MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ MOULOUD MAMMERI, TIZI-OUZOU

FACULTÉ DES SCIENCES
DÉPARTEMENT DE MATHÉMATIQUES

MEMOIRE DE MAGISTER(ECOLE DOCTORALE)

SPECIALITE : MATHEMATIQUES
OPTION : STATISTIQUE

Présenté par

Khadidja BOUDANE

Sujet :

ESTIMATION DE LA FONCTION DE RISQUE CONDITIONNELLE POUR DES DONNÉES MARKOVIENNES

Devant le jury d'examen composé de :

Hocine FELLAG	Professeur	UMMTO	Président
Abderahmane YOUSFATE	Professeur	UDL	Rapporteur
Lynda ATIL	Maître de conférences B	UMMTO	Examinatrice
Mohand Arezki BOUDIBA	Maître de conférences A	UMMTO	Examinateur

Année universitaire 2012/2013

Remerciements

Tout d'abort, je tiens à remercier le Bon Dieu pour m'avoir illuminée et menée jusqu'ici. Nombreus sont ceux qui m'ont aidé, encouragé, soulagé, tout au long de ces années, et je ne saurais leur exprimer mes remerciements autant que le souhaiterais ; Merci pour leur présence, leur discussions, leur conseils et suggestions éclairées.

En premier lieu,je tiens à remercier chaleureusement Monsieur Abderahmane YOUSFATE, Professeur à UDL d'avoir accepté d'être mon directeur de memoire. Je le remercier pour sa disponibilité, sa compétence, son précieux soutien, ses conseils judicieux et la confiance dont il m'a fait part lors de la réalisation de ce travail.

Je souhaite adresser mes sincères remerciements aux membres de mon jury de soutenance : Monsieur Hocine FELLAG, Professeur à UMMTO de m'avoir accordé l'honneur de présider mon jury de soutenance, M^{me} Lynda ATIL, Maître de conférences à UMMTO et M^{r} Mohand Arezki BOUDIBA, Maître de conférences à UMMTO, d'avoir accepté la tâche d'évaluer, en qualité d'examinateurs, mon travail.

Mes profonds et mes plus grands remerciements vont aux membres de ma famille : aux deux personnes qui me sont les plus chères au monde ; mes parents, a mes soeurs et frères. Merci pour être toujours à coté de moi. Merci du fond du coeur.

Table des matières

Introduction générale

Problématique

Au cours des dernières années, la branche de la statistique consacrée à l'étude des variables fonctionnelles a connu un réel essor tant en terme de dévelopements théoriques que de diversification des domaines d'application.

L'estimation de la fonction de hasard est un problème intéressant qui apparaît dans l'analyse statistique des durées de vie, notamment l'étude de la survie. L'analyse des durées de vie est un domaine de la statistique qui étudie l'apparition d'un évènement au cours du temps. Pour ce faire, il est nécessaire de disposer du temps de suivi de tous les individus, ainsi que du moment auquel l'évènement est produit. Ce qui est particulier avec ce type d'étude, c'est la présence des données censurées (données incomplètes) pour les sujets chez qui l'évènement d'intérêt est non observé. L'analyse des durées de vie est donc particulièrement utile pour étudier plusieurs types d'évènements, notamment des pannes d'équipement, de tremblements de terre, des divorces et évidemment, des décès.

Le présent travail traite de l'analyse statistique non paramétrique et paramétrique des durées de vie. Il s'intéresse à la durée jusqu'à l'apparition d'un évènement d'intérêt, comme la durée de vie avant un décès dû à une certaine cause (cancer, maladie infectieuse, accident de la route,...), la durée de réponse à un traitement, la durée avant le développement d'une pathologie particulière, etc...

L'estimation du taux de hasard (fonction de risque) est une question importante en statistique. Ce sujet peut être abordé sous plusieurs angles selon la complexité du problème posé : présence éventuelle de censure dans l'échantillon observé (phénomène courant dans les applications médicales par exemple), présence éventuelle de dépendance entre les variables observées (phénomène courant dans les applications sismologiques ou économétriques) ou bien présence de variables explicatives.

L'analyse de survie trouve des applications en science actuarielle, démographie, épidémiologie, recherche médicale, analyse de fiabilité et beaucoup d'autres champs. Les exemples des

durées de dérangement incluent les vies des composants de machine en fiabilité industrielle, les durées des grèves ou périodes du chômage dans les sciences économiques, les temps pris par des individus pour accomplir une tâche spécifique dans l'expérimentation psychologique et les longueurs des voies d'un plat photographique dans la physique de particules. Dans la recherche médicale, si le point final est la mort d'un patient, les données résultant sont littéralement des vies. Cependant, des données d'une forme semblable peuvent être obtenues quand le point final n'est pas mortel. Les exemples des vies dans la recherche clinique incluent le temps de début du traitement au soulagement d'une douleur, et le temps de début du traitement à la répétition des symptômes et étudient une maladie infectieuse, le temps de début de l'infection au début de la maladie.

Les modèles multi-états constituent une alternative intéressante pour modéliser des données de type mesures répétées. D'un point de vue théorique, l'objectif de ce document est d'étudier des méthodes d'estimation statistiques pour les modèles multi-états.

Modélisation

Les modèles multi-états sont considérés comme une généralisation des modèles de survie ils ne cessent de connaître un intérêt croissant. Ils sont caractérisés par un processus stochastique à espace d'état fini pour décrire un phénomène. L'utilisation de processus se fait pour représenter les différents états successivement occupés à chaque temps d'observation. Par exemple, en épidémiologie, ils permettent de représenter l'évolution d'un patient à travers les différents stades d'une maladie. Après définition des différents stades(états), les modèles multi-états permettent d'étudier de nombreuses dynamiques complexes. L'étude de ces modèles consiste à analyser les forces de passage (intensités de transition) entre les différents états.

La popularité et la richesse des modèles de survie, en particulier du modèle de Cox, dessert l'utilisation de ces modèles dans le domaine appliqué.

Dans les modèles multi-états les plus simples, l'information sur l'état présent renseigne sur les états précédents : par exemple, les modèles progressifs (Hougaard [1999]), les modèles à risques compétitifs (Huber-Carol et Pons [2004], Andersen et al. [1993]), ou encore les modèles de survie qui représentent le cas le plus simple avec uniquement deux états, (Therneau et Grambsch [2000]). Cependant, dès que le modèle comprend des états réversibles (c'est-à-dire que certains évènements sont récurrents), il devient nécessaire de faire des hypothèses sur l'histoire de l'individu. Les modèles de type Markovien sont très utiles, car ils supposent que l'information sur les états précédents est résumée par l'état présent. Le terme de modèle multi-états regroupe de nombreuses problématiques biostatistiques.

On pourra se référer, par exemple, aux travaux de Andersen et Keiding [2002], Hougaard [1999], Andersen et al.[1993] et Commenges [1999] qui font le point sur l'état de l'art dans ce domaine.

Dans ces modèles de type Markovien, les intensités de transition entre les états peuvent dépendre de différentes échelles de temps.

Dans certaines applications, la durée du suivi n'est pas toujours l'échelle de temps le mieux adaptée. En effet, le temps calendaire et l'âge peuvent également être considérés comme échelle de temps principale. Par exemple, le temps calendaire peut être adapté quand on considère le risque de contracter une maladie qui a une incidence variant beaucoup, comme l'infection par le VIH dans les années 80. Le choix entre les échelles de temps dépend de ce qui est le plus important dans une application donnée.

La durée de vie d'intérêt est alors modélisée par une variable aléatoire positive X dont on veut estimer la loi. Dans la pratique, il est courant qu'on ne puisse pas observer X directement. C'est le cas, par exemple, quand un individu quitte l'étude en cours avant la survenue de l'évènement d'intérêt qui est supérieure à la durée passée dans l'étude.

La fonction de survie S(t) intègre l'ensemble des évènements observés avant t et décrit mal la dynamique instantanée du processus.

La dynamique de ce processus peut s'exprimer sous la forme d'une fonction de risque instantanée, traduisant le risque de présenter l'évènement sur un intervalle de temps infinitésimal, conditionnellement au fait de ne pas l'avoir présenté auparavant. Cette fonction de risque peut être paramétrable (exprimable sous forme d'une formule mathématique). C'est le cas du modèle exponentiel (qui suppose un risque instantané constant au cours du temps) et du modèle de Weibull.

On peut aussi exprimer cette fonction d'une manière non paramétrique (sans faire d'hypothèse sur son allure au cours du temps). Le plus souvent, dans ce cas, on estime la fonction de risque instantanée $\lambda(t)$ par un estimateur de Kaplan-Meier :

Pour chaque temps t_i, la proportion d'évènements observés est $\lambda(t_i) = m_i/n_i$ où m_i est le nombre d'évènements observés en t_i et n_i le nombre de sujets exposés au risque d'évènement juste avant t_i.

Concernant les modèles statistiques proprement dits, trois approches sont possibles : paramétrique, non-paramétrique et semi-paramétrique.

L'approche paramétrique stipule l'appartenance de la loi de probabilité réelle des observations à une classe particulière de lois, qui dépendent d'un certain nombre(fini)de paramètres. 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 la méthode paramétrique est l'inadéquation pouvant exister entre le phénomène étudié et le modèle retenu.

L'approche non-paramétrique ne nécessite aucune hypothèse quant à la loi de probabilité réelle des observations, et c'est là son principal avantage. Il s'agit dés lors d'un problème d'estimation fonctionnelle, cela implique, par exemple, la fonction de survie, qui est continue, sera estimée par une fonction discontinue. 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 appartient à un espace de dimension infinie.

L'approche semi-paramétrique est une sorte de compromis entre les deux approches précédentes. La loi de probabilité réelle des observations est supposée appartenir à une classe de lois pour partie dépendant de paramètres, et pour partie s'écrivant sous forme de fonction non-paramétrique. Cette approche est très répandue en analyse de la survie, notamment au travers du modèle de régression de Cox (1972).

Objectifs

Le travail présenté dans ce mémoire a pour objectifs d'etudier l'estimation paramétrique, non-paramétrique et semi-paramétrique de la fonction de hasard conditionnelle lorsque les observations dépendantes et la covariable est dans un espace semi-métrique de dimension éventuellement infinie.

L'estimation de la fonction de hasard joue un rôle très important en statistique. Elle est utilisée dans l'analyse de risque ou pour l'étude des phénomènes de survie dans de nombreux domaines tels (médecine, géographique, fiabilité, ...).

L'approche adoptée dans ce travail pour l'estimation de fonction de risque conditionnelle est une synthèse d'une série de plusieurs recherches (Ferraty et al. 2008, M.C.Iglesias Pérez et Manteiga W.González.2003, Ouhbi Brahim et Limnios Nikolaos. 1999,...)

Organisation du mémoire

Nous présentons notre travail dans cinq chapitres.

Le premier chapitre est constituée par le recensement bibliographique réalisé sur les données de survie, à savoir : les différentes notations qui se présentaient dans les ouvrages

de référence, certains résultats, les définitions et les notations qui paraissaient essentiels à la compréhension de la théorie des données de survie.

Dans le deuxième chapitre nous avons rassemblé le bagage nécessaire pour effectuer une étude des modèles Markovienne, tout en donnant quelques notions et définitions sur les processus de Markov et sa généralisation. Ce type de modèles est appliqué à plusieurs cas notamment dans le domaine médical.

Dans le chapitre trois, nous commençons, tout d'abord par présenter le modèle de Cox et sa généralisation, par la suite, nous passons à l'étude des résultats existants pour l'estimation des paramètres de ce modèle.

Le quatrième chapitre présente un aperçu sur les différentes approches et méthodes d'estimation, comprenant précisément l'estimation d'intensité de transition.

Le dernier chapitre synthétise les résultats obtenus dans le cas d'une application sur des données médicales.

Chapitre 1

Analyse de survie et modèles multi-états

L'objet de ce chapitre est de motiver et d'introduire la notion de durée de vie (d'analyse de survie) ainsi que l'approche statistique au travers de laquelle nous allons étudier ce type de données. Les modèles multi-états sont une généralisation naturelle des modèles de survie. La théorie des modèles de Markov a été développée depuis longtemps mais les applications restaient rares. Dans ce premier chapitre, nous présentons les outils de modélisation utilisés en analyse de données de vie et nous introduisant le problème de la censure qui affecte ce type de données.

1.1 Motivations

En analyse de survie, on s'intéresse à un groupe d'individus associes à un évènement d'intérêt, souvent appelé échec ou mort, survenant après une durée appelée durée de vie ou donnée de survie. Les modèles de durée sont appropriés dès lors que les phénomènes étudiés sont représentés par des variables aléatoires positives. Des exemples classiques sont la panne de composants électroniques en fiabilité industrielle, la fin d'une grève ou d'une période de chômage en économie, en expérimentation psychologique ou, en médecine... . Dans la plupart des cas, l'évènement d'intérêt symbolise la transition d'un état à un autre.

Pour déterminer précisément la survenue de l'échec, il est nécessaire de définir sans ambiguïté l'origine des temps et le terme d'échec, ainsi que de choisir un échelle de temps. L'origine des temps n'est pas nécessairement le même pour tous les individus et doit être définie précisément pour chacun d'entre eux. Dans la plupart des cas, le temps 0 est choisi comme étant le moment d'une transition.

L'analyse des donnée de survie s'intéresse au temps de survenue d'un évènement précis dans un ou plusieurs groupes d'individus.

Un modèle de survie est un modèle multi-état ne comportant que deux états avec une seule transition possible de l'état 0 à l'état 1. L'objectif est alors de modéliser l'intensité de transition entre l'état 0 et l'état 1.

Par exemple, lors de la mesure d'un âge, l'origine des temps est la naissance pour un essai de traitement médical, l'origine naturelle des temps est le début du traitement, mais en ce qui concerne l'évolution d'une maladie, la date de contamination n'est en général pas connue, on peut considérer le moment du diagnostic comme origine. Cela peut paraître une alternative convenable même si cela entraînera des approximations par la suite. En ce qui concerne l'échelle des temps, le cas le plus fréquent est une mesure horaire, mais d'autres possibilités existent, comme le kilométrage d'un véhicule ou le temps cumulé d'utilisation d'un système.

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

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

1.2 Concepts de base et notation

L'analyse statistique des données de survie étudie les lois d'instants d'occurrence d'évènements à partir d'observations de durées et éventuellement de variables explicatives faites d'une manière discrète ou continue dans le temps.

Ainsi, nous désignons par T une variable aléatoire continue positive définie sur un espace probabilisé (Ω, A, P) associée à une durée de vie (c'est à dire une v.a à valeurs dans $\mathbb{R}$) passée dans un certain état, l'origine des temps étant prédéfinie. Dans le domaine médical

cet évènement peut être la mort ou la guérison, dans le domaine économique, la perte d'un emploi et en fiabilité, l'instant de première panne. Par la suite, nous notons F la fonction de répartition de la durée de vie T.

$$F(t) = Pr(T \leq t); t > 0$$

Les durées de survie sont des variables aléatoires positives, de distribution, le plus souvent dissymétrique rendant difficile leur description par les distributions théoriques usuelles. Par exemple, la distribution normale qui joue un rôle important en statistique ne peut pas être utilisée en analyse de survie car les variables étudiées sont supposées ne prendre que des valeurs positives. Toutefois, l'étude des durées de survie utilise les fonctions classiques permettant de décrire les variables aléatoires continues.

La loi de cette durée peut aussi être caractérisée par l'intermédiaire d'autres fonctions faciles à interpréter et qui de plus s'introduisent naturellement dans divers calculs, parmi ces fonctions, les plus importantes sont : La fonction de densité, la fonction de répartition, la fonction de survie, la fonction de hasard de la durée de vie.

Pour faciliter la description de ces fonctions, nous supposons que T est continu

-Caractérisation de la loi de durée

Définition 1.2.1. La fonction de densité de probabilité de T, notée f(t) est définie par :

$$f(t) = \lim_{\Delta t \to 0^+} P\frac{t \leq T < t + \Delta t}{\Delta t}$$

Définition 1.2.2. La fonction de répartition F(t) est la probabilité de décéder entre 0 et t :

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

La fonction de répartition est croissante telle que $\lim_{t\to 0^+} F(t) = 0$ et $\lim_{t\to\infty} F(t) = 1$.

Définition 1.2.3. :On appelle fonction de survie S la probabilité que la durée de vie T soit supérieur à un temp t :

$$\forall t \in \mathbb{R}, S(t) = P(T > t) = 1 - F(t)$$

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

$$\forall t \in \mathbb{R}, S(t) = \int_t^\infty f(t)dt$$

Définition 1.2.4. La fonction de risque (parfois appelé aussi fonction de hasard, taux de hasard, taux de défaillance ou taux de survie)au point t s'interprète comme la probabilité instantanée de sortir de l'état que l'on observe (vie, chômage, cohabitation, etc.)à la date t, sachant que le sujet est encore dans cet état en t, soit :

$$\lambda(t) = \begin{cases} \lim_{\Delta_t \to 0} \frac{P(t<T\leq t+\Delta_t \backslash T>t)}{\Delta_t} & \text{si } t>0 \text{ et tel que } P(T>t)>0 \\ +\infty & \text{sinon} \end{cases}$$

La fonction de risque peut avoir des formes différentes mais est nécessairement positive sur $\mathbb{R}$.

Supposons maintenant que T soit une variable continue, on observe alors que :

$$\begin{aligned} \forall t \in \mathbb{R}, \lambda(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{d}{dt} ln(S(t)) \end{aligned}$$

La fonction de hasard caractérise la loi de T du fait de la relation

$$S(t) = \exp(-\int_0^t \lambda(u)du)$$

cette formule conduit d'ailleurs à introduire le hasard cumulé

Définition 1.2.5. La fonction de risque cumulée $\Lambda(t)$ est définie par

$$\Lambda(t) = \int_0^t \lambda(s)ds = -\ln(S(t));$$

Il s'ensuit que

$$S(t) = \exp(-\Lambda(t)).$$

La distribution de la durée de vie ou du temps dans l'état 0, T, peut être décrite par l'une des fonctions f, F, S, λ ou Λ. Chacune d'entre elle est caractéristique de la distribution de T. Cependant, la fonction de risque $\lambda(t)$ est la plus intéressante car elle décrit le futur immédiat du sujet et reflète des différences entre les modèles souvent moins explicites avec les fonctions de répartition et de survie.

– **Fonction de survie conditionnelle :**

$$\begin{aligned} S(t\backslash t_0) &= Pr(T > t + t_0 \backslash T > t_0) \\ &= \frac{S(t+t_0)}{S(t_0)} \\ &= \exp(-\int_{t_0}^t \lambda(u)du), \forall t > t_0 \end{aligned}$$

– **Espérance de durée de vie restante :** Finalement, définissons la durée de vie moyenne restante. Si à la date t l'individu est encore dans un état donné et si T désigne sa date de sortie de cet état, sa durée de vie résiduelle est : T-t. Prenant l'espérance, nous obtenons :

$$r(t) = E(T - t \backslash T > t).$$

si $\lim_{n\to+\infty} uS(u) = 0$, alors $r(t) = \frac{1}{S(t)} \int_t^{+\infty} S(u)du$

– **Démonstration :**

$$\begin{aligned} r(t) &= E(T - t \backslash T > t) \\ &= \frac{E(T-t).\mathbf{1}(T > t)}{Pr(T > t)} \\ &= \frac{1}{S(t)} \left\{ \int_t^{+\infty} uf(t)du - tS(t) \right\} \\ &= \frac{1}{S(t)} \left\{ [-uS(u)]_t^{+\infty} + \int_t^{+\infty} S(u)du - tS(t) \right\} \end{aligned}$$

Si $\lim_{u\to+\infty} uS(u) = 0$ alors : $r(t) = \frac{1}{S(t)} \int_t^{+\infty} S(u)du$

Exemples. 1.1 : Nous pouvons donc caractériser la loi de la durée T par une fonction de risque instantané constante :

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

où, λ est une constante strictement positive.

On obtient ainsi les fonctions définies ci-dessus pour $t \in \mathbb{R}$:

$$f(t) = \lambda e^{-\lambda t}, \quad F(t) = 1 - e^{-\lambda t}, \quad S(t) = e^{-\lambda t}, \quad \Lambda(t) = \lambda t, \quad r(t) = 1/\lambda = E(t)$$

La variable T suit donc une loi exponentielle de paramètre λ.

C'est la distribution à risque constant ou sans mémoire, il est équivalent de dire que le logarithme de la fonction de survie : $\ln(S)$, est linéaire.

Exemples. 1.2 : La généralisation de Weibull pour la loi exponentielle à la loi du même nom en introduisant un nouveau paramètre, de manière à ce que la fonction de risque soit la suivante :

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

où, λ et α sont deux constantes strictement positives.

Le paramètre λ donne l'amplitude de la fonction de risque, et la position de α par rapport à 1 définit la monotonie de la fonction de risque : si $\alpha = 1$, on retrouve la fonction de risque constant et donc la loi exponentielle, si $\alpha > 1$, λ_T est croissante (respectivement décroissante) dans le temps.

Grâce à l'expression de la fonction de risque, on obtient les expressions suivantes pour $t \in \mathbb{R}^+$:

$$f(t) = \lambda\alpha t^{\alpha-1}e^{-\lambda t^\alpha}, F(t) = 1 - e^{-\lambda t^\alpha}, S(t) = e^{-\lambda t^\alpha}, \Lambda(t) = \lambda t^\alpha$$

La loi de Weibull est très largement utilisée dans les domaines industriel (fiabilité) et biomédical (analyse de durée de vie).

L'estimation de la fonction de survie ne peut malheureusement pas être employée lorsque nous sommes en présence du temps de survie censurés. Il existe par contre deux méthodes non-paramétriques très connues pour l'estimation de S(t) en présence de censure, c'est-à-dire la méthode actuarielle et la méthode de Kaplan-Meier (aussi appelée méthode du produit-limite).

1.3 Données incomplètes (censurées)

L'analyse des durées de vie pose des problèmes particulier dus au fait que les observations des durées de vie sont le plus souvent censurées. Une des caractéristiques des données de survie est l'existence d'observations incomplètes. Par exemple, dans les enquêtes épidémiologiques, les données sont souvent recueillies de façon incomplète.

La censure et la troncature font partie des processus générant ce type de données. Elles doivent être prises en compte dans l'écriture de la vraisemblance. Nous parlerons de donnée censurées lorsque la durée de survie n'est connue que lorsqu'elle est limitée par une durée limitée d'observation. Par exemple, dans le cas dit de censure à droite, seul le minimum entre la durée de survie et une durée limite supérieure d'observation est connu, ainsi que l'indicateur exprimant que la durée de survie a été censurée ou non.

La censure à gauche correspond au cas où l'individu a déjà subi l'évènement avant qu'on ne l'observe. Dans ce cas, la seule information dont on dispose est que ces durées sont

inférieures à une certaine valeur. C'est un cas moins répandu car, en général, les critères d'inclusion des enquêtes exigent d'étudier des sujets qui n'ont pas subi l'évènement. La censure par intervalles. Les deux phénomènes ci-dessus sont conjugués. On Considère trois types de censure, censure de type I, censure de type II, censure aléatoire.

Une observation est dite tronquée si elle est conditionnelle à un autre évènement. On dit que la variable T de durée de vie ou de temps de séjour dans l'état 0 est tronquée si T n'est observable que sous une certaine condition dépendant de la valeur de T. On note Z_i la variable aléatoire (v.a) positive associée à la durée de vie d'un individu i ($1 \leq i \leq n$), où n est la taille de l'échantillon. Les n observations disponibles T_1,...,T_n, sont alors définies par

$$T_i = \min(Z_i, C_i) \quad et \quad \delta_i = \mathbb{I}_{[Z_i<C_i]}$$

où, C_i est une v.a. positive, dite variable de censure. Les v.a. T_i et C_i sont i.i.d., en outre C_i et Z_i sont supposées indépendantes entre elles. On observe, aussi, s'il y a eu censure ou pas : pour chaque individu i, on connait la valeur de la variable indicatrice de non censure, δ_i. Nassiri Abdelhak, Delecroix Michel, Bonneu Michel (2000) [63]

-Vraisemblance dans un modèle de survie censuré

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

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

La variable de censure C est supposée indépendante de la variable T et on suppose que la distribution de T est entièrement définie par la connaissance d'un paramètre θ de dimension finie (à estimer par MV), on dit que la loi de la censure C est non informative. On note $f_C(.)$, $F_C(.)$, $S_C(.)$ les densité, fonction de répartition et fonction de survie de la variable C.

Les observations sont donc des réalisations de $T = \min(Z; C)$ et de l'indicatrice de censure $\delta = \mathbb{I}_{[Z<C_i]}$. Notons $(T_i; \delta_i)_{i\in\{1,...,n\}}$ un échantillon des variables (T ;C). L'estimation de θ_0 à partir des observations peut être effectuée par la méthode du maximum de

vraisemblance.

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

$$\begin{aligned} L_n(\theta) &= \Pi_{i=1}^n (f_{Z;\theta}(T_i)S_C(T_i))^{\delta_i}(S_{Z;\theta}(T_i)f_C(T_i))^{1-\delta_i} \\ &= \Pi_{i=1}^n \lambda_{Z;\theta}(T_i)^{\delta_i} S_{Z;\theta}(T_i) S_C(T_i)^{\delta_i} f_C(T_i)^{1-\delta_i} \end{aligned}$$

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

$$\Pi_{i=1}^n \lambda_{Z;\theta}(T_i)^{\delta_i} S_{Z;\theta}(T_i)$$

1.4 Les modèles multi-états de type Markovien

Les modèles multi-états ne cessent de connaître un intérêt croissant. Ces modèles utilisent la notion d'« état » et de processus pour décrire un phénomène. La notion de processus est utilisée pour représenter les différents états successivement occupés à chaque temps d'observation. L'étude de ces modèles consiste à analyser l'intensité de transition entre les différents états.

On s'intéresse ici à un modèle multi-état de survie ne comportant que deux états avec une seule transition possible de l'état 0 à l'état 1. Cette fonction est aussi appelée fonction de risque instantané de décès, ou intensité de transition entre les états 0 et 1.

1.4.1 Modèle de Markov à deux états

Rappelons que pour une chaîne de Markov du premier ordre, l'état de la variable $X(t)$ à l'instant t ne dépend que de son état observé au dernier instant.

Un processus de Markov à deux états (homogène ou non-homogène) $\{X(t); t \in \mathcal{F}\}$ à temps continu et à espace d'états fini $S = \{0, 1\}$ est complètement défini par

– Son vecteur des probabilités initiales, noté P_0 tel que

$$P_0[j] = Pr(X(0) = j), j = 0, 1$$

avec $\sum_{j=0}^{1} P(X(0) = j) = 1$,

– Sa matrice de probabilités de transition : $P(t, t+s) = (p_{ij}(t, t+s))_{i,j}$ telle que

$$p_{ij}(t, t+s) = Pr(X(t+s) = j \backslash X(s) = i) \quad \forall s, t \in \mathcal{F} \quad et \quad i, j \in S$$

Ainsi, nous avons quatre situations :

$$p_{00}(t, t+s) = Pr(X(t+s) = 0 \backslash X(s) = 0)$$

$$p_{01}(t, t+s) = Pr(X(t+s) = 1 \backslash X(s) = 0)$$

$$p_{10}(t, t+s) = Pr(X(t+s) = 0 \backslash X(s) = 1)$$

$$p_{11}(t, t+s) = Pr(X(t+s) = 1 \backslash X(s) = 1)$$

Le processus peut rester dans un état ou transiter vers un autre état à chaque temps. Cela se traduit par la propriété suivante :

$$\sum_{j} P_{ij}(t, t+s) = 1; \quad j = 0, 1.$$

Où p_{ij} est la probabilité d'aller à l'état j sachant qu'on se trouve à l'état i.

$$P(t, t+s) = \begin{pmatrix} p_{00}(t, t+s) & p_{01}(t, t+s) \\ p_{10}(t, t+s) & p_{11}(t, t+s) \end{pmatrix}$$

on en déduit pour tout $t, s > 0$:

$$P_{ij}(0, t+s) = \sum_{k} p_{ik}(0, t) p_{kj}(t, t+s) \tag{1.1}$$

Que l'on peut écrire sous forme matricielle :

$$P(0, t+s) = p(0, t) p(t, t+s) \tag{1.2}$$

Cette relation est appelée équation de Chapman-Kolmogorov.

Le paramètre d'intérêt en analyse de survie dans les modèles multi-états est l'intensité de transition (**ou fonction de risque instantanée**) $\alpha_{ij}(t)$, $i, j \in S$. Elle est définie, pour

$i \neq j$ par :

$$\alpha_{ij}(t) = \lim_{\Delta t \to O^+} \frac{P_{ij}(t, t+\Delta t)}{\Delta t} \tag{1.3}$$

Notons que $\alpha_{ij}(t)\Delta t$ représente la probabilité que le processus passe dans l'état j entre les temps t et $t + \Delta t$ sachant que le processus est dans l'état i au temps t. La fonction $\alpha_{ij}(t)$ représente donc la vitesse de transition de i vers j au temps t. Si, $i = j$, $\alpha_{ii}(t)$ est défini comme suite :

$$\alpha_{ii}(t) = -\sum_{j \neq i} \alpha_{ij}(t) \quad et \quad \sum_{j} \alpha_{ij}(t) = 0 \tag{1.4}$$

1.4.2 Modèle de type Markovien homogène

Dans les applications, le processus markovien peut être considéré comme homogène : Les intensités de transitions ne varient pas dans le temps. L'équation de Chapman-Kolmogorov peut alors s'écrire :

$$P(t+s) = p(t)p(s) \tag{1.5}$$

donc

$$\begin{aligned} \frac{dP(s)}{ds} &= \lim_{\Delta s \to 0^+} (P(s+\Delta s) - P(s))/\Delta(t-s) \\ &= P(t-s) \lim_{\Delta s \to 0^+} (P(\Delta(t-s) - I)/\Delta(t-s) \\ &= P(t-s)Q \end{aligned} \tag{1.6}$$

Où, I est la matrice identité et la matrice Q est la matrice des intensités de transition :

$$Q = \begin{pmatrix} \alpha_{00} & \alpha_{01} \\ \alpha_{10} & \alpha_{11} \end{pmatrix}$$

Dans un modèle de Markov homogène, les intensités de transition ne dépendent pas du temps. La solution de l'équation différentielle est :

$$P(t) = \exp(Qt) \tag{1.7}$$

Il peut être utile de décomposer la matrice Q en $Q = BDB^{-1}$, où D est la matrice diagonale des valeurs propres de la matrice des intensités de transition et B est la matrice des vecteurs propres correspondants. La distribution du temps d'attente dans l'état i est définie par une loi exponentielle :

$$P_{ii}(u) = \exp(-\alpha_{ii}u) \tag{1.8}$$

Ces distributions des temps de séjour données par la diagonale de la matrice P(t) sont dites sans mémoire

1.4.3 Modèle de type Markovien non-homogène

Dans un modèle de Markov non-homogène, la mesure d'intensité cumulée est un autre paramètre qui permet de définir un processus de Markov. C'est une matrice de fonctions de dimension 2×2 ; notée $A = \{A_{hj}\}_{h,j}$,tel que,

$$A_{hh}(t) = -\sum_{j \neq h} A_{hj}(t). \quad h, j \in S$$

A_{hj} est la fonction d'intensité cumulée pour les transitions de l'état h vers l'état j, alors que A_{hh} est l'opposée de la fonction d'intensité cumulée pour les transitions qui quittent l'état h.

Les équations différentielles de Kolmogorov définissent le lien entre la matrice de probabilité de transition et la matrice d'intensité cumulée

– équation de Kolmogorov :

$$\frac{\partial P(s,t)}{\partial t} = P(s,t)A(dt),$$

$$\frac{\partial P(s,t)}{\partial t} = A(ds)P(s,t),$$

Proposition 1.4.1. *Soient A un processus croissant et C un processus prévisible. Alors, pour tout t,*

$$\int_0^t C(s)dA(s) = \int_0^t C(s)A(ds),$$

est une variable aléatoire.

Si $A(t)$ est un processus croissant alors,

$$\begin{aligned} A(t) &= \int_0^t dA(s) \\ &= \int_0^t A(ds) \quad avec \quad A(0) = 0 \end{aligned}$$

les fonctions $A_{hj}(.)$ sont supposées absolument continues, c'est-à-dire qu'il existe des **fonctions d'intensité** α_{hj} tel que

$$A_{hj}(t) = \int_0^t \alpha_{hj}(u)du$$

où,$\alpha_{hj}(.)$ est déterministe.

Les fonctions $\alpha_{hj}(.)$ sont appelées les intensités de transition et sont définies par

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t+\Delta t)}{\Delta t}, \quad h \neq j,$$

$$\alpha_{hh}(t) = -\sum_{h \neq j} \alpha_{hj}(t), \quad h = 0, 1.$$

– **Cas particulier : données de survie**

« L'évènement étudié » dans l'analyse des données de survie est le passage irréversible entre deux états fixés. Le premier état est généralement nommé « vivant » et l'état absorbant est communément appelé « décès ». Le terme « décès » représente un changement d'état irréversible qui peut aussi bien représenter la mort d'un individu, l'apparition d'une maladie, ou encore une panne de machine...

Le modèle de survie peut être considéré comme un modèle de Markov non-homogène particulier comportant deux états avec une seule transition possible. Le processus est Markovien dans le sens où le passé du processus se résume à l'état présent. Le processus ponctuel de comptage associé est un processus dont les marques sont : état 0 (vivant), état 1 (décès). La matrice des probabilités de transition associée est définie par

$$P(t, t+s) = \begin{pmatrix} p_{00}(t, t+s) & p_{01}(t, t+s) \\ 0 & 1 \end{pmatrix}$$

Et la matrice des intensités cumulées par

$$A(t) = \begin{pmatrix} A_{00}(t) & A_{01}(t) \\ 0 & 0 \end{pmatrix}$$

Dans ces modèles de type Markovien, les intensités de transition entre les états peuvent dépendre de différentes échelles de temps, en particulier,

– La durée du suivi (temps depuis l'inclusion dans l'étude),

– Le temps depuis la dernière transition (durée dans l'état présent). Il existe plusieurs possibilités pour définir les intensités de transition $\alpha(t;d)$, où t représente la durée du suivi et d la durée passée dans l'état. Lorsque $\alpha(t;d) = \alpha$; le modèle est dit homogène par rapport au temps t. Lorsque $\alpha(t;d) = \alpha(t)$ le modèle est dit non-homogène.

1.4.4 Les modèles multi-états

Nous avons donc présenté dans la section précédente le modèle de survie classique, pouvant être représenté comme un modèle à deux états. Dans cette section, nous généralisons ce dernier modèle aux modèles multi-états avec plus de deux états. Il existe plusieurs cas découle des modèles multi-états qui résument bien l'utilisation des modèles multi-états. Nous présentons dans cette section ces différents modèles.

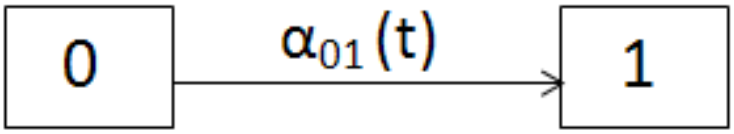

FIG. 1.1 – Modèle de survie

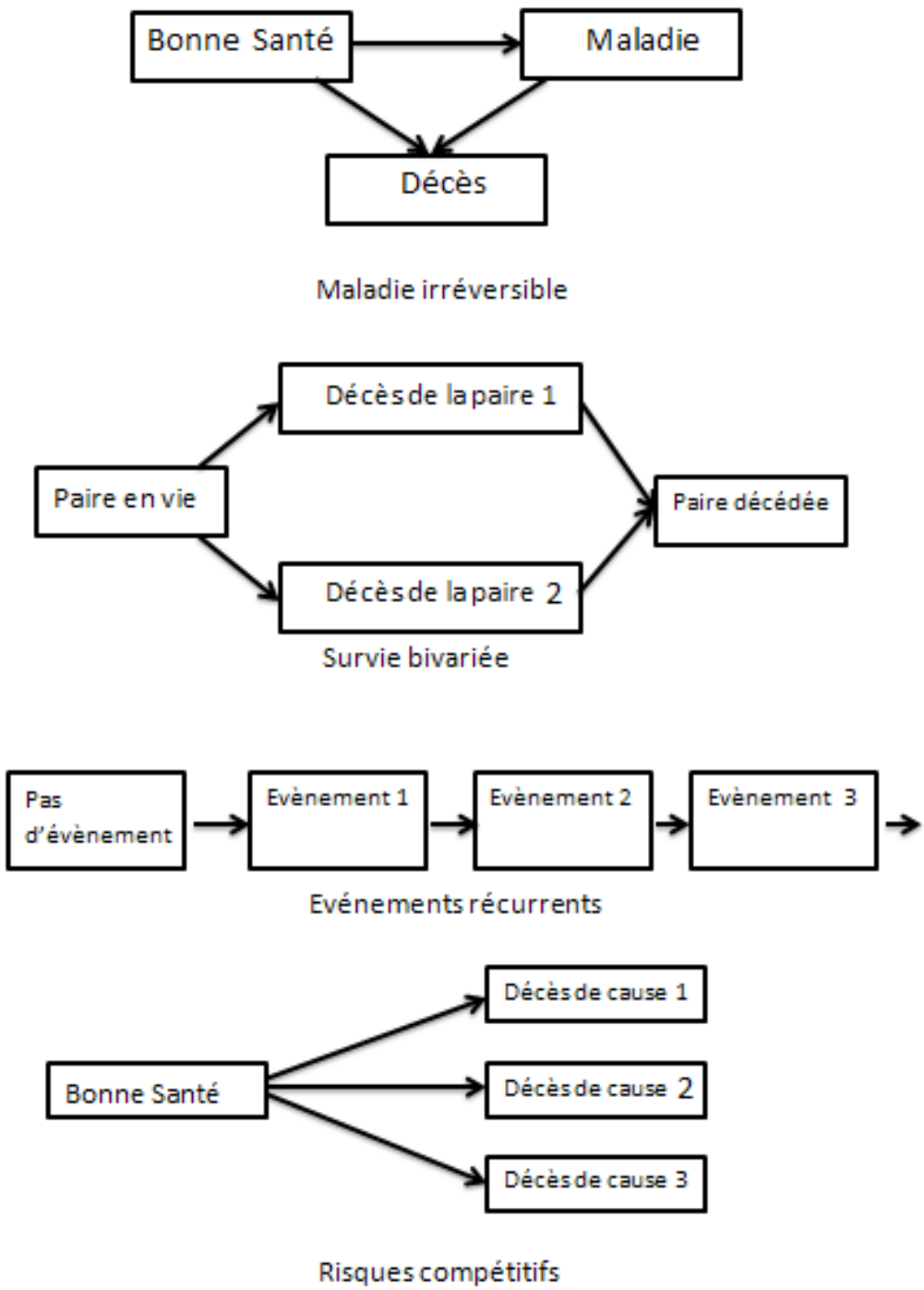

FIG. 1.2 – Structures de modèles multi-états

Plusieurs stratégies sont possibles pour l'estimation direct de la fonction de hasard. Nous nous intéressons principalement dans ce chapitre à une approche exploratoire des modèles à travers l'étude d'un estimateur paramétrique (On peut supposer que la variable de durée suit une loi de probabilité donnée, par exemple une loi exponentielle, une loi de Weibull...). On peut alors écrire la vraisemblance de l'échantillon observé, et estimer ses paramètres par maximisation, on peut aussi introduire dans le modèle des covariables qui déterminent la valeur de certains paramètres(voir chapitre 3).Par la suite on va présenté une partie de la spécification de la loi de la durée, on parlera de modèle semi-paramétrique ou non paramétrique.

1.5 Estimation paramétrique

L'estimation des fonctions de hasard doit a priori s'effectuer sur des populations homogènes. Si la population regroupe des catégories dont les lois de durées sont différentes, le risque est en effet de conclure faussement à une décroissance de la fonction de hasard.

Il existe plusieurs catégories de familles paramétriques. Les plus courantes sont les familles à hasard proportionnel et les familles à hasard accéléré.

Dans les familles à hasard proportionnel, la fonction de hasard a pour forme générale :

$$h(t) = h_0(t)\phi(X, \beta).$$

$h_0(t)$ est appelé "la fonction de hasard de base", et $\phi(X, \beta)$ est une fonction positive des exogènes X, β étant un vecteur de paramètres.

Dans les familles à hasard accéléré, la fonction de hasard a pour forme générale :

$$h(t, X, \beta) = h_0[t \exp(X\beta)] \exp(X\beta).$$

Cette écriture permet d'écrire simplement les modèles à durée de vie accélérée sous la forme :

$$\log T = -X\beta + \log T_0$$

Cette écriture peut faire penser à un modèle de régression linéaire, où $\log T_0$ jouerait le rôle de la perturbation.

L'approche paramétrique peut être utilisée pour modéliser la distribution de survie. Ce type d'estimation ayant retenu une forme de distribution donnée cherche à en estimer les paramètres. Un terme correctif pouvant prendre en compte l'effet de variables exogènes ou covariables.

La méthode d'estimation paramétrique repose sur une estimation des lois des temps de séjour par des fonctions paramétriques.

L'une des difficultés d'estimation des modèles de durées est l'impossibilité d'appliquer les modèles de régression habituels, sauf dans des cas très particuliers. on pouvait penser à écrire un modèle de la forme :

$$\log T = X\beta + U_i$$

où, U est une perturbation. Mais les moindres carrées ordinaires ne sont généralement pas convergents, sauf dans le cas où les données observées ne sont pas censurées. La méthode utilisée est donc presque toujours le maximum de vraisemblance.

1.5.1 Ecriture de la vraisemblance dans les modèles de durée

Supposons que, dans le cas d'un échantillon de taille N, soient des durées observées, complètes ou censurées, t_i pour chaque individu $i = 1, ..., N$. Cela revient à disposer, en plus de la valeur de t_i, d'une variable indicatrice de censure C_i, telle que $C_i = 1$ si la durée t_i est censurée, et 0 sinon.

La vraisemblance du modèle s'écrit alors :

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

En effet, la probabilité qu'une durée soit censurée en t_i, donc supérieure où égale à t_i est la valeur de la survie $S(t_i)$.

La log-vraisemblance a donc pour forme

$$\log L = \sum_{i=1}^{n} c_i \log f(t_i) + \sum_{i=1}^{n} (1 - c_i) \log S(t_i).$$

Cette expression peut se simplifier en utilisant la relation $\lambda(t_i) = f(t_i)/S(t_i)$, ce qui donne

$$\log L = \sum_{i=1}^{n} c_i \log \lambda(t_i) + \sum_{i=1}^{n} \log S(t_i).$$

Lorsque l'on spécifie une forme particulière pour λ et donc pour S, avec éventuellement introduction de variables exogènes, on obtient simplement la valeur de la fonction à maximiser en calculant $\log \lambda(t_i)$ et $\log S(t_i)$

1.5.2 Cas particulier d'un modèle de Weibull

Dans le cas d'un modèle de durée simple(sans sélection endogène), la log-vraisemblance s'écrit donc :

$$\log L = \sum_{i=1}^{n} c_i \log \lambda(t_i) + \sum_{i=1}^{n} \log S(t_i),$$

où, C_i est la variable indicatrice de censure. Dans le cas d'un modèle de Weibull à hasard proportionnel, le hasard s'écrit :

$$\lambda(t_i) = \alpha(\exp(\acute{x_i}\beta))t_i^{\alpha-1},$$

où, $\acute{x_i}$ est le vecteur ligne des valeurs prises par les variables exogènes pour l'individu i. La survie a pour forme :

$$S(t_i) = \exp[-\exp(\acute{x_i}\beta)t_i^{\alpha}].$$

La log-vraisemblance vaut donc :

$$\log L = \sum_{i=1}^{n} c_i[\log \alpha + \acute{x_i}\beta + (\alpha - 1)\log t_i] - \sum_{i=1}^{n}(\exp(\acute{x_i}\beta))t_i^{\alpha}].$$

Les dérivées partielles de la log-vraisemblance par rapport à α et β valent :

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} c_i[\frac{1}{\alpha} + \log t_i] - \sum_{i=1}^{n} \exp(\acute{x_i}\beta)t_i^{\alpha} \log t_i$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} c_i\acute{x_i} - \sum_{i=1}^{n} \acute{x_i} \exp(\acute{x_i}\beta)t_i^{\alpha}.$$

1.6 Estimation non-paramétrique :Kaplan-Meier

Ce type d'estimation vise à approximer l'une ou plusieurs des différentes fonctions caractérisant la distribution observée (F ou λ le plus souvent) sans faire d'hypothèse sur celle-ci.

L'estimateur de Kaplan Meier est très simple à calculer, et généralise la notion de répartition empirique en tenant compte des données censurées à droite. C'est pourquoi il sert généralement de base à toute étude sur les durées. Il peut en effet guider le choix d'une forme paramétrique particulière. Rappelons qu'il doit être calculé pour des populations homogènes.

Pour comprendre le principe du calcul, plaçons-nous dans le cas où il n'y a pas de censure. Alors la survie en t peut être simplement estimée par :

$\hat{S}(t) = 1 - \hat{F}(t)$ où $\hat{F}(t) = n_t/N$,

avec n_t : nombre de durées inferieures à t et N : nombre total d'observations.

On peut remarquer que la fonction de survie estimée peut s'écrire simplement comme un produit de probabilités conditionnelles. Dans le cas simple sans censure et où on n'observe qu'une seule fois chaque valeur de durée, que l'on notera dans l'ordre croissant $t_0, t_1, ..., t_N$, avec $t_0 = 0$. On a alors

$$S(t) = P(T > t) = \prod_{t_i \leq t} P(T > t_i / T > t_{i-1}) = \prod_{j<i} (1 - \alpha_i),$$

où, α_i est la probabilité instantanée de sortir en t_j (l'équivalent de la fonction de hasard en temps discret). Cette probabilité α_i vaut alors $1/(N - j + 1)$, puisqu'on observe une sortie en j parmi les $N - (j - 1)$ personnes qui survivent juste après t_{j-1}. Ces $N - (j - 1)$ personnes sont appelées, par référence aux données médicales, l'ensemble à risque en t_j.

Si maintenant certaines durées sont censurées à droite, on va reprendre la même idée, mais en adaptant la notion d'ensemble à risque en t_j. Il sera cette fois défini comme le nombre r_j d'observations ni sorties, ni censurées avant t_j. Alors l'estimateur de α_j s'écrira $1/r_j$, et la survie sera estimée par $\prod_{j<i}(1 - 1/r_j)$.

Dans le cas le plus général où l'on peut observer un nombre d_j supérieur à 1 de sorties

à chaque date j, l'estimateur de Kaplan Meier pour le hasard à la date j sera d_j/r_j, et celui de la survie sera :

$$\hat{S}(t_j) = \prod_{t_j<t}(1 - d_j/r_j).$$

Notons également que l'on peut l'utiliser pour estimer une durée moyenne puisque l'espérance de la durée peut généralement s'écrire :

$$E(T) = \int_0^\infty uf(u)du = \int_0^\infty S(u)du,$$

on peut utiliser l'estimateur suivant :

$$\bar{T} \quad = \quad \sum_{i=1}^{I}(t_i - t_{i-1})\hat{S}(t_i), \tag{1.9}$$

I étant le nombre de durées différentes observées. La durée moyenne ne sera donc la moyenne empirique que s'il n'y a pas de censure.

-Le taux de hasard défini précédemment(section 1.2) admet trois formulations équivalentes, notées :

$\lambda_l(t), l = 1, 2, 3$. Pour tout $t \in \mathbb{R}^+$ tel que $F(t) < 1$:

1.

$$\lambda_1(t) \quad = \quad \frac{f(t)}{S(t)}.$$

Où, $S(t) = 1 - F(t)$ est la fonction de survie de T ;

2.

$$\lambda_2(t) \quad = \quad \frac{g(t)}{1 - F^*(t)},$$

où, g(.) est la densité de la mesure $\nu(.)$ définie par $\nu(.) = prob[(T_i) \in A) \cap (\delta_i = 1)]$, pour tout borélien A de $\mathbb{R}^+$.$g(.)$ correspond donc à la densité des données non censurées. En revanche, $F^*(.)$ est la fonction de répartition des T_i

3.

$$\lambda_3(t) = \frac{-d[\ln S(t)]}{dt}.$$

Plusieurs estimateurs naturels non-paramétriques du taux de hasard ont été définis à partir de ces trois formulations.

Concernant la définition (2), il est clair que l'échantillon fournit les observations nécessaires à la construction d'un estimateur à noyau de la densité $g(.)$ et d'un estimateur de $F^*(.)$ (on prend la fonction de répartition empirique $F_n^*(.)$ de T_i). On en déduit l'estimateur classique suivant

$$\hat{\lambda}_2(t) = \frac{\hat{g}(t)}{1 - F_n^*(t)}$$

où, $\hat{g}(t) = \frac{1}{n.h} \sum_{i=1}^{n} \delta_i . K(\frac{T_i - t}{h})$ pour un noyau $K(.)$ de Parzenrosenblatt et une fenêtre h. Cet estimateur a été proposé par Blum et Susarla [1980].

En revanche, dans le cas des deux autres formulations (1) et (3), nous ne disposons pas des n observations de T pour en estimer simplement la densité $f(t)$ et la fonction de survie $S(t)$. En pratique, pour estimer cette fonction de survie, on utilise l'estimateur de Kaplan-Meier [1958] défini par :

$$\hat{S}_{KM}(t) = \prod_{T_i < t} \left(\frac{n - r_i}{n - r_i + 1} \right)^{\delta_i}$$

Où, r_i représente le rang de l'observation T_i dans l'échantillon. Le procédé d'estimation de (1) et (3) consiste donc à approximer globalement la loi des Z_i par une loi discrète définie sur l'ensemble des observations des T_i, par les probabilités p_i déduites de $\hat{S}_{KM}(t)$ Soit

$$P_i = \frac{\delta_i}{n - r_i + 1} \prod_{j/t_j < t_i} \left(\frac{n - r_j}{n - r_j + 1} \right)^{\delta_i}$$

On définit alors les deux estimateurs respectifs aux formulations (1) et (3) par :

$$\hat{\lambda}_1(t) = \frac{\hat{f}(t)}{\hat{S}_{KM}(t)}$$

$$\hat{\lambda}_3(t) = \frac{1}{h} \sum_{i=1}^{n} \frac{\delta_i}{n - r_i + 1} K \left(\frac{T_i - t}{h} \right)$$

Où $\hat{f}(t) = \frac{1}{h}\sum_{i=1}^{n} p_i K\left(\frac{T_i - t}{h}\right)$. L'estimateur $\hat{\lambda}_1(t)$ a été notamment étudié par Földes et al [1981], et $\hat{\lambda}_3(t)$ par Tanners et Wong [1983].

L'estimateur de Kaplan Meier a de bonnes propriétés : Il est en effet biaisé à distance finie, mais convergent et de loi asymptotique connue (Normale). Il est donc possible d'utiliser les tests asymptotiques habituels.

1.7 Estimation semi-paramétrique : le modèle de Cox

Un modèle de régression de Cox fait partie de la famille des modèles à risques multiplicatifs ou à risque proportionnel, permettant d'ajuster les intensités de transition en fonction de la valeur des covariables.

La méthodologie de la vraisemblance partielle de Cox permet d'estimer l'effet des covariables dépendantes et indépendantes du temps, sans hypothèse concernant la distribution de base.

Cette section introduit le modèle de Cox et une de ses généralisations que nous allons étudier plus en détail,

1.7.1 Présentation générale

Le modèle de régression à risque proportionnel proposé par Cox en 1972 pour étudier la relation entre le temps d'apparition d'un évènement et un ensemble de covariables en présence de censure est, sans conteste, le modèle le plus utilisé pour l'analyse des données de survie. Il suppose cependant, comme tout modèle de régression multiple, plus d'observations que des variables non fortement corrélées entres elles.

Le modèle de régression de Cox est un des modèles les plus utilisés pour la modélisation de l'influence de covariables sur des données de survie et il a été étudié dans de nombreux ouvrages (Cox & Oakes, 1984 ; Fleming & Harrington, 1991 ; Therneau & Grambsch, 2000 ; Martinussen et Scheike, 2002), car il est d'usage simple et efficace.

Une méthode d'estimation semi-paramétrique concerne les modèles à hasard proportionnels présentés dans la partie 1.7.2 avec la spécification suivante pour la fonction de hasard :

$$\lambda(t) = \exp(Xb)\lambda_0(t),$$

Où λ_0 est la fonction de hasard de base. Elle repose sur la maximisation de la « vraisemblance partielle » de Cox. Elle présente en outre l'avantage de ne pas contraindre les variables explicatives à être constantes au cours du temps.

1.7.2 Vraisemblance partielle de Cox

Reprenons la situation la plus simple où l'on observe autant de durées que d'individus et où il n'y a pas de censure, on ordonne les valeurs des I durées différentes observées : $t_1 < t_2 < ... < t_I$. Soit comme précédemment $r(t_i)$ l'ensemble à risque en t_i. La probabilité pour que ce soit l'individu j de $r(t_i)$ qui sorte en t_i vaut :

$$\frac{\lambda_0(t)\exp(X_j b)}{\sum_{k\in r(t_i)}\lambda_0(t)\exp(X_k b)}$$

Le dénominateur est la probabilité qu'une sortie ait lieu en t_i au sein de l'ensemble à risque. Il vaut la somme des probabilités de sortie de tous les individus de cet ensemble. L'expression se simplifie puisque $\lambda_0(t_i)$ figure dans le dénominateur et le numérateur, et elle vaut finalement :

$$\frac{\exp(X_j b)}{\sum_{k\in r(t_i)}\exp(X_k b)}$$

Sachant que c'est l'individu j qui sort à la date i. La vraisemblance partielle de Cox est le produit de ces probabilités pour l'ensemble des sorties.

$$L(b) = \prod_{i=1}^{I} \frac{\exp(X_{j_i} b)}{\sum_{k\in r(t_i)}\exp(X_k b)}$$

S'il n'y a pas de censure, elle s'interprète comme la vraisemblance de la statistique de rang associée aux durées. L'estimateur semi-paramétrique de b va être obtenu en maximisant la log-vraisemblance partielle par rapport à b au moyen d'une méthode itérative. L'estimateur obtenu converge presque sûrement vers b et est asymptotiquement normal. C.Cases, S.Lollivier[12], A.Morau[54].

1.7.3 Estimation de la fonction de hasard de base

De préférence généralement, estimer directement la fonction de survie. Dans le modèle de Cox, la fonction de survie s'écrivait :

$$S(t) = [S_0(t)]^{\exp(Xb)}.$$

L'estimation de la fonction de survie de base se présenté en deux étapes, on estime b par une maximisation de vraisemblance partielle, ensuite, on remplace b par son estimation, et on maximise la vraisemblance par rapport à S_0

Cette stratégie revient à estimer la fonction de survie de base par :

$$\hat{S}_0(t) = \prod_{t_i<t} \hat{\alpha}_i$$

avec

$$\hat{\alpha}_i \exp(X_i \hat{b}) = 1 - \frac{\exp(X_i \hat{b})}{\sum_{k \in r(t_i)} \exp(X_k \hat{b})}$$

.

La fonction de hasard intégrée $\hat{\Lambda}_0(t) = \int_0^t \lambda_0(u)du$ est alors simplement estimé par $-\log(\hat{S}_0(t))$ si on considère un modèle simple avec pour seule variable exogène une constante s'écrivant :

$$\lambda(t) = \exp(b)\lambda_0(t)$$

Alors la fonction de hasard intégrée sera :

$$\Lambda(t) = \exp(b)\Lambda_0(t).$$

Chapitre 2

Processus Markovien(généralisation)

Dans ce chapitre, nous nous intéressons aux processus Markovien à temps continu et à espace d'états discret, consulter Karlin et Taylor (1975)[45] par exemple.

Nous commençons ici par une rappele des notions sur les processus qui seront nécessaires par la suite. Cette notion de processus est indispensable dans les modèles multi-états.

2.1 Définitions et propriétés

Processus

soit(Ω, A, P) un espace probabilisé, τ l'espace des temps et S l'espace d'états. Un processus stochastique $X(t), t \in \tau$ est une application définie par :

$$X : \tau * \Omega \to S$$

$$(t, w) \mapsto X(t, w)$$

tel que pour tout $t \in \tau$ la fonction $w \mapsto X(t, w)$ est une variable aléatoire sur (Ω, A, P) et à valeur dans S et pour un w donné la fonction $t \mapsto X(t, w)$ est la trajectoire du processus.

Propriété Markovienne

Un processus stochastique vérifie la propriété de Markov si et seulement si la distribution conditionnelle de probabilité des états futurs, étant donné l'instant présent, ne dépend que de ce même état présent et pas des états passés. Cette propriété est résumée par l'équation suivante :

$$P(X(t+h) = r | X(s) = x(s), s \leq t) = P(X(t+h) = r | X(t) = x(t)).$$

Un processus de Markov $\{X(t); t \in \tau\}$ à temps continu et à espace d'états fini est un processus dont l'évolution future $\{X(t); t \geq s\}$ ne dépend pas de son passé si son état à l'instant s est connu.

Définition 2.1.1. (Chaîne de Markov)

Un **processus de Markov** à temps continu et à espace d'états fini $S = \{1, ..., k\}$ est complètement défini par

1. Son vecteur des probabilités initiales, noté $\mathbf{P}_0$ tel que

$$P_0[j] = Pr\{X(0) = j\}, \quad j = 1, ..., k$$

avec $\sum_{j=1}^{p} P\{X(0) = j\} = 1$,

2. Sa matrice de probabilités de transition : $P(s,t) = (p_{ij}(s,t))_{i,j}$ tel que

$$p_{ij}(s,t) = Pr\{X(t) = j \backslash X(s) = i\} \quad \forall s, t \in \tau \quad et \quad i, j \in \mathcal{S}.$$

Avec $P(s,s) = Id$, et $\sum_{j=1}^{k} p_{ij}(s,t) = 1$ et $0 \leq s \leq t$.

Les probabilités de transition d'un processus Markovien vérifient la relation suivante, $\forall i, j \in S = \{1, ..., k\} \quad et \quad \forall 0 < s < u < t$,

$$P_{ij}(s,t) = \sum_{k \in S} p_{ik}(s,u) p_{kj}(u,t), \tag{2.1}$$

Cette propriété est appelée équation de Chapman-Kolmogorov. Sous forme matricielle, l'équation (2.1) s'écrit

$$P(s,t) = P(s,u)P(u,t) \quad \forall s \leq u \leq t.$$

Les intensités de transition sont d'autres paramètres qui permettent de définir un processus de Markov.

Le paramètre d'intérêt en analyse de survie dans les modèles multi-états est l'intensité de transition (ou fonction de risque instantané) $\alpha_{ij}(t)$. Qu'elle est définie, pour $i \neq j$ par :

$$\alpha_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t}, i \neq j$$

$$\alpha_{ii}(t) = -\sum_{i \neq j} \alpha_{ij}(t), i = 1, ..., k.$$

Notons que $\alpha_{ij}(t)\Delta t$ représente la probabilité que le processus passe dans l'état j entre les temps t et $t + \Delta t$ sachant que le processus est dans l'état i au temps t. La fonction $\alpha_{ij}(t)$ représente donc la vitesse de transition de i vers j au temps t

2.2 Processus Markovien homogène

Un processus de Markov est homogène si la probabilité de transition de l'état i vers j est définie par

$$\begin{aligned} p_{ij}(s,t) &= Pr\{X(t)=j \backslash X(s)=i\} \\ &= p_{ij}(0,t-s), \\ &= P(t-s). \end{aligned}$$

Les probabilités de transition dépendent uniquement du temps entre deux transitions. Dans ce cas particulier, les intensités de transition du processus sont indépendantes du temps, pour tout $i \neq j$

$$\alpha_{ij}(s) = \lim_{\Delta t \to 0} \frac{p_{ij}(s, s+\Delta t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{p_{ij}(\Delta t)}{\Delta t} = \alpha_{ij}.$$

À l'aide de l'équation de Chapman-Kolmogorov (2.1), on obtient l'équation différentielle suivante :

$$\frac{\partial P(0,t)}{\partial t} = P(0,t)Q.$$

Où Q est la matrice des intensités de transition.

Sachant que $P_{ii}(0) = 1$ et $P_{ij}(0) = 0$, la solution de cette équation est donnée par,

$$P(0,t) = \exp(Q \times t). \tag{2.2}$$

2.3 Modèle de Markov homogène

Les processus de Markov homogènes peuvent être utilisés pour modéliser l'évolution d'une maladie par exemple. L'hypothèse d'homogénéité permet d'avoir une définition simplifiée des probabilités de transition à partir des intensités de transition

2.3.1 Vraisemblance

Soit $\{X(t), t \in F\}$ un processus de Markov homogène à espace d'états fini $S = \{1, ..., k\}$ chaque individu se déplace indépendamment entre les états suivant le processus X.

La contribution de l'individu h à la vraisemblance est

$$l_h = P_0[x_{h,0}] \times \prod_{j=1}^{n_h} P_{x_{h,j-1}, x_{h,j}}(T_{h,j} - T_{h,j-1})$$

La vraisemblance totale est le produit des contributions individuelles,

$$L = \prod_{h=1}^{n} l_h \tag{2.3}$$

La vraisemblance dépend alors uniquement des intensités de transition. La méthode du maximum de vraisemblance permet d'obtenir les estimations des paramètres.

2.3.2 Prise en compte de covariable

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

On considère X=(X_1,....,X_p)' le vecteur de p variables explicatives réelles associée à la durée de survie T. On dispose de covariables sur chaque individu. Ainsi, l'approche classique de modélisation de l'effet des covariables sur une donnée de survie est de modéliser la fonction de risque conditionnelle aux covariables comme une fonction de celles-ci. Deux classes de modèles généraux sont utilisées pour cette modélisation :

Les modèles à risque additif ne seront pas abordés ici, le lecteur pourra se référer à D.Y. Lin, Zhiliang ying [42] ...

Le modèle peut être étendu de manière simple, en considérant un modèle de régression à intensités proportionnelles (Cox [1972]). Les intensités de transition peuvent s'écrire :

$$\alpha_{ij}(X) = \alpha_{ij0} \exp(\acute{\beta}_{ij} X).$$

Avec X un vecteur de covariables indépendantes du temps de dimension s, β_{ij} un vecteur de s coefficients de régression et α_{ij0} l'intensité de transition de base associée à la transition de l'état i vers l'état j. En effet, les estimations des intensités de transition sont toujours positives quelles que soient les valeurs de X et de β_{ij}. De plus, ce modèle fournit des résultats en terme de risques relatifs qui sont facilement interprétables (comme dans

le modèle de Cox à risques proportionnels).

-Les modèles à hasard proportionnel :

La forme générale de la fonction de hasard pour ce type de modèle s'écrit :

$$\lambda(t) = \varphi(X, b)\lambda_0(t)$$

$\lambda_0(t)$ est appelé "fonction de hasard de base". L'effet des variables explicatives consiste à multiplier par un facteur d'échelle ce hasard de base. Le plus souvent, on adopte la convention :

$$\varphi(X, b) = \exp(Xb)$$

Ce qui revient à postuler un facteur d'échelle multiplicatif. Parmi les modèles à hasard proportionnel, on peut citer :

– la loi exponentielle pour laquelle la fonction de hasard de base est constante. Cela signifie qu'à n'importe quelle date, la probabilité de changer d'état est la même. C'est la raison pour laquelle on dit fréquemment sur le modèle exponentiel qu'il est « sans mémoire » (le processus sous-jacent est Markovien). Ses caractéristiques sont les suivantes :

$$\lambda(t) = \exp(Xb)$$

$$S(t) = \exp[-t\exp(Xb)]$$

$$f(t) = \exp(Xb)\exp[-t\exp(Xb)]$$

et l'espérance de la durée s'écrit :

$$E(T/X) = \exp(-Xb)$$

Notons que la fonction de répartition peut aussi s'écrire :

$$F(t) = 1 - \exp[-\exp(\log(t) + Xb)]$$

de sorte que le modèle exponentiel peut également s'interpréter comme un modèle à durée de vie accélérée

Le modèle peut être adapté afin de prendre en compte des covariables dépendantes du temps. En effet, la vraisemblance est le produit des contributions associées à chaque transition observée dans la base. Le terme $P_{ij}(t-s\backslash X)$, peut être remplacé par $P_{ij}(t-s\backslash X(s))$ en supposant que les valeurs des covariables dépendantes du temps restent constantes entre les deux temps consécutifs s et t

2.4 Modèle semi-Markovien homogène

Les processus semi-Markoviens constituent alors une alternative intéressante puisqu'ils intègrent dans la définition du modèle les lois de temps de séjour dans l'état. Un processus semi-Markovien dont les temps de séjour suivent des lois exponentielles devient un processus Markovien homogène. Les modèles semi-Markoviens généralisent ainsi les modèles Markoviens dans le sens où ils permettent de définir explicitement les lois des temps de séjour dans les états.

Les modèles semi-Markoviens commencent à être utilisés dans plusieurs domaines. En épidémiologie, Huber-Carol et Pons [2004] ont appliqué ces modèles à la transplantation cardiaque, Heutte et al.[2001] ont modélisé l'évolution d'un patient atteint du VIH, alors que Dabrowska et al. [1994] ont étudié la greffe de moelle osseuse. Ils sont aussi appliqués en fiabilité par (Perez-Ocon et Torres-Castro [2002]), dans les sciences sociales pour la recherche d'emploi, par exemple (Vassiliou et Papadopoulou [1992]), et en finance, (Janssen et al. [1997]).

Définition 2.4.1. En l'absence de covariables, on observe pour chaque individu le couple $(T, X) = (T_n, X_n) : n \geq 0$, où $0 = T_0 < T_1 < ... < T_n$ sont les temps consécutifs d'entrée dans les états $X_0, X_1, ..., X_n \in E$, avec $X_{p+1} \neq X_p, \forall p \geq 0$. n représente le numéro de la transition. Pour faire le lien avec les processus de comptage, nous noterons pour une seule réalisation du processus (un individu) :

$$\tilde{N}_{ij}(t) = \sum_{n\geq 1} I\{T_n \leq t, X_n = j, X_{n-1} = i\} \ \forall i, j \text{ tels que } i \neq j$$

où $\tilde{N}_{ij}(t)$ représente le nombre de transitions $i \rightarrow j$ observées dans l'intervalle de temps $[0, t]$.

où $\tilde{N}(t) = \sum_{i,j} \tilde{N}_{ij}(t)$ est le nombre total de transitions observées dans $[0, t]$. Ainsi, l'état occupé par le processus au temps t, X(t) sera maintenant noté $X_{\tilde{N}(t)}$. Les séquences $X = X_n, n \geq 0$ forment une chaîne de Markov. Les probabilités de transition associées à cette chaîne sont définies par :

$$P_{ij} \quad = \quad P(X_{n+1} = j | X_n = i) \tag{2.4}$$

- Si l'état i n'est pas un état absorbant, alors
  $P_{ij} > 0$, si $i \neq j$
  $P_{ij} = 0$, si $i = j$.
- Sinon :
  $P_{ij} = 0$ si $i \neq j$
  $P_{ij} = 1$ si $i = j$

le processus (T, X) est dit semi-Markovien si la distribution des temps de séjour $(T_{n+1}-T_n)$ satisfait la condition suivante :

$$P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_0, T_0, .., X_n, T_n) \quad = \quad P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n) \tag{2.5}$$

sachant la séquence des états X, les temps de séjour $T_1, T_2-T_1, T_3-T_2, ...$ sont indépendants.

Parallèlement à l'analyse de survie on note :

1. La fonction de répartition :

$$F_{ij}(x) \quad = \quad P(T_{n+1} - T_n \leq x | X_{n+1} = j, X_n = i) \tag{2.6}$$

2. La fonction de survie :

$$S_{ij}(x) \quad = \quad 1 - F_{ij}(x) = P(T_{n+1} - T_n > x | X_{n+1} = j, X_n = i) \tag{2.7}$$

3. La fonction de densité :

$$f_{ij}(x) \quad = \quad \lim_{dx \to 0^+} P(x < T_{n+1} - T_n < x + dx | X_{n+1} = j, X_n = i)/dx \tag{2.8}$$

4. La fonction de risque :

$$\lambda_{ij}(x) \quad = \quad \lim_{dx \to 0^+} P(x < T_{n+1} - T_n < x + dx | T_{n+1} - T_n \geq x, X_{n+1} = j, X_n = i)/dx \tag{2.9}$$

5. La fonction de risque cumulé :

$$\Lambda_{ij}(x) \quad = \quad \int_0^x \lambda_{ij}(u) du \tag{2.10}$$

D'après le théorème de Bayes et les définitions notées précédemment nous pouvons préciser la condition des modèles semi-Markoviens. pour $i \neq j$

$$P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n = i)$$

$$\begin{aligned} &= P(T_{n+1} - T_n \leq x | X_{n+1} = j, X_n = i) P(X_{n+1} = j | X_n = i) \\ &= F_{ij}(x) P_{ij} \end{aligned} \tag{2.11}$$

Par le théorème des probabilités totales :

$$\begin{aligned} F_{i.}(x) &= P(T_{n+1} - T_n \leq x | X_n = i) \\ &= \sum_j P(T_{n+1} - T_n \leq x, X_{n+1} = j | X_n = i) \\ &= \sum_{j \neq i} F_{ij}(x) P_{ij}. \quad (puisque \quad P_{ii} = 0) \end{aligned} \tag{2.12}$$

Donc la fonction de survie marginale sera :

$$\begin{aligned} S_{i.}(x) &= 1 - F_{i.}(x) \\ &= \sum_{j \neq i} S_{ij}(x) P_{ij} \end{aligned} \tag{2.13}$$

La fonction de densité marginale :

$$\begin{aligned} f_{i.}(x) &= \partial F_{i.}(x)/\partial x \\ &= \sum_{j \neq i} P_{ij} f_{ij}(x) \end{aligned} \tag{2.14}$$

-Par définition, la fonction de risque instantanée de i vers j du processus Semi-Markovien, correspond à la probabilité du processus à transiter juste après le temps x vers l'état j, sachant qu'il est dans l'état i depuis une durée x :

$$\begin{aligned} \alpha_{ij}(x) &= \lim_{\Delta x \to 0^+} P(x \leq T_{n+1} - T_n < x + \Delta x, X_{n+1} = j | T_{n+1} - T_n \geq x, X_n = i)/\Delta x \\ &= P(X_{n+1} = j | X_n = i)/P(T_{n+1} - T_n \geq x | X_n = i) \\ &\times \lim_{dx \to 0^+} P(x \leq T_{n+1} - T_n < x + \Delta x | X_{n+1} = j, X_n = i)/\Delta x \end{aligned} \tag{2.15}$$

donc

$$\alpha_{ij}(x) = P_{ij} f_{ij}(x)/S_{i.}(x) \quad \text{avec} \quad i \neq j \text{ et } \alpha_{ii}(x) = -\sum_{j \neq i} \alpha_{ij}(x).$$

Cette fonction de risque du processus Semi-Markovien, $\alpha_{ij}(x)$, ne doit pas être confondue avec la fonction de risque de la loi des temps de séjour, $\lambda_{ij}(x)$, définie en (2.9). D'après la formule précédente :

$$\begin{aligned} \sum_{i \neq j} \alpha_{ij}(t) &= \sum_{j \neq i} P_{ij} f_{ij}(t)/S_{i.}(t) \\ &= (S_{i.}(t))^{-1} \sum_{j \neq i} P_{ij} f_{ij}(t) \\ &= (S_{i.}(t))^{-1} f_{i.}(t) \\ &= \alpha_{i.}(t) \end{aligned} \tag{2.16}$$

avec $\alpha_{i.}(t)$ la fonction de risque marginale sur j.

2.4.1 La probabilité de transition

Cette sous-section présente la définition et le calcul des probabilités de transition du processus Semi-Markovien.

Le processus Semi-Markovien $Z = \{Z_t; t \in R^+\}$peut être aussi défini par des processus de comptage $Z_t = X_{\tilde{N}_t}$. Donc La probabilité que le processus transite de l'état i à l'état j est défini par :

$$\begin{aligned} p_{ij}(l, l+t) &= P(Z(l+t) = j | Z(l) = i) \\ &= P(X_{\tilde{N}(l+t)} = j | X_{\tilde{N}(l)} = i) \\ &= P(Z(t) = j | Z(0) = i) \\ &= p_{ij}(t), i, j \in E. \end{aligned} \tag{2.17}$$

Et on note que le premier état k est apparu au temps x, la probabilité que le processus soit égal à j au temps $t(t > x)$ s'écrit directement :

$$\begin{aligned} P(Z(t) = j | Z(x) = k) &= P(Z(t-x) = j | Z(0) = k) \\ &= p_{kj}(t-x) \end{aligned} \tag{2.18}$$

Cette probabilité est déterminée par l'équation suivante.

$$p_{ij}(t) = \sum_{k=1}^{r} \int_0^t P_{ik} f_{ik}(x) p_{kj}(t-x) dx + \delta_{ij} \sum_{l \neq i}^{r} P_{il} S_{il}(t) \tag{2.19}$$

2.4.2 Vraisemblance

Soit un échantillon de taille n et k un individu de cet échantillon. On note les $T_{k,1} = 0 < T_{k,2} < ... < T_{k,m_k}$ les temps d'entrée dans les différents états, $X_{k,1}, X_{k,2}, ..., X_{k,m_k}$. Étudions le dernier temps d'observation de l'individu h, noté T_{k,m_h}. Il peut correspondre à une nouvelle transition, ou alors à une censure. Ces deux cas sont classiques en analyse de données de survie :

1-Si l'individu choisi est dans l'état i et transite ensuite dans l'état j après un temps x, alors la contribution pour cette portion de chemin est :

$$\lim_{\Delta x \to 0^+} P(x < T_{n+1} - T_n < x + \Delta x, X_{n+1} = j | X_n = i)/\Delta x = \alpha_{ij}(x) S_{i.}(x) = P_{ij} f_{ij}(x)$$

2-L'observation est censurée à droite, autrement dit, le processus reste dans l'état i jusqu'au temps de séjour x, sa contribution s'exprime donc en terme de survie :

$$P(T_{n+1} - T_n > x | X_n = i) = S_{i.}(x).$$

La vraisemblance peut alors s'écrire comme le produit de toutes les contributions :

$$\nu = \prod_{k \in nc} \left[\prod_{r=1}^{m_k} \{ P_{X_{k,r-1}, X_{k,r}} f_{X_{k,r-1}, X_{k,r}}(T_{k,r} - T_{k,r-1}) \} \right] \tag{2.20}$$

$$\times \prod_{k \in c} \left[\prod_{r=1}^{m_k - 1} P_{X_{k,r-1}, X_{k,r}} f_{X_{k,r-1}, X_{k,r}}(T_{k,r} - T_{k,r-1}) S_{X_{k,m_k-1}}(T_{k,m_k} - T_{k,m_k-1}) \right] \tag{2.21}$$

Avec les deux types d'individus, censurée à droite (c) ou non-censurée (nc).

2.5 Processus Markovien non-homogène

Soit $\{X(t), t \in \mathcal{T} = [0, \tau]\}$ un processus de Markov non-homogène (à temps continu) à espace d'états fini $S = \{1, ..., k\}$ sur $(\Omega, A, \mathbb{P})$. $X(t)$ représente l'état du processus au temps t.

Définition 2.5.1. Un processus de Markov à temps continu est complètement défini par

1. .Son vecteur des probabilités initiales, notées P_0 tel que

$$\mathbf{P}_0[j] = \mathbb{P}\{X(0) = j\}, \quad j = 1, ..., k.$$

avec $\sum_{j=1}^{k} P\{X(0) = j\} = 1$

2. Sa matrice de probabilités de transition entre les instants s et t : $P(s,t) = \{p_{hj}(s,t)\}_{h,j}$ tel que

$$p_{hj}(s,t) \quad = \quad \mathbb{P}\{X(t) = j \backslash X(s) = h\}, \forall \quad 0 \leq s \leq t, \forall \quad h, j \in S,$$

avec $\sum_{j=1}^{k} p_{hj}(s,t) = 1$ pour tout h et $0 \leq s \leq t$

La mesure d'intensité cumulée est un autre paramètre qui permet de définir un processus de Markov. C'est une matrice de fonctions de dimension $k \times k$; notée $A = \{A_{hj}\}_{h,j}$, tq

$A_{hh}(t) = -\sum_{j \neq h} A_{hj}(t)$, pour tout t.

A_{hj} est la fonction d'intensité cumulée pour les transitions de l'état h vers l'état j, alors que A_{hh} est l'opposée de la fonction d'intensité cumulée pour les transitions qui quittent l'état h. Les équations différentielles de Kolmogorov définissent le lien entre la matrice de probabilité de transition et la matrice d'intensité cumulée

– équation « forward » de Kolmogorov

$$\frac{\partial P(s,t)}{\partial t} = P(s,t)A(dt).$$

– équation « backward » de Kolmogorov

$$\frac{\partial P(s,t)}{\partial s} = A(ds)P(s,t).$$

Si $A(t)$ est un processus croissant alors,

$$\begin{aligned} A(t) &= \int_0^t dA(s) \\ &= \int_0^t A(ds). \end{aligned}$$

les fonctions $A_{hj}(.)$ sont supposées absolument continues, c'est-à-dire qu'il existe des fonctions d'intensité α_{hj} tel que

$$A_{hj}(t) = \int_0^t \alpha_{hj}(u)du.$$

$\alpha_{hj}(.)$ est déterministe. Les fonctions $\alpha_{hj}(.)$ sont appelées les intensités de transition et sont définies par

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t+\Delta t)}{\Delta t}, \quad h \neq j,$$

$$\alpha_{hh}(t) = -\sum_{h \neq j} \alpha_{hj}(t), \quad h = 1, ..., k$$

Le temps de séjour dans l'état h suit une loi continue de fonction de risque $-\alpha_{hh}(.)$. La probabilité de quitter l'état h, sachant une transition vers $j \neq h$ au temps t, est donnée par $-\alpha_{hj}(t)/\alpha_{hh}(t)$.

2.5.1 Caractéristiques de la censure à droite

Les mécanismes de censure à droite sont décrits dans le cadre des données de survie. En effet, les mêmes principes s'appliquent aux processus de Markov.

Dans le cas de l'analyse de la survie, les données observées sont :

$$(\tilde{T}_i, D_i, i = 1, ..., n)$$

Avec

- $\tilde{T}_i = min(T_i; U_i)$, le temps d'observation ;
- T_i est la date de survenue de l'événement chez l'individu i ;
- U_i la date de censure correspondante ;
- $D_i = 1_{\{\tilde{T}_i = T_i\}}$, un indicateur de censure.

Quand l'évènement se produit, T_i est « réalisée » ($D_i = 1$). Quand il ne se produit pas (individu étant perdu de vue ou bien exclu vivant), c'est U_i qui est « réalisée » ($D_i = 0$).

– **Censure non aléatoire de type I :**

Les observations pour chaque individu sont arrêtées à un temps fixé u commun à tous, $i.e$

$$\begin{cases} T_i = min(T_i, u) \\ D_i = \mathbf{1}_{\{T_i \leq u\}} \end{cases}$$

Ce mécanisme de censure est couramment utilisé pour tester la durée de vie de n objets identiques sur un intervalle d'observation fixé [0,u] :

– **Censure aléatoire de type I :**

Les observations pour chaque individu sont :

$$\begin{cases} T_i = min(T_i, U_i) \\ D_i = \mathbf{I}_{\{T_i \leq U_i\}} \end{cases}$$

Où, Ui est un temps de censure aléatoire indépendant de T_i : Si de plus, les (U_i)i=1,...,n sont i.i.d. Cette censure est l'une des plus utilisées pour l'analyse de données de survie.

– **Censure aléatoire de type II :**

$$\begin{cases} T_i = min(T_i, T_{(r)}) \\ D_i = \mathbf{I}_{\{T_i \leq T_{(r)}\}} \end{cases}$$

où r est un entier fixé, $1 \leq r \leq n.T(1), ..., T(r), ..., T(n)$ est la statistique d'ordre, $i.e.T_{(r)}$ correspond au $r^{ième}$ temps de décès.

Notons, que dans le cas des processus de Markov, la censure peut dépendre des conditions initiales. Par exemple, on pourrait avoir une censure aléatoire avec des distributions différentes suivant l'état de l'individu au temps 0.

Cela correspond à un mécanisme de censure aléatoire de type I, type de censure que nous allons traiter dans la suite de ce travail.

2.5.2 Modèle avec un état de censure

Considérons un modèle de Markov à trois états avec retour, les intensités de transition de l'état h vers l'état j sont

$$\lambda_{hj}(t|.) = \lim_{\Delta t \to 0} \frac{P(X(t+\Delta t) = j | X(t) = h, .)}{\Delta t}, \quad h, j = 1, 2, 3, \quad h \neq j. \tag{2.22}$$

Dans ces modèles l'hypothèse de censure indépendante peut se définir par

$$\lambda_{hj}(t|C_h > t) \quad = \quad \lambda_{hj}(t), \quad h, j = 1, 2, 3, \quad h \neq j. \tag{2.23}$$

Où C_h correspond au temps de censure pour un individu dans l'état h si on suppose que les covariables déterminent le processus d'évènement, alors si elles déterminent aussi la censure : le processus d'évènement et la censure seront dépendants par l'intermédiaire des covariables. L'objectif est alors d'étudier les risques de censure afin de montrer un lien entre la censure et évènement.

Soit V(t) les covariables qui prédisent le processus d'évènement au temps t et soit $\{\bar{V}(t) = V(x); 0 \leq x \leq t\}$. Le risque de censure à partir de l'état h au temps t ne dépend plus du possible temps d'évènement non observé T_h, où l'évènement est une transition à partir de l'état h, ce qui peut aussi s'écrire :

$$\lambda_{hC}(t|\bar{V}(t), T_h, T_h > t) \quad = \quad \lambda_{hC}(t|\bar{V}(t), T_h > t), \quad h = \{1, 2, 3\}. \tag{2.24}$$

Chapitre 3

Fonction de risque conditionnelle

Dans de nombreuses situations pratiques, on peut disposer d'une variable explicative X et la question devient celle de l'estimation du taux de hasard conditionnel.

L'objectif de ce chapitre est d'étudier un modèle de hasard conditionnel dans lequel la variable explicative X n'est pas nécessairement réelle ou multidimensionnelle mais seulement supposée être à valeurs dans un espace abstrait F muni d'une semi-métrique d. Comme dans tout problème d'estimation non-paramétrique, la dimension de l'espace F joue un rôle important dans les propriétés de concentration de la variable X. Ainsi, lorsque cette dimension n'est pas nécessairement finie, la fonction de risque conditionnelle est définie par

$$\lambda(t|x) \quad = \quad \lim_{\Delta t \to 0} \frac{P(T \leq t + \Delta t | T > t, X)}{\Delta t} \tag{3.1}$$

Qui s'écrit aussi à partir de la densité conditionnelle $f(.|x)$ et de la fonction de répartition conditionnelle $F(.|x)$ (au travers de la fonction de survie conditionnelle

$S(.|x) = 1 - F(.|x))$ de t sachant x, sous la forme

$$\lambda(t|x) \quad = \quad \frac{f(.|x)}{S(.|x)} \tag{3.2}$$

1. **Introduction de variables exogènes (covariables) :**

   L'estimation des fonctions de hasard doit a priori s'effectuer sur des populations homogènes. Si la population regroupe des catégories dont les lois de durées sont différentes, le risque est en effet de conclure faussement à une décroissance de la fonction de hasard. Le mécanisme qui mène à ce biais est connu sous le nom de "mover-stayer".

   Pour éviter ce risque de mauvaise interprétation, il est possible de partager l'échantillon observé en sous-échantillons(ou strates) les plus homogènes possibles. Procéder ainsi

suppose qu'il reste dans chaque sous-échantillon suffisamment d'individus pour que l'estimateur conserve de bonnes propriétés asymptotiques. On peut aussi spécifier une forme paramétrique particulière dans laquelle les paramètres s'expriment en fonction de variables exogènes (covariables).
Il existe plusieurs catégories de familles paramétriques qui permettent de procéder ainsi. Les plus courantes sont les familles à hasard proportionnel et les familles à hasard accéléré.

Dans les familles à hasard proportionnel, la fonction de hasard a la forme générale donnée précédemment (chapitre I,1.7.2). La fonction de hasard de base peut être estimé par la méthode du maximum de vraisemblance en spécifiant une forme paramétrique particulière, ou bien par une méthode non paramétrique (on parle alors d'une estimation semi-paramétrique pour λ ou l'intensité α (voir le détail d'une méthode : modèle de Cox)).
Dans les familles à hasard accéléré, la fonction de hasard a pour forme générale :

$$\lambda(t, X, \beta) = \lambda_0[t \exp(X\beta)] \exp(X\beta).$$

Les variables exogènes ont alors un effet de paramètre d'échelle sur les durées : tout se passe comme si la durée T d'un individu de la "catégorie" X s'écrivait $T_0 \exp(-X\beta)$, où, T_0 serait la durée de vie de la catégorie de référence. Tout se passe donc comme si le temps avançait plus ou moins rapidement pour les différents types d'individus. Cette écriture permet d'écrire simplement les modèles à durée de vie accélérée sous la forme :

$$\log T = -X\beta + \log T_0.$$

Cette écriture peut faire penser à un modèle de régression linéaire, où, $\log T_0$ jouerait le rôle de la perturbation.

2. **Les données censurées :**

   Une des particularités les plus fréquentes des données de durée est qu'elles sont rarement parfaitement observées. La période d'observation est en effet souvent trop courte pour mesurer les durées les plus longues. On parle alors d'observations censurées. Le type de censure le plus fréquent est ainsi la "censure à droits".

Il existe différents types de censure (voir chapitre II). Il est en général assez simple de tenir compte de la censure si elle intervient de manière indépendante du mécanisme de sortie, c'est-à-dire si la loi des durées censurées est bien la même que celle des durées non censurées.

3. **Vraisemblance dans un modèle de survie censuré avec covariables**

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

Notons $f_{Z\backslash X;\theta}(.)$, $F_{Z\backslash X;\theta}(.)$, $S_{Z\backslash X;\theta}(.)$, $\lambda_{Z\backslash X;\theta}(.)$, $\Lambda_{Z\backslash X;\theta}(.)$, représentent respectivement la densité, fonction de répartition, fonction de survie, fonction de risque instantanée et fonction de risque cumulée de la variable durée de vie Z, sous la loi $P_{\theta,X}$. On suppose que la loi de X, de densité f_X, ne dépend pas du paramètre θ. De la même façon que dans le cas où, les covariables n'interviennent pas, on considère que la loi de Z conditionnelle à X est indépendante de la loi de C conditionnelle à X, et que la loi de C est non informative pour le paramètre θ. Avec les mêmes notations que précédemment, la vraisemblance associée à l'échantillon $(T_i, \delta_i, X_i)_{i\in\{1,\dots,n\}})$ s'écrit, grâce à la formule de Bayes, sous la forme

$$L_n(\theta) = \Pi_{i=1}^n (f_{Z\backslash X;\theta}(T_i) S_{C\backslash X}(T_i))^{\delta_i} (S_{Z\backslash X;\theta}(T_i) f_{C\backslash X}(T_i))^{1-\delta_i} f_X(X_i)$$

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

$$\Pi_{i=1}^n \lambda_{Z\backslash X;\theta}(T_i)^{\delta_i} S_{Z\backslash X;\theta}(T_i)$$

Lorsque la fonction de hasard de base n'est pas spécifiée sous une forme paramétrique, le modèle est semi-paramétrique. La fonction de hasard s'écrit alors :

$$\lambda(t|x) = \lambda_0(t).\exp(\acute{x}\beta), \forall t$$

Le modèle à hasard proportionnel correspondant est appelé modèle de Cox.

Le modèle de Cox est un modèle classique en analyse de survie et a été largement étudié. Ici, on se concentre sur son utilisation dans le cadre de la mesure du risque d'estimation.

3.1 Estimation semi-paramétrique : le modèle de Cox

Elle concerne les modèles à hasard proportionnels présentés précédemment et a été introduit par Cox (1972). Il a été décrit par la formulation de la fonction de risque instantané de la donnée de survie. Le modèle de régression de Cox est un des modèles les plus utilisés pour la modélisation de l'influence de covariables sur des données de survie (Cox et Oakes, 1984...). Il définit une famille de lois conditionnelles de la donnée de survie, sachant un vecteur de variables explicatives.

Notons T^0 la variable aléatoire durée de vie du modèle et $\lambda_{T^0|X}$ sa fonction de risque instantanée connaissant X, un p-vecteur de covariables. Alors le modèle de régression de Cox est défini par la relation suivante pour tout $t \in R^+$:

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

Où $\beta_0 \in R^p$ est un paramètre de régression vectoriel et λ_0 la fonction de risque instantané de base définie sur R^+.

En effet, notons X_1 et X_2 deux vecteurs de covariables indépendantes du temps, alors le rapport des risques instantanés est constant par rapport à t :

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

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

Dans cette configuration, on voit que le modèle de Cox comprend différents modèles paramétriques usuels, comme le modèle exponentiel et le modèle de Weibull (obtenus avec les fonctions de risque instantané respectives $\lambda_0 = \lambda$ et $\lambda_0(t) = \lambda\alpha t^{\alpha-1}$, la fonction de risque

cumulé conditionnelle aux covariables X pour $t \in R^+$ définie par $\Lambda_{T^0|X}(t) = \int_0^t \lambda_{T^0|X}(s)ds$ elle peut s'écrire dans le modèle de Cox

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

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

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

On introduit une variable de censure C à valeurs positives indépendante de la durée de vie T_0 conditionnellement à la variable explicative X. Notons $f_{C|X}$ la fonction densité et $S_{C|X}$ la fonction de survie de la variable censure conditionnelles à X. Nous nous intéressons au cas de la censure à droite :

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

Le problème statistique consiste à estimer les paramètres inconnus du modèle à partir de l'observation de n-échantillon $(T_i, \Delta_i, X_i)_{1 \leq i \leq n}$. Le paramètre d'intérêt du modèle est le vecteur β permettant d'expliquer la nature de l'influence des covariables.

3.1.1 Méthode du maximum de vraisemblance partielle

Soit X un vecteur aléatoire de densité $f_X(x, \theta)$, où, θ est un vecteur paramètre (ϕ, β). Dans certains problèmes, il suffira de maximiser la vraisemblance en $\theta = (\phi, \beta)$ conjointement et d'utiliser la partie appropriée de la matrice de covariance complète de l'estimateur du maximum de vraisemblance θ pour l'inférence sur β.

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

$$f_{X;\theta}(x) \quad = \quad f_{W|V;\theta}(w|v) f_{V;\theta}(v) \tag{3.3}$$

dans certains modèles compliqués, un des facteurs du membre de droite de (3.3) peut ne pas contenir ϕ et être utilisé directement pour l'inférence sur β. L'autre facteur dépendant dans la majorité des cas de ϕ et β à la fois, de l'information sera perdue lors de l'utilisation d'une seule partie de la vraisemblance, mais le gain en simplicité est susceptible de

compenser une certaine perte d'efficacité.

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

$$\begin{aligned} f_{X;\theta}(x) &= f_{V_1,W_1,...,V_L,W_L;\theta}(v_1 w_1, ..., v_L, w_L) \\ &= \prod_{i=1}^{L} f_{W_l|V_1,W_1,...,V_l;\theta}(w_l|v_1 w_1, ..., v_l, w_l) \\ &\times \; f_{V_l|V_1,W_1,...,V_{l-1},W_{l-1};\theta}(w_l|v_1 w_1, ..., v_{l-1}, w_{l-1}) \\ &= (\prod_{l=1}^{L} f_{W_l|Q;\theta}(w_l|q_l))(\prod_{l=1}^{L} f_{V_l|P_l;\theta}(w_l|p_l)) \end{aligned} \tag{3.4}$$

$P_1 = \emptyset, Q_1 = V_1$ et pour $l \in \{2, ..., L\}$, $P_l = (V_1, W_1, ..., V_{l-1}, W_{l-1})$ et $Q_l = (V_1, W_1, ..., W_{l-1}, V_{l-1})$.

Quand le premier terme du produit de (3.4) ne dépend que de β, il est appelé vraisemblance partielle pour β basée sur W.

La vraisemblance pour l'estimation de $\theta_0 = (\beta_0, \Lambda_0$ relative au n-échantillon (T_i, Δ_i, X_i) égale à

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

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

$$L_n(\theta) = \prod_{i=1}^{n} \lambda(T_i)^{\Delta_i} e^{\Delta_i \beta' X_i} \exp(-e^{\beta' X_i} \Lambda(T_i)) \tag{3.5}$$

L'estimation du paramètre fonctionnel Λ_0 pose problème, pour cela Cox a donc proposé d'estimer le paramètre fini-dimensionnel β_0 à partir d'une vraisemblance partielle obtenue grâce au principe décrit précédemment.

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

Reprenons le cas où l'on a ordonné les valeurs des L durées différentes observées : $T_1 < ... < T_L$ et où, il n'y a pas de censeure ainsi réordonnées et $(X_1, ..., X_L)$ les covariables associées. Soit m_l le nombre de censures intervenant dans l'intervalle $[T_l, T_{l+1}[$.

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

Cox propose d'ignorer, dans le cadre de l'hypothèse de censure non informative et de l'hypothèse sur la loi de X, le terme $\prod_{i=1}^{n} f_{v_i|p_i;\theta}(w_i|p_i)$ apportant peu d'information sur β et donc de baser l'inférence pour β sur la vraisemblance partielle

$$\prod_{l=1}^{L} \mathbb{P}(W_l = l|Q_l, (X_k)_k, \beta) \quad = \quad \prod_{i=1}^{n} \frac{e^{\beta' X_i}}{\sum_{j=1}^{n} e^{\beta' X_j} 1_{T_i \leq T_j}} \tag{3.6}$$

S'il n'y a pas de censeure, elle s'interprète comme la vraisemblance de la statistique de rang associée aux durées. L'estimateur semi-paramétrique de β va être obtenu en maximisant la log-vraisemblance partielle par rapport à β au moyen d'une méthode itérative.

L'estimateur obtenu converge presque surement vers β et est asymptotiquement normal.

3.1.2 Estimation de la fonction de risque cumulé de base Λ_0

La maximisation de la vraisemblance partielle (3.6) ne permet pas d'estimer la fonction de risque cumulé de base, puisque Λ n'apparaît pas dans la formule de la vraisemblance partielle. Plusieurs méthodes ont été proposées pour l'estimation de Λ_0, la plus souvent retenue étant celle proposée par Breslow (1972,1974) généralisant l'estimateur de Nelson. Si le paramètre de régression β_0 a été estimé par $\widehat{\beta_n}$, Breslow propose d'estimer Λ_0 par $\widehat{\Lambda_n}$ donné par

$$\widehat{\Lambda_n}(t) \quad = \quad \sum_{i=1}^{n} \frac{\Delta_i 1_{T_i \leq t}}{\sum_{j=1}^{n} e^{\widehat{\beta_n}' X_j} 1_{T_i \leq T_j}} = \int_0^t \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} \exp(\widehat{\beta_n}' X_j) Y_j(s)} dN_i(s). \tag{3.7}$$

Pour cela, on prendre comme estimateurs de Λ_0 les fonctions en escalier ayant des sauts aux instant T_i non censurés, c'est-à-dire tels que $\Delta_i = 1$. Il s'agit donc de remplacer dans la vraisemblance (3.5) la fonction Λ par une telle fonction et de remplacer les valeurs $\lambda(T_i)$ par les sauts de cette fonction en T_i, quand $\Delta_i = 1$, puis de maximiser cette vraisemblance en

dimension finie : les paramètres de l'estimation du modèle sont le paramètre de régression β et les valeurs des sauts de la fonction en escalier estimant Λ_0 aux instants T_i tels que $\Delta_i = 1$. Il se vérifie que les estimateurs obtenus par cette méthode sont l'estimateur du maximum de vraisemblance partielle (3.6) pour β et l'estimateur de Breslow (3.7) pour Λ.

Chapitre 4

Etude des transitions

Le but de ce chapitre est de montrer les liens qui existent entre la théorie des fonctions de risque et des processus stochastiques. Les processus stochastiques utilisés sont des processus possédant la propriété de Markov.

Ce chapitre est consacré à la modélisation paramétrique des temps de séjour dans chaque état du modèle défini précédemment. L'observation des deux états sera considérée comme un modèle de survie classique. Nous rappelons ici quelques concepts de base de l'analyse de survie. Mais pour plus de détail (voir chapitre 1, ou [48],[23],[60],[8]...)

Cas d'un modèle semi-Markovien homogène

4.1 Estimation paramétrique des temps de séjour

4.1.1 Généralité

L'estimation paramétrique repose sur une estimation des lois des temps de séjour par des fonctions paramétriques. Rappelons la définition des fonctions de risque des temps d'attente dans les états,

$$\begin{aligned}\tilde{\alpha}_{ij}(d) &= \lim_{\Delta d \to 0} \frac{1}{\Delta d} Pr(d < X_{n+1} \leq d + \Delta d | X_{n+1} > d, J_n = i, J_{n+1} = j) \\ &= \begin{cases} \alpha_{ij}(d) & \text{si } J_n = i \text{ et } X_{n+1} > d, \\ 0 & \text{sinon} \end{cases}\end{aligned}$$

L'estimation paramétrique va supposer que les fonctions de risque $\alpha_{ij}(.)$ appartiennent à une classe de fonctions paramétriques. Les fonctions $S_{ij}(.)$ et $f_{ij}(.)$ correspondantes respectivement aux fonctions de survie et de densité associées aux fonctions de risque $\alpha_{ij}(.)$

peuvent s'écrire à partir de $\alpha_{ij}(.)$.

$$\begin{aligned}
\frac{\partial S_{ij}(d)}{\partial d} &= -\lim_{\Delta d\to 0}\frac{1}{\Delta d}Pr(d < X_{n+1} \leq d+\Delta d | J_n = i, J_{n+1} = j) \\
&= -\lim_{\Delta d\to 0}\frac{1}{\Delta d}Pr(d < X_{n+1} \leq d+\Delta d | X_{n+1} > d, J_n = i, J_{n+1} = j) \\
&\times \; Pr(X_{n+1} > d | J_n = i, J_{n+1} = j) \\
&= -S_{ij}(d) \times -\lim_{\Delta d\to 0}\frac{1}{\Delta d}Pr(d < X_{n+1} \leq d+\Delta d | J_n = i, J_{n+1} = j) \\
&= -S_{ij}(d) \times \alpha_{ij}(d).
\end{aligned}$$

La résolution de cette équation sachant que $S_{ij}(0) = 1$, donne

$$S_{ij}(d) = \exp(-\int_0^d \alpha_{ij}(u)du). \tag{4.1}$$

Comme $S_{ij}(.) = 1 - F_{ij}(.)$ et comme f_{ij} est la densité de $F_{ij}(.)$, on peut écrire

$$f_{ij}(d) = -\frac{\partial S_{ij}(d)}{\partial d} = S_{ij}(d)\alpha_{ij}(d). \tag{4.2}$$

4.1.2 Modèle semi paramétrique

Les modèles semi paramétriques sont des modèles qui permettent d'étudier les effets des covariables sur la distribution de T. L'avantage avec ces modèles est qu'on n'introduit pas d'hypothèses sur les fonctions de densité ou de risque instantané, mais font des hypothèses sur la manière dont les covariables vont influencer le déroulement du phénomène temporel.

Dans les modèles semi paramétriques, on distingue habituellement deux classes, les modèles à temps accélérés et les modèles à risque proportionnel. Les modèles à risque proportionnel expriment les effets multiplicatifs des covariables sur la fonction de risque. Cependant, les modèles à temps accélérés expriment des effets dilatoires des covariables sur le temps. Pour la suite, nous utiliserons les modèles à risque proportionnel et plus précisément le modèle de "Cox" car il est d'usage simple et efficace.

4.2 Modèle à risque proportionnel

Les modèles à risque proportionnel expriment les effets multiplicatifs des covariables sur la fonction de risque.

En effet, l'utilisation de covariables permet de prendre en compte l'hétérogénéité de la population et d'obtenir des résultats adaptés aux caractéristiques des patients. L'utilisation de cette méthode paramétrique permet d'incorporer des covariables dans la modélisation

des fonctions de risque des temps d'attente. Nous utiliserons à cet effet, un modèle à risques proportionnels (Cox [1972]).

Considérons $(J_n, S_n)_{n>0}$ un processus semi-Markovien. Les covariables sont introduites dans les fonctions de risque des temps d'attente dans les états. La chaîne de Markov $(J_n)_{n>0}$ ne dépend pas du vecteur de covariables, et ainsi, la chaîne conserve la probabilité de transition $p_{ij} = Pr(J_{n+1} = j | J_n = i)$. Les fonctions de risque du processus semi-Markovien ne dépendent pas des covariables. Les fonction d'intensité $\tilde{\alpha}_{ij}(t)$ pour un processus de comptage $N(t)$ peuvent s'écrire

$$\tilde{\alpha}_{ij}(t - S_{N(t^-)}) \quad = \quad 1_{\{J_{N(t^-)}=i\}} \tilde{\alpha}_{ij}(t - S_{N(t^-)}).$$

Où, t est le temps

$X_{ij}(t)$ est le vecteur des covariables associé à la transition de l'état i vers l'état j. Dans ce qui suit les covariables sont fixées au cours du temps : $X_{ij}(t) = X_{ij}$. On suppose que toutes les covariables respectent l'hypothèse de la proportionnalité. Pour chaque transition, on inclut les covariables par un modèle de Cox [18] de la manière suivante :

$$\tilde{\alpha}_{ij}(t - S_{N(t^-)}) \quad = \quad 1_{\{J_{N(t^-)}=i\}} \alpha_{ij,0}(t - S_{N(t^-)}) \exp(\beta_{ij}^T X_{ij}), \tag{4.3}$$

Avec β_{ij} le vecteur des coefficients de régression associés à X_{ij} et $\alpha_{ij,0}(.)$ est le risque de base.

$\forall i, j \in E$, $\alpha_{ij}(d, X) = \alpha_{ij,0}(d) e^{\beta_{ij}^T X_{ij}}$, d'après les équations(2)et(3) la fonction de survie correspondantes donnée par

$$\begin{aligned} S_{ij}(d, X) &= \exp(-\int_0^d \alpha_{ij}(u) du) \\ &= \exp(-\int_0^d \alpha_{ij}(u) e^{\beta_{ij}^T X_{ij}} du) \\ &= S_{ij,0}(d)^{e^{\beta_{ij}^T X_{ij}}} \end{aligned} \tag{4.4}$$

où, $S_{ij,0}(d) = \exp(-\int_0^d \alpha_{ij,0}(u) du)$ et la fonction densité

$$\begin{aligned} f_{ij}(d, X) &= S_{ij}(d, X) \alpha_{ij}(d, X) \\ &= \alpha_{ij,0}(d) e^{\beta_{ij}^T X_{ij}} S_{ij,0}(d)^{e^{\beta_{ij}^T X_{ij}}} \end{aligned} \tag{4.5}$$

4.3 Modélisation paramétrique

La modélisation paramétrique consiste à estimer les fonctions de risque des temps d'attente dans les états par des fonctions paramétriques. On considère que $\alpha_{ij} = h_{ij}(t, \theta)$, où h_{ij}

fonction paramétrique. L'estimation de $\alpha_{ij}(.)$ consiste à estimer le vecteur de paramètres θ_{ij}. D'aprés la vraisemblance totale et la fonction de survie en peut écrire

$$L = \prod_{h=1}^{n} \left\{ \prod_{k}^{N_h} PJ_{h,k-1}J_{h,k} f J_{h,k-1}J_{h,k}(X_{h,k}) \times \left[\sum_{j=1}^{s} PJ_{h,N_h} j S_{J_{h,N_h}} j(U_h) \right]^{\delta_h} \right\} \quad (4.6)$$

À partir des équations présenté précédemment, la vraisemblance (4.6) peut s'écrire en fonction des paramètres p_{ij} et θ_{ij}. L'estimation des paramètres se fait ensuite par maximisation de la vraisemblance. On obtient ainsi les estimations $\hat{p}_{ij}$ des probabilités de la chaîne de Markov et les estimations $\hat{\alpha}_{ij}(.)$ des fonctions de risque des temps de séjour. On en déduit les estimateurs $\hat{f}_{ij}()$ et $\hat{S}_{ij}(.)$ respectivement des fonctions de densité et de survie.

Donc, il est possible de déduire les estimateurs $\hat{\lambda}_{ij}(.)$ des intensités du processus semi-Markovien par la formule suivante

$$\hat{\lambda}_{ij}(.) = \frac{\hat{p}_{ij}\hat{f}_{ij}(d)}{\sum_{j=1}^{s} \hat{p}_{ij}\hat{S}_{ij}(d)} \quad (4.7)$$

En réalité cette quantité est intéressante parce qu'elle représente la probabilité instantané de changé d'état au temps $t + \Delta t$ sachant qu'on est resté pendant un temps t dans l'état avant transition, sans subir l'évènement d'intérêt. On a vu que dans le modèle de Markov homogène cette quantité appelée aussi " taux de transition " ne dépendait pas du temps, cependant, dans le modèle semi-Markov homogène elle dépend du temps.

Remarque : Le modèle par lequel on introduit les covariables est intéressant car les coefficients sont interprétés comme des risques relatifs

Dans les études de survie, les familles de fonctions les plus couramment utilisées pour modéliser les risques sont les lois exponentielles, les lois de Weibull et les lois de Weibull généralisées. Dans notre cas nous utiliserons loi de weibull qui généralise la loi exponentielle.

-Extension d'une loi de Weibull

Pour la modélisation des données de survie, la loi de Weibull permet de prendre en compte une évolution monotone du risque instantané au cours du temps. On utilise la loi de Weibull pour modéliser le risque alors,

$\forall i, j \in E, i \neq j$,

$$\alpha_{ij}(d) = \nu_{ij} \left(\frac{1}{\sigma_{ij}} \right)^{\nu_{ij}} d^{\nu_{ij}-1}, \forall d \geq 0, \forall \nu_{ij} > 0, \forall \sigma_{ij} > 0.$$

Pour $\nu_{ij} = 1$, on obtient la distribution exponentielle.

La fonction de risque avec covariables d'un modèle à risque proportionnel s'écrit,

$$\alpha_{ij}(d, X) = \nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1}\exp(\beta_T^{ij} X_{ij}).$$

Et la fonction de survie est

$$\begin{aligned} S_{ij}(d, X) &= S_{ij}(d)^{e^{\beta_{ij}^T X_{ij}}} \\ &= \left[\exp(-\int_0^d \nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} u^{\nu_{ij}-1} du)\right]^{e^{\beta_{ij}^T X_{ij}}} \\ &= \left[\exp(-(\frac{d}{\sigma_{ij}})^{\nu_{ij}})\right]^{e^{\beta_{ij}^T X_{ij}}} \end{aligned}$$

Où, $S_{ij}(d)$ la fonction de survie associée à une loi de Weibull sans covariable. D'après (4.5), la densité correspondante est

$$\begin{aligned} f_{ij}(d, X) &= S_{ij}(d, X)\alpha_{ij}(d, X) \\ &= \left[\exp(-(\frac{d}{\sigma_{ij}})^{\nu_{ij}})\right] \nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1}\exp(\beta_{ij}^T X_{ij}). \end{aligned}$$

***Remarque* 4.3.1.** Si $\nu_{ij} = 1$, on retrouve les fonctions associées à la loi exponentielle avec covariables :

$$\alpha_{ij}(d, x) = \frac{1}{\sigma_{ij}}\exp(\beta_{ij}^T X_{ij})$$

$$S_{ij}(d, x) = \exp(-\frac{d}{\sigma_{ij}})^{e^{\beta_{ij}^T X_{ij}}}$$

$$f_{ij}(d, x) = \frac{1}{\sigma_{ij}}\exp(\beta_{ij}^T X_{ij})\exp(-\frac{d}{\sigma_{ij}})^{e^{\beta_{ij}^T X_{ij}}}$$

4.4 Cas d'un modèle de Markov non-homogène

Nous faisons l'hypothèse d'un modèle de Markov non-homogène. Des estimateurs des intensités de transition ont été proposés par Aalen(Aalen,1978 ; Aalen, jhoansen,1978 ; Andersen, Borgan, Gill, Keiding. 1993).

4.4.1 Estimation non-paramétrique

On considère l'échantillon $X_1(.), ..., X_n(.)$de processus de Markov indépendants et identiquement distribués à espace d'états fini $S = \{1, ..., k\}$. Le processus $X_i(.)$ associé à l'individu i représente l'état de l'individu au temps t. On pose $\mathcal{T} = [0, \tau]$, où τ est la date de point. On définit également, pour $i \in \{1, ..., n\}$

– Les processus de comptage $N_{hji}(t)$ qui comptent le nombre de transitions de l'état h vers l'état j dans $[0, t]$ pour l'individu i,

$$N_{hji}(t) = card\{s \leq t : X_i(s^-) = h, X_i(s) = j\}, \quad \forall h, j = 1, ..., k, h \neq j$$

-$Y_{hi}(t)$ qui est un indicateur pour savoir si X_i est dans l'état h juste avant le temps t,

$$Y_{hi}(t) = \mathbf{1}_{\{X_i(t-)=h\}}, h = 1, ..., k$$

– Le processus de comptage $N_{hj+}(t)$ qui compte le nombre total de transitions de l'état h vers l'état j dans $[0, t]$ (pour toute la population),

$$N_{hj+}(t) = \sum_{i=1}^{n} N_{hji}(t), \forall h, j = 1, ..., k, h \neq j.$$

–$Y_{h+}(t)$ renseigne sur le nombre total de personne « à risque » dans l'état h juste avant l'instant t,

$$Y_{h^+}(t) = \sum_{i=1}^{n} Y_{hi}(t), h = 1, ..., k.$$

Proposition 4.4.1. *Le processus $N_{hji}(t)$ satisfait aux conditions d'un modèle à intensité multiplicative, i.e. $\forall i = 1, ..., n; \forall h, j = 1, ..., k, h \neq j$,*

$$\lambda_{hij}(t) \quad = \quad \alpha_{hji}(t) Y_{hi}(t)$$

La population étant supposée homogène, $\alpha_{hji}(t) = \alpha_{hj}(t)$ pour tout i, ainsi

$$\lambda_{hij}(t) \quad = \quad \alpha_{hj}(t) Y_{hi}(t) \tag{4.8}$$

La vraisemblance associée au processus de comptage $N = \{N_{hji}(t), i = 1, ..., n; h, j \in S, h \neq j\}$ conditionnellement aux données initiales

$$
\begin{aligned}
l &= p_{t\in T}\left\{\prod_i\prod_{h\neq j}(\alpha_{hj}(t)Y_{hi}(t))^{\Delta N_{hji}(t)}\left(1-\sum_i\sum_{h\neq j}\alpha_{hj}(t)Y_{hi}(t)dt\right)^{1-\sum_i\sum_{h\neq j}\Delta N_{hji}(t)}\right\}\\
&= p_{t\in T}\left\{\prod_{h\neq j}(\alpha_{hj}(t))^{\Delta N_{hj+}(t)}\left(1-\sum_{h\neq j}\alpha_{hj}(t)Y_{h+}(t)dt\right)^{1-\sum_{h\neq j}\Delta N_{hj+}(t)}\right\}.
\end{aligned}
$$

$N_{hj+}(t)$ est un processus de comptage ayant pour intensité

$$\lambda_{hj}(t)=\alpha_{hj}(t)Y_{h+}(t), h\neq j.$$

-Estimation des intensités cumulées

Un estimateur des intensités cumulées $A_{hj}(t)=\int_0^t\alpha_{hj}(u)du$ est obtenu par Nelson en 1972 pour des données censurées et par Aalen en 1978 dans le cadre des processus de comptage.

Définition 4.4.1. L'estimateur de Nelson-Aalen des fonctions d'intensité cumulée est défini par

$$\hat{A}_{hj}(t) = \int_0^t\frac{J_h(u)}{Y_{h+}(u)}dN_{hj+}(u), \forall h\neq j, \tag{4.9}$$

où $J_h(t)=\mathbf{1}_{\{Y_{h+}(t)>0\}}$.

Notons quand $Y_{h+}(s)=0$, $J_h(s)\backslash Y_{h+}(s)$ est interprété comme étant 0

Proposition 4.4.2. *Un estimateur de la variance de $\hat{A}_{hj}(t)$ est*

$$\sigma^2_{hj}=\int_0^t\frac{J_h(u)}{(Y_{h+}(u))^2}dN_{hj+}(u), \forall h\neq j.$$

Proposition 4.4.3. *(Andersen et al.1993), $\hat{A}_{hj}(t)$ est un estimateur*

– *Biaisé, tel que*

$$E\left(\hat{A}_{hj}(t)\right)-A_{hj}(t) = -\int_0^t\alpha_{hj}(u)P(Y_{h+}(u)=0)du$$

– Uniformément consistant,

– Et asymptotiquement normal, tel que

$$\left(\sqrt{n}\left(\hat{A}_{hj}(t)-A_{hj}(t)\right);h\neq j\right)\xrightarrow{l}(U_{hj};h\neq j),$$

4.4.2 Estimation semi-paramétrique

Cette section, présente un modèle de régression permettant d'ajuster les intensités de transition en fonction de la valeur des covariables. Les estimateurs de Nelson-Aalen et de Aalen-Johansen peuvent être étendus au cas d'un modèle où chaque intensité de transition suit un modèle de régression à intensités proportionnelles de type Cox. La méthodologie de la vraisemblance partielle de Cox permet d'estimer l'effet des covariables dépendantes et indépendantes du temps.

Définitions et notations

On considère l'échantillon $X_1(.), ..., X_n(.)$ de processus de Markov indépendants et identiquement distribués à espace d'états fini $S = \{1, ..., k\}$. Soit τ la date de point. Soient également, pour $i \in \{1, ..., n\}$

– $N_{hji}(t)$ qui compte le nombre de transitions de l'état h vers l'état j dans $[0, t]$ pour l'individu i ;

– $Y_{hi}(t) = 1_{\{X_i(t-)=h\}}$, 1 si l'individu i est à risque dans l'état h juste avant le temps t, 0 sinon ;

– $Z_i = (Z_{1i}, ..., Z_{pi})$, le vecteur de covariables de dimension p.

– $N_{hj+}(t) = \sum_{i=1}^{n} N_{hji}(t)$ et $Y_{h+}(t) = \sum_{i=1}^{n} Y_{hi}(t)$; les processus agrégés.

–$\beta_{hj} = (\beta_{hj,1}, ..., \beta_{hj,p})$, le vecteur de dimension p des coefficients de régression ;

– $N = \{N_{hji}, h, j \in S, h \neq j; i = 1, ..., n\}$; le processus de comptage multivarié associé aux n individus ;

– $\lambda = \lambda_{hji}, \quad h, j \in S, h \neq j; i = 1, ..., n\}$, le processus d'intensité par rapport à la filtration $\mathcal{T}_t$.

Le modèle considéré se caractérise par une structure multiplicative des processus d'intensité individuelle

$$\lambda_{hij}(t) = Y_{hi}(t)\alpha_{hji}(t; Z_i),$$

où, α_{hji} spécifie la dépendance avec les covariables Z_i. De plus, il est supposé que les intensités de transition α_{hji} suivent un modèle semi-paramétrique à risque multiplicatif, c'est-à-dire,

$$\alpha_{hji}(t; Z_i) \quad = \quad \alpha_{hj0}(t) \exp(\beta_T^{hj} Z_i), \quad h, j \in S, \quad h \neq j, \quad i = 1, ..., n \qquad (4.10)$$

où, $\alpha_{hj0}(t)$ est l'intensité de transition de base associée à la transition de l'état h vers l'état j. Plus précisément, $\alpha_{hj0}(.)$ est la fonction de risque des sujets pour lesquels toutes les covariables explicatives sont nulles. Ce modèle est dit semi-paramétrique du fait de la

présence, dans la définition des intensités de transition, d'une partie paramétrique (la partie de régression $\exp(\beta_T^{hj} Z_i)$) et d'une partie non-paramétrique (le risque de base $\alpha_{hj0}(.)$).

De plus, le modèle (4.10) est dit à risques proportionnels car, par définition, quels que soient deux individus (1 et 2), le rapport des intensités de transition ne varie pas au cours du temps

$$\frac{\alpha_{hj1}(t)}{\alpha_{hj2}(t)} = \exp(\beta_T^{hj}(Z_1 - Z_2)) \tag{4.11}$$

Les intensités de transition sont donc proportionnelles.

Notons que tous les modèles à structure multiplicative, c'est-à-dire les modèles où les intensités de transition sont séparables en deux termes dont l'un dépend uniquement du temps et l'autre non (par exemple (4.10)), ont cette propriété. Ce sont des modèles à risques proportionnels. Dans ces modèles, le rapport des intensités de transition représente un risque relatif à l'instant t des sujets de caractéristique Z_1 par rapport aux sujets de caractéristique Z_2.

Le logarithme de l'intensité de transition est une fonction linéaire de Z_i,

$$\log \alpha_{hji}(t) - \log \alpha_{hj0}(t) = \beta_T^{hj} Z_i.$$

En effet, l'hypothèse de log-linéarité suppose que le risque relatif est constant pour une augmentation d'une unité quelle que soit la valeur de la covariable explicative. C'est une hypothèse qu'il convient de vérifier ou tout au moins d'avoir à l'esprit quand on utilise ce modèle de régression. Par exemple, si l'on considère l'âge comme variable explicative continue et que l'on étudie une maladie qui touche essentiellement les personnes âgées,

Cette relation log-linéaire est souvent utilisée dans la littérature (Cox [1972],[16]), car elle permet d'avoir des intensités définies positives quelle que soit la valeur des coefficients de régression. De plus, les résultats obtenus sont bien connus des cliniciens et sont facilement interprétables.

Estimation des intensités de base

On suppose les fonctions $\alpha_{hj0}(.)$ sont positives et que $A_{hj0}(t) = \int_0^t \alpha_{hj0}(u)du < \infty, \forall h \neq j$.

Rappelons que la vraisemblance associée à un processus non censuré N^* est

$$l^*(\beta) = \prod_{t \leq \tau} \prod_{i=1}^{n} \prod_{h \neq j} \left\{ (dA_{hji}(t) Y_{hi}(t))^{\Delta N^*_{hji}(t)} \right\} \times \exp\left[- \sum_{h \neq} \sum_{i=1}^{n} \int_0^\tau Y_{hi}(t) dA_{hji}(t) \right].$$

La vraisemblance associée à un processus censuré (censure à droite indépendante) a une forme identique à la vraisemblance complète : Cette vraisemblance d'un processus censuré N est appelée la vraisemblance partielle. Dans le cadre d'un modèle semi-paramétrique multiplicatif, la vraisemblance partielle s'écrit

$$\begin{aligned} l(\beta) &= \prod_{t\leq\tau}\prod_{i=1}^{n}\prod_{h\neq j}\left\{(dA_{hj0}(t)Y_{hi}(t)\exp(\beta_T^{hj}Z_i))^{\Delta N_{hji}(t)}\right\} \\ &\times \exp\left[-\sum_{h\neq}\int_0^\tau S_{hj}^0(\beta,u)dA_{hj0}(u)\right] \end{aligned} \tag{4.12}$$

Avec

$$S_{hj}^0(\beta,t) = \sum_{i=1}^{n}\exp(\beta_{hj}^T Z_i)Y_{hi}(t).$$

La maximisation de la vraisemblance (4.12) par rapport à $\Delta A_{hj0}(.)$ conduit à

$$\Delta\hat{A}_{hj0}(t) = \frac{\Delta N_{hj+}(t)}{S_{hj}^0(\beta,t)}$$

Proposition 4.4.4. *Pour β fixé, $A_{hj0}(t) = \int_0^t \alpha_{hj0}(u)du$ est estimé par l'estimateur de Breslow (Breslow [1974]),*

$$\hat{A}_{hj0}(t) = \int_0^t \frac{J_h(u)}{\sum_{i=1}^n \exp(\beta_{hj}^T Z_i)Y_{hi}(t)} dN_{hj+}(u), \tag{4.13}$$

Avec $J_h(u) = 1_{\{Y_{h+}(t)>0\}}$.

4.4.3 Estimation des coefficients de régression

Dans (4.12), en remplaçant $A_{hj0}(t)$ par son estimation obtenue en (4.13), la vraisemblance partielle devient,

$$\begin{aligned} l(\beta) &= \prod_t \prod_i \prod_{h\neq j} \left\{ \left(d\hat{A}_{hj0}(t) Y_{hi}(t) \exp(\beta_{hj}^T Z_i) \right)^{\Delta N_{hji}(t)} \right\} \\ &\times \exp\left[-\sum_{h\neq j} \int_0^\tau S_{hj}^{(0)}(\beta, u) d\hat{A}_{hj0}(u) \right] \\ &= \prod_t \prod_i \prod_{h\neq j} \left\{ \left[\frac{Y_{hi}(t) \exp(\beta_{hj}^T Z_i)}{S_{hj}^{(0)}(\beta, t)} \right]^{\Delta N_{hji}(t)} [J_h(u) dN_{hj+}(u)]^{\Delta N_{hji}(t)} \right\} \\ &\times \exp\left[-\sum_{h\neq j} \int_0^\tau J_h(u) dN_{hj+}(u) \right] \\ &= l_{Cox}(\beta) \times \prod_t \prod_{i=1} \prod_{h\neq j} [J_h(u) dN_{hj+}(u)]^{\Delta N_{hji}(t)} \times \exp\left[-\sum_{h\neq j} \int_0^\tau J_h(u) dN_{hj+}(u) \right], \end{aligned}$$

Avec

$$l_{Cox}(\beta) = \prod_t \prod_{i=1} \prod_{h\neq j} \left[\frac{Y_{hi}(t) \exp(\beta_{hj}^T Z_i)}{S_{hj}^{(0)}(\beta, t)} \right]^{\Delta N_{hji}(t)} \tag{4.14}$$

Par définition, $l_{Cox}(\beta)$ est la vraisemblance partielle de Cox. Cette vraisemblance est introduite par Cox (1972).

Considérons la fonction de log-vraisemblance partielle de Cox,

$$\log l_{Cox}(\beta) = \sum_i \sum_{h\neq j} \int_0^\tau \left[\beta_{hj}^T Z_i - \log S_{hj}^{(0)}(\beta, t) \right] dN_{hji}(t).$$

Les vecteurs scores (dérivées de la Log-Vraisemblance par rapport à β_{hj}) sont donnés par

$$\begin{aligned} U_{hj}(\beta) &= \frac{\partial \log l_{Cox}(\beta)}{\partial \beta_{hj}} \\ &= \sum_i \int_0^\tau \left[Z_i - \frac{S_{hj}^{(1)}(\beta, t)}{S_{hj}^{(0)}(\beta, t)} \right] dN_{hji}(t), \end{aligned}$$

Avec

$$S_{hj}^{(1)}(\beta, t) = \frac{\partial S_{hj}(\beta, t)}{\partial S_{hj}} = \sum_{i=1}^n Y_{hi}(t) Z_i \exp(\beta_{hj}^T Z_i).$$

Proposition 4.4.5. *L'estimateur $\hat{\beta}$ du maximum de la vraisemblance partielle de Cox vérifie*

$$U_{hj}(\hat{\beta}) = 0.$$

De plus,

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightarrow^l N(0, nI^{-1}(\hat{\beta})),$$

Où $\mathcal{I}(\beta)$ est la matrice d'information de Fisher :

$$\begin{aligned} I(\beta) &= -\frac{\partial^2 \log l_{Cox}(\beta)}{\partial \beta^2} \\ &= \sum_{h \neq j} \int_0^\tau \frac{\left[S_{hj}^2(\beta,t) S_{hj}^0(\beta,t)\right] - \left[S_{hj}^{(1)}(\beta,t)\right]^2}{\left(S_{hj}^{(0)}(\beta,t)^2\right)} dN_{hj+}(t), \end{aligned}$$

Avec

$$S_{hj}^{(2)}(\beta,t) = \sum_{i=1}^{n} Y_{hj}(t)(Z_i)^2 \exp(\beta_{hj} Z_i).$$

4.5 Estimation des probabilités de transition

À partir des estimateurs donnée précédemment, la matrice des probabilités de transition s'obtient en suivant la démarche permettant d'obtenir l'estimateur de Aalen-Johansen.

Proposition 4.5.1. *Un estimateur de la matrice des probabilités de transition est donné par le produit intégral*

$$\hat{P}(s,t|Z_0) = \mathcal{P}_{u \in]s,t[} \left(Id + d\hat{A}(u|Z_0) \right), s \leq t \leq \tau,$$

Où, Z_0 valeur des covariables pour un individu. Avec $\hat{A} = \left\{ \hat{A}_{hj} \right\}_{hj}$,

$$\begin{aligned} d\hat{A}_{hj}(t|Z_0) &= d\hat{A}_{hj0}(t|Z_0) \exp(\hat{\beta}_{hj}^T) \\ &= \frac{J_h(t) \times \exp(\hat{\beta}_{hj}^T)}{\sum_{i=1}^n \exp(\hat{\beta}_{hj}^T) Y_{hi}(t)} \times \Delta N_{hj+}(t), h \neq j, \end{aligned}$$

et

$$d\hat{A}_{hh}(t|Z_0) = -\sum_{j \neq h} d\hat{A}_{hj}(t|Z_0), h = 1, ..., s$$

4.6 Cas particulier : données de survie

Le modèle de survie est un modèle à deux états où, une seule transition est possible, comme étant décrit dans la page 19. La méthodologie présentée précédemment comprend ainsi le cas particulier des données de survie.

En considérant l'espace d'états $\{0, 1\}$, la vraisemblance partielle de Cox (4.14) s'écrit

$$l_{Cox}(\beta) = \prod_t \prod_{i=1} \left[\frac{Y_{0i}(t) \exp(\beta_{01}^T Z_i)}{S_{01}^{(0)}(\beta, t)} \right]^{\Delta N_{01i}(t)},$$

Où, $N_{01i}(t)$ vaut 1 si l'individu i passe de l'état 0 à l'état 1 au temps t et 0 sinon, β_{01} représente le coefficient de régression associé à la transition vers l'état 1. Cette expression est bien la vraisemblance obtenue par Cox pour des données de survie. De même, l'estimateur (4.13) des intensités cumulées de base généralise l'estimateur de Breslow introduit en 1974 pour des données de survie,

$$\hat{A}_{010}(t) = \int_0^t \frac{J_0(u)}{\sum_{i=1}^n \exp(\beta_{01}^T Z_i) Y_{0i}(t)} dN_{01+}(u),$$

Où, $\hat{A}_{010}(t)$ représente l'intensité cumulée de base associée au « état 1 ». Le modèle de Cox à risques proportionnels qui est couramment utilisé représente ainsi un cas particulier de la méthodologie présentée précédemment.

Chapitre 5

Application

On applique dans ce chapitre quelques formules et méthodes citées dans les quatre chapitres précédents. L'objet de ce chapitre est de réanalyser un jeu de données. Pour illustrer notre travail, nous présentons dans ce chapitre un exemple d'application simple qui a été également utilisé pour détailler les travaux de [22] et [63]. Cette application repose sur des données pour un test clinique (la leucémie myélogène aiguë(AML)) et des données de vasculopathie de l'allogreffe cardiaque (CAV). Il faut discrétiser le problème au niveau des variables. Nous devons modéliser les fonctions de survie et de "risque instantané" de changement d'état pendant le temps à l'évènement.

Nous commençons par l'étude des données d'AML. La discrétisation du problème est citée dans l'exemple 1.

5.1 Exemple d'application.1

Les données présentées dans le tableau 5.1 sont des résultats préliminaires d'un test clinique pour évaluer l'efficacité de la chimiothérapie d'entretien pour la leucémie myélogène aiguë (AML).

L'étude a été entreprise par Embury et al.(1977) à l'Université de Stanford. Après avoir atteint un état de rémission grâce à un traitement par chimiothérapie, les patients impliquer par l'étude ont été assignées aléatoirement à deux groupes. Le premier groupe a reçu la chimiothérapie d'entretien, alors que le second non. L'objectif de l'épreuve était de voir si entretien chimiothérapique a prolongé le temps jusqu'à la rechute ou pas.

Tableau 5.1 : Les données de l'étude d'entretien AML. A + indique une valeur censuré.

Group	Longueur de rémission complète (en semaines)
Maintenu	9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+
Non maintenu	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

L'objectif ici est d'apprendre des méthodes pour modéliser et analyser des données de survie.

5.1.1 Méthodes non paramétriques

Nous commençons par des méthodes non paramétriques d'inférence concernant la fonction de survie.

$$S(t) = P(T > t)$$

et la fonction de risque cumulatif

$$\Lambda(t) \quad = \quad \int_0^t \lambda(u)du = -\log\left(S(t)\right) \tag{5.1}$$

-La valeur moyenne

$$E(T) \quad = \quad \int_0^\infty S(t)dt. \tag{5.2}$$

-L'objectif de cette section est de

1. Savoir calculer l'estimateur de survie de Kaplan-Meier .
2. Estimer la fonction de risque et la fonction de risque cumulatif.
3. Tracer la courbe de K-M.
4. Réaliser l'analyse non paramétrique de la fonction S en utilisant la commande survifit.

Estimateur de Kaplan-Meier de survie

La fonction de survie peut être exprimée comme :

$$S(t) = P(T > t) = \prod_{y(i)\leq t} p_i.$$

Les estimations de p_i et de q_i sont :

$\widehat{q_i} = \frac{d_i}{n_i}$ et $\widehat{p_i} = 1 - \widehat{q_i} = \left(\frac{n_i - d_i}{n_i}\right)$

-Donc l'estimateur de Kaplan-Meier de la fonction de survie est :

$$\widehat{S}(t) = \prod_{y(i)\leq t} \widehat{p_i} = \prod_{y(i)\leq t} \left(\frac{n_i - d_i}{n_i}\right) = \prod_{i=1}^{k} \left(\frac{n_i - d_i}{n_i}\right)$$

Où, $y(k) \leq t < y(k+1)$

Avec :

n_i=vivants(est non censurée)juste avant $y(i)$

d_i=mort au temps $y(i)$

p_i=p(survivre à travers I_i| vivant au début I_i) =$p(T > y(i)|T > y(i-1))$

$q_i = 1 - p_i$=p(mourir dans I_i| vivant au début I_i)

-L'estimation de la fonction de risque dans l'intervalle $t_i \leq t < t_{i+1}$:

$$\widehat{\lambda}(t) = \frac{d_i}{n_i(t_{i+1} - t_i)} \tag{5.3}$$

Ceci ressemble au type d'estimation de K-M. Il estime le taux de décès par unité de temps dans l'intervalle $[t_i; t_{i+1})$.

-Les estimations de $\Lambda(.)$, la fonction de risque cumulatif au temps t :

1.Construit avec K-M :

$$\widehat{\Lambda}(t) = -\log \prod_{y(i)\leq t} \left(\frac{n_i - d_i}{n_i}\right), \tag{5.4}$$

$$\widehat{var}(\widehat{\Lambda}(t)) = \sum_{y(i)\leq t} \frac{d_i}{n_{i(n_i - d_i)}} \tag{5.5}$$

2. Estimateur de Nelson-Aalen(1972, 1978) :

$$\widetilde{\Lambda}(t) = \sum_{y(i)\leq t} \frac{d_i}{n_i}, \tag{5.6}$$

Nous considérons ici les données d'AML présentées dans le Tableau 5.1

Sous le logiciel R, pour exécuter une analyse de survie, il est nécessaire d'installer la bibliothèque d'analyse de survie et des données. La commande de R est :

```
>library(survival)
>Data(aml)
```

Nous traitons d'abord ces données comme si il n'y avait aucune observation censurées. La fonction empirique de survie, dénotée par $S_n(t)$, est défini par :

$$S_n(t) = \frac{\{t_i > t\}}{n}$$

Où, t_i une valeur observée commandée

Les commandes (en R) pour trouver les résultats d'estimation sont données comme suite :

Programme 1

```
> Surv(aml$time,aml$status) # Objet de surv
> km.fit <- survfit(Surv(time,status)~1,type="kaplan-meier",
+ data=aml)
> km.fit
> summary(km.fit) # survival est l'estimation de S(t)
> H.hat <- -log(km.fit$surv);
> op <- par(mfrow=c(2,1))
> plot(km.fit,xlab="temps jusqu'à la rechute(en semaines)",
+ ylab="proportion sans rechute",col='red')
> mtext("Courbe de survie de K-M",3,line=-1,cex=0)
> plot(km.fit, lty=2:3, fun="cumhaz",
+ xlab="Temps",ylab=" Hazard cumulative")
> par(op)
> abline(h=0)
```

$S_n(t)$ est la proportion de patients toujours en rémission après t semaines. Les résultats obtenus d'après l'étude des donnée d'aml sont présentés comme suit

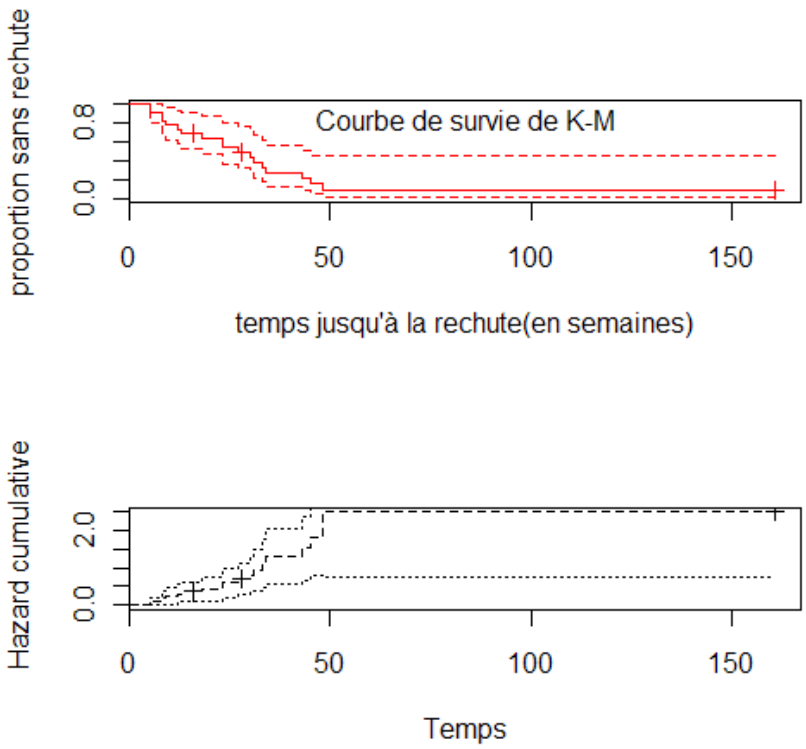

FIG. 5.1 – courbe des estimateurs de Kaplan-Meier

La Fig.5.1 représente l'estimateur de Kaplan-Meier de la fonction de survie et la fonction de hasard cumulatif.

-La mort est le mot générique pour l'évènement d'intérêt. Dans l'étude d'AML, une « rechute » (fin de période de rémission)= « mort »

-Une cohorte est un groupe de personnes qui sont suivies au cours de l'étude.

-Les personnes à risque au début de l'intervalle t_i sont les personnes qu'ont survécu (pas morte, perdu ou retiré)de l'intervalle précédent t_{i-1}.

La courbe KM est une fonction en escalier droit continu qui quitte seulement à une observation non censurée. Le + sur la courbe KM représente la probabilité de survie à un temps censuré. Notez que la courbe KM ne saute pas à zéro et le temps d'une survie plus grande (161 +) est censuré.

Nous ne pouvons pas estimer S(t) au-delà de t = 48. Certains se réfèrent $\hat{S}(t)$ en tant que fonction de survie défectueuse.

Programme 2

```
> km.fit <- survfit(Surv(time,status)~1,type="kaplan-meier",
+ data=aml)
```

```
> km.fit
> summary(km.fit) # survival est l'estimation de S(t)
> H.hat <- -log(km.fit$surv);
> t=km.fit$time
> for(i in 1:15){
+ a[i]=t[i+1]-t[i];
+ h.sort.of <- km.fit$n.event / (km.fit$n.risk *a);}
```

L'algorithme fournit les valeurs de $\widehat{\lambda}(t)|$ où, $\widehat{h}(t)$ est la fonction de risque estimé. Le Tableau 5.2 représente ces valeurs.

Tableau 5.2

Time	n.risk	n.event	$\widehat{S}(t)$	$\widehat{\lambda}(t)$	$\widehat{\Lambda}(t)$	lower 95% CI	upper 95% CI
5	23	2	0.9130	0.02898	0.09097178	0.8049	1.000
8	21	2	0.8261	0.09523	0.19105524	0.6848	0.996
9	19	1	0.7826	0.01754	0.24512246	0.6310	0.971
12	18	1	0.7391	0.05555	0.30228087	0.5798	0.942
13	17	1	0.6957	0.01961	0.36290549	0.5309	0.912
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

5.1.2 Méthodes paramétriques

-L'estimateur de maximum de vraisemblance (MLE)

– **Likelihood :**

$$\begin{aligned} L(\lambda) &= \prod_u f(y_i|\lambda).\prod_c S(y_i|\lambda) \\ &= \prod_u \lambda \exp(-\lambda y_i) \prod_c \exp(-\lambda y_i) \\ &= \lambda^{n_u} \exp(-\lambda \sum_u y_i) \exp(-\lambda \sum_c y_i) \\ &= \lambda^{n_u} \exp(-\lambda \sum_{i=1}^{n} y_i) \end{aligned}$$

– **Log-likelihood :**

$$\log L(\lambda) = n_u \log(\lambda) - \lambda \sum_{i=1}^{n} y_i$$

- **MLE :**
  $\widehat{\lambda} = \frac{n_u}{\sum_{i=1}^{n} y_i}$ et $var_a(\widehat{\lambda}) = \left(-E\left(-\frac{n_u}{\lambda^2}\right)\right)^{-1}$,
- Le MLE de la fonction de survie $S(t) = \exp(-\lambda(t))$:
  $\widehat{S}(t) = \exp(-\widehat{\lambda}(t))$

```
> library(survival)
> # Ajustement exponentiel(Exponential fit)
> attach(aml)
> exp.fit <- survreg(Surv(aml$time,aml$status)~1,dist="weib",
+ scale=1)
> exp.fit
Coefficients:
(Intercept)
   3.628776
Scale fixed at 1 Loglik(model)= -83.3  n= 23
```

l'Intercept =3.628776, ce qui égale $\widehat{\mu} = -\log(\widehat{\lambda}) = \log(\widehat{\theta})$. Les commandes suivantes produisent un C.I. à 95% pour la moyenne θ

```
> coeff <- exp.fit$coeff # muhat
> var <- exp.fit$var
> thetahat <- exp(coeff) # exp(muhat)
> thetahat
> C.I.mean1 <- c(thetahat,exp(coeff-1.96*sqrt(var)),
+ exp(coeff+1.96*sqrt(var)))
> names(C.I.mean1) <- c("mean1","LCL","UCL")
> C.I.mean1
```

Médiane estimé avec un C.I. à 95%

```
> medhat <- predict(exp.fit,type="uquantile",p=0.5,se.fit=T)
> medhat1 <- medhat$fit[1]
> medhat1.se <- medhat$se.fit[1]
> exp(medhat1)
> C.I.median1 <- c(exp(medhat1),exp(medhat1-1.96*medhat1.se),
+ exp(medhat1+1.96*medhat1.se))
> names(C.I.median1) <- c("median1","LCL","UCL")
> C.I.median1
```

Estimation de S(t)

```
> muhat <- exp.fit$coeff
> time.u <- time[status == 1]
> nu <- length(time.u)
> scalehat <- rep(exp(muhat),nu)
> Shat <- 1 - pweibull(time.u,1,scalehat)
> LCL <- exp(log(Shat)*exp(1.96/sqrt(nu)))
> UCL <- exp(log(Shat)*exp(-1.96/sqrt(nu)))
> C.I.Shat <- data.frame(time.u,Shat,LCL,UCL)
> round(C.I.Shat,5)
```

Les résultats obtenus d'après les commandes notées précédemment sont résumé dans ces Tableaux

Tableau.5.3 : Intervalles de confiance préférés à 95% pour la moyenne et la médiane d'un modèle de survie exponentiel.

Paramètre	Estimation	I.C.95%
Moyenne	37.66667	[23.73140 ,59.78483]
Mediane	26.10854	[16.44935, 41.43969]

$$\widehat{S}(t) = \exp(-\widehat{\lambda}(t) = \exp(-0.0265t).$$

Tableau 5.4 : résultats de l'estimation de S.

time.u	$\widehat{S}$	LCL	UCL
9	0.78746	0.69798	0.85318
13	0.70813	0.59489	0.79505
13	0.70813	0.59489	0.79505
18	0.62010	0.48718	0.72792
23	0.54301	0.39896	0.66646
.	.	.	.
.	.	.	.
.	.	.	.

5.2 Exemple d'application.2

L'objectif de cette section est donc d'estimer les intensités de transition entre les différents états.

Les données sont spécifiées comme une série d'observations, groupée par patient. Au minimum il devrait y avoir une forme de données avec des variables indiquant :

- Le temps de l'observation.
- L'état observé du processus.

Si les données ne contiennent pas le numéro d'identification du sujet, alors on assume que toutes les observations sont du même sujet.

L'identification du sujet n'est pas obligatoirement numérique, mais les données doivent être groupées par sujet. Un exemple d'ensemble de données, pris du contrôle d'un ensemble de receveur de greffe de cœur, est fourni par la commande msm. Sharples et al [56] ont étudié la progression de vasculopathie d'allogreffe cardiaque (CAV). Cet ensemble de données est disponible à la session courante de R avec la commande

```
> library("msm")
> data("cav")
> cav
> statetable.msm(state, PTNUM, data = cav)
```

La base de données est constituée de 614 individus, ce qui représente 2816 observations.

PTNUM est l'identificateur de sujet. Approximativement, tous les ans après une greffe cardiaque, chaque patient subit une angiographie, pour diagnostiquer la CAV. Le résultat de l'essai est dans l'ensemble $1, 2, 3, 4$ représentant respectivement cav-Absente, cav-Moyenne, cav-Grave et mort donne la période de l'essai(test) depuis la greffe du cœur.

D'autres variables incluent l'âge (âge à l'écran), le dage (âge de distributeur), le sexe (0=male, 1=femelle), le pdiag (diagnostic primaire, ou la raison de la greffe - IHD représente la maladie cardiaque ischémique, l'IDC représente la cardiomyopathie dilatée idiopathique), le cumrej (nombre cumulatif des épisodes de rejet), et firstobs, un indicateur qui égale 1 quand l'observation correspond à la greffe du patient (première observation), et 0 quand l'observation correspond à une angiographie postérieure.

Ainsi il y avait les 148 décès CAV-Absente, les 48 décès des décès de l'état 2, et 55 de l'état 3. À seulement quatre occasions, il y avait une observation de CAV grave suivie d'une observation sans CAV.

Le modèle est défini par la figure(5.2). La structure étudiée sera ainsi définie par quatre états transitives(voir fig5.2).

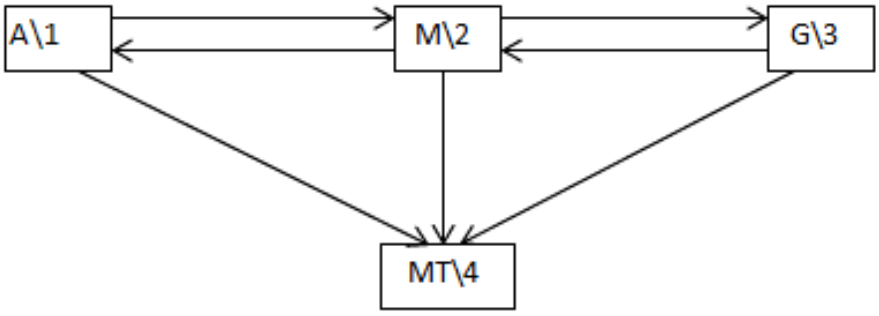

FIG. 5.2 – Modèle multi état représentant tous les parcours possibles d'un individu

Soit $X(t)$ l'état occupé par un individu au temps t, si les individus se déplacent parmi les état 1 à 4 selon un processus de Markov à temps continu $\{X(t), t \in t\}$ qui a comme états possibles $\{A, M, G, MT\}$. Alors les transitions entre les états sont traduites par les intensités de transition (hasard instantané) :

$$q_{rs} = \lim_{\Delta t \to 0^+} P[X(t + \Delta t) = s | X(t) = r]/\Delta t$$

L'objectif est de calculer la matrice Q des intensités de transition qui représente le risque instantané de déplacement de l'état r ver s :

$$Q = \begin{pmatrix} -(q_{12} + q_{14}) & q_{12} & 0 & q_{14} \\ q_{21} & -(q_{21} + q_{23} + q_{24}) & q_{23} & q_{24} \\ 0 & q_{32} & -(q_{32} + q_{34}) & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

5.2.1 Model 1 : Modèle sans covariable

On voit ici un modèle simple sans covariable, les commandes utilisées :

```
> twoway4.q <- rbind(c(0, 0.25, 0, 0.25), c(0.166, 0, 0.166, 0.166),
+ c(0, 0.25, 0, 0.25), c(0, 0, 0, 0))
> rownames(twoway4.q) <- colnames(twoway4.q) <- c("A","M","G","MT")
> cav.msm <-msm(state ~ years, subject = PTNUM, data = cav,
+  qmatrix = twoway4.q,death = 4)
> cav.msm
```

On a $q_{rr} = -\sum_{h \neq j} q_{rs} = -0.5$ pour r=1,2,3, et $q_{12} = q_{14} = q_{32} = q_{34} = 0.25$, $q_{21} = q_{23} = q_{24} = 0.166$, les résultat de l'estimation de maximum de vraisemblance(MLE) et les intervalles de confiance à 95% sont présentés dans la table suivante.

Tableau5.5 : Estimateur des intensités du Markov

passages	estimateur d'intensité	IC à95%
q_{11}	-0.1702	[-0.1901,-0.1524]
q_{12}	0.1277	[0.1112 , 0.1466]
q_{14}	0.0425	[0.0341 , 0.0530]
q_{21}	0.2244	[0.1670 , 0.3016]
q_{22}	-0.6062	[-0.7068,-0.5199]
q_{23}	0.3406	[0.2714 , 0.4273]
q_{24}	0.0412	[0.0119 , 0.1423]
q_{32}	0.1312	[0.0800 , 0.2152]
q_{33}	-0.4361	[-0.5517,-0.3447]
q_{34}	0.3049	[0.2368 , 0.3925]

De l'état acceptable, la vitesse de transition vers une CAV-moyenne est trois fois supérieure à celle vers la mort, mais ne transite jamais vers CAV-grave sans passer par CAV-moyenne.

De la matrice d'intensité estimée, nous voyons que les patients sont trois fois aussi probables pour développer des symptômes, que meurent sans symptômes (première rangée). Après début de la maladie (état 2), la progression vers les symptômes graves (l'état 3) est de 50% plus rapide que le rétablissement, et la mort de l'état CAV-grave est rapide.

5.2.2 Model 2 : Modèle de Markov avec covariable

Maintenant nous avons une matrice d'intensité $Q(z)$ qui dépend d'un vecteur de covariable z. Pour chaque entrée de $Q(z)$, l'intensité de transition pour l'individu i au temps d'observation j est :

$$q_{rs}(z_{ij}) = q_{rs}^{(0)} \exp(\beta_{rs}^T z_{ij}).$$

Nous considérons un modèle avec juste un covariable, sexe femelle. Hors des 622 receveurs de greffe, 535 sont masculins et 87 sont féminins. Par défaut, tous les effets linéaires de covariate β_{rs} sont initialisés à zéro.

```
> cavsex.msm <- msm(state ~ years, subject = PTNUM,
+ data = cav, qmatrix = twoway4.q, death = 4,
+ covariates = ~sex)
> cavsex.msm
> op<-par(mfrow = c(1,2))
> plot.survfit.msm(cav.msm, main = "cav.msm: sans covariables",
```

```
+ mark.time = FALSE)
> plot.survfit.msm(cavsex.msm, main = "cavsex.msm: avec covariables",
+ mark.time = FALSE)
> par(op)
```

Les résultats de l'estimation sont donnés comme suit :

Tableau.5.6 :

Passages	Estimateur d'intensité	$\hat{\beta}$	IC à95%
$q_{12}(z)$	0.1308	0.5338549	[0.1138, 0.1505)]
$q_{14}(z)$	0.04175	1.2387824	[0.03333, 0.05229]
$q_{21}(z)$	0.2429	0.9832775	[0.1817, 0.3248]
$q_{23}(z)$	0.3794	1.5641957	[0.3016, 0.4774]
$q_{24}(z)$	0.05876	1.7957093	[0.0251, 0.1375]
$q_{32}(z)$	0.1748	2.1707943	[0.1028, 0.2974]
$q_{34}(z)$	0.3065	1.9544936	[0.238 , 0.3947]

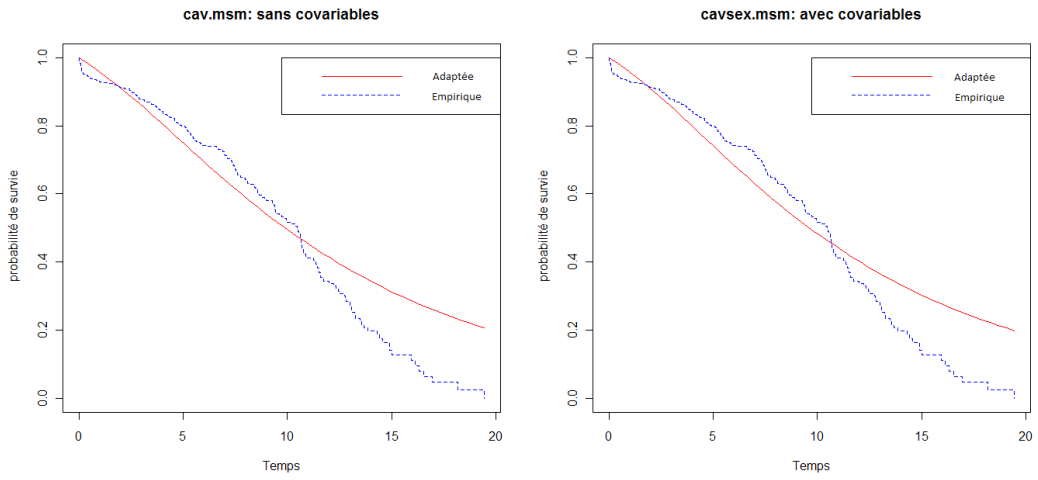

FIG. 5.3 – Comparaison de survie observée et adaptée pour les données de CAV.

La figure.5.3 représente un état particulier avec l'estimation non paramétrique, courbes de Kaplan-Meier avec les deux modèles citées précédemment.

La fonction (hazard.msm) donne l'estimation des rapports de risque correspondant à chaque effet de covariable sur les intensités de transition. 95% limites de confiance sont calculées en assumant la normalité du log-effet. Par exemple, pour le modèle défini précédemment avec le sexe féminin comme covariable, les rapports de risque de tableau

montre que la seule transition sur laquelle l'effet du sexe est significatif au niveau de 5% est la transition 1-2.

Dans les études des maladies chronique, une utilisation importante des modèles multi-états est en prévoyant la probabilité de survie pour des patients dans les états de plus en plus graves de la maladie. Ceci peut être obtenu directement à partir de la matrice de probabilité de transition P(t).

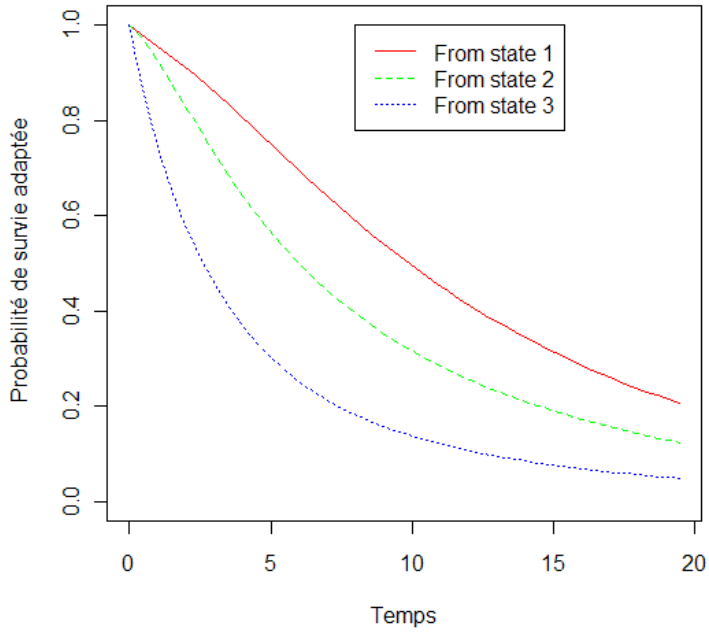

FIG. 5.4 – probabilité de survie pour les données de CAV.

Ceci montre que la probabilité de survie de dix ans avec CAV-Grave est approximativement 0.1, par opposition à 0.3 avec CAV-Moyenne et 0.5 sans CAV. Avec CAV-Grave la probabilité de survie diminue très rapidement à 0.3 en cinq premières années après greffe.

5.2.3 Modèle avec état de censure

On considère un modèle comprend les trois états notés précedement (états transitoires) avec un état absorbant correspondant à ce que nous appellerons l'état « censure »(état de mort). L'intérêt de ce modèle est de permettre une estimation des risques de censure et d'observer si les covariables ont un impact sur ces risques. Le tableau 5.7 donne les estimations, des coefficients de régression.

Tableau5.7 : Modèle semi-paramétrique pour les risques de censure(Transition 1→ C)

Covariable	$\hat{\beta}$	(ec)	P
sex	5.371e-01	2.672e-01	0.0445
age	0.01039	0.00360	< 0.01
pdiagIDC	-0.1068	0.2156	0.6202
pdiagIHD	-0.1023	0.2155	0.6349
pdiagOther	0.9842	0.6146	0.1093
.	.	.	.
.	.	.	.

(ec) estimations des écarts-types.

(P) p avec le test de Wald pour $H_0 : \beta = 0$:

Tableau5.8 : Modèle semi-paramétrique pour les risques de censure(Transition 2→ C)

Covariable	$\hat{\beta}$	(ec)	P
sex	-0.5187	0.2508	0.0386
age	3.561e-02	5.760e-03	< 0.01
pdiagIDC	3.726e-02	4.200e-01	0.9293
pdiagIHD	4.465e-01	4.164e-01	0.2836
pdiagOther	-1.284e+01	1.577e+03	0.9935
.	.	.	.
.	.	.	.

(ec)estimations des écarts-types.

(P) p avec le test de Wald pour $H_0 : \beta = 0$:

Tableau5.9 : Modèle semi-paramétrique pour les risques de censure (Transition 3 → C)

Covariable	$\hat{\beta}$	(ec)	P
sex	-0.5644	0.2760	0.0409
age	-1.09394	0.42053	< 0.01
pdiagIDC	0.1372	0.4591	0.765087
pdiagIHD	0.4222	0.4564	0.354944
pdiagOther	2.0583	1.0978	0.060805
.	.	.	.
.	.	.	.

(ec)estimations des écarts-types.

(P) p avec le test de Wald pour $H_0 : \beta = 0$:

Ces résultats montrent l'effet des covariables sur les risques de censure. À partir de l'état CAV-A, les patients agés de $<= 50ans$ ont un risque qui augmente par rapport au

diagnostic primaire. De même à partir de l'état CAV-M, mais le sex diminue ce risque. À propos du risque de censure, a partir de l'état CAV-G, il semble que les patients agés de $<= 50ans$ diminuent ce risque. Cependant, on peut l'expliquer par le fait que le phénomène de censure à partir d'un état de (CAV-A, CAV-M, CAV-G) sont de nature différente.

Conclusion et perspectives

Au terme de ce travail, après avoir posé quelques points de repère dans le cadre de l'estimation. Nous avons, dans un premier temps, étendu les méthodes et les démarches existantes dans le cas d'un modèle d'analyse de la survie ou de la durée. Par ailleurs nous avons développé et interprété un modèle décrivant la dynamique d'un processus de Markov à temps continu. Nous pouvons donc dire que ce travail a été motivé par l'estimation des passages entre les différents états d'un modèle de Markov, à partir des données de surveillance. L'application du modèle à des problèmes réels a été développée principalement dans le dernier chapitre.

Pour bien présenter l'utilité des méthodes développées au cours de ce travail, nous les avons appliquées à un jeu de données réels médicale. Grâce aux données de l'étude de (CAV)et à l'approche multi-états, nous avons pu obtenir des estimations des probabilités possiblement dans les quatre états. L'étude statistique de ces données est devenue essentielle pour une meilleure compréhension des maladies et une amélioration du suivi. Les modèles de type Markovien répondent à cette problématique et constituent un outil important pour l'analyse de ces données. Par ailleurs, nous avons vu l'intérêt de ce type d'approches en médecine, où le temps passé dans un certain état de santé constitue un facteur important de l'évolution future des patients.

La théorie des processus de comptage fournit un cadre formel à de nombreuses problématiques complexes. L'utilisation de cette théorie semble essentielle pour le développement théorique des méthodes liées aux modèles multi-états. Il nous semble également important de continuer à développer des programmes et des algorithmes plus détaillés afin de faciliter l'utilisation des modèles multi-états. Il sera ainsi intéressant de mettre en place des logiciels permettant d'ajuster certaines méthodes d'estimation pour ce type de modèle.

Bibliographie

[1] Aalen,O.O. *(1978).Nonparametric inference for a family of counting processes. 6,701-726.Ann.Statist.*

[2] Aalen O. O. et Johansen S. *(1978).An empirical transition matrix for non-homogeneous Markov chains based on censored observations. vol. 5. pages 141–150.Scandinavian Journal of Statistics.*

[3] Anaes *(Septembre 2004), Recommandations pour le suivi médical des patients asthmatiques adultes et adolescents. Recommandations pour la pratique clinique de l'Agence National d'Accréditation et d'Evaluation en Santé. URL http ://www.anaes.fr.*

[4] Andersen P. K., Borgan., Gill R. D. et Keiding N. *(1993).Statistical models based on counting processes. Vol.38,P 894-904. Springer-Verlag, New York.*

[5] Andersen P. K. et Keiding N. *(2002). Multi-state models for event history analysis. vol.11,n2. pages 91–115.Statistical Methods in Medical Research.*

[6] Antoniadis Anestis, Sapatinas Theofanis. *(2007)Estimation and inference in functional mixed-effects models* tom 51. 4793-4813.Elsevier B.V. Computational Statistics

[7] Avram Florin. *(2010)Les Probabilités du Bonheur, et les Processus de Markov et de Levy dans l'Actuariat, Files d'attente et Fiabilité.Florida State University.Mémoire de Master.*

[8] Blum J.,Susarla V. *(1980)Maximal Deviation Theory of Density and Failure Rate Function Estimates Based on Censored Data, Multivariate Analysis V, Ed. by P. Krishnaiah,213-222, North-Holland, Amsterdam.*

[9] Boudreau.C and Lawless.J.F. *(2010)Survival Analysis Based on the Proportional Hasards Model and Survey Data. 506- 516. Canadian Journal of Statistics.*

[10] Bouzebda Salim et Elhattab Issam *(2009)A Strong Consistency of a Nonparametric Estimate of Entropy under Random Censorship.*

[11] Breslow, N. and J. Crowley. *(1974)A large sample study of the life table and productlimit estimates under random censorship. 2, 437–453. Ann. Statist.*

[12] Cases.C, Lollivier.S. *(1993)L'économétrie des modèles de durée avec SAS.Document de travail CREST,9344BIS.43-84.Actes des journées de méthodologie statistique.*

[13] Chang I-Shou, Hsiung Chao A. and Wu Su-Mei *(2000)Estimation in a proportional hasard model for semi-markov counting processs.1257-1266.Statistica Sinica 10*

[14] Commenges D. *(1999). Multi-State Models in Epidemiology. vol.5. pages 315–327.Lifetime Data Analysis.*

[15] Core Team R Development *(2010)R : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0,URL http ://www.R-project.org/.*

[16] Cox D. R. *(1972). Regression models and life tables (with discussion). vol. 34. pages 187–220.J Royal Statistical Soc B.*

[17] Cox, D. R. and Oakes, D. *(1984). Analysis of Survival Data, Chapman and Hall, London.*

[18] Dabrowska D. M.,Sun G. et Horowitz M. M. *(1994). Cox regression in a markov renewal model : an application to the analysis of bone transplan data. vol.89. pages 867–877.Journal of the American Statistical Association.*

[19] Daniel Block.A and Lawrence M. Leemis. *(JUNE 2008)Parametric Model Discrimination for Heavily Censored Survival Data.VOL.57,NO. 2,248-259.IEEE Transations on reliability*

[20] Delsol Laurent . *(2008). Régression sur variable fonctionnelle, Tests de structure et Applications. :thèse de doctorat.l'Université Paul Sabatier Toulouse 3.*

[21] Dixon Stephanie.N, Darlington Gerarda.A, Desmond Anthony.F. *(2011)A competing risks model for correlated data based on the subdistribution hazard.Volume.17.Computational Statistics*

[22] Embury, S.H., Elias, L., Heller, P.H., Hood, C.E. Greenberg, P.L.,et Schrier, S.L. *(1977).Remission maintenance therapy in acute myelogenous leukemia. 126,267-272.Western Journal of Medicine.*

[23] Fan Jianqing . *(2005)A Selective Overview of Nonparametric Methods in Financial Econometrics.vol. 20, 317–337.J Statistical Science.Institute of Mathematical Statistics.*

[24] Ferraty.F and Vieu.P. *(2006).Nonparametric Functional Data Analysis. Theory and Practice. Springer Series in Statistics. Springer, New York*

[25] Ferraty.F, Rabhi.A, and Vieu.P. *(2008)Estimation non-paramétrique de la fonction de hasard avec variables explicatives fonctionnelles.* 53. 1-18.Rev. Roumaine math. Pures appl.

[26] Fleming.T.et Harrington.D. *(1991).Counting Processes and Survival Analysis.New York : Wiley.*

[27] Figini Silvia . *(2009)Penalized models to estimate customer survival.vol. 11.141-150.International Journal of Risk Assessment and Management-Int J Risk Assess Manag.*

[28] Földes A., Rejto L., Winter B. *(1981). Strong Consistency Properties of Nonparametric Estimations for Randomly Censored Data,II, Estimations of Density and Failure Rate , Period. Math. Hungar, 12.*

[29] Fox John . *(2002)Cox Proportional-Hazards Regression for Survival Data Appendix to An R and S-PLUS Companion to Applied Regression. Journal :computational statistics and data analysis.DOI 20*

[30] Fox John , Sanford Weisberg. *(2010)Cox Proportional-Hasard Regression for Survival Data in R. J Computational statistics.*

[31] Garcia Tanya.P,Ma Yanyuan, Yin Guosheng. *(2011)Efficiency improvement in a class of survival models through model-free covariate incorporation*

[32] Goret Sheila.M., Pocock Stuart.J., Kerr Gillian.R. *(1984). Regression Models and Non-proportional Hazards in the Analysis of Breast Cancer Survival.vol.33, No.2,pp.176-195.Appl. Statist.*

[33] Gürler Ülkü et Wang Jane-Ling . *((1993))Nonparametric estimation of hasard functions and their derivatives under truncation model. Vol.45, No. 2, 249-264.Ann.Inst. Statist. Math.*

[34] Heller.G. *(2010)Proportional hazards regression with interval censored data using an inverse probability weight.Volume.17,Issue 3,1380-7870. Journal :Lifetime Data Analysis.Springer*

[35] Herndon II James Emmett *(April 1988). A parametric survival model which generates monotonic and non-monotonic hasard functions and incorporates time-dependent covariables*

[36] Heutte.N,Huber.C et Pons O. *(2001).Semi-Markovian models appied to aids with censoring. vol 21.pages 3369–3382. Statistics in Medicine*

[37] Ho Weang Kee. Henderson Robin, Philipson Peter.M. *((2010))Tests for Hazard Transformation.Stat Biosci 2 : 41–64.*

[38] Horová Ivana , Pospisil Zdenek et Zelinka Jiri . *(2007)Semiparametric Estimation of Hazard Function for Cancer Patients.Volume 69, Part 3, pp. 494-513.The Indian Journal of Statistics.*

[39] Hougaard P. *(1999). Multi-State Models : a Review. vol. 5.pages 239–264.Lifetime Data Analysis.*

[40] Huber-Carol C.et Pons O. *(Mai 2004),Independent competing risks versus a general semi-Markov model :application to heart transplant data. Prépublication.URL http ://www.math-info.univ-paris5.fr/map5/publis/PUBLIS04/huber-2004-13.pdf.*

[41] Iglesias Pérez.M.C and González Manteiga.W. *(2003)bootstrap for the conditionl distribution function with truncated and censored data.vol. 55, 331–357*

[42] Jackson C. H. *(Janvier 2005), Multi-state Markov models in continuous time. Package MSM, R Foundation for Statistical Computing.URL http ://www.R-project.org.*

[43] Janssen J., Manca R. et Volpe E. *(1997). Markov and semi-Markov options pricing models with arbitrage possibility. vol.13, n° 2.pages 103–113.Applied Stochastic Models and Data Analysis.*

[44] Kaplan EL et Meier P. *(1958)Non-parametric estimation from incomplete observations.5 :457-481. Journal of the American Statistical Association.*

[45] Karlin S and Taylor HM. A *(1975)first course in stochastic processes,chapter 4. Academic Press, second edition.*

[46] Kessler Mathieu and Sqrensen Michael. *(2005)On Time-Reversibility and Estimating Functions for Markov Processes.Volume 8. p 95–107.Springer in its journal Statistical Inference for Stochastic Processes*

[47] Klein.J.P and Moeschberger.M.L. *(2003)Survival Analysis : Techniques for Censored and Truncated Data. .Springer, New York,2ème Edition.2*

[48] Laksaci Ali, Mechab Boubakeur. *(2008)Estimation non-paramétrique de la fonction de hasard avec variable explicative fonctionnelle : cas des données spatiales* (55). 1-18. Rev.Roumaine Math.Pures Appl.

[49] Läuter Henning et Liero Hannelore. *(2004)Nonparametric Estimation and Testing in Survival Models.1613-3307. Institut für Mathematik der Universität Potsdam*

[50] Ledauphin.S, Pommeret.D et Qannari.E.M. *(2007)Application des chaines de Markov pour le suivi de la dégradation de produits alimentaires.tome 148,n3.Journal de la Société Française de Statistique.*

[51] Lin.D.Y, ying Zhiliang.(1997) *Additive Hasards Regression Models for Survival Data*, .123, 185-198.Springer-Verlag.

[52] Lortet-Zuckermann.M.C. *Tests de l'hypothèse d'une chaîne de Markov. Application à la succession des diverses sortes d'explorations de SS Cyg.Vol.29, p.205. Journal : Annales d'Astrophysique.*

[53] Martinussen, T. et Scheike.T.H. *(2002). Efficient estimation in additive hazards regression with current status data. 89 : 649-658. Biometrika.*

[54] Morau A. *(1989). Econométrie des variables de durée. Note Département recherche. N.123 /G 305.*

[55] Nassiri Abdelhak, Delecroix Michel, Bonneu Michel : *(2000)Annales d'économie et de statistique.Estimation non-paramétrique du taux de hasard :application à des durées de chomage censurées à droite.* Annales d'économie et de statistique . N°. (58). 216-232.Published by : L'INSEE / GENES

[56] Nelson, W.B. *(1972). A short life test for comparing a sample with previous accelerated test results.14, 175–185. Technometrics*

[57] Ouhbi Brahim and Limnios Nikolaos. *(1999)Nonparametric Estimation for Semi-Markov Processes Based on its Hazard Rate Functions.vol.2, issue 2.151-173.Statistical Inference for Stochastic Processes.Springer Online.*

[58] Perez-Ocon R. et Torres-Castro I. *(2002). A reliability semi-Markov model involving geometric processes. vol. 16,n° 2. pages 157–170.Applied Stochastic Models in Business and industry.*

[59] Peto.R. *(1973).Empirical survival curves for interval censored data. 22,86-91.Appl. Statist.*

[60] Ramsay.J.O, Silverman.B.W. *(April 2005).Functional Data Analysis Second Edition.Springer Series in Statistics*

[61] Roxström.A, Ducrocq.V, Strandberg E. *(2002)Survival analysis of longevity in dairy cattle on a lactation basis.P 305-318.Institut national de la recherche agronomique.*

[62] Saint-Pierre Philippe , Gurie Agathe : *(2009).Fonction de survie bivariée de variables censurées à droite et à gauche.publié dans "41èmes Journées de Statistique, SFdS, Bordeaux.*

[63] Sharples LD *(1993).Use of the Gibbs Sampler to Estimate Transition Rates Between Grades of Coronary Disease Following Cardiac Transplantation. N 12, P 1155-1169.Statistics in Medicine.*

[64] Sharples LD, Jackson CH, Parameshwar J, Wallwork J, Large SR *(2003)."Diagnostic Accuracy of Coronary Angiography and Risk Factors for Post-Heart-Transplant Cardiac Allograft Vasculopathy." Transplantation, 76(4),679-682.*

[65] Sommen Cecile. *(2009).Modèles pour l'estimation de l'incidence de l'infection par le VIH en France à partir des données de surveillance VIH et SIDA : thèse de doctorat.université bordeaux 2*

[66] Spierdijk Laura . *(2005)Nonparametric Conditional Hazard Rate Estimation :A Local Linear Approach.*

[67] Taibi-Hassani.S., Youndjé.E. *(2003)Validation croisée pour l'estimateur lissé de la fonction de hasard : cas des données censurées. tom 51,p. 73-86.Revue de statistique appliquée.*

[68] Tanners M.A., Wong W.H. *(1983). The Estimation of the Hazard Rate Function from Right-Censored Data by the Kernel Method, Vol. 11, n° 3.The Annals of Statistics.*

[69] Therneau T.M. et Grambsch P. M. *(2000). Modeling Survival Data : Extending the Cox Model. Springer – Statistics for Biology and Health.*

[70] Van Keilegom Ingrid et Veraverbek Noel *(2001)Hasard rate estimation in nonparametric regression with censored data.Vol.53,No.4,730-745.Ann.Inst.Statist.Math.*

[71] Vassiliou P.C. et Papadopoulou A.A. *(1992). Non homogeneous semi-Markov systems and maintainability of the state sizes. vol. 29.pages 519–534.Journal of Applied Probability.*

[72] Viallon Vivian. *(2006). Processus empiriques, estimation non paramétrique et données censurées : thèse de doctorat.université paris 6.*

[73] Wand.M.P, Devroye.L. *(1993)How easy is a given density to estimate?.Computational Statistics et Data Analysis. vol 16.311-323*

[74] Wang Jane-Ling . *(2003)Smoothing Hazard Rates Contribution to the Encyclopedia of Biostatistics*

[75] Youndjé É,Sarda P et Vieu P. *(1996)Optimal smooth hasard estimates. Vol 5, N.2, 374-379, Test.*

[76] Zhao Yichuan, Jinnah Ali. *(2011)Inference for Cox's regression models via adjusted empirical likelihood.Volume 26,issue 4, p 1-12.Springer in its journal Computational Statistics*