables associé au temps de survie $X_{ij}$ de l'individu $j$ du groupe $i$. Supposons que pour cet individu, on dispose des données de survie $(T_{ij}, \delta_{ij}, Z_{ij}, u_i)$ qui sont des v.a qui sont $i.i.d$, avec $\delta_{ij} = \mathbb{I}_{\{X_{ij} \leq C_{ij}\}}$ l'indicatrice de censure, $C_{ij}$ le temps de censure et $T_{ij} = \min(X_{ij}, C_{ij})$ son temps de survie observé. La fonction de vraisemblance correspondante est donnée par :

$$L_n(\beta, \alpha, H_0(.)) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \left\{ h_{ij}\left(T_{ij}|u_i, Z_{ij}\right)^{\delta_{ij}} S_{ij}\left(T_{ij}|u_i, Z_{ij}\right) \right\} \prod_{i=1}^{n} g(u_i), \tag{4.2}$$

avec $S(t) = \exp(-H_0(t))$, la fonction de survie et $h(t|Z, u)$, le taux de hasard conditionnellement au vecteur de covariables $Z$ et de fragilité $u_i$ avec $u_i$ partagée par le groupe $i$. Les fragilités suivent une loi Gamma dont la densité donnée par

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1} \exp(-\alpha u)}{\Gamma(\alpha)}, \tag{4.3}$$

avec

$$\Gamma(\alpha) = \int_0^{+\infty} \exp(-u) u^{\alpha-1} du.$$

On considère un modèle à risques propotionnels de Cox avec fragilité partagée,

$$h_{ij}(t|Z_{ij}, u_i) = h_o(t) u_i \exp(\beta^\top Z_{ij}), \tag{4.4}$$

où $h_0(t)$ est le taux de risque de base et $\beta$ un vecteur des paramètres d'intérêt. Soit $H_0(t) = \int_0^t h_o(\mu)\, d\mu$ le taux de risque cumulé. En remplaçant $h_{ij}(.)$ par sa valeur, la vraisemblance (équation 4.2) devient :

$$L_n(\beta, \alpha, H_0(.)) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} H_0(T_{ij})^{\delta_{ij}} \exp(\beta^\top Z_{ij}) u_i^{\delta_{ij}} \exp\left\{-H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u_i\right\} \prod_{i=1}^{n} g(u_i). \tag{4.5}$$

La vraisemblance des données observées, après intégration de l'expression (4.5) par rapport à $u_1, ..., u_n$, est donnée par :

$$\begin{aligned} \int_{u_1}^{+\infty} \dots \int_{u_n}^{+\infty} & \prod_{i=1}^{n} \prod_{j=1}^{J_i} \left\{ H_0(T_{ij})^{\delta_{ij}} \exp(\beta^\top Z_{ij}) u_i^{\delta_{ij}} \exp\left[-H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u_i\right] \right\} \\ & \times \prod_{i=1}^{n} g(u_i)\, du_n \dots\, du_1 \\ & = \prod_{i=1}^{n} \prod_{j=1}^{J_i} H_0(T_{ij})^{\delta_{ij}} \exp(\beta^\top Z_{ij}) \times A, \end{aligned}$$

où

$$\begin{aligned} A &= \int_{u_1}^{+\infty} \dots \int_{u_n}^{+\infty} \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J_i} u_i^{\delta_{ij}} \exp\left[ - H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u_i \right] \right\} \prod_{i=1}^{n} g(u_i)\, du_n \dots\, du_1 \\ &= \int_{u_1}^{+\infty} \dots \int_{u_n}^{+\infty} \prod_{i=1}^{n} \left\{ u_i^{\sum_{j=1}^{J_i} \delta_{ij}} \exp\left[ - \sum_{j=1}^{J_i} H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u_i \right] \prod_{i=1}^{n} g(u_i) \right\} du_n \dots\, du_1. \end{aligned} \tag{4.6}$$

En remplaçant (4.3) dans (4.6), on a

$$A = \prod_{i=1}^{n} \underbrace{\int_{u_i}^{+\infty} u_i^{(A_i+\alpha)-1} \exp\left\{ - \left[ \sum_{j=1}^{J_i} H_0(T_{ij}) \exp(\beta^\top Z_{ij}) + \alpha \right] u_i \right\} du_i} \times \prod_{i=1}^{n} \frac{\alpha^\alpha}{\Gamma(\alpha)}$$

où $A_i = \sum_{j=1}^{J_i} \delta_{ij}$.

En posant $k = \Big[\sum_{j=1}^{J_i} H_0(T_{ij}) \exp(\beta^\top Z_{ij}) + \alpha\Big]$, alors on obtient

$$A = \prod_{i=1}^{n} \int_{u_i}^{+\infty} (ku_i)^{(A_i+\alpha)-1} \exp(-(ku_i))\, d(ku_i) \frac{1}{(k)^{A_i+\alpha}} \prod_{i=1}^{n} \frac{\alpha^\alpha}{\Gamma(\alpha)}$$

$$A = \prod_{i=1}^{n} \Gamma(A_i+\alpha) \frac{1}{\Big[\sum_{j=1}^{J_i} H_0(T_{ij}) \exp(\beta^\top Z_{ij}) + \alpha\Big]^{A_i+\alpha}} \prod_{i=1}^{n} \frac{\alpha^\alpha}{\Gamma(\alpha)}.$$

Ainsi la vraisemblance des données observées peut s'écrire :

$$L_n(\beta, \alpha, H_0(.)) = \prod_{i=1}^{n} \frac{\alpha^\alpha \prod_{j=1}^{J_i} H_0(T_{ij})^{\delta_{ij}} \exp(\beta^\top Z_{ij})\delta_{ij}}{\Gamma(\alpha)\Big[\sum_{j=1}^{J_i} H_0(T_{ij}) \exp(\beta^\top Z_{ij}) + \alpha\Big]^{A_i+\alpha}} \Gamma(A_i+\alpha). \tag{4.7}$$

Par conséquent, la log-vraisemblance de l'equation (4.7) est donnée par :

$$\begin{aligned} \ell_n(\beta, \alpha, H_0(.)) = \sum_{i=1}^{n} \Bigg\{ & \alpha \log \alpha + \sum_{j=1}^{J_i} \big[\beta^\top Z_{ij}\delta_{ij} + \delta_{ij} \log H_0(T_{ij})\big] + \log \Gamma(A_i+\alpha) \\ & - \log \Gamma(\alpha) - (A_i+\alpha) \log \Big[\sum_{j=1}^{J_i} H_0(T_{ij}) \exp(\beta^\top Z_{ij}) + \alpha\Big] \Bigg\}. \end{aligned} \tag{4.8}$$

[36] ont formulé la vraisemblance profilée en considérant une modélisation non informative et non paramétrique du taux de risque cumulé $H_0(.)$ . $H_0(T_{ij})$ a des sauts de taille $\rho_l$ à des durées obsevés $\tilde{T}_l$. Alors

$$\begin{aligned} H_0(T_{ij}) &= \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \le T_{ij}\}} \\ h_0(T_{ij}) &= \prod_{l=1}^{N} \rho_l^{\mathbb{I}_{\{\tilde{T}_l \le T_{ij}\}}}, \end{aligned} \tag{4.9}$$

où $\tilde{T}_l, l = 1, ..., N$ sont des instants observés non censurés. En remplaçant (4.9) dans (4.8) et en dérivant le résultat obtenu par rapport à $\rho_k$, on a

$$\begin{aligned} \frac{\partial \ell_n(\beta, \alpha, H_0(.))}{\partial \rho_k} = & \sum_{i=1}^{n} \sum_{j=1}^{J_i} \delta_{ij} \mathbb{I}_{\{\tilde{T}_k \le Z_{ij}\}} \frac{1}{\rho_k} \\ & - \sum_{i=1}^{n} (A_i+\alpha) \frac{\sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \mathbb{I}_{\{\tilde{T}_k \le T_{ij}\}}}{\alpha + \sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \le T_{ij}\}}}, \quad k = 1, ...N. \end{aligned} \tag{4.10}$$

Supposons qu'il n'y ait pas des événements simultannés ("ties") pour les différents groupes d'individus. Alors la solution de l'équation (4.10) vérifie l'équation

$$\frac{1}{\rho_k} = \sum_{i=1}^{n} \frac{(A_i + \alpha) \sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \mathbb{I}_{\{\tilde{T_k} \leq T_{ij}\}}}{\alpha + \sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T_l} \leq T_{ij}\}}}, \quad k = 1, ...N. \tag{4.11}$$

La valeur de $\rho_k$ dans (4.11) est obtenue numériquement par l'algorithme décrit dans la section 4.3.

### 4.2.1 Estimateur Group Lasso dans le modèle à risque proportionnel de Cox avec fragilité

La fonction objective à minimiser pour l'estimateur group Lasso dans le modèle de Cox avec fragilité est

$$Q_n(\beta, \lambda_n) = -\frac{1}{n}\ell_n(\alpha, \beta, H_0(.)) + \lambda_n \sum_{j=1}^{K} \sqrt{p_j} \|\beta_{(j)}\|_2 \tag{4.12}$$

où $Q_n(\beta, \lambda_n)$ est une fonction objective qui est convexe à minimiser par rapport au paramètre d'intérêt $\beta$ avec un paramètre de régularisation $\lambda_n$ donné. Ce paramètre de régularisation contrôle le degré de pénalisation. $\ell_n(\beta, \alpha, H_0(.))$ est la log-vraisemblance définie en (4.8). Le paramètre d'intérêt $\beta$ est un vecteur $\beta_{(j)}(j = 1, 2, ..., K)$ de $K$ composantes correspondants aux $K$ groupes de covariables respectifs. Le terme $\sqrt{p_j}$ ajuste la variation de taille d'un groupe de covariable et $\|.\|_2$ est la norme euclidienne.

L'estimateur Group Lasso dans le modèle à risque proportionnel de Cox avec fragilité est défini par :

$$\hat{\beta}_n(\lambda_n) = \arg\min_{\beta} \left\{ -\frac{1}{n}\ell_n(\alpha, \beta, H_0(.)) + \lambda_n \sum_{j=1}^{K} \sqrt{p_j} \|\beta_{(j)}\|_2 \right\}. \tag{4.13}$$

Cet estimateur n' est pas obtenu explicitement en général en raison de la non-différentiabilité de la norme euclidienne. Par conséquent, nous utilisons une procédure itérative pour résoudre ce problème de minimisation. En fonction de la valeur du paramètre optimal de régularisation $\lambda_n$, les coefficients estimés pour un groupe de paramètres donné sont tels que $\hat{\beta}_{(j)} = 0$ pour toutes les composantes ou $\hat{\beta}_{(j)} \neq 0$ pour toutes les composantes. Cela est dû à la non-différentiabilité de la fonction racine carrée au point 0 $(\beta_{(j)} = 0)$. Si la taille de tous les groupes est égale à 1, la méthode se réduit au Lasso standard.

### 4.2.2 Paramètre optimal de lissage $\lambda_n$ et sélection de modèle

Il est d'une importance capitale de disposer une méthode automatique de sélection du paramètre optimal de régularisation $\lambda_n$ permettant de contrôler le degré de pénalisation en fonction d'un critère spécifique, tel que le critère d'information Akaike (AIC) [3] défini par

$$AIC(\lambda_i) = \log\left(\ell_n(\hat{\beta})\right)/n + d(\lambda_i)/n, \tag{4.14}$$

le critère de l'information de Bayes (BIC) [102] défini par

$$BIC(\lambda_i) = \log\left(\ell_n(\hat{\beta})\right)/n + \log(n)d(\lambda_i)/n \tag{4.15}$$

où $d(\lambda_i)$ est le degré de liberté de l'estimateur pour un $\lambda_i$ donné [84] et la validation croisée (CV) [28] détaillée ci-dessous. Il n'existe pas de moyen simple ou universellement reconnu comme meilleur moyen pour déterminer la valeur optimale de $\lambda_n$. En général, la valeur sélectionnée est basée sur l'optimisation d'une fonction, généralement appelée "fonction de perte". Si $\lambda_i = 0$, l'estimateur Group Lasso coincide avec l'estimateur du maximum de vraisemblance qui est inadquat en grande dimension. Si $\lambda_i \to \infty$, la procédure group Lasso ne sélectionne aucun groupe de variables, car toutes les composantes de $\beta$ sont estimées à zéro. Donc $\lambda_i = 0,\ \lambda_i \to \infty$ sont inadéquats. L'estimateur group Lasso sélectionne d'autant plus de groupes de variables explicatives que $\lambda_n$ est petit, et plus $\lambda_i$ est grand, plus les composantes de $\beta$ sont contraintes à être nulles. L'objectif est de déterminer une valeur de $\lambda_n$ qui permet de sélectionner les groupes de variables pertinentes et ainsi d'améliorer les performances en prédiction du modèle. Toutes ces trois méthodes citées permettent de trouver la valeur optimale de $\lambda_n$ suivant le choix du critère donné et la méthode de la validation croisée est la plus couramment utilisée. Il consiste à se donner une grille de valeurs de $\lambda = \lambda_{\min}, ..., \lambda_{\max}$. Pour chaque $i \in \{\min, ..., \max\}$ on répète les étapes suivante

1. Partitionner l'ensemble des individus en k groupes $G_1, ..., G_k$
2. Pour chaque $j \in \{1, ..., k\}$, l'estimation des paramètres se fait sur $G_j^c$, c'est-à-dire l'échantillon initial diminué du groupe $j$ d'individus et chacune des k groupes ($G_j$) est utilisée comme jeu de test.
3. Ensuite, on calcule une estimation de l'erreur de prédiction définie par

$$CV(\lambda_i) = -\frac{1}{k}\sum_{j=1}^{k} \ell\left(\hat{\beta}_{(k-j)}(\lambda_i)\right) \tag{4.16}$$

où $\hat{\beta}_{(k-j)}(\lambda_i)$ est l'estimateur de $\beta$ à une valeur de $\lambda_i$ sur l'échantillon initial diminué d'un sous ensemble $j$ des données. Un sous ensemble $j$ des données est considéré comme un échantillon test et $k-j$ données est considéré comme échantillon d'apprentissage. $\ell(.)$ est la log-vraissemblance partielle pour l'échantillon d'apprentissage. Le paramètre optimal de régularisation est celui qui minimise l'erreur de prédiction. La validation croisée est recommandée lorsque l'objectif de l'analyse est la prédiction [74] mais elle pose en général un problème majeur du coût de calcul élevé. L'adaptation d'un modèle de risques proportionnels pénalisés est une opération de calcul lourde, en particulier si le modèle doit être ajusté plusieurs fois pour chaque valeur de $\lambda$ que nous voulons évaluer. Dans cette thèse, le choix de $k$ est fixé à $10$.

## 4.3 Algorithme proposé

Pour minimiser (4.12), on suit la procedure suivante : on scinde (4.8) en deux pseudo log-vraisemblances. L'une principalement dépendante de $\beta$ :

$$\ell_n^{(\beta)}(\beta, \alpha, H_0(.)) \equiv \sum_{i=1}^{n}\sum_{j=1}^{J_i} \beta^\top Z_{ij}\delta_{ij} - \sum_{i=1}^{n}(A_i+\alpha)\log\left\{\sum_{j=1}^{J_i} H_0(T_{ij})\exp(\beta^\top Z_{ij}) + \alpha\right\} \tag{4.17}$$

et l'autre principalement dépendante de $\alpha$:

$$\begin{aligned}\ell_n(\beta, \alpha, H_0(.)) \equiv \sum_{i=1}^{n}\Bigg\{&\alpha\log\alpha + \log\Gamma(A_i+\alpha) - \log\Gamma(\alpha)\\ &- (A_i+\alpha)\log\Big[\sum_{j=1}^{J_i} H_0(T_{ij})\exp(\beta^\top Z_{ij}) + \alpha\Big]\Bigg\}.\end{aligned} \tag{4.18}$$

Etant donné que le terme de pénalité dans (4.12) dépend seulement de $\beta$, minimiser (4.12) par rapport de $\beta$ est équivalente à minimiser

$$-\frac{1}{n}\ell_n^{(\beta)}(\beta, \alpha, H_0(.)) + \lambda_n\sum_{j=1}^{K}\sqrt{p_j}\|\beta_{(j)}\|_2 \tag{4.19}$$

Nous parcourons les groupes de paramètres et minimisons (4.19) en gardant tous les groupes de paramètres sauf le groupe actuel fixe. L'algorithme de Block Coordinate Gradient Descent (BCGD) est appliqué pour résoudre le problème d'optimisation convexe non lisse en (4.19) [131]. Cet algorithme serait également implémenté pour optimiser l'équation (4.18). Cependant, l'équation (4.18) implique

que les dérivés du premier et second ordres de la fonction gamma peuvent ne pas exister pour certaines valeurs de $\alpha$. Pour contourner cette difficulté, nous utilisons une approche similaire à celle de [36] en utilisant une grille discrète de valeurs possibles pour le paramètre de fragilité $\alpha$ et en recherchant le minima de (4.18) comme [87] l'ont suggéré.

On note $Q_{\lambda_n}(\beta) = -\frac{1}{n}\ell_n^{(\beta)}(\beta, \alpha, H_0(.)) + \lambda_n \sum_{j=1}^{K} \sqrt{p_j}\|\beta_{(j)}\|_2$ une fonction objective pénalisée à minimiser et $\nabla Q_{\lambda_n}(\beta)$ son gradient à evaluer en $\beta$

**Tableau** 4.1 – L'algorithme Block Co-ordinate Gradient (BCGD)

| Etapes | Algorithme |
|---|---|
| 1. | Pour $j = 1, ..., K$<br>on choisit $\hat{\beta}_{(j)}^{(0)}$ comme valeur initiale. |
| 2. | Pour la $m^{ieme}$ itération, $\hat{\beta}_{(j)}^{(m+1)} \leftarrow \hat{\beta}_l^{(m)} - \gamma_n \nabla Q_{\lambda_n}(\hat{\beta}_l^{(m+1)})$ avec $m = 0, 1, 2, ...$ et $\gamma_n > 0$ le pas calculé en suivant la règle d'Armijo |
| 3. | Pour chaque $j$, on repète l'étape 2 jusqu'à la convergence |

Avec BCGD au Tableau (4.1), on propose l'algorithme suivant pour résoudre le problème en (4.12).

| Etapes | Algorithme |
|---|---|
| 1. | Pour $j = 1, ..., K$<br>on choisit $\hat{\beta}_{(j)}^{(0)}, \hat{\alpha}_{(j)}^{(0)}, \hat{\rho}_{j,k}^{(0)}$, k=1,...,N comme valeurs initiales. |
| 2. | Pour la $m^{ieme}$ itération, $\hat{\rho}_{j,k}^{(m+1)}$ est mis à jour en (4.11) avec $m = 0, 1, 2, ...$ et ça calcule $\hat{H}_0^{(m+1)}$ en (4.9) |
| 3. | Puisque $\hat{H}_0^{(m+1)}(.)$ est connu, on peut alors minimiser (4.18) par rapport à $\left(\hat{\beta}_{(j)}^{(m+1)}\right)$ en utilisant l'algorithme BCGD |
| 4. | Puisque $\left(\hat{H}_0^{(m+1)}(.), \hat{\beta}_{(j)}^{(m+1)}\right)$ sont connus, on peut alors minimiser (4.19) par rapport à $\left(\hat{\alpha}_{(j)}^{(m+1)}\right)$ comme indiqué plus haut |
| 5. | Pour chaque $j$, on repète létape 2 à 4 jusqu'à la convergence |

## 4.4 La consistance théorique de la méthode proposée

On considère l'estimateur pénalisé du vecteur $\beta$ des paramètres :

$$\hat{\beta}_n(\lambda_n) = \arg\min_{\beta} \left\{ -\frac{1}{n}\ell_n(\alpha, \beta, H_0(.)) + \lambda_n \sum_{j=1}^{K} \sqrt{p_j}\|\beta_{(j)}\|_2 \right\}$$

On note $\beta^0$ la vraie valeur du paramètre du modèle $\beta = (\alpha, \beta, H_0(.))$. $\forall \varepsilon > 0$, nous voulons montrer que $\mathbb{P}\left\{\|\hat{\beta}_n(\lambda_n) - \beta^0\| < \varepsilon\right\} \to 1$ quand $n \to \infty$.

En utilisant la théorie des processus de comptage $N_i$ abordée dans la section 1.3.2 et détaillée dans la partie annexe A.1 de cette thèse, on considère la fonction de log-vraisemblance partielle de Cox sur l'intervalle $[0, t]$ :

$$\log(L_{Cox}(\beta, t)) = \sum_{i=1}^{n} \int_0^t \left[\beta^\top Z_i - \log S^{(0)}(\beta, u)\right] \, dN_i(u), \tag{4.20}$$

avec

$$S^{(0)}(\beta, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t) \exp\left(\beta^\top Z_i\right).$$

Le vecteur score

$$U(\beta, t) = \frac{\partial \log(L_{Cox}(\beta, t))}{\partial \beta}$$

peut s'écrire

$$\begin{aligned} U(\beta, t) &= \sum_{i=1}^{n} \int_0^t \left[Z_i - \mathbb{E}(\beta, u)\right] \, dN_i(u) \\ &= \sum_{i=1}^{n} \int_0^t \left[Z_i - \mathbb{E}(\beta, u)\right] \, dM_i(u), \end{aligned}$$

où

$$\mathbb{E}(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}$$

et

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

$[Z_i - \mathbb{E}(\beta, u)]$ étant un vecteur de processus prévisible, $U(\beta, t)$ est une somme de $n$ martingales vectorielles, et est donc lui-même une martingale. La suite de martingale $M^{(n)}(t) = n^{-1/2}U(\beta, t)$ vérifie les conditions d'application du théorème de Rebolledo. En appliquant le théorème de Rebolledo (cf. annexe A.3.1), nous déduisons

de la loi du processus limite $M^\infty = M^\tau$ la nouvelle formulation de la proposition 1.6.1 de façon suivante :

**Proposition 4.4.1.** *Soit $\hat{\beta}$ l'estimateur de $\beta$ i.e la quantité vérifiant*

$$U(\hat{\beta}, \tau). \tag{4.21}$$

*Soit $\widehat{\beta}$ l'estimateur du maximum de vraisemblance de $\beta$ i.e. la quantité vérifiant :*

$$U(\widehat{\beta}) = 0$$

*Alors*

$$\widehat{\beta} \xrightarrow{\mathcal{L}} \mathcal{N}(\beta_\prime, I^{-1}(\widehat{\beta}))$$

*$I^{-1}$ est l'inverse de la matrice d'information de* **Fisher** *(notée $\Sigma$) basée sur la vraisemblance partielle.*

$$\begin{aligned} I(\beta) &= \frac{\partial^2 \log(L_{Cox}(\beta))}{\partial \beta^2} \\ &= -\sum_{i=1}^{n} \int_0^\tau \left[ \frac{S^{(2)}(\beta, s)}{S^{(0)}(\beta, s)} - \mathbb{E}(\beta, s)^{\otimes 2} \right] dN_i(u), \end{aligned} \tag{4.22}$$

*avec*

$$S^{(2)}(\beta, s) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\left(\beta^\top Z_i(t)\right) Z_i^{\otimes 2}.$$

$I^{-1}$, l'inverse de la matrice d'information de **Fisher** , fournit une estimation de la variance de $\hat{\beta}$.
Pour pouvoir démontrer la consistance de l'estimateur $\hat{\beta}_n(\lambda_n)$ de $\beta$, on fait d'abaord un rappel sur des conditions de régularité (A)-(D) de [5]. Ces conditions sont supposées être vérifiées tout au long de cette section.
CONDITIONS:

A. (interval fini): $\int_0^1 h_0(t)dt < \infty$
B. (Stabilité asymptotique). Il existe un voisinage $\mathscr{B}$ de la vraie valeur $\beta^0$ et des fonctions constante, vectorielle et matricielle $s^{(0)}, s^{(1)}$ et $s^{(2)}$ definies sur $\mathscr{B} \times [0, 1]$ telles que, pour $j = 0, 1, 2$

$$\sup_{t \in [0,1], \beta \in \mathscr{B}} \| S^{(j)}(\beta, t) - s^{(j)}(\beta, t) \| \xrightarrow{\mathbb{P}} 0.$$

C. (Condition de Lindeberg). Il existe $\delta > 0$ tel que

$$n^{-\frac{1}{2}} \sup_{i,t} |Z_i(t)| Y_i(t) \mathbb{I}_{\{\beta_0^\top Z_i(t) > -\delta |Z_i(t)|\}} \xrightarrow{\mathbb{P}} 0$$

D. (Conditions de régularité asymptotique). Soient $\mathscr{B}, s^{(0)}, s^{(1)}$ et $s^{(2)}$ comme définis dans la condition B et on définit $e = \frac{s^{(1)}}{s^{(0)}}$ et $v = \frac{s^{(2)}}{s^{(0)}} - e^{\otimes 2}$. Pour tout $\beta \in \mathscr{B}, t \in [0,1]$ :

$$s^{(1)}(.,t) = \frac{\partial}{\partial \beta} s^{(0)}(\beta, t), s^{(2)}(.,t) = \frac{\partial^2}{\partial \beta^2} s^{(0)}(\beta, t)$$

$s^{(0)}(.,t), s^{(1)}(.,t)$ et $s^{(2)}(.,t)$ sont des fonctions continues de $\beta \in \mathscr{B}$, uniformément en $t \in [0,1], s^{(0)}, s^{(1)}$ et $s^{(2)}$ sont bornées sur $\mathscr{B} \times [0,1]$; $s^{(0)}$ est borné en déhors de zéro sur $\mathscr{B} \times [0,1]$, et $\Sigma = \int_0^t v(\beta_0, t) s^{(0)}(\beta_0, t) h_0(t) dt$ est défini positive.

On note que les conditions de l'existence de dérivées partielles de $s^{(0)}, s^{(1)}$ et $s^{(2)}$ sont verifiées par $S^{(0)}, S^{(1)}$ et $S^{(2)}$ et que $\Sigma$ est automatiquement défini semi positif. De plus, l'interval $[0,1]$ utilisé dans ces conditions peut être remplacé partout par $\{t : h_0(t) > 0\}$.

**Théorème 4.4.2.** *(consistance) Supposons que $(Z_{ij}, X_{ij}, C_{ij}, u_i)$ sont des v.a qui sont $i.i.d.$ $X_{ij}$ et $C_{ij}$ sont conditionnellement indépendants sachant $Z_{ij}$ et $u_i$, $i = 1, ..., n, \quad j = 1, ..., J_i$. Sous les conditions de régularité (A)-(D) dans [5], si $\lambda_n \to 0$ quand $n \to \infty$, alors $\hat{\beta}_n(\lambda_n)$ est un estimateur consistant de $\beta^0$.*

***Preuve.*** En s'appuyant sur le théorème 5.7 de [115] avec une approche légèrement différente, le théorème 4.4.2 peut être démontré comme suit : montrons d'abord que $Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n)$.
Posons $\Delta_n = Q_n(\beta_n, \lambda_n) - Q_n(\beta_n^0, \lambda_n)$

$$\begin{aligned}
\Delta_n &= -\frac{1}{n}\left(\ell_n(\beta) - \ell_n(\beta^0)\right) + \sum_{j=1}^{K} \lambda_n \sqrt{p_j} \left(\|\beta_{(j)}\| - \|\beta_{(j)}^0\|\right) \\
&\geq -n^{-1/2} \left(n^{-1/2} \frac{\partial}{\partial \beta}\Big(\ell_n(\beta^0)\Big)\right)^\top \left(\beta - \beta^0\right) + \left(\beta - \beta^0\right)^\top \left(n^{-1/2} \frac{\partial^2}{\partial \beta^2}\Big(\ell_n(\beta^0)\Big)\right) \left(\beta - \beta^0\right) \\
&\quad + n^{-1} o_p\left(\|\beta - \beta^0\|^2\right) - \sum_{j=1}^{K} \lambda_n \sqrt{p_j} \left(\|\beta_{(j)}\| - \|\beta_{(j)}^0\|\right) \\
&\geq -n^{-1} O_p(1) \|\beta - \beta^0\| + \left(\beta - \beta^0\right)^\top \left(\Sigma + o_p(1)\right) \left(\beta - \beta^0\right) \\
&\quad + n^{-1} o_p\left(\|\beta - \beta^0\|^2\right) - \lambda_n \sum_{j=1}^{K} \sqrt{p_j} \left(\|\beta_{(j)}\| - \|\beta_{(j)}^0\|\right)
\end{aligned}$$

Puisque $\lambda_n \to 0$ quand $n \to \infty$ alors $Q_n(\beta_n, \lambda_n) - Q_n(\beta_n^0, \lambda_n) \geq (\beta - \beta^0)^\top (\Sigma + o_p(1)) (\beta - \beta^0)$ et la quantité à droite de l'équation est positive puis $\Sigma$ est positif. $Q_n(\beta_n, \lambda_n)$ est non-vide et borné inférieurement par $Q_n(\beta_n^0, \lambda_n)$ et par conséquent, il admet un minimum local. Puisque $Q_n(\beta_n, \lambda_n)$ est concave, son minimum local est un minimum global.

$$Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n).$$

Pour tout $\varepsilon$ positif

$$\left\{ \sup_{\beta : \|\beta - \beta^0\| = a} Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n) \right\} \subseteq \left\{ \|\hat{\beta}_n(\lambda_n) - \beta^0\| < \varepsilon \right\}$$

$$\Rightarrow \mathbb{P}\left\{ \|\hat{\beta}_n(\lambda_n) - \beta^0\| < \varepsilon \right\} \geq \mathbb{P}\left\{ \sup_{\beta : \|\beta - \beta^0\| = a} Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n) \right\}$$

Alors $\mathbb{P}\left\{ \|\hat{\beta}_n(\lambda_n) - \beta^0\| < \varepsilon \right\} \to 1$. □

## 4.5 Applications

Avec l'avènement de la biologie moléculaire qui étudie la relation entre la génétique et les maladies tel que le cancer ainsi que des mesures prises pour contrôler des multiples polluants ou agents pathogènes dans l'air et dans l'eau, il devient possible pour les chercheurs et les agents de la santé publique de générer des vastes bases de données en grande dimension où le nombre de covariables $p$ est supérieur à la taille de l'échantillon $n$. Des méthodes statistiques sont nécessaires pour traiter et analyser cette base de données [99]. Dans le cas de l'épidémiologie génétique, les chercheurs sont en mesure d'identifier les gènes qui agissent le long de voies identiques ou similaires et de regrouper ces gènes afin de comprendre la relation entre ces gènes et l'éclosion de la maladie et enfin de pouvoir calculer le risque cumulatif.

Dans le cadre de l'évaluation de l'exposition à la maladie, les scientifiques du domaine de la santé environnementale comprennent maintenant que les polluants rejetés par plusieures sources de pollution contribuent aux mêmes morbidités. Les exemples incluent de nombreux produits chimiques qui sont contenus dans la fumée de tabac, les émissions des véhicules et les cheminées industrielles. Les personnes souffrant de diarrhée peuvent par exemple être coinfectées par plusieurs

agents pathogènes se trouvant dans l'eau et la compréhension de la nature de l'émergence de la maladie peut être améliorée au fur et à mesure que la science de l'exposition à l'eau se développe. La médecine personnalisée a ouvert la porte à la santé publique personnalisée car de nouvelles informations peuvent être recueillies au niveau individuel. L'utilisation de la méthode du group Lasso avec fragilité au niveau du groupe d'individus dans l'analyse de survie, serait préférable afin de pouvoir comprendre l'effet de la maladie ainsi que des facteurs clés contribuant à l'exposition à une telle maladie. La préférence de la méthode group Lasso qui met tous les groupes de covariables non significatives dans le modèle à zéro est de produire des modèles sparses (c'est-à-dire les modèle ayant un nombre important de coefficients nuls) qui renvoient à des sources de pollution plutôt qu'à des expositions chimiques ou biologiques de l'individu. Cette application pourrait s'appliquer dans le cas d'études d'utilisation du sol, d'évaluation des risques liés aux zones industrielles et d'études d'impact sur l'environnement de nouveaux projets de construction. Le groupe Lasso dans le modèle à risques proportionnels de Cox avec fragilité fera partie du nouveau paradigme de l'évaluation des risques qui englobe les expositions cumulatives [25]. Comme applications en épidémiologie génétique, le test de l'association et complexité génétique deviennent de moins en moins couteux. De grandes bases de données de cohortes deviendront disponibles pour établir plus efficacement des associations entre les marqueurs génétiques et épigénétiques et les effets sur les maladies. Comme indiqué précédemment, le groupe Lasso avec la fragilité au niveau du groupe permet d'intégrer des voies et des mécanismes communs à l'analyse, tout en incluant un terme de fragilité pour prendre en compte la susceptibilité ou de la résilience non mesurée existant dans les sous populations.

## 4.6 Exemples d'application

### 4.6.1 Les données simulées

Des données ont été simulées avec une taille d'échantillon total $m = \sum_{i=1}^{n} J_i$ (où $n$ est le nombre de groupes d'individus et $J_i$ est la taille de l'échantillon du groupe $i$ ) fixé arbitrairement à $100$ et le nombre de covariables $p$ fixé arbitrairement à $200$. La taille de groupes d'individus (par rapport à la fragilité) et de covariables (par rapport aux groupes de covariables) ont été tous fixés arbitrairement à 10. Ceci peut varier en fonction de la taille des données. On a simulé la matrice

$Z_{ij}$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$ de données avec $Z_i \overset{i.i.d}{\to} \mathcal{N}(0, 1)$. On a aussi simulé la matrice de variance-covariance $\Sigma_{i,j} = \rho^{|i-j|}$ où $\rho = 0.5$. En pratique, l'hypothèse de constance du taux de risque n'est pas souvent vérifiée. La forme la plus générale de la fonction de risque est donnée par la distribution de Weibull, qui est caractérisée par deux paramètres positifs : le paramètre d'échelle $(\lambda > 0)$ et le paramètre de forme $(\nu > 0)$. La fonction de risque de base est donnée par

$$h_0(t) = \lambda \nu t^{\nu-1}$$

et le temps de survie $X$ pour un modèle de Cox avec une fragilité gamma partagée $U$ est donnée par

$$X = \left( \frac{\log(G) \exp(-\beta^\top Z)}{\lambda U} \right)$$

avec $G = \mathbb{P}(X > t|Z) = \exp\left[-\exp\left(\beta^\top Z\right) \lambda X^\nu\right]$ et $G \rightsquigarrow Uniform[0, 1]$ et $U \rightsquigarrow Gamma(\alpha, \alpha)$. En tenant compte de l'état de censure, nous avons simulé des temps de censure à partir de la distribution exponentielle: $C \rightsquigarrow Exponential(3)$. Le temps de défaillance observé pour chaque observation est le minimum entre son temps de survie $T$ et son statut de censure $C$. L'algorithme décrit en (4.3) a été utilisé pour sélectionner le paramètre optimal $\lambda_n$ de régularisation qui minimise le critère de la validation croisée (CV). La performance du group Lasso dans le modèle à risques proportionnels de Cox avec fragilité est comparée à celle du group SCAD et du groupe MCP. La figure 4.2 montre un exemple de chemin de régularisation pour le group Lasso, le group SCAD et le group MCP, respectivement.

**Tableau** 4.2 – Résumé sur la sélection de modèles pour les données simulées.

| **Méthode utilisée** | **# de** $\beta \neq 0$ | **# de groupes** $\beta \neq 0$ | $\lambda_n$ |
|---|---|---|---|
| Group Lasso(n=100, p=200) | 80 | 8 | 0.0294 |
| Group SCAD(n=100, p=200) | 20 | 2 | 0.1079 |
| Group MCP(n=100, p=200) | 10 | 1 | 0.1255 |

A partir du tableau 4.2, le group Lasso sélectionne 8 groupes à une valeur de paramètre de régularisation égale à 0.0294 pendant que le group SCAD et group MCP pénalisent fortement les groupes de variables et ne sélectionnent que 2 groupes à une valeur de paramètre de régularisation égale à 0.1079 et un seul groupe à une valeur de paramètre de régularisation égale 0.1255 respectivement.

Les figures 4.1-4.4 comparent la perfomance de trois méthodes sur 100 simulations sur une grille de valeurs de paramètres de régularisation, erreur de la vali-

dation croisée et la valeur de $R^2$, respectivement (rappellons-nous qu'il s'agit d'un ensemble de données simulées). Quelques tendances sommaires apparaissent. Notamment pour cette simulation, le groupe Lasso a tendance à choisir la plus petite valeur de paramètre de régularisation, centrée autour de 0,03 comparé à 0.09 pour le group SCAD et 0.10 pour le group MCP.

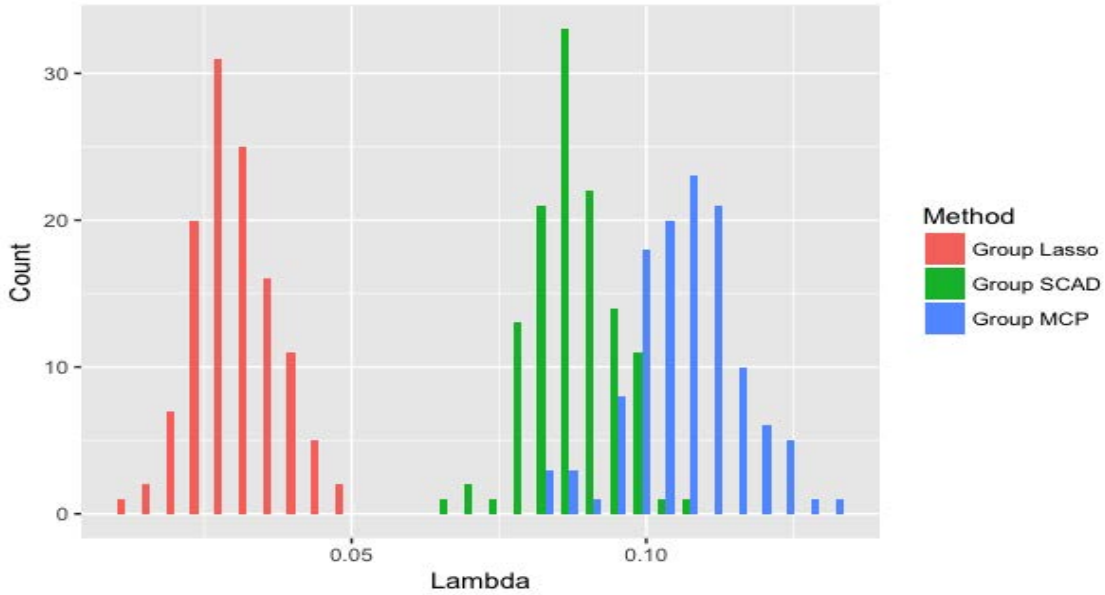


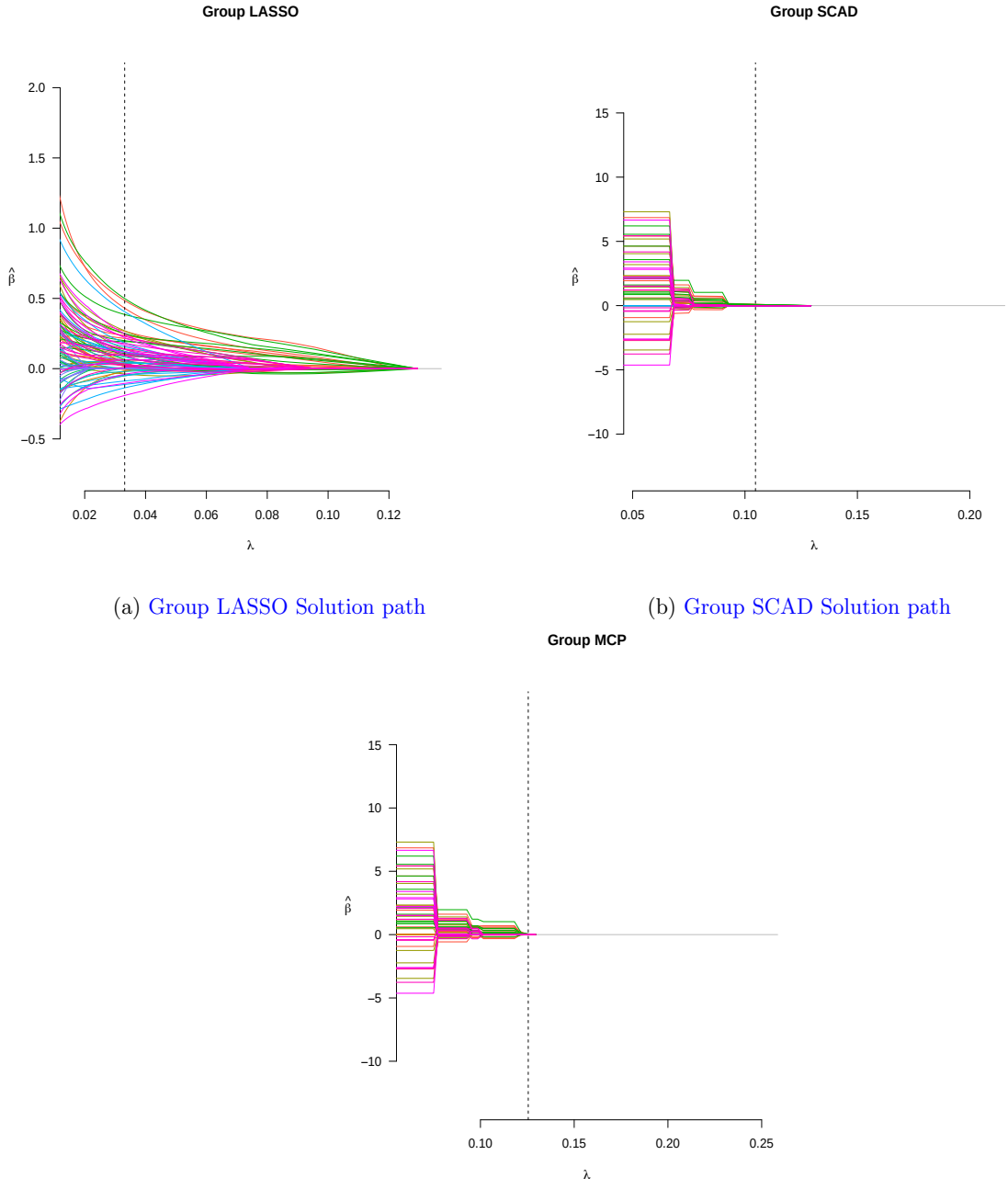

**Figure** 4.1 – Distribution du paramètre de régularisation pour chacune des trois méthodes sur 100 simulations.

Si l'on considère l'erreur de validation croisée, les résultats sont presque similaires, le group Lasso n'affiche qu'une performance légèrement meilleure (139 pour le groupe Lasso, 151 pour le groupe SCAD et 156 pour le groupe MCP) dans cet example de simulations.

La performance de $R^2$ pour le group Lasso est significativement meilleure, se situant en moyenne autour de 0,18 par rapport à 0.05 pour le group SCAD et 0.03 pour le group MCP.

### 4.6.2 Exemple sur des données réelles

Dans cette section, nous appliquons la méthode group Lasso dans le modèle à risques proportionnels de Cox avec fragilité pour la sélection du modèle. Nous comparons sa performance avec celle du group SCAD et du group MCP aux données génétiques en grande dimension de l'Institut National de Cancerologie des




(a) Group LASSO Solution path

(b) Group SCAD Solution path

(c) Group MCP Solution path

**Figure** 4.2 – Chemin de régularisation pour: Group Lasso, Group SCAD, Group MCP pour les données simulées

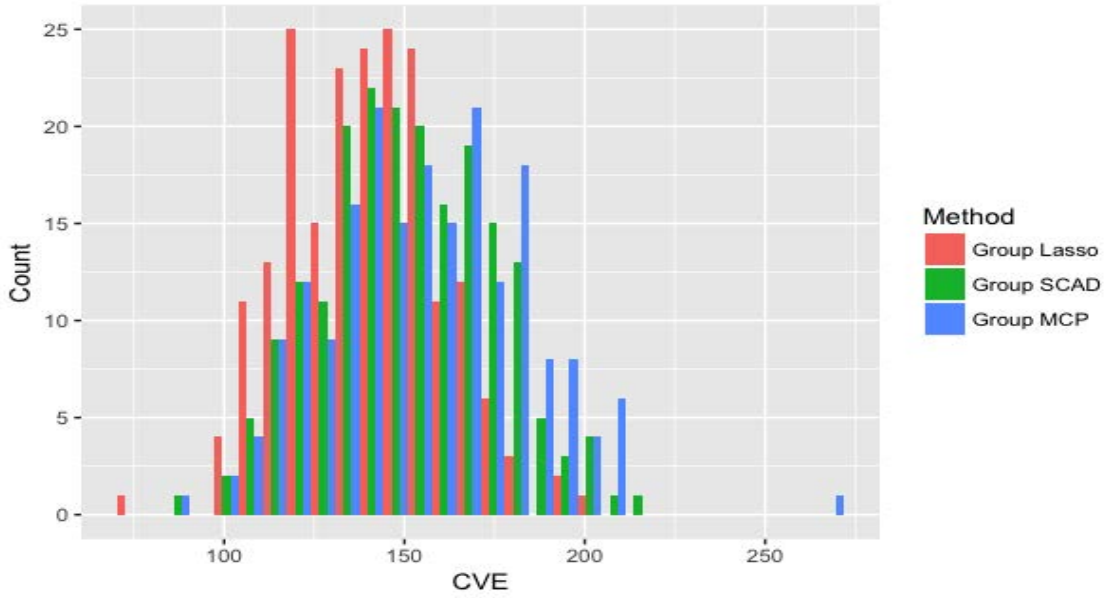



**Figure** 4.3 – Distribution de l'erreur de la validation croisée pour chacune des trois méthodes sur 100 simulations.

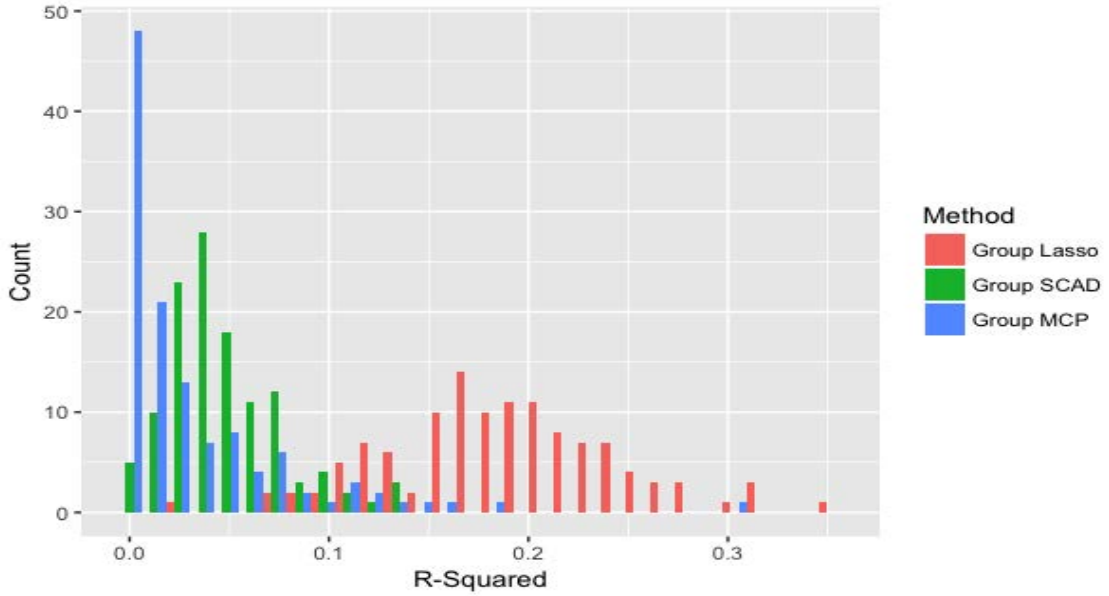



**Figure** 4.4 – Distribution de $R^2$ pour chacune des trois méthodes sur 100 simulations.

Etats Unis sur 470+ patients ayant la lymphome à cellules B diffus (DLBCL) recevant un traitement standard avec rituximab plus cyclophosphamide, doxorubicine, vincristine et prednisolone (R-CHOP). La fragilité gamma est appliquée pour tenir compte de l'hétérogénéité en fonction des facteurs de risque identifiés dans les données démographiques des patients. La classification hiérarchique a été utilisée

pour déterminer les groupes d'expression des génétiques, car on dispose de peu d'information sur ces gènes et leurs voies d'expression pertinentes pour le DLBCL. Les patients dont les données démographiques ou génétiques étaient incomplètes ont été exclus, ce qui a donné 470 profils de patients à utiliser dans le modèle. L'ensemble de cette base de données est 54634 gènes; en raison du coût de calcul computationnel elevé, 5 000 gènes ont été sélectionnés pour le modèle et ont été regroupés selon une classification hiérarchique en 10 groupes, comparables à ceux de l'exemple sur les données simulées en 4.6.1. Pour déterminer la fragilité du groupe, huit indices de susceptibilité au risque de LDCLB ont été sélectionnés à partir des données démographiques. En fonction de la présence ou de l'absence de ces caractéristiques, sept groupes ont été constitués et des fragilités partagées ont été assignées par ordre de risque non mesuré, du plus faible au plus élevé, en fonction de leurs données démographiques.

La figure 4.6 montre un exemple de chemin de régularisation pour le group Lasso, le group SCAD et le group MCP, respectivement. Le group Lasso a tendance à sélectionner plus de groupes de variables à une plus petite valeur de paramètre de régularisation par rapport au group SCAD et au group MCP.

**Tableau** 4.3 – Resumé sur la sélection de modèle sur un exemple des données DLBCL

| **Méthode utilisée** | **# de** $\beta \neq 0$ | **# de groupes** $\beta \neq 0$ | $\lambda_n$ |
|---|---|---|---|
| Group Lasso(n=470, p=5000) | 636 | 6 | 0.0303 |
| Group SCAD(n=470, p=5000) | 636 | 6 | 0.0316 |
| Group MCP(n=470, p=5000) | 278 | 3 | 0.0372 |

A partir du tableau 4.3, le group Lasso sélectionne 6 groupes à une valeur de paramètre de régularisation égale à 0.0303 , le groupe SCAD sélectionne 6 groupes à une valeur de paramètre de régularisation égale à 0.0316 et le groupe MCP sélectionne 3 groupes à une valeur de paramètre de régularisation égale à 0,0372.

Les figures 4.5-4.8 comparent la performance des trois méthodes sur 100 simulations sur une grille de valeurs de paramètre de régularisation, des erreurs de validation croisée et des $R^2$, respectivement. Quelques tendances sommaires apparaissent. Notamment pour ces simulations, le group Lasso et le groupe SCAD choisissent la plus petite valeur de paramètre de régularisation centrée autour de 3.5 $\times 10^{-2}$ comparé à 3.9 $\times 10^{-2}$ pour le group MCP.

La performance de $R^2$ pour le groupe Lasso est comparable à celle du groupe SCAD, les deux ayant une valeur moyenne de $R^2$ d'environ 4,5$\times 10^{-3}$ comparée à

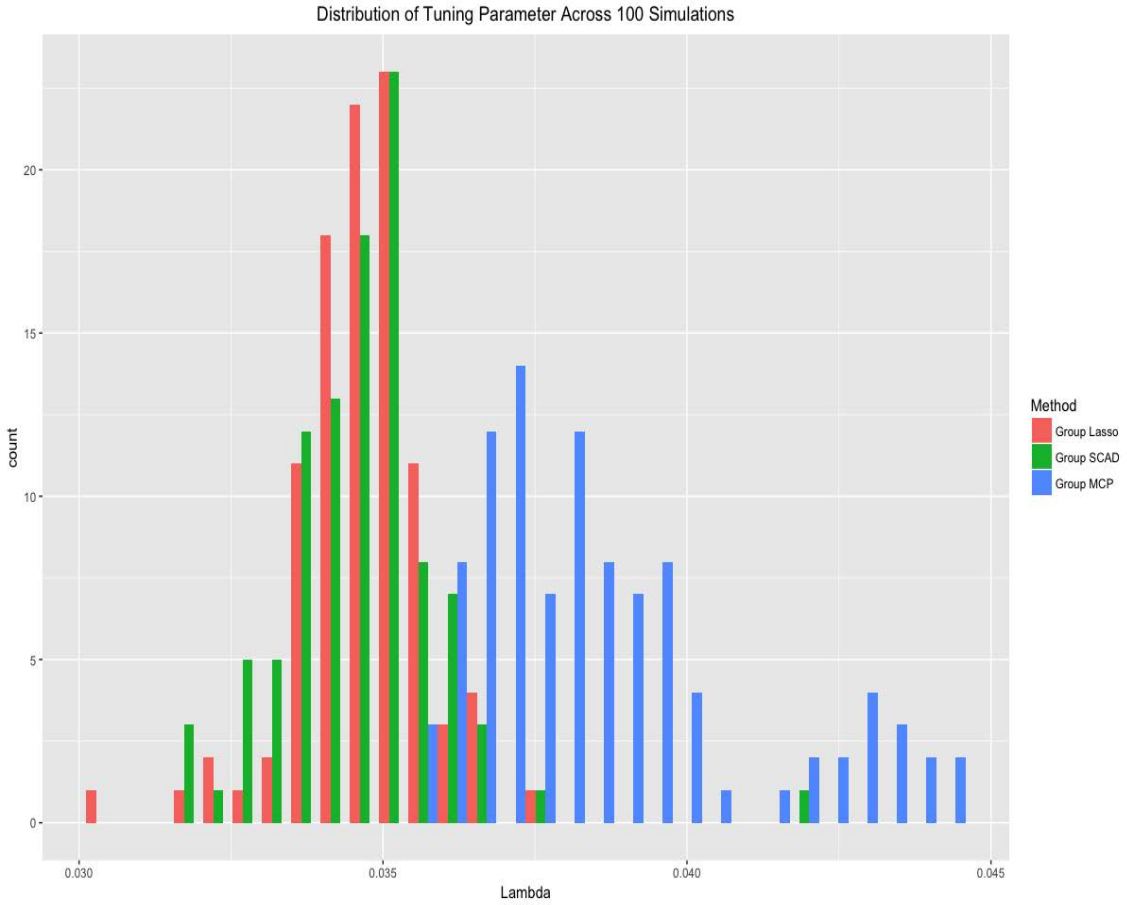
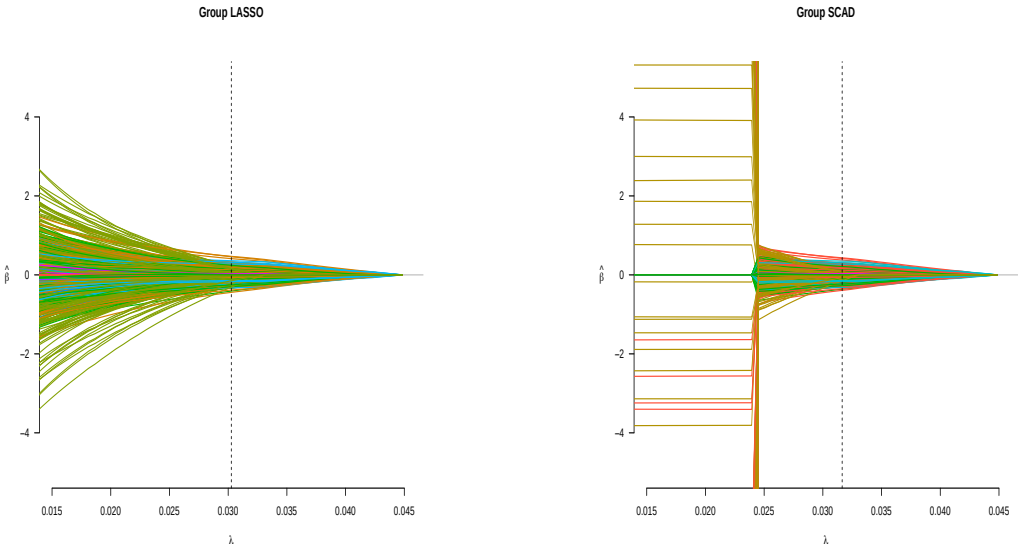



**Figure** 4.5 – Distribution de paramètre de régularisation pour les trois méthodes sur 100 simulations

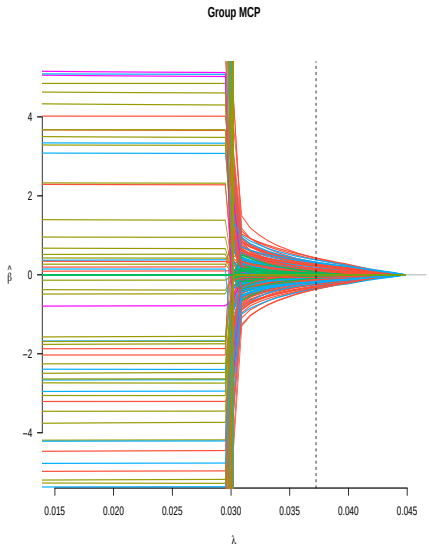
une valeur de $2,0 \times 10^{-3}$ pour le group MCP.

Si l'on considère l'erreur de validation croisée (figure 4.7), les résultats pour le group Lasso et le group SCAD sont également similaires, avec des valeurs autour de 1459, le groupe MCP étant légèrement inférieur à une valeur moyenne de 1463. En resumé, pour cet exemple des données réeles, on peut dire que la méthode group Lasso et la méthode SCAD ont les mêmes performances.




(a) Group LASSO Solution path

(b) Group SCAD Solution path




(c) Group MCP Solution path

**Figure** 4.6 – Chemin de régularisation pour la méthode : Group Lasso, Group SCAD, Group MCP pour un example de données réelles

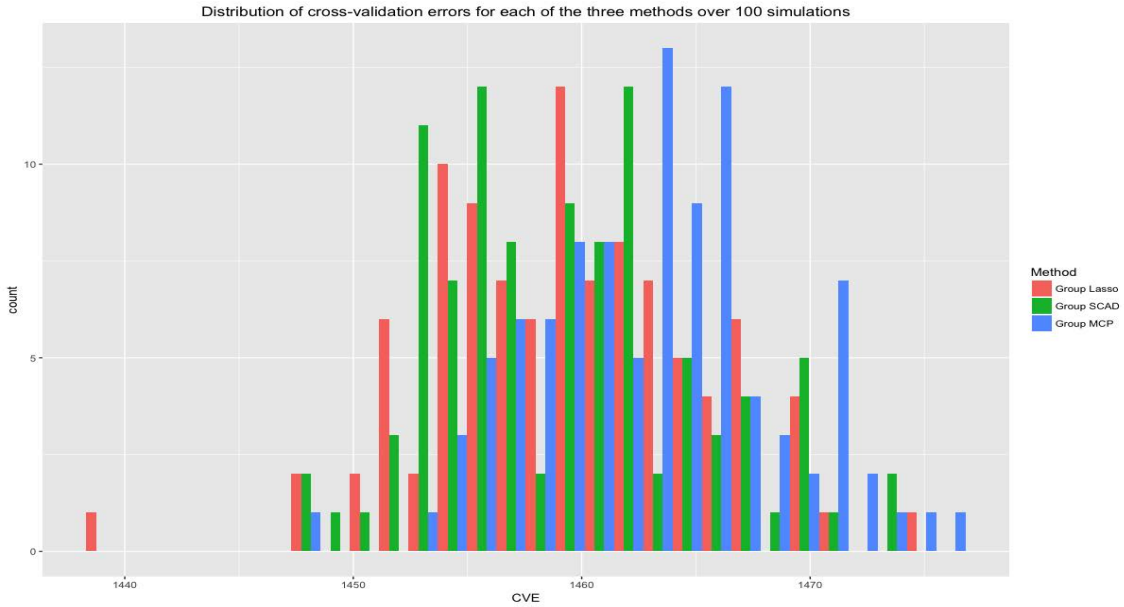



**Figure** 4.7 – Distribution de l'erreur de la validation croisée pour les trois méthodes sur 100 simulations

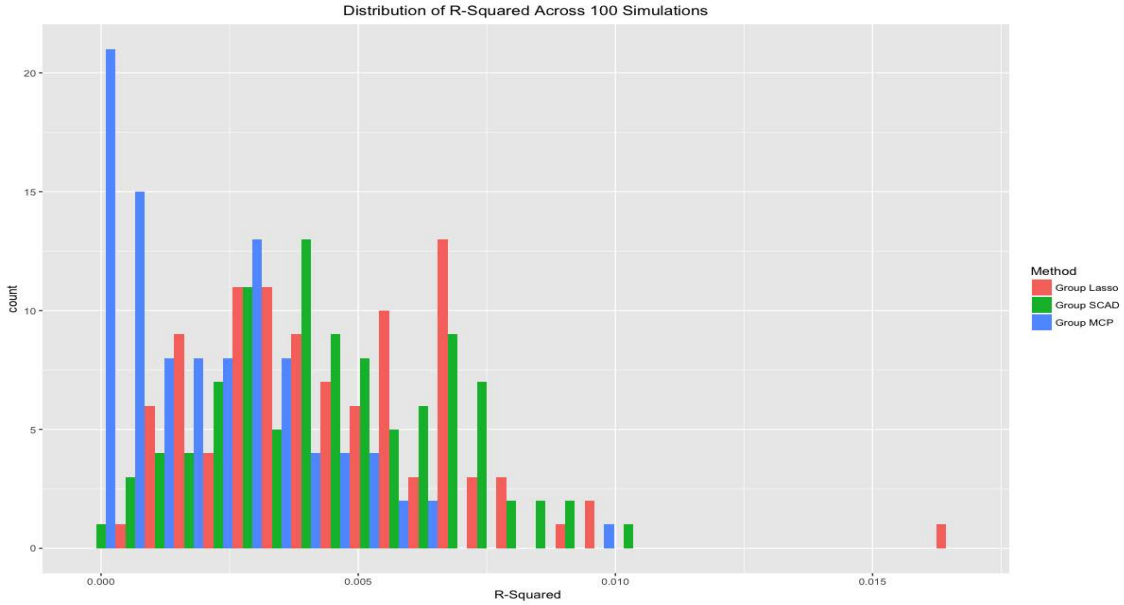



**Figure** 4.8 – Distribution de valeurs $R^2$ pour les trois méthodes sur 100 simulations

## 4.7 Conclusion

La performance relativement faible de toutes les méthodes de sélection des variables de groupe, y compris le group Lasso sur l'exemple de données du monde réelles en 4.6.2, met en évidence des problèmes dans ce domaine pour relever ces

défis. Le manque de données informatives sur la structure des groupes de gènes et leurs voies d'entrée rend la description et la mise en oeuvre de leurs sélections de groupes moins réalistes et affaiblit l'utilité des données sur les résultats. De plus, le concept de fragilité partagé par un groupe d'individus, un concept constituant une hétérogénéité non mesurée entre des groupes d'individus, est difficile à quantifier et exige un certain degré de connaissance sur des différences entre les groupes, même si elle n'est pas bien mesurée. L'application de méthode du group Lasso en tenant compte de la fragilité sur une base de données qui n'a pas été conçus de façon prospective pour recueillir les informations pertinentes afin de mettre en évidence les structures de simularité et de regroupement constitue un obstacle majeur pour une analyse rigoureuse de durée de survie.

Des limitations de cette méthode chevauchent des limitations du Lasso standard. La méthode group Lasso demeure une méthode de pénalisation qui ne convient pas à toutes les études et circonstances et qui est parfois dépassée par la régression Ridge, la régression LARS et régression garrotte non négative [130]. Même si la méthode group LASSO et la fragilité partagée par un groupe d'individus font des ajustements pour tenir compte des effets de regroupement, cette méthode exige des connaissances de base sur des données qui ne sont pas disponibles pour de nombreuses bases de donnés pour de nommbreux probèmes de recherche. Les recherches futures devraient continuer à élucider ces scénarios et à rendre certaines bases de données plus faciles à implémenter la méthode group LASSO. Pour un $\lambda_n$ approprié, tous les coefficients de certains groupes seront mis à zéro simultannément. Cependant, le group Lasso ne produit pas de parcimonie au sein d'un groupe, ce qui serait préférable. Par exemple, si les prédicteurs sont des gènes, il serait utile d'identifier les gènes particulièrement " importants " dans les voies d'intérêt. Pour ce, [41] ont discuté la méthode "group Lasso Sparse" dans le modèle de régression linéaire où ils ont introduit à la fois les deux pénalités $l_1$ et $l_2$. Ils ont discuté de la "sparsité" et d'autres propriétés de régularisation de l'ajustement optimal de ce modèle et ont montré que la méthode a un effet désirable de "sparsité" de groupe et à l'intérieur du groupe. Même si le group Lasso est une méthode attrayante pour la sélection de variables, puisqu'il tient compte de la structure de données groupes, il n'est généralement pas consistent pour la sélection et peut également sélectionner des groupes qui ne sont pas importants dans le modèle [121]. Pour améliorer les résultats de la sélection, les chercheurs ont proposé une méthode Adaptive group Lasso qui est aussi une généralisation de la méthode Adaptive Lasso.

Dans ce contexte, dans le chapitre 5, nous avons généralisé cette méthode en adaptant la méthode " Adaptive group Lasso" dans le modèle à risques proportionnels de Cox avec fragilité, pour optimiser la sélection des variables groupées.

CHAPITRE 5

# Méthode du group Adaptive Lasso dans le modèle de Cox avec fragilité


## Résumé

*Ce chapitre présente l'intérêt de la méthode Adaptive group Lasso pour généraliser la méthode group Lasso dans la sélection des variables groupées. Quelques pénalités classiques sélectionnant des variables individuelles seront introduites. Ces pénalités nous permettent d'élaborer clairement la formulation mathématique des méthodes de sélection de variables groupées, particulièrement la méthode Adaptive group Lasso dans le modèle de Cox avec fragilité. La consistance et la sparsité (c'est-à-dire une combinaison comportant un nombre important de coefficients nuls dans le modèle) de la méthode Adaptive group Lasso seront présentées. Ensuite, l'algorithme "Group coordinate descent" sera utilisé pour comparer les performances de la méthode proposée avec des méthodes concurrentes à savoir : la méthode group Lasso, la méthode group SCAD et la méthode group MCP.*

## Sommaire

## 5.1 Introduction

La méthode group-Lasso agit comme le Lasso au niveau des groupes. Pour un paramètre $\lambda$ approprié, tous les coefficients d'un groupe seront mis à zéro simultanément. Cependant, le group Lasso ne produit pas de parcimonie au sein d'un groupe. Il est important de noter que si la taille de chaque groupe $j$ de variables notée $p_j$ est égale à 1, alors le Lasso par groupe se réduit au Lasso. De plus, le Lasso par groupe ne possède pas les propriétés oracles puisqu'il pénalise tous les groupes de la même façon, ce qui résulte en une estimation biaisée des coefficients significativement différents de zéro. Afin d'éviter ce problème, [119] ont proposé la méthode Adaptive group-Lasso pour la régression linéaire.

## 5.2 Méthodes d'estimation pénalisée

La pénalité la plus couramment utilisée est la pénalité $l_1$ connue aussi sous le nom de la méthode Lasso (Least Absolute Shrinkage and Selection Operator) proposée pour la première fois par [111] et définie comme

$$\mathcal{P}_{\lambda}^{\text{lasso}}(\beta) = \lambda|\beta| \ \ \forall \ \ \beta \in \mathbb{R}, \ \lambda > 0 \tag{5.1}$$

et la pénalité "Hard-thresholding" proposée par [7] définie comme

$$\mathcal{P}_{\lambda}^{\text{hard}}(\beta) = \lambda^2 - (|\beta| - \lambda)^2 \, \mathbb{I}_{\{|\beta| \leq \lambda\}}, \tag{5.2}$$

où $\lambda$ est le paramètre de régularisation et $\mathbb{I}_{\{.\}}$ est la fonction indicatrice. Toutefois, ces deux pénalités ne remplissent pas simultanément les conditions mathématiques nécessaires : des conditions de non-biais, de sparsité et de continuité. Pour améliorer les propriétés de la pénalité $l_1$ et de la pénalité "Hard-thresholding", [35] ont proposé la fonction de pénalité différentielle continue connue sous le nom de pénalité "Smoothly Clipped Absolute Deviation (SCAD)" définie comme suit :

$$\mathcal{P}_{\lambda,\gamma}^{\text{scad}}(\beta) = \begin{cases} \lambda\beta, & \text{si } \beta \leq \lambda \\ \dfrac{\gamma\lambda - 0.5(\beta^2 + \lambda^2)}{\gamma - 1}, & \text{si } \lambda < \beta \leq \lambda\gamma \\ \dfrac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)}, & \text{si } \beta > \gamma\lambda. \end{cases} \tag{5.3}$$

Sa dérivée première est définie comme :

$$\mathcal{P}_{\lambda,\gamma}^{\text{scad}}(\beta)' = \lambda \left\{ \mathbb{I}_{\{|\beta| \leq \lambda\}} + \frac{(\gamma\lambda - \beta)_+}{(\gamma - 1)\lambda} \mathbb{I}_{\{\beta > \lambda\}} \right\} \tag{5.4}$$

$$= \begin{cases} \lambda, & \text{si } \beta \leq \lambda \\ \dfrac{\gamma\lambda - \beta}{\gamma - 1}, & \text{si } \lambda < \beta \leq \lambda\gamma \\ 0, & \text{si } \beta > \gamma\lambda. \end{cases} \tag{5.5}$$

où le paramètre de régularisation $\lambda > 0$ permet de gérer le compromis entre la qualité d'ajustement du modèle et la parcimonie. Le paramètre $\gamma > 2$ est un paramètre de rétrécissement qui détermine l'ordre de la pénalité sur la fonction. D'un point de vue bayésien, [35] confirment que pour un $\gamma \approx 3,7$, l'on observe des résultats satisfaisants dans de nombreux problèmes de sélection de variables. Dans le même contexte, [133] ont également proposé une sélection de variables non biaisée sous la pénalité concave minimale (MCP) prenant $\lambda \geq 1$ et $\gamma > 1$ pour la fonction de pénalité définie sur $[0, \infty)$. La pénalité de la MCP est définie comme suit :

$$\mathcal{P}^{\text{mcp}}_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta - \dfrac{\beta^2}{2\gamma}, & \text{si } \beta \leq \lambda\gamma \\ \dfrac{1}{2}\gamma\lambda^2, & \text{si } \beta > \lambda\gamma. \end{cases} \tag{5.6}$$

Sa dérivée première par rapport à $\beta$ est donnée par :

$$\mathcal{P}^{\text{mcp}}_{\lambda,\gamma}(\beta)' = \begin{cases} \lambda - \dfrac{\beta}{\gamma}, & \text{si } \beta \leq \lambda\gamma \\ 0, & \text{si } \beta > \lambda\gamma. \end{cases} \tag{5.7}$$

Un inconvénient de ces types de sélection de variables est qu'ils ne sélectionnent que des variables individuelles et ne peuvent pas traiter adéquatement des variables catégorielles comme le sexe, les antécédents familiaux et d'autres variables cliniques. En combinant des covariables continues (comme l'expression des gènes) et catégorielles (comme le sexe et les antécédents familiaux), il est possible d'obtenir de meilleures prédictions [77]. Un autre inconvénient associé à la sélection de variables individuelles est que les voies biologiques affectées par l'expression génétique et les mutations ne peuvent être prises en compte pour étudier des maladies complexes et multifactorielles. Par conséquent, il serait raisonnable de sélectionner des groupes de gènes apparentés plutôt que des gènes individuels [76] et de les ajuster au sein des groupes en fonction des connaissances disponibles. Le modèle du groupe lasso [128, 10, 20, 8] surmonte ces problèmes dans un contexte de régression linéaire en introduisant une extension appropriée de la pénalité.

$$\mathcal{P}^{\text{glasso}}_{\lambda}(\beta) = \lambda \sum_{j=1}^{K} \|\beta_j\|. \tag{5.8}$$

Le paramètre d'intérêt $\beta$ est décomposé en $K$ vecteurs $\beta_j, j = 1, 2, ..., K$ correspondant respectivement aux $K-$groupes des covariables. $\|.\|$ est la norme euclidienne. [68] ont étendu cette approche aux données de survie. Le group Lasso pénalise chaque facteur de la même manière que le Lasso standard en utilisant le paramètre de régularisation $\lambda$ pour chaque facteur indépendamment de son importance relative. Dans un cadre de régression typiquement linéaire, il a été démontré qu'une telle pénalité excessive appliquée aux variables pertinentes peut dégrader l'efficacité de l'estimation [35] et affecter la consistance de la sélection [129, 136]. La méthode group Lasso présente donc les mêmes inconvénients. Pour surmonter des telles limitations dans le cas du modèle de régression linéaire, [119] ont proposé la pénalité adaptive group Lasso qui est définie comme

$$\mathcal{P}_{\lambda_j}^{\text{Aglasso}}(\beta) = \sum_{j=1}^{K} \lambda_j \|\beta_j\| \tag{5.9}$$

avec $\lambda_j$ le paramètre de régularisation spécifique pour un groupe $j$ et il est approximé dans la section 5.4. La méthode Adaptive group Lasso permet aux différents facteurs d'être associés aux différents paramètres de régularisation. Une telle flexibilité produit à son tour des différents degrés de pénalisation. Intuitivement, si une très forte pénalisation est appliquée aux coefficients des variables non pertinentes et une faible pénalisation est utilisée pour les coefficients des variables pertinentes, un estimateur avec une meilleure efficacité peut être obtenu. Motivés par l'extension de la méthode group Lasso aux données de survie [68], nous considérons la méthode Adaptive group Lasso dans l'analyse de survie de Cox avec une fragilité partagée discutée dans [32] et [60].

## 5.3 Estimateur Adaptive Group Lasso dans le modèle à risque proportionnel de Cox avec fragilité

Supposons qu'on a $n$ groupes d'individus et que le $i^{eme}$ groupe est constitué de $J_i$ individus avec une fragilité partagée $u_i$. Soit $Z_{ij}$ un vecteur de covariables associé au temps de survie $X_{ij}$ de l'individu $j$ du groupe $i$. Supposons que pour cet individu, on dispose des données de survie $(T_{ij}, \delta_{ij}, Z_{ij}, u_i)$ qui sont des v.a qui sont $i.i.d$ avec $\delta_{ij} = \mathbb{I}_{\{X_{ij} \leq C_{ij}\}}$ l'indicatrice de censure, $C_{ij}$ le temps de censure et $T_{ij} = \min(X_{ij}, C_{ij})$

son temps de survie observé. La fonction de vraisemblance correspondante est donnée par :

$$L_n(\beta, \alpha, H_0(.)) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \left\{ h_{ij}\Big(T_{ij}|u_i, Z_{ij}\Big)^{\delta_{ij}} S_{ij}\Big(T_{ij}|u_i, Z_{ij}\Big) \right\} \prod_{i=1}^{n} g(u_i), \tag{5.10}$$

avec $S(t) = \exp(-H_0(t))$, la fonction de survie et $h(t|Z, u)$, le taux de hasard conditionnellement au vecteur de covariables $Z$ et de la fragilité $u_i$ où $u_i$ partagée par un groupe $i$. Les fragilités suivent une loi Gamma avec une densité donnée par :

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1} \exp(-\alpha u)}{\Gamma(\alpha)}.$$

On considère un modèle à risques proportionnels de Cox avec fragilité partagée,

$$h_{ij}(t|Z_{ij}, u_i) = h_0(t) u_i \exp(\beta^\top Z_{ij}), \tag{5.11}$$

avec $h_0(t)$ le taux de risque de base et $\beta$ un vecteur de paramètres d'intérêt. Soit $H_0(t) = \int_0^t h_0(\mu)\, d\mu$ le taux de risque cumulé. En remplaçant $h_{ij}(.)$ par sa valeur (équation 5.11), la fonction de la vraisemblance (équation 5.10) devient

$$L_n\left(\beta, \alpha, H_0(.)\right) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} h_0(T_{ij})^{\delta_{ij}} \exp(\beta^\top Z_{ij}) u_i^{\delta_{ij}} \exp\{-H_0(T_{ij}) \exp(\beta^\top Z_{ij}) u_i\} \prod_{i=1}^{n} g(u_i). \tag{5.12}$$

Partant de la même procédure que dans le chapitre 4, la vraisemblance profilée est donnée par :

$$\begin{aligned} \ell_n\left(\beta, \alpha, H_0(T_{ij})\right) \equiv & \sum_{i=1}^{n} \sum_{j=1}^{J_i} \delta_{ij} \left\{ \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \leq T_{ij}\}} \log \rho_l \right\} \\ & - \sum_{i=1}^{n} (A_i + \alpha) \log \left\{ \alpha + \sum_{j=1}^{J_i} \exp\left(\beta^\top Z_{ij}\right) \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \leq T_{ij}\}} \right\} \end{aligned} \tag{5.13}$$

où $A_i = \sum_{j=1}^{J_i} \delta_{ij}$ et $H_0(.)$ un taux de risque cumulé où $H_0(T_{ij})$ a des sauts de taille $\rho_l$ à des durées obsevées $\tilde{T}_l, , l = 1, ..., N$. C'est-à-dire

$$\begin{aligned} H_0(T_{ij}) &= \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \leq T_{ij}\}} \\ h_0(T_{ij}) &= \prod_{l=1}^{N} \rho_l^{\mathbb{I}_{\{\tilde{T}_l \leq T_{ij}\}}} \end{aligned} \tag{5.14}$$

En remplaçant (5.14) dans (5.13) et en dérivant le résultat obtenu rapport à $\rho_k$, on a

$$\frac{\partial \ell_n(\beta, \alpha, H_0(.))}{\partial \rho_k} = \sum_{i=1}^{n}\sum_{j=1}^{J_i} \delta_{ij} \mathbb{I}_{\{\tilde{T}_k \leq Z_{ij}\}} \frac{1}{\rho_k} - \sum_{i=1}^{n}(A_i + \alpha) \frac{\sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \mathbb{I}_{\{\tilde{T}_k \leq T_{ij}\}}}{\alpha + \sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \leq T_{ij}\}}}, \quad k = 1, ...N \tag{5.15}$$

Supposons qu'il n'y ait pas des événements simultanés ("ties") pour les différents groupes d'individus. La solution de l'équation (5.15) vérifie l'équation

$$\frac{1}{\rho_k} = \sum_{i=1}^{n} \frac{(A_i + \alpha) \sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \mathbb{I}_{\{\tilde{T}_k \leq T_{ij}\}}}{\alpha + \sum_{j=1}^{J_i} \exp(\beta^\top Z_{ij}) \sum_{l=1}^{N} \rho_l \mathbb{I}_{\{\tilde{T}_l \leq T_{ij}\}}}, \quad \text{pour} \quad k = 1, ...N \tag{5.16}$$

Pour simplifier les notations, on note $\ell_n(\beta, \alpha, H_0(.)) = \ell_n(\beta)$. L'estimateur Adaptive group Lasso dans le modèle à risque proportionnel de Cox avec fragilité est défini comme :

$$\mathcal{Q}(\beta) = -\ell_n(\beta) + \mathcal{P}_{\lambda_j}^{\text{Aglasso}}(\beta) \tag{5.17}$$

où $\mathcal{Q}(\beta)$ est une fonction objective qui est convexe à minimiser par rapport au paramètre d'intérêt $\beta$ avec un paramètre de régularisation $\lambda_j$ donné. Ce paramètre de régularisation contrôle le degré de pénalisation pour chaque group $j$ des variables. $\ell_n(\beta)$ est la log-vraisemblance définie dans (5.13) et $\mathcal{P}_{\lambda_j}^{\text{Aglasso}}(\beta)$ est la pénalité adaptive group Lasso définie dans l'équation (5.9). La vraie valeur du paramètre $\beta$ est notée $\beta^0$. Sans perte de généralité, nous supposons que les coefficients des $S \leq K$ premiers groupes de covariables sont différents de zéro et que les autres sont nuls. En prenant la dérivée première de la fonction objective $\mathcal{Q}(\beta, \lambda_n)$, les conditions d'optimalité de Karush-Kuhn-Tucker (KKT) pour trouver le minimum de $\mathcal{Q}(\beta, \lambda_n)$ sur un ensemble de $\beta$, peuvent être établies comme suit :

$$\begin{cases} \beta_j, & \forall \tilde{\beta}_j = 0 \\ -\frac{1}{n}\dot{\ell_{n,j}}(\beta) + \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|} = 0, & \forall \beta_j \neq 0, \tilde{\beta}_j \neq 0 \\ \|\frac{1}{n}\dot{\ell_{n,j}}(\beta)\| \leq \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|}, & \forall \beta_j = 0, \tilde{\beta}_j \neq 0. \end{cases} \tag{5.18}$$

En plus de la première condition, la troisième condition de la KKT en (5.18) force certaines composantes de $\beta$ à être nulles. La troisième condition du KKT sera principalement utilisée pour montrer la sparsité de la méthode. Dans ce chapitre, nous

comparons la méthode Adaptive group Lasso avec d'autres pénalités au niveau du groupe inclus group Lasso, group MCP et group SCAD. De la même formulation que [61], on peut définir la pénalité group SCAD comme

$$\mathcal{P}^{\text{gscad}}(\beta) = \sum_{j=1}^{K} \mathcal{P}^{\text{scad}}\Big(\sqrt{p_j}\|\beta_j\|; \lambda, p_j\gamma\Big) \tag{5.19}$$

et la pénalité group MCP comme

$$\mathcal{P}^{\text{gmcp}}(\beta) = \sum_{j=1}^{K} \mathcal{P}^{\text{mcp}}\Big(\sqrt{p_j}\|\beta_j\|; \lambda, p_j\gamma\Big) \tag{5.20}$$

avec $\mathcal{P}^{\text{scad}}(.)$, $\mathcal{P}^{\text{mcp}}(.)$ la pénalité SCAD et la pénalité MCP définies respectivement en (5.3) et en (5.6). Le multiplicateur $\sqrt{p_j}$ sur $\|\beta_j\|$ normalise la taille des groupes. Ceci signifie que les groupes de petite taille ne seront pas surpondérés par les groupes de grande taille. Le multiplicateur $p_j$ pour $\gamma$ rend le niveau de régularisation par groupe proportionnel à sa taille. Ainsi, l'interprétation de $\gamma$ reste la même que dans le cas où toutes les tailles de groupe sont égales à 1.

## 5.4 Paramètre de régularisation

Afin de pouvoir implémenter toutes ces méthodes d'estimation pénalisée citées ci-dessus, on doit décider des valeurs des paramètres de régularisation $(i.e., \lambda_j)$. Traditionnellement, la validation croisée (CV) ou la validation croisée généralisée (GCV) ont été largement utilisées. Cependant, ces méthodes ne sont pas directement adaptées pour la méthode Adaptive group Lasso. En effet, il y a trop de paramètres de régularisation à calculer ce qui exige un coût elevé de calcul. Une solution simple à ce défi informatique est de considérer, comme dans [136, 120, 134],

$$\lambda_j = \lambda_n \|\tilde{\beta}_j\|^{-\delta}$$

avec $\tilde{\beta}_j$ un estimateur consistent connu de $\beta_j$ et $\delta$ une constante positive prédéfinie. Tout au long de ce chapitre, $\delta$ sera fixé à 1. Ensuite, le problème de départ de sélection des K-paramètres de régularisation $(\lambda_1, ..., \lambda_K)$ se réduit simplement à un problème univarié pour $\lambda$. Par la suite, toute méthode de sélection du paramètre de régularisation appropriée peut être utilisée. Deux critères de sélection du modèle seront utilisés pour sélectionner le paramètre de régularisation. Le critère d'information Bayesienne [102] est défini comme :

$$BIC(\lambda_n) = \log(\ell_n(\hat{\beta}))/n + \log(n)d(\lambda_n)/n,$$

avec $d(\lambda_n)$ le degré de liberté de l'estimateur pour un $\lambda$ donné [84] et celui de la validation croisée détaillé par [90].

## 5.5 Algorithme de calcul

Pour aborder la sélection des variables par la méthode Adaptive group Lasso, group Lasso, group SCAD et group MCP, dans ce chapitre, nous décrivons l'algorithme "group coordinate descent" utilisé. C'est une extension de l'algorithme "coordinate descent". L'algorithme "coordinate descent" peut être appliqué soit aux pénalités convexes, Lasso et adaptive Lasso, soit aux pénalités non convexes, SCAD et MCP [39, 40, 14]. L'algorithme "coordinate descent" optimise la fonction objective par rapport à un seul paramètre $\beta_j$ associé à une variable donnée et fait défiler tous les paramètres jusqu'à ce que la convergence soit atteinte tandis que l'algorithme "group coordinate descent" optimise la fonction objective par rapport à un paramètre $\beta_j$ associé à un groupe des variables, et fait défiler tous les groupes jusqu'à la convergence [128]. Cet algorithme a été utilisé par [15] dans les cas de l'estimation pénalisée dans le modèle linéaire non convexe et régression logistique et [42] dans le modèle à risques compétitifs. L'algorithme "coordinate descent" dans le tableau 5.1 et décrit dans [39, 40, 14], présente les mises à jour de paramètres pour Lasso/Adaptive Lasso, SCAD et MCP.

L'algorithme "coordinate descent" du tableau 5.1 a été étendu sur la sélection par groupe [122, 15, 42] dans le tableau 5.2

### 5.5.1 L'algorithme de Group Coordinate Descent pour la méthode Adaptive group Lasso

On note $X_j = (x_{j_1}, ..., x_{jp_j})^\top$ une $(n \times p_j)-$ matrice design de variables groupées dans un groupe $j$ et $\eta = X_j\beta_j$. On définit le vecteur gradient $u = \partial\ell(\beta)/\partial\eta$ et la matrice Hessienne $H = \partial^2\ell(\beta)/\partial\eta\partial\eta^\top$. Le vecteur pseudo réponse est $Y = \eta + H^{-1}u$. En utilisant le développement de Taylor du second ordre, la log-vraisemblance négative peut être approximée par $-\ell(\beta) \approx 1/2\big(Y-\eta\big)^\top H\big(Y-\eta\big)$. Ainsi, à chaque étape d'itération, nous devons minimiser la fonction objective

$$\mathcal{Q}(\beta) = 1/2\big(Y-\eta\big)^\top H\big(Y-\eta\big) + \mathcal{P}_{\lambda_j}^{\text{Aglasso}}(\beta) \tag{5.21}$$

**Tableau** 5.1 – Mise à jour des paramètres dans l'algorithme "Coordinate descent"

| Methode | Mise à jour |
|---|---|
| Lasso/Adaptive Lasso<br>$\theta_j$ est le poids adaptif | $\beta_j^{m+1} \leftarrow \mathcal{S}(z_j, \theta_j\lambda) = \begin{cases} z_j - \theta_j\lambda, & \text{si } z_j > \theta_j\lambda \\ 0, & \text{si } \lvert z_j\rvert \leq \theta_j\lambda \\ z_j + \theta_j\lambda, & \text{si } z_j < \theta_j\lambda \end{cases}$ |
| SCAD | $\beta_j^{m+1} \leftarrow \mathcal{S}^{\text{Scad}}(z_j, \lambda, \gamma) = \begin{cases} \mathcal{S}(z_j, \lambda), & \text{si } \lvert z_j\rvert \leq 2\lambda \\ \frac{\mathcal{S}\left(z_j, \lambda\gamma/(\gamma-1)\right)}{1-1/(\gamma-1)}, & \text{si } 2\lambda < \lvert z_j\rvert \leq \lambda\gamma \\ z_j, & \text{si } \lvert z_j\rvert > \lambda\gamma \end{cases}$ |
| MCP | $\beta_j^{m+1} \leftarrow \mathcal{S}^{\text{Mcp}}(z_j, \lambda, \gamma) = \begin{cases} \frac{\mathcal{S}(z_j, \lambda)}{1-1/\gamma}, & \text{si } \lvert z_j\rvert \leq \lambda\gamma \\ z_j, & \text{si } \lvert z_j\rvert > \lambda\gamma \end{cases}$ |

**Tableau** 5.2 – Mise à jour des paramètres dans l'algorithme "Group Coordinate descent"

| Methode | Mise à jour |
|---|---|
| Group Lasso/Adaptive Lasso | $\beta_j^{(m+1)} \leftarrow \mathcal{S}\left(\lVert z_j\rVert, \sqrt{p_j}\theta_j\lambda\right)\frac{z_j}{\lVert z_j\rVert}, \quad \theta_j = \lVert\tilde{\beta}_j\rVert^{-1}$ |
| Group SCAD | $\beta_j^{(m+1)} \leftarrow \mathcal{S}^{\text{Scad}}\left(\lVert z_j\rVert, \sqrt{p_j}\lambda\right)\frac{z_j}{\lVert z_j\rVert}$ |
| Group MCP | $\beta_j^{(m+1)} \leftarrow \mathcal{S}^{\text{Mcp}}\left(\lVert z_j\rVert, \sqrt{p_j}\lambda\right)\frac{z_j}{\lVert z_j\rVert}$ |

Pour éviter le coût de calcul, nous remplaçons $H$ par une matrice diagonale $D$ avec des entrées $h(\eta)_i = -\partial^2\ell/\partial\eta_i^2, \quad i = 1, ..., n$, ainsi $Y = \eta + D^{-1}u$. On définit

$$z_j = n^{-1}X_j^\top D(u + X_j\beta_j^{(m)}).$$

En se servant du tableau 5.2, nous pouvons résumer et donner l'algorithme complet pour minimiser la quantité en (5.17) sur un ensemble de paramètres $\beta$ dans les étapes suivantes

**Tableau** 5.3 – Algorithme "Group Coordinate descent"

| Etape | Algorithme |
|---|---|
| **1.** | On obtient $\tilde{\beta}$ en minimisant la log-vraisemblance négative non pénalisée $-\ell(\beta)$ dans (5.17) par l'algorithme New-Raphson. |
| **2.** | Initialisation $\hat{\beta}_j = 0$ pour l'itération $m = 1$ |
| **3.** | On calcule $\eta,\ u,\ D$ et $Y$ en se basant sur les valeurs actuelles de $\hat{\beta}_j^{(m)}$ |
| **4.** | Minimisation de l'Eq (5.21) par rapport à la valeur de $\hat{\beta}_j^{m+1}$ de l'algorithme "group coordinate descent" dans le tableau 5.2 |
| **5.** | On pose $m = m + 1$. On repète **l'étape 3** et **l'étape 4** pour tous les groupes jusqu'à la convergence. |

## 5.6 La consistance théorique et la sparsité de la méthode

### 5.6.1 consistance

Considérons l'estimateur pénalisé du vecteur $\beta$ des paramètres :

$$\hat{\beta}_n(\lambda_n) = \arg\min_{\beta} \left\{ -\frac{1}{n}\ell_n(\alpha, \beta, H_0(.)) + \lambda_n \sum_{j=1}^{K} \frac{\sqrt{p_j}}{\|\tilde{\beta}_j\|}\|\beta_j\| \right\}$$

On note $\beta^0$ la vraie valeur du paramètre du modèle $\beta = (\alpha, \beta, H_0(.))$. $\forall \varepsilon > 0$, nous voulons montrer que $\mathbb{P}\left\{\|\hat{\beta}_n(\lambda_n) - \beta_n^0\| < \varepsilon\right\} \to 1$ lorsque $n \to \infty$.

**Théorème 5.6.1.** *(Consistance) Supposons que* $(Z_{ij}, X_{ij}, C_{ij}, u_i)$ *sont des v.a qui sont* $i.i.d$*,* $X_{ij}$ *et* $C_{ij}$ *sont conditionnellement indépendants sachant* $Z_{ij}$ *et* $u_i$*,* $i = 1, ..., n, \quad j = 1, ..., J_i$*.Sous les conditions de régularité (A)-(D) en [5], si* $\lambda_n \to 0$ *quand* $n \to \infty$*, alors* $\hat{\beta}_n(\lambda_n)$ *est un estimateur consistant de* $\beta^0$*.*

***Preuve.*** La preuve du Théorème 5.6.1 est similaire à celle du Théorème 5.7 de [115] avec une approche légèrement différente. Il suffit de montrer que $\forall \varepsilon > 0$

$$\mathbb{P}\left\{ \sup_{\beta:\|\beta-\beta^0\|=\varepsilon} Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n) \right\} \to 1, \tag{5.22}$$

ce qui implique qu'avec une probabilité qui tend vers 1, $Q_n(\beta_n, \lambda_n)$ a un minimum local dans une boule $\{\beta : \|\beta - \beta^0\| < \varepsilon\}$ pour $\lambda_n$ donnée. De plus $\mathbb{E}\{Q_n(\beta_n^0, \lambda_n\} \to$

$\mathbb{E}\{n^{-1}\ell_n(\beta)\}$ quand $\lambda_n \to 0$ et $\mathbb{E}\{n^{-1}\ell_n(\beta)\}$ est une fonction strictement convexe avec un minimiseur global $\beta^0$. Pour montrer (5.22), nous montrons tout d'abord que $Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n)$. Puisque $\tilde{\beta}$ est un estimateur consistent de $\beta^0$, on a

$$\mathbb{P}\left\{\|\tilde{\beta}_j\|\mathbb{I}_{\{\beta_j^0 \neq 0\}} > \frac{1}{2}\|\beta_j^0\|\right\} \to 1$$

et il existe une constante $C > 0$ telle que

$$\min_{j=1}^{S}\|\tilde{\beta}_j\| > \frac{1}{2}\min_{j=1}^{S}\|\beta_j^0\| > C, \quad S \leq K \ (S \textbf{ premiers facteurs}).$$

Posons $\Delta_n = Q_n(\beta_n, \lambda_n) - Q_n(\beta_n^0, \lambda_n)$. Alors,

$$\begin{aligned}
\Delta_n &= -\frac{1}{n}\left(\ell_n(\beta) - \ell_n(\beta^0)\right) + \sum_{j=1}^{K}\frac{\lambda_n\sqrt{p_j}}{\|\tilde{\beta_j}\|}\left(\|\beta_j\| - \|\beta_j^0\|\right) \\
&\geq -n^{-1/2}\left(n^{-1/2}\frac{\partial}{\partial\beta}\Big(\ell_n(\beta^0)\Big)\right)^{\top}\left(\beta - \beta^0\right) + \left(\beta - \beta^0\right)^{\top}\left(n^{-1/2}\frac{\partial^2}{\partial\beta^2}\Big(\ell_n(\beta^0)\Big)\right)\left(\beta - \beta^0\right) \\
&\quad + n^{-1}o_p\left(\|\beta - \beta^0\|^2\right) + \lambda_n\sum_{j=1}^{S}\frac{\sqrt{p_j}}{\|\tilde{\beta_j}\|}\left(\|\beta_j\| - \|\beta_{(j)}^0\|\right) \\
&\geq -n^{-1}O_p(1)\|\beta - \beta^0\| + \left(\beta - \beta^0\right)^{\top}\left(\Sigma + o_p(1)\right)\left(\beta - \beta^0\right) \\
&\quad + n^{-1/2}o_p\left(\|\beta - \beta^0\|^2\right) - \lambda_n\sum_{j=1}^{S}\frac{\sqrt{p_j}}{c}\left(\|\beta_{(j)}\| - \|\beta_{(j)}^0\|\right)
\end{aligned}$$

Puisque $\lambda_n \to 0$ quand $n \to \infty$ alors

$$Q_n(\beta_n, \lambda_n) - Q_n(\beta_n^0, \lambda_n) \geq \left(\beta - \beta^0\right)^{\top}\left(\Sigma + o_p(1)\right)\left(\beta - \beta^0\right) \geq 0,$$

car $\Sigma$ est positif. Puisque $Q_n(\beta_n, \lambda_n) \geq Q_n(\beta_n^0, \lambda_n)$, nous avons,

$$\left\{\sup_{\beta:\|\beta-\beta^0\|=\varepsilon} Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n)\right\} \subseteq \left\{\|\hat{\beta}_n(\lambda_n) - \beta_n^0\| < \varepsilon\right\}, \quad \varepsilon > 0.$$

D'où

$$\mathbb{P}\left\{\|\hat{\beta}_n(\lambda_n) - \beta_n^0\| < \varepsilon\right\} \geq \mathbb{P}\left\{\sup_{\beta:\|\beta-\beta^0\|=a} Q_n(\beta_n, \lambda_n) > Q_n(\beta_n^0, \lambda_n)\right\},$$

ce qui implique que

$$\mathbb{P}\left\{\|\hat{\beta}_n(\lambda_n) - \beta_n^0\| < \varepsilon\right\} \to 1$$

□

### 5.6.2 Sparsité de la méthode

L'une des propriétés attrayantes de la méthode Adaptive group Lasso par rapport au group Lasso est sa "sparsité", c'est-à-dire une combinaison comportant un nombre important de coefficients nuls dans le modèle. Nous adoptons les notations suivantes $\beta^0 = (\beta_1^0, \beta_2^0)$, avec $\beta_1^0 = \Big((\beta_1^0)^\top, ..., (\beta_S^0)^\top\Big)^\top$ comme un vecteur de $S \leq K$ premiers facteurs. Nous supposons que ces $S$ facteurs sont pertinents ($\|\beta_j\| \neq 0$ pour $j \leq S$), et $\beta_2^0 = \Big((\beta_{S+1}^0)^\top, ..., (\beta_K^0)^\top\Big)^\top$ un vecteur de tous les facteurs non pertinents ($\|\beta_j\| = 0$ pour $j > S$). Une fois de plus, nous notons $\hat{\beta}_{n,1}(\lambda_n)$ et $\hat{\beta}_{n,2}(\lambda_n)$ leurs estimateurs Adaptives group Lasso associés. Le théorème suivant montre que si l'Adaptive group Lasso choisit de manière consistente le vrai modèle, alors avec une probabilité tendant vers 1, tous les coefficients associés aux variables non pertinents doivent être estimés exactement comme zéro, c'est-à-dire, $\mathbb{P}\Big\{\hat{\beta}_{n,2}(\lambda_n) = 0\Big\} \to 1$

**Théorème 5.6.2.** *(Sparsité) On suppose que $n\lambda_n \to \infty$ lorsque $n \to 0$, dans les conditions du Théorème 5.6.1, avec une probabilité qui tend vers 1. Alors l'estimateur Adaptive group Lasso $\hat{\beta}_n(\lambda_n)$ vérifie les conditions de sparsité : $\hat{\beta}_{n,2}(\lambda_n) = 0$.*

***Preuve.*** Puisque la troisième condition de KKT établie dans léquation (5.18) est une condition nécessaire et suffisante pour minimiser la fonction objective, on a $\Big\{\hat{\beta}_{n,2}(\lambda_n) = 0\Big\}$ quand $\|\frac{1}{n}\dot{\ell_{n,j}}(\beta)\| \leq \frac{\lambda\sqrt{p_j}}{\|\tilde{\beta}_j\|}$. Il suffit de démontrer que

$$\mathbb{P}\left\{\|n^{-\frac{1}{2}}\dot{\ell_{n,j}}\Big(\hat{\beta}_n(\lambda_n)\Big)\| \geq \frac{\sqrt{n}\lambda_n\sqrt{p_j}}{\|\tilde{\beta}_j\|}\right\} \to 0, n \to 0.$$

En appliquant le développement de Taylor, on a

$$\begin{aligned} n^{-\frac{1}{2}}\dot{\ell_{n,j}}\big(\hat{\beta}_n(\lambda_n)\big) &\triangleq n^{-\frac{1}{2}}\frac{\partial}{\partial\beta}\ell_{n,j}\big(\hat{\beta}_n(\lambda_n)\big) \\ &= n^{-\frac{1}{2}}\dot{\ell_{n,j}}\big(\beta^0\big) + \frac{1}{n}\Big(\frac{\partial}{\partial\beta}\ell_{n,j}\big(\beta^\star\big)\Big)'\sqrt{n}\Big(\hat{\beta}_n(\lambda_n) - \beta^0\Big) \\ &= \mathcal{O}_p(1) - \mathcal{O}_p(1)\mathcal{O}_p(1) \\ &= \mathcal{O}_p(1), \end{aligned} \tag{5.23}$$

avec $\beta^\star$ sur le segment entre $\hat{\beta}_n(\lambda_n)$ et $\beta^0$. D'autre part,

$$\begin{aligned} \frac{\sqrt{n}\lambda_n\sqrt{p_j}}{\|\tilde{\beta}_j\|} &= \sqrt{n}\lambda_n\sqrt{p_j}/\mathcal{O}(n^{-\frac{1}{2}}) \\ &= n\lambda_n\sqrt{p_j}\mathcal{O}(1) \to \infty, \end{aligned} \tag{5.24}$$

puisque $n\lambda_n \to 0$ quand $n \to 0$. En comparant l'équation (5.23) et (5.24), on voit clairement que lorsque $n \to 0$,

$$\mathbb{P}\left\{\|n^{-\frac{1}{2}}\dot{\ell}_{n,j}\left(\hat{\beta}_n(\lambda_n)\right)\| \geq \frac{\sqrt{n}\lambda_n\sqrt{p_j}}{\|\tilde{\beta}_j\|}\right\} \to 0.$$

C'est-à-dire

$$\mathbb{P}\left\{\hat{\beta}_{n,2}(\lambda_n) = 0\right\} \to 1.$$

□

## 5.7 Résultats et Discussion

Pour implementer la méthode proposée, dans ce chapitre, nous appliquons des méthodes de sélection de variables groupées : méthode Adaptive group Lasso avec fragilité partagée par groupe d'individus dans la sélection du modèle. Ensuite, nous comparons ses performances avec celles du group SCAD et du group MCP à un exemple d'une base de données réelles sur la survie au cancer du sein du Programme de l'Institut national de surveillance, d'épidémiologie et des résultats (SEER) des Etats-Unis. Les variables supplémentaires et non significatives ont été supprimées si l'on savait sans doute qu'elles n'avaient aucune incidence biologique ou démographique sur l'issue du cancer. Après le processus de nettoyage des données, on est resté avec 236 observations contenant 44760 micropuces de ADN regroupées en 100 groupes en fonction de caractéristiques biologiques ou démographiques communes. Pour de raison de la forte corrélation observée dans l'ensemble de données, nous avons utilisé l'analyse en composantes principales pour enlever la corrélation de données.

La figure 5.1 montre les chemins de régularisation pour la méthode Adaptive group Lasso, le group LASSO, le group SCAD et le group MCP, respectivement. En se basant sur la validation croisée généralisée (GCV), la méthode group Adaptive group Lasso tend à sélectionner le plus petit nombre de groupes de variables à une valeur de paramètre de régularisation plus petite par rapport aux autres groupes de variables sélectionnées. Ceci nous garantit à un certain niveau le choix du modèle parcimonieux par la méthode Adaptive group Lasso pour cet exemple de données réelles. Le paramètre de régularisation $\lambda$ passe de $\lambda = 0$ lorsque le modèle est excessivement grand à la valeur maximale de $\lambda_{max}$ lorsque tous les coefficients pénalisés sont à 0. Pour cet exemple, nous nous sommes rendus compte

qu'en ce qui concerne les critères de sélection du modèle BIC, presque toutes les 4 méthodes ne sélectionnent aucun groupe de variables. La valeur de rétrécissement $\gamma$ pour la méthode group MCP et la méthode group SCAD a été fixée à $2.7$ et à $3.7$ respectivement, conformément à la recommandation par défaut suggérée dans [42].

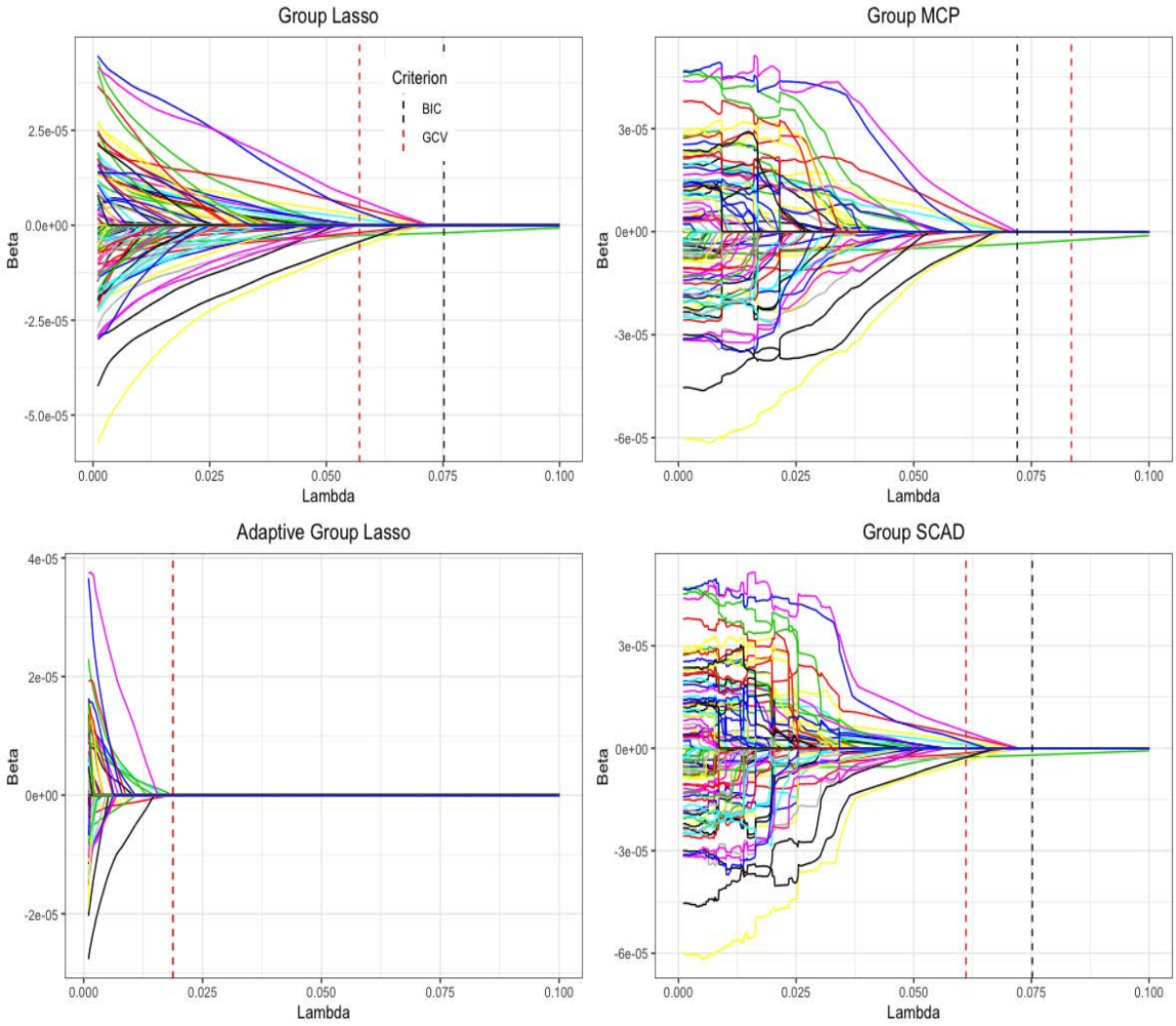



**Figure** 5.1 – Chemin de régularisation des méthode group lasso, Adaptive group Lasso, group MCP, et group SCAD

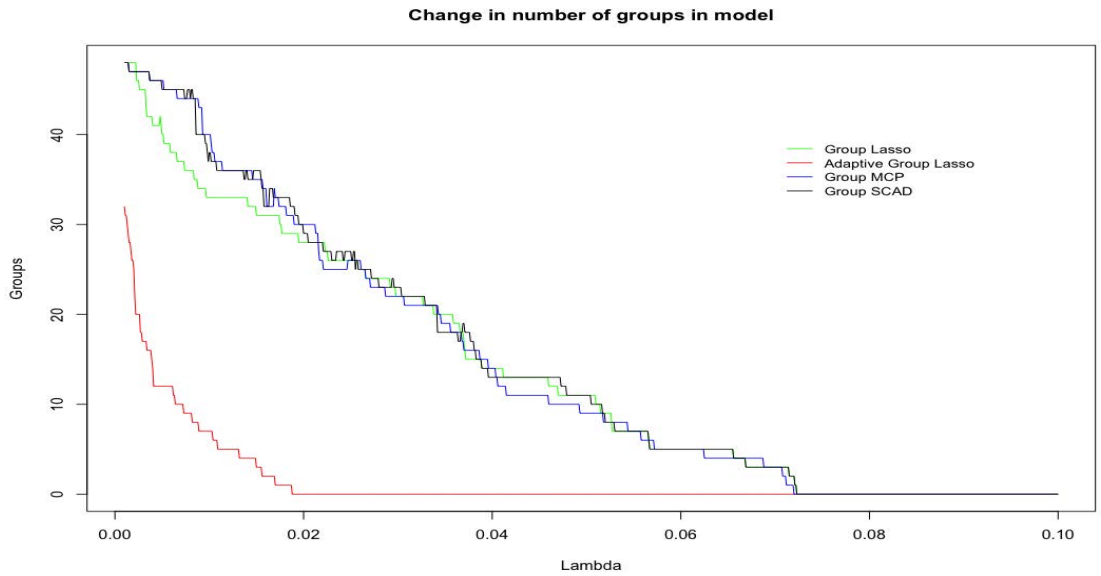



**Figure** 5.2 – Taille du modèle par rapport au paramètre de régularisation $\lambda$ des méthode group lasso, Adaptive group Lasso, group MCP, et group SCAD

La figure 5.2 complète la figure 5.1. Il montre la variation à laquelle un groupe de variables est sélectionné par la méthode adaptive group Lasso, group Lasso, group MCP et group SCAD. En général, la méthode adaptive group Lasso sélectionne des modèles avec beaucoup plus petit nombre de variables et s'approche de la taille réelle du modèle (à la valeur optimale de paramètre de régularisation) beaucoup plus rapidement que la méthode group Lasso, group MCP et group SCAD.

Pour évaluer la performance de ces 4 méthodes pour cet exemple des données, nous évaluons la précision du modèle par l'erreur quadratique moyenne (RMSE). La figure 5.2 montre que pour les petites valeurs du paramètre de régularisation, les 4 méthodes de régularisation par groupe fonctionnent de manière similaire. Cependant, suivant l'augmentation de la valeur de ce paramètre, la méthode adaptive group Lasso tend à avoir la plus petite valeur de RMSE par rapport au group Lasso, group MCP et group SCAD. En comparant le group MCP et le group SCAD, les deux sont presque identiques en termes de précision de prédiction tandis que le group Lasso a une meilleure précision en comparant avec le group MCP et le group SCAD.

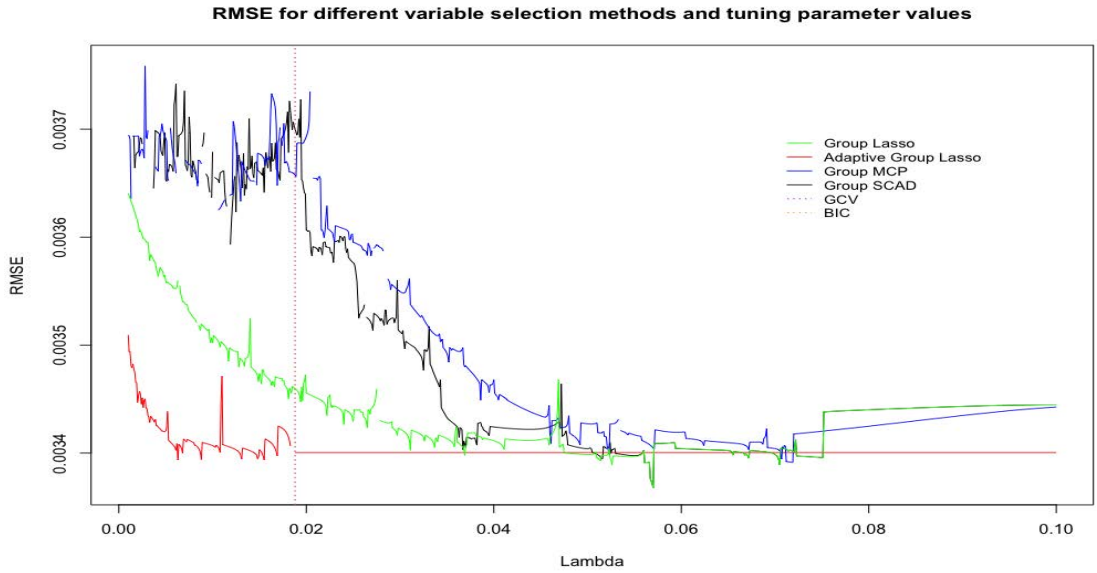



**Figure** 5.3 – Evaluation de l'erreur de prédiction

## 5.8 Conclusion et perspectives

Dans ce chapitre, nous avons proposé la méthode Adaptive group Lasso pour la sélection des variables groupées dans le modèle de Cox avec fragilité. La méthode adaptive group Lasso est un moyen alternatif relativement puissant par rapport au group Lasso dans les problèmes impliquant la sélection de variables groupées. Cependant, il y a beaucoup à faire dans ce domaine. La performance relativement faible de toutes les méthodes de sélection des variables groupées, y compris la méthode Adaptive group Lasso sur un exemple d'une base des données réelles met en évidence la principale question dans ce domaine, à savoir comment relever ces défis. Le manque de connaissances avancées sur la structure des groupes de gènes et leurs voies d'entrée rend la description et la sélection de variables groupées moins réalistes et affaiblit l'utilité des données sur les résultats. Dans le cadre de travaux futurs, nous aimerions appliquer la méthode adaptive group Lasso à différents ensembles de données, y compris des exemples simulés, afin de généraliser les résultats de cette méthode.

# Conclusion générale et Perspectives

## 5.9 Conclusion générale

Dans cette thèse, nous nous sommes intéressés aux problèmes de sélection de variables en grande dimension en utilisant des méthodes d'estimation pénalisée pour des données de survie censurées. Nous avons mis l'accent sur la méthode group Lasso dans le modèle de Cox avec fragilité et nous avons généralisé cette approche.

En effet les données massives sont de plus en plus populaires et de plus en plus utilisées dans la recherche statistique. Les chercheurs recueillent des milliers de données pour mieux comprendre et mieux expliquer divers phénomènes. Par exemple, dans le domaine biomédical, les scientifiques collectent de l'information sur une maladie complexe comme le Cancer pour mieux saisir les interactions entre les gènes et mieux cibler les dysfonctionnements survenant dans un réseau de voies biologiques (les gènes faisant partie d'un même groupe et interagissant ensemble) responsable à cette maladie. Nos objectifs principaux lors de l'analyse des données de survie dans le modèle de régression à risques proportionnels de Cox, étaient d'identifier les facteurs de risque d'une maladie, de comparer les réponses à des traitements, d'estimer les probabilités de survenue d'un événement (décès, rechute, etc.) chez un individu identifié par un vecteur donné de variables explicatives. Nous avons dû faire face à trois grands défis.

Tout d'abord, dans des jeux de données de grandes dimensions, plusieurs gènes ne sont pas informatifs. C'est-à-dire qu'il y a des gènes qui n'ont pas de lien avec la maladie. Ceci nous conduit à la singularité d'une matrice de covariables de dimension $n \times p$ contenant les expressions génétiques où le nombre de colonnes est beaucoup plus grand que le nombre de lignes ($p >> n$). Pour remédier à ce problème, dans cette thèse, nous avons présenté et appliqué différentes méthodes de régularisation qui ont été développées dans le modèle de Cox pour sélectionner un modèle parcimonieux en éliminant les informations non pertinentes.

Un second défi, est que la plupart des modèles de survie supposent que la population étudiée est homogène. Cela suppose que le risque de décès est le même pour tous les individus de cette population. Cependant ce n'est pas toujours le cas, les individus ont des propensions différentes sous l'effet d'un traitement ou l'influence de certaines covariables. Dans de nombreuses situations, il est impossible de mesurer toutes les covariables pertinentes liées à la pathologie d'intérêt et donc d'introduire ces covariables comme effets fixes dans le modèle. Les raisons sont souvent économiques ou liées à la méconnaissance de l'importance de certaines covariables sur la pathologie étudiée. Par ailleurs, lorsque l'étude porte par exemple, sur des patients issus de même famille ou un événement à répétition sur un même patient, la présence d'un facteur caché liant les unités statistiques d'un même groupe peut être suspectée. Ainsi, l'approche de la fragilité est un concept de modélisation statistique qui vise à représenter l'hétérogénéité dans une population étudiée provoquée par les covariables non mesurées. Les modèles de fragilités ont été ainsi discutés dans cette thèse. Un accent particulier a été mis sur les modèles à fragilité gamma partagée. L'aspect méthodologique, algorithmique et illustration des propos sur différentes bases de données ont été présentés.

Malgré les méthodes de régularisation développées durant ces dernières années autour des problèmes de singularité de la matrice de covariables et la réduction de dimension dans les modèles de survie, ces méthodes ne prennent pas en compte la structure de dépendance dans les données. Par exemple dans le cas de variables catégorielles, l'estimateur Lasso sélectionne les indicatrices des modalités et non le groupe d'indicatrices, i.e la variable dans sa totalité. Ceci nous amène au troisième défi. Pour remédier à tous ces défis relevés, dans cette thèse, nous avons proposé la méthode group Lasso dans la sélection des variables dans le modèle de Cox avec fragilité pour les données en clusters. La consistance théorique de la méthode proposée est établie. Cette méthode est appropriée pour répondre à différentes questions de recherche allant du domaine de la génétique à la pollution atmosphérique. L'analyse de données simulées et des données réelles montre des performances prometteuses pour le group Lasso par rapport à d'autres méthodes, y compris le group SCAD et le group MCP.

L'utilisation de cette méthode encourage tous les coefficients d'un même groupe à être simultanément nuls. Cependant, le group Lasso ne produit pas de parcimonie au sein d'un groupe. Afin de généraliser ce travail de thèse, nous nous sommes intéressés à la méthode Adaptive group Lasso dans le modèle de Cox pour les données en clusters. Nous avons aussi pris en compte l'hétérogénéité non observée

dans les groupes d'individus. La consistance théorique, la sparsité de la méthode, l'algorithme utilisé pour résoudre ce problème d'optimisation ont été presentés. Cette méthode fournit une sélection parcimonieuse à l'intérieur des groupes. L'application de cette méthode sur une base de données de Cancer de sein montre la bonne performance de prédiction de la méthode Adaptive group Lasso par rapport au group Lasso, group SCAD et le group MCP.

## 5.10 Perspectives

Nous terminons ce manuscrit en donnant quelques perspectives et prolongements naturels de ce travail de thèse.

La normalité asymptotique est une propriété plus utile dans la statistique asymptotique. Elle consiste en effet, pour une suite d'estimateurs $(\hat{\beta}_n)_{n\geq 1}$ i.i.d de loi $P$, à comprendre si et comment obtenir une loi asymptotique du genre $\sqrt{n}\left(\hat{\beta}_n - \beta^0\right)$ converge en loi vers une gaussienne. Avec $\hat{\beta}$ un estimateur supposé consistant (convergeant en probabilité vers $\beta^0$). Dans ce travail de thèse, nous regrettons de ne pas avoir traité cette propriété des estimateurs obtenus pour la méthode group Lasso et sa généralisation (méthode Adaptive group Lasso) dans le modèle de Cox avec fragilité. Cette propriété fait partie de nos premières perspectives dans le futur.

Le choix de la distribution de l'effet aléatoire reste toujours un problème majeur dans les modèles de fragilité. Dans cette thèse, nous nous sommes intéressés particulièrement à la fragilité gamma. Dans nos travaux futurs, nous allons nous intéresser à l'investigation et la modélisation de problèmes faisant intervenir d'autres distributions de l'effet aléatoire comme : Positive Stable, Log-normale. Nous pourrons également nous intéresser au cas où des effets aléatoires individuels (et non plus seulement propres aux clusters) sont présents dans la modélisation.

Dans les techniques de régularisation et de sélection de variables via la vraisemblance pénalisée, la complexité du modèle et le taux de rétrécissement appliqué aux coefficients de régression sont fortement liés au choix des paramètres de régularisation. Dans le chapitre 5 de ce travail de thèse, nous avons remarqué que seule la validation croisée pouvait sélectionner les paramètres de régularisation optimaux et par conséquent sélectionner des modèles parcimonieux (c'est-à-dire les modèles qui expliquent les phénomènes avec un minimum de variables explicatives). Les critères AIC (4.14) et BIC (4.2.2) ont été caractérisés par un grand taux

de rétrécissement et par conséquent, ils ne sélectionnent presque aucun groupe de variable. Nous comptons dans le futur de chercher sous quelles conditions, les critères d'information AIC ou BIC peuvent être de bonnes alternatives pour choisir les paramètres de régularisation optimaux, afin de sélectionner des variables pertinentes.

# Bibliographie


[1] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, pages 701–726, 1978. 25, 27

[2] Odd O Aalen. Effects of frailty in survival analysis. *Statistical Methods in Medical Research*, 3(3):227–243, 1994. 6

[3] H Akaike. Theory and an extension of the maximum likelihood principal. In *International symposium on information theory. Budapest, Hungary: Akademiai Kaiado*, 1973. 4, 101

[4] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012. 17, 27, 33, 34, 156

[5] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982. 44, 105, 106, 129

[6] Per Kragh Andersen, John P Klein, Kim M Knudsen, and René Tabanera y Palacios. Estimation of variance in cox's regression model with shared gamma frailties. *Biometrics*, pages 1475–1484, 1997. 61

[7] Anestis Antoniadis. Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2):97, 1997. 121

[8] Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001. 122

[9] Anestis Antoniadis, Piotr Fryzlewicz, and Frédérique Letué. The dantzig selector in cox's proportional hazards model. *Scandinavian Journal of Statistics*, 37(4):531–552, 2010. 6

[10] Sergey Bakin et al. Adaptive regression and model selection in data mining problems. 1999. 122

[11] Theodor Adrian Balan and Hein Putter. frailtyem: An r package for estimating semiparametric shared frailty models, 2017. 63

[12] Udo Baron, Ivana Turbachova, Alexander Hellwag, Florian Eckhardt, Kurt Berlin, Ulrich Hoffmüller, Paul Gardina, and Sven Olek. Dna methylation analysis as a tool for cell typing. *Epigenetics*, 1(1):56–61, 2006. 4

[13] R Bellmann. Dynamic programming princeton university press. *Princeton, NJ*, 1957. 2

[14] Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011. 127

[15] Patrick Breheny and Jian Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and computing*, 25(2):173–187, 2015. 127

[16] Leo Breiman et al. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383, 1996. 4

[17] Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, pages 45–57, 1975. 34

[18] Karl W Broman and Terence P Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):641–656, 2002. 5

[19] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. 81

[20] T Tony Cai. Discussion of "regularization of wavelet approximations"(by a. antoniadis and j. fan). *J. Am. Statist. Ass*, 96:960–962, 2001. 122

[21] Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007. 81

[22] George Casella, F Javier Girón, M Lina Martínez, Elias Moreno, et al. Consistency of bayesian procedures for variable selection. *The Annals of Statistics*, 37(3):1207–1228, 2009. 5

[23] David Clayton and Jack Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A (General)*, 148(2):82–108, 1985. 70

[24] Richard J Cook and Jerald Lawless. *The statistical analysis of recurrent events*. Springer Science & Business Media, 2007. 6, 55, 56

[25] National Research Council et al. *Science and decisions: advancing risk assessment*. National Academies Press, 2009. 108

[26] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. 5, 6, 39, 40, 41, 43, 55

[27] David R Cox and David Oakes. Analysis of. *Survival Data*, 1984. 39

[28] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978. 101

[29] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 60, 63

[30] Jean Deshayes and Dominique Picard. Lois asymptotiques des tests et estimateurs de rupture dans un modèle statistique classique. In *Annales de l'IHP Probabilités et statistiques*, volume 20, pages 309–327, 1984. 45

[31] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(32):375, 2000. 1

[32] Luc Duchateau and Paul Janssen. *The frailty model*. Springer Science & Business Media, 2007. 6, 55, 65, 123

[33] Vincent Ducrocq and G Casella. A bayesian analysis of mixed survival models. *Genetics Selection Evolution*, 28(6):505, 1996. 74

[34] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004. 86

[35] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001. 3, 4, 81, 84, 85, 121, 122, 123

[36] Jianqing Fan, Runze Li, et al. Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99, 2002. 6, 99, 103

[37] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011. 5, 35, 75

[38] Johan Fosen, Ørnulf Borgan, Harald Weedon-Fekjær, and Odd O Aalen. Dynamic analysis of recurrent event data using the additive hazard model. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 48(3):381–398, 2006. 39

[39] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007. 87, 127

[40] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010. 82, 87, 127

[41] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. 117

[42] Zhixuan Fu, Chirag R Parikh, and Bingqing Zhou. Penalized variable selection in competing risks regression. *Lifetime data analysis*, 23(3):353–376, 2017. 127, 133

[43] Stéphane Gaïffas, Agathe Guilloux, et al. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012. 6

[44] Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007. 87

[45] RD Gill. Discussion of the paper by d. clayton and j. cuzick. *Journal of the Royal Statistical Society A*, 148:108–109, 1985. 60

[46] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014. 80, 81

[47] Major Greenwood. The" errors of sampling" of the survivorship tables. *Reports on public health and medical subjects*, 1926. 31

[48] Major Greenwood and G Udny Yule. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society*, 83(2):255–279, 1920. 54

[49] Jiang Gui and Hongzhe Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008, 2005. 6, 96

[50] Guang Guo and German Rodriguez. Estimating a multivariate proportional hazards model for clustered data using the em algorithm, with an application to child survival in guatemala. *Journal of the American Statistical Association*, 87(420):969–976, 1992. 60

[51] David D Hanagal. *Modeling survival data using frailty models*. Chapman and Hall/CRC, 2011. 6, 55, 72

[52] Todd C Handy. *Event-related potentials: A methods handbook*. MIT press, 2005. 4

[53] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5(Oct):1391–1415, 2004. 86

[54] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009. 6

[55] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976. 80

[56] Arthur E Hoerl and Robert W Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970. 4, 81

[57] Philip Hougaard. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396, 1986. 56, 70

[58] Philip Hougaard. Frailty models for survival data. *Lifetime data analysis*, 1(3):255–273, 1995. 56, 70

[59] Philip Hougaard. Analysis of multivariate survival data. 2000. 77

[60] Philip Hougaard. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012. 123

[61] Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012. 97, 126

[62] Jian Huang, Tingni Sun, Zhiliang Ying, Yi Yu, and Cun-Hui Zhang. Oracle inequalities for the lasso in the cox model. *Annals of statistics*, 41(3):1142, 2013. 6

[63] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009. 97

[64] Gareth M James, Peter Radchenko, and Jinchi Lv. Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):127–142, 2009. 87

[65] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12(Oct):2777–2824, 2011. 97

[66] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011. 22

[67] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958. 43

[68] Jinseog Kim, Insuk Sohn, Sin-Ho Jung, Sujong Kim, and Changyi Park. Analysis of survival data with group lasso. *Communications in Statistics-Simulation and Computation*, 41(9):1593–1605, 2012. 6, 96, 123

[69] Yongdai Kim, Sunghoon Kwon, and Hosik Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13(Apr):1037–1057, 2012. 5, 80

[70] John P Klein. Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, pages 795–806, 1992. 60

[71] Inge Riis Korsgaard and Anders Holst Andersen. The additive genetic gamma frailty model. *Scandinavian Journal of Statistics*, 25(2):225–269, 1998. 70

[72] Balaji Krishnapuram, Lawrence Carin, Mario AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005. 87

[73] Sarah Lemler et al. Oracle inequalities for the lasso in the high-dimensional aalen multiplicative intensity model. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 981–1008. Institut Henri Poincaré, 2016. 6

[74] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273, 2006. 102

[75] Wenbin Lu and Hao H Zhang. Variable selection for proportional odds model. *Statistics in medicine*, 26(20):3771–3781, 2007. 6

[76] Shuangge Ma and Jian Huang. Combining clinical and genomic covariates via cov-tgdr. *Cancer informatics*, 3:117693510700300015, 2007. 122

[77] Shuangge Ma, Xiao Song, and Jian Huang. Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics*, 8(1):60, 2007. 122

[78] Torben Martinussen and Thomas H Scheike. *Dynamic regression models for survival data*. Springer Science & Business Media, 2007. 5, 40, 55

[79] Torben Martinussen and Thomas H Scheike. Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, 36(4):602–619, 2009. 6

[80] Anne Marie McCarthy, Brad Keller, Despina Kontos, Leigh Boghossian, Erin McGuire, Mirar Bristol, Jinbo Chen, Susan Domchek, and Katrina Armstrong. The use of the gail model, body mass index and snps to predict breast cancer among women with abnormal (bi-rads 4) mammograms. *Breast Cancer Research*, 17(1):1, 2015. 80

[81] CA McGilchrist and CW Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47(2):461–466, 1991. 65

[82] Clyde A McGilchrist. Reml estimation for survival models with frailty. *Biometrics*, pages 221–225, 1993. 65

[83] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. 6, 87, 96

[84] John E Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *Advances in neural information processing systems*, pages 847–854, 1992. 101, 127

[85] Susan A Murphy. Asymptotic theory for the frailty model. *The Annals of Statistics*, pages 182–198, 1995. 60

[86] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972. 27

[87] Gert G Nielsen, Richard D Gill, Per Kragh Andersen, and Thorkild IA Sørensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, pages 25–43, 1992. 60, 103

[88] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007. 6, 96

[89] Jørgen Holm Petersen. An additive frailty model for correlated life times. *Biometrics*, pages 646–661, 1998. 70, 71

[90] Richard R Picard and R Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984. 127

[91] Andrew Pickles and Robert Crouchlev. Generalizations and applications of frailty models for survival and event data. *Statistical Methods in Medical Research*, 3(3):263–278, 1994. 70

[92] Russell A Poldrack, Jeanette A Mumford, and Thomas E Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011. 4

[93] Rolando Rebolledo. Central limit theorems for local martingales. *Probability Theory and Related Fields*, 51(3):269–286, 1980. 155

[94] Samuli Ripatti and Juni Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, 2000. 65

[95] Virginie Rondeau, Daniel Commenges, and Pierre Joly. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime data analysis*, 9(2):139–153, 2003. 67, 76

[96] Virginie Rondeau, Laurent Filleul, and Pierre Joly. Nested frailty models using maximum penalized likelihood estimation. *Statistics in medicine*, 25(23):4036–4052, 2006. 74

[97] Virginie Rondeau and Juan R Gonzalez. Frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer methods and programs in biomedicine*, 80(2):154–164, 2005. 67

[98] Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008. 32

[99] Maral Saadati, Jan Beyersmann, Annette Kopp-Schneider, and Axel Benner. Prediction accuracy and variable selection for penalized cause-specific hazards models. *Biometrical Journal*, 60(2):288–306, 2018. 107

[100] Narayan Sastry. A nested frailty model for survival data, with an application to the study of child survival in northeast brazil. *Journal of the American Statistical Association*, 92(438):426–435, 1997. 74

[101] Mark D Schluchter. Methods for the analysis of informatively censored longitudinal data. *Statistics in medicine*, 11(14-15):1861–1870, 1992. 22

[102] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978. 4, 80, 101, 126

[103] Mark R Segal. Microarray gene expression data with linked survival phenotypes: diffuse large-b-cell lymphoma revisited. *Biostatistics*, 7(2):268–285, 2005. 6, 96

[104] Dari Shalon, Stephen J Smith, and Patrick O Brown. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome research*, 6(7):639–645, 1996. 4

[105] Jun Shao. An asymptotic theory for linear model selection. *Statistica sinica*, pages 221–242, 1997. 5

[106] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003. 87

[107] GR Shorack and JA Wellner. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. In *Empirical Processes with Applications to Statistics*. Wiley, 1986. 156

[108] Insuk Sohn, Jinseog Kim, Sin-Ho Jung, and Changyi Park. Gradient lasso for cox proportional hazards model. *Bioinformatics*, 25(14):1775–1781, 2009. 6, 96

[109] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013. 65

[110] Terry M Therneau, Patricia M Grambsch, and V Shane Pankratz. Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1):156–175, 2003. 65, 67

[111] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 2, 3, 4, 6, 81, 82, 121

[112] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997. 6, 96

[113] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. 6

[114] Jean Claude Utazirubanda, Tomás M. León, and Papa Ngom. Variable selection with group lasso approach: Application to cox regression with frailty model. *Communications in Statistics-Simulation and Computation*, pages 1–21, 2019. 9

[115] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 106, 129

[116] Hans C van Houwelingen, Tako Bruinsma, Augustinus AM Hart, Laura J van't Veer, and Lodewyk FA Wessels. Cross-validated cox regression on microarray gene expression data. *Statistics in medicine*, 25(18):3201–3216, 2006. 87

[117] James W Vaupel, Kenneth G Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979. 54

[118] Pierre JM Verweij and Hans C Van Houwelingen. Penalized likelihood in cox regression. *Statistics in medicine*, 13(23-24):2427–2436, 1994. 3

[119] Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286, 2008. 121, 123

[120] Hansheng Wang, Guodong Li, and Chih-Ling Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78, 2007. 126

[121] Fengrong Wei, Jian Huang, and Hongzhe Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21(4):1515, 2011. 117

[122] Fengrong Wei and Hongxiao Zhu. Group coordinate descent algorithms for nonconvex penalized regression. *Computational Statistics & Data Analysis*, 56(2):316–326, 2012. 127

[123] Andreas Wienke. *Frailty models in survival analysis*. Chapman and Hall/-CRC, 2010. 6, 55, 71

[124] Andreas Wienke, Niels V Holm, Axel Skytthe, and Anatoli I Yashin. The heritability of mortality due to heart diseases: a correlated frailty model applied to danish twins. *Twin Research and Human Genetics*, 4(4):266–274, 2001. 72

[125] Tong Tong Wu, Kenneth Lange, et al. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008. 87

[126] Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005. 5

[127] Anatoli I Yashin, James W Vaupel, and Ivan A Iachine. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical population studies*, 5(2):145–159, 1995. 70, 71

[128] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. 6, 86, 96, 122, 127

[129] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. 123

[130] Ming Yuan and Yi Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007. 117

[131] Sangwoon Yun, Paul Tseng, and Kim-Chuan Toh. A block coordinate gradient descent method for regularized convex separable optimization and covariance selection. *Mathematical programming*, 129(2):331–355, 2011. 102

[132] Per-Henrik Zahl. Frailty modelling for the excess hazard. *Statistics in medicine*, 16(14):1573–1585, 1997. 72

[133] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010. 3, 84, 85, 122

[134] Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007. 6, 126

[135] Hao Helen Zhang, Wenbin Lu, and Hansheng Wang. On sparse estimation for semiparametric linear transformation models. *Journal of multivariate analysis*, 101(7):1594–1606, 2010. 6

[136] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. 3, 83, 123, 126

[137] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005. 3, 83, 84, 86

ANNEXE A

# Processus de comptage

## A.1 Processus aléatoires

On s'intéresse à la modélisation des occurences dans le temps d'événements aléatoires. On étudiera particulièrement le cas où les événements ont lieu en temps continu et sont de nature discrète. On considère donc un intervalle de temps continu. $\mathcal{T} = [0, \tau]$.

**Définition A.1.1.** *Soit* $(\Omega, \mathcal{A}, \mathbb{P})$ *un espace probabilisé. Une* **filtration** *(ou histoire) est une famille* $\{\mathcal{F}_t : t \in \mathcal{T}\}$ *de tribus, telle que*

$$\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{A}, \quad \forall 0 \leq s \leq t,$$

**Définition A.1.2.** *Un processus* $X(t), t \in \tau$ *est intégrable si*

$$\sup_{t \in \mathcal{T}} \mathbb{E}(|X_t|) < \infty.$$

*Un processus* $X(t), t \in \tau$ *est de carré intégrable si*

$$\sup_{t \in \mathcal{T}} \mathbb{E}(X_t^2) < \infty.$$

**Définition A.1.3.** *Un processus* $M(t), t \in \mathcal{T}$ *intégrable et adapté à une filtration* $\{\mathcal{F}_t : t \in \mathcal{T}\}$ *est une* **martingale** *(resp. une sur-martingale, resp. une sous-martingale) si*

$$\forall (s,t),\ 0 \leq s \leq t,\ \mathbb{E}\left(M(t)|\mathcal{F}_s\right) = M(s)\ \textit{p.s}\ (\textit{resp.} \leq, \textit{resp.} \geq).$$

**Propriété A.1.4.** *Soit* $M(t)$ *une martingale par rapport à une filtration* $\{\mathcal{F}_t : t \in \mathcal{T}\}$. *Alors*

*(i)* $\mathbb{E}\{M(t)|\mathcal{F}_{t-}\} = M(t-)$, *où* $\mathcal{F}_{t-} = \underset{s<t}{\cup} \mathcal{F}_s$.

*(ii)* $\mathbb{E}\{dM(t)|\mathcal{F}_{t-}\} = 0$.

**Définition A.1.5.** *Un* **temps d'arrêt** $T$ *par rapport à* $\mathcal{F}_t$ *est une* $v.a$ *prenant ses valeurs dans* $[0,\infty]$ *telle que* $\{T \leq t\} \in \mathcal{F}_t,\ \forall t \in [0,\infty]$.
*Un* **processus arrêté** $X^T$ *est défini par*

$$X_t^T(w) = \begin{cases} X_t^T(w), & \textit{si } T(w) > t \\ X_T(w), & \textit{sinon}. \end{cases}$$

**Définition A.1.6.** *Un processus stochastique* $X$ *est dit* $\mathcal{F}_t-$ **prévisible** *si*

*(i) comme fonction de* $(t,w) \in \mathcal{T} \times \Omega \mapsto \mathbb{R}$*, il est mesurable par rapport à la tribu sur* $\mathcal{T} \times \Omega$ *engendrée par les processus adaptés et continus à gauche. ou si*

*(ii)* $X(T)$ *est* $\mathcal{F}_t-$ *mesurable pout tout point d'arrêt* $T$.

**Définition A.1.7.** *Soit* $(\Omega, \mathcal{A}, \mathbb{P})$ *un espace probabilisé et* $\mathcal{F}_t$ *une filtration. Le processus* $X(t), t \in \mathcal{T}$ *est dit* $\mathcal{F}_t-$**adapté** *si* $\forall t, X(t)$ *est* $\mathcal{F}_t-$ *mesurable.*

**Proposition A.1.8.** *Soit* $X$ *un processus* $\mathcal{F}_t-$ *prévisible alors pour* $t, X_t$ *est* $\mathcal{F}_t-$ *mesurable.*

**Proposition A.1.9.** *Tout processus adapté à* $\mathcal{F}_t$ *est continu à gauche et* $\mathcal{F}_t-$ *prévisible.*

**Définition A.1.10.** *Un processus adapté* $M_t, t \in \mathcal{T}$ *tel que* $M_0 = 0$ *est une* **martingale locale** *si il est continu et il existe une suite croissante* $T_n, n \in \mathbb{N}$*, de temps d'arrêt tendant vers* $+\infty$ *telle que* $M_t^{T_n}$ *soit une martingale.*

## A.2 Processus de comptage

**Définition A.2.1.** *Un* **processus de comptage** $N(t)$ *est un processus continu à droite avec une limite à gauche, adapté, nul en zéro, croissant et ayant des sauts d'amplitude 1.* $N(t)$ *compte le nombre d'événements d'intérêt qui se sont produits avant* $t$.

**Proposition A.2.2.** *Soit* $N(t)$ *un processus de comptage. Il existe un processus* $\Lambda(t)$ *prévisible, croissant, continu à droite et nul en zéro tel que*

$$M(t) = N(t) - \Lambda(t), \quad t \geq 0 \tag{A.1}$$

*soit une martingale.* $\Lambda(t)$ *est appelé* **compensateur** *de* $N(t)$*, ou encore* **processus d'intensité cumulée.**

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

*où $\lambda(s)$ est aussi unprocessus prévisible, appelé* **intensité** *du processus ponctuel. Elle est définie par*

$$\lambda(s) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{P}\big(N(s+\varepsilon) - N(s) \geq 1 | \mathcal{F}_s\big).$$

*$\mathcal{F}_s$ est l'ensemble des événements observables à l'instant $s$.*

Pour une loi de probabilité donnée et une filtration donnée, (A.1) est connue sous la décomposition de **Doob-Meyer** et elle est unique. On utilise souvent, quand elle existe, la repr'esentation différentielle de (A.1) :

$$dM(t) = dN(t) - \lambda(t)dt, \quad t \geq 0. \tag{A.2}$$

En appliquant propriété A.1.4, on a

$$\mathbb{E}(dN(t)|\mathcal{F}_t) = \lambda(t)dt.$$

Mais

$$\mathbb{E}(dN(t)|T \geq t) = \mathbb{P}\left(t \leq T < t + dt | T > t\right) \text{ et } \mathbb{E}(dN(t)|T < t) = 0.$$

Par définition de risque et l'indicateur de risque, nous avons

$$\lambda(t) = Y(t)h(t).$$

## A.3 Théorème Limite Centrale

**Théorème A.3.1.** *[93] -Si $M_n$ est une suite de martingales, et si*

*(i) $\langle M_n \rangle_t$ converge en probabilité vers $v_t$ déteministe,*

*(ii) $\forall \varepsilon \ \exists M_{n,\varepsilon}$ suite de martingales telle qu'aucune différence $M_n - M_{n,\varepsilon}$ n'ait une amplitude supérieure à $\varepsilon$,*

*alors $M_n(t)$ a une limite $M(t)$ de processus croissant $v_t$ et $M(t)$ est un processus gaussien :*

$$\frac{M_n(t)}{v_t} \xrightarrow{\mathcal{L}} \mathcal{N}(0.1)$$

## A.4 Produit infini (ou intégral

**Définition A.4.1.** *Soit $X(s)$ un processus càdlag, nul en $0$, et à variation bornée. On obtient une mesure en posant*

$$X(]s,t]) = X(t) - X(s)$$

*Soit une partition* $t_0 = s < t_1 < \cdots < t_n = t$. *Son* **pas** *est*

$$|\delta| = \sup_i |t_i - t_{i-1}|.$$

*On appelle* **produit infini** *(ou* **produit intégral***)*

$$\begin{aligned} \mathcal{P}_s^t(1+dX) &= \mathcal{P}_{]s,t]}(1+dX) \\ &= \lim_{|\delta|\to 0} \prod_{i=1}^{n} [1 + X\left(]t_{i-1}, t_i]\right)] \end{aligned}$$

*qui est indépendante de la suite des* $(\delta)$.

**Propriété A.4.2.** *-Si* $X(t)$ *est continue, alors*

$$\mathcal{P}_{]s,t]}(1+dX) = \exp(X(t)).$$

**Théorème A.4.3.** *(***Duhamel***) [4], pg 90.*
*Soient* $Y = \mathcal{P}_{]s,t]}(1+dX)$ *et* $Y' = \mathcal{P}_{]s,t]}(1+dX')$*, alors*

$$Y(t) - Y'(t) = \int_{s\in[0,t]} \mathcal{P}_{[0,s)}(1+dX)\left(X(ds) - X'(ds)\right)\mathcal{P}_{(s,t]}(1+dX').$$

*Si* $Y'(t)$ *est régulière, alors*

$$\begin{aligned} \frac{Y(t)}{Y'(t)} - 1 &= \int_{s\in[0,t]} \mathcal{P}_{[0,s)}(1+dX)\left(X(ds) - X'(ds)\right)\left[\mathcal{P}_{[0,s]}(1+dX')\right]^{-1}, \\ &= \int_0^t \frac{Y(s-)}{Y(s)}\left[X(ds) - X'(ds)\right]. \end{aligned}$$

**Théorème A.4.4. Glivenko-Cantelli** *[107]*$-pg.304-305$. *-Si* $M_n$ *est une suite de martingales, et si*

*(i ) Pour une suite* $(X_n)$ *de vecteurs aléatoires à valeurs dans* $\mathbb{R}^k$ *indépendants, de loi* $F$*, les répartitions empiriques* $\bar{F}_n(w,.)$ *où*

$$\bar{F}_n(w,.) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(w)$$

*convergent étroitement vers* $F$ *pour presque tout* $w$.

*(ii) Pour* $k = 1$*, la convergence des fonctions de répartition est p.s uniforme : pour presque tout* $w$,

$$\sup_x \left|\bar{F}_n(w,x) - F(x)\right| \longrightarrow 0, \quad n \longrightarrow \infty.$$

# Résumé & Abstract

**Résumé :** En épidémiologie, on est amené à prendre en compte simultanément le rôle de plusieurs facteurs de risque dans la survenue d'une maladie. Le modèle de régression à risques proportionnels de Cox est l'un des modèles le plus populaire dans l'analyse des données de survie censurées. Il permet d'exprimer le risque instantané de survenue de l'événement en fonction des facteurs explicatifs. Cependant, ce modèle ne répond pas à certains besoins émergents dans la majorité des cas en statistiques appliquées. Les données observées sont souvent de grande taille, c'est-à-dire de dimension $p$ très grand devant $n$ ($p >> n$). Par exemple en analyse de survie d'une maladie génétique, pour chaque individu, on dispose d'un grand nombre de gènes, qui est parfois nettement plus élevé que le nombre d'individus dans l'échantillon. Dans cette thèse, nous avons réalisé un état de l'art exhaustif sur les méthodes d'estimation pénalisée, souvent utilisées pour faire de la sélection de variables en grande dimension dans les modèles de durées de vie censurées. Cet état de l'art concerne les aspects méthodologiques, algorithmiques, computationnels et théoriques des méthodes proposées.

Cet état de l'art commence par une présentation des théories et méthodologies existant sur l'intérêt de la sélection de variables dans les modèles de durées de vie censurées. Nous nous sommes également intéressés à la littérature développée sur les modèles de survie avec fragilités en particulier des modèles à risque multiplicatif avec hypothèse d'une distribution gamma pour la variable de fragilité. Ensuite, nous avons présenté les méthodes de régularisation basées sur l'estimation pénalisée pour la sélection des variables dans le modèle de régression semi-paramétrique à risques proportionnels de Cox.

Dans un second temps, nous avons adapté la méthode du "group-Lasso" au modèle de Cox pour des données en clusters ou "groupées" en tenant compte de la fragilité propre à chaque cluster. Les aspects algorithmiques ont été traités et des simulations ont été réalisées pour étudier le comportement des estimateurs obtenus. Ensuite, la performance de la méthode group Lasso a été comparée avec celle des autres méthodes concurrentes : group-MCP et group-SCAD. La consistance des estimateurs obtenus par cette méthode a été établie théoriquement. Le travail a été ensuite généralisé en implémentant la méthode de l' "Adaptive group-lasso" dans le modèle de Cox avec fragilité pour des données en clusters. Chacune des méthodologies étudiées a été confrontée aux différents jeux de données réelles, issus en grande partie du domaine biomédical. Les résultats montrent que la méthode group Lasso et la méthode Adaptive group Lasso dans le modèle de Cox avec fragilité gamma partagée, sont en général plus performantes en matière de

prédictions par rapport aux autres méthodes concurrentes à savoir : la méthode group SCAD et la méthode group MCP. Néanmoins, la connaissance a priori de structures de groupes de variables reste un problème majeur dans la modélisation des données en clusters.

**Mots clés** : Survie, group Lasso, fragilité

**Abstract :** In epidemiology, the role of several risk factors in the occurrence of a disease must be taken into account simultaneously. Cox's proportional hazard model is one of the most popular models in the analysis of censored survival data. It allows the instantaneous risk of the event occurring to be expressed in terms of explanatory factors. However, this model does not meet some emerging needs in applied statistics in the most of the cases. The observed data are often large in size, i.e. high dimensional situation (p » n). For example, in survival analysis of a genetic disease, for each individual, there is a large number of genes, which is sometimes significantly higher than the sample size. In this thesis, we have carried out an exhaustive litterature review on penalized estimation methods, often used to select high dimensional variables in censored lifetime models. This litterature review concerns with the methodological, algorithmic, computational and theoretical aspects of the proposed methods.

This general review starts with a presentation of existing theories and methodologies on the interest of variable selection in censored lifetime models. We were also interested in the literature developed on survival models with frailty, in particular multiplicative risk models with shared gamma frailty. Next, we presented the regularization methods based on the penalized estimation for variables selection in the Cox semi-parametric proportional hazard model.

In a second step, we adapted the "group-Lasso" method to the Cox model for clustered or "grouped" data, taking into account the shared frailty of each cluster. Algorithmic aspects have been investigated and simulations were carried out to study the behaviour of the obtained estimators. Then, the performance of the Lasso group method was compared with that of the other competing methods: group-MCP and group-SCAD. The consistency of the estimators obtained by this method has been theoretically established. The work was then generalized by implementing the "Adaptive group-lasso" method in the Cox model with frailty for clustered data. Each of the methodologies studied was applied to different real and simulated dataset, mainly from biomedical research area.

The results show that the group Lasso method and the Adaptive group Lasso method in the Cox model with shared gamma frailty are generally more predictive than the other competing methods, namely the SCAD group method and the MCP group method. Nevertheless, a prior knowledge of variable group structures remains a major problem in cluster data modeling.

**Key words** : survival, Lasso group, frailty