N° d'ordre :

REPUBLIQUE ALGERIENNE DEMOCRATIQUE & POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR & DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DJILLALI LIABES
FACULTE DES SCIENCES EXACTES
SIDI BEL ABBES

THESE DE DOCTORAT DE 3ème CYCLE

Présentée par : BENCHEKOR Asma
Domaine : Mathématiques Informatiques
Filière : Mathématiques
Intitulé de la formation : Calcul stochastique-Statistiques et Application

Intitulée

Estimation semi-paramétrique de la fonction de risque dans le modèle de Cox pour un processus semi-markovien

Soutenue le : 21/06/2017
Devant le jury composé de :

Président : (OUAHAB Abdelghani, Professeur, UDL - SBA)
Examinateurs : BOUCHENTOUF Amina Angelika, MCA, UDL – SBA
MADANI Fethi, MCA, UMT – Saida
TEBBOUNE Fethallah, MCA, UMT – Saida
TERBECHE Mekki, Professeur université d'Oran1
Directeur de thèse : YOUSFATE Abderrahmane, Professeur, UDL - SBA
Co-Directeur de thèse : ///

Année universitaire : 2016/2017

Remerciements

En premier lieu, je remercie le bon Dieu, tout puissant, de m'avoir donné la force pour survivre, ainsi que l'audace pour dépasser toutes les difficultés.

Je tiens tout d'abord à exprimer toute ma reconnaissance à Mr **YOUSFATE Abderrahmane**, mon directeur de thèse qui est à l'origine de ce travail. C'est un honneur pour moi de travailler avec lui et je ne peux qu'admirer son talent. Je lui suis vivement reconnaissante, non seulement parce qu'il a accepté de me prendre en thèse, mais aussi parce qu'il a partagé ses idées avec moi et, surtout, il m'a transmis sa passion de la recherche et la motivation nécessaire pour mener à bien ce travail. Je le remercie aussi de son aide morale et scientifique durant les moments difficiles de cette thèse. Ses conseils et ses critiques constructifs, ses encouragements et son optimisme m'ont permis de lever bien des doutes. Merci encore pour les conférences auxquelles vous m'avez encouragé à participer.

Je suis extrêmement reconnaissante à Mr **OUAHHAB Abdelghani** pour l'intérêt qu'il a donné pour mon travail, pour ses questions et ses précieux conseils lors de mes exposés dans les séminaires hebdomadaire de mathématiques. Je vous remercie d'avoir accepté de juger ce travail en tant que président de jury.

Je tiens bien entendu à exprimer ma sincère reconnaissance à mon premier repère Mr **TEBBOUNE Fethallah** qui a accepté, malgré toutes ces occupations, de me faire l'honneur de participer à ce jury. Je profite aussi de cette occasion pour vous témoigner ma gratitude et ma profonde reconnaissance pour votre soutien et vos conseils au cours de mon cursus universitaire. Merci mon professeur et mon premier maitre de recherche.

Je remercie vivement Mr **TERBECHE Mekki**, et Mr **MADANI Fethi**, pour la confiance dont ils me font preuve en faisant parties de ce jury.

Mes remerciements les plus sincères et ma reconnaissance à Melle **BOUCHENTOUF Amina Angelika** non seulement d'accepter de juger mon travail, mais aussi pour son soutien, et ses précieux conseils qui m'ont été d'une grande utilité tout au long de mon parcours.

A ma chère famille ; merci de m'avoir soutenue et encouragée toutes ces longues années d'étude. Je remercie Maman et Papa de m'avoir toujours soutenue durant ces années d'éloignement dans mes études et dans mes choix. Merci d'avoir été à l'écoute lorsque j'en avais besoin et d'avoir cru en mes capacités. Tous les mots du monde ne sauraient exprimer l'immense amour que je vous porte. Que Dieu tout puissant vous garde et vous procure santé, bonheur et longue vie.

Ce travail n'aurait pu voir le jour sans le soutien constant et la présence de mes chèrs **Nesrine Rania** et **Mohamed Ghanem** ils ont rendu les jours difficiles moins pénibles ; votre soutien pour moi une source de courage et de confiance. Je vous souhaite toute la réussite et tout le bonheur du monde.

Je dis un grand merci du fond du coeur à tous mes amis pour leur soutient moral. Je suis fier de vous avoir comme amis.
A ma chère **BOUTEFAL Zohra** merci pour ton amour, ton amitié. Tu étais toujours là pour me soutenir, m'aider et m'écouter.Nchallah notre amitié reste à jamais.

Je ne pourrais clôturer ces remerciements sans me tourner vers **Sid Ahmed DIFI** pour ses conseils avisés et ses encouragements. Je n'ai pas de mots assez forts pour t'exprimer ma gratitude. Tu m'as toujours soutenue lors des moments souvent difficiles, cette année 2016, pleine d'événements et d'émotions ! est particulière pour moi merci infiniment.

A tous ceux qui ont participé de près ou de loin à la réalisation de ce travail.

Dédicace

Soumia

Nous avions ensemble fait tant de choses. Et voilà que maintenant tu nous quittes sans retour. Avec toi, j'ai partagé tant de projets et tant d'espoirs. Il y a tant de choses encore que nous aurions voulu faire ensemble.Mais cela semble s'arrêter aujourd'hui ; et ce n'est plus ensemble que nous allons réaliser ce que tu espérais. On a rêvé de célébrer ce jour ensemble mais comme un mur, la mort nous sépare, de toi, je te promet de faire mon meilleur et de continuer le parcours que tu as commencé. Que dieu l'accueille dans son vaste paradis.

Résumé

Bien que les modèles de durées aient été utilisés depuis le XVII^{e} siècle, ce n'est qu'à partir du milieu du XX^{e} siècle qu'un intérêt particulier est donné à ces modèles.

Dans cette étude, après présentation de différentes approches équivalentes pour aborder un modèle de durée ; nous mettons l'accent particulièrement sur la fonction de risque avec ou sans censure tout en introduisant le modèle de durée "multi-états" sous inférence semi-markovienne. Le cas classique de modèle de durée devient alors un cas particulier, à savoir un processus à deux états dont l'un est transient et l'autre absorbant.

Après avoir mis en exergue, les approches classiques paramétrique, non paramétrique et semi-paramétrique, nous traitons le cas de modèle de Cox sous inférence semi-markovienne non-homogène forte et sous l'hypothèse que les variables de durée soient à supports bornés. Le problème est traité aussi bien dans le cas continu que dans le cas discret. Le nombre d'états considéré dans cette étude est fini ou infini dénombrable. Cette approche commence à intéresser plusieurs auteurs ; nous en citons notamment les travaux de Korolyuk et al. ainsi que ceux de Limnios et al..

Abstract

Although the duration models have been used since the XVII th century, particular interest is given to these model only until the middle of the XXth century.

In this study, after presenting different equivalent approaches to study duration model ; we focus on the risk function particularly with or without censorship and with introducing the multi-state duration model under semi-Markovian inference. The classical case of time model then becomes a special case, namely a two-state process of which one is transient and the other one absorbing state.

After highlighting the classical parametric, nonparametric and semi-parametric approaches, we treat the Cox model case under strong nonhomogeneous semi-Markovian inference and under the assumption that the duration variables have bounded supports. The problem isstidied both in the continuous case and in the discrete case. The number of states considered in this study is finite or infinite countable. The use of the Cox model on semi-Markov processes begins to interest several authors ; we mention in particular the work of Korolyuk et al. as well as those of Limnios et al.

Publication

– Benchekor, A., Yousfate, A. (2017). A strong Markov model for a discrete time inhomogeneous Semi-Markov Process with bounded phase staying. International Journal of Statistics and Economics. Vol 18, Issue 2

Communications dans des séminaires internationaux

– Bellaouedj, A., Benchekor, A., Yousfate, A. (2014). On non-parametric estimation of a functional Markov chain transition operator. Journées de Statistique. Biskra, 06-07 mai 2014.

– Benchekor, A., Yousfate, A. (2015). Processus semi-markoviens à temps discret et à durées de transitions bornées. Théorie et applications, dimacos'2015, Sidi Bel Abbès, 15-19 novembre 2015.

– Benchekor, A., Yousfate, A. (2016). Processus semi-markoviens à durée de transitions bornée. RAMA10, Ouargla, 07-11 février 2016.

– Benchekor, A., Yousfate, A. (2016). On a canonical decomposition of an inhomogeneous Semi-Markov Process with bounded phase staying. Discrete time case. Marakech, 3ème Conférence de probabilités et statistique de Marrakech, 25-28 avril 2016.

Communications dans des séminaires nationaux

– Benchekor, A. (Mars 2012). Modèle de survie avec covariable(s) Modèle de survie avec covariable(s). Le séminaire de mathématique et d'informatique de l'Université de Sidi Bel Abbès

– Benchekor, A. (juin 2013). Modèle semi-markovien homogène. Le séminaire de mathématique et d'informatique de l'Université de Sidi Bel Abbès

– Benchekor, A. (Janvier 2014). Modèle de survie et Package survival Modèle de survie et Package survival. Le séminaire de mathématique et d'informatique de l'Université de Sidi Bel Abbès

– Benchekor, A. (décembre 2015). Modèle de cox généralisé. Le séminaire de mathématique et d'informatique de l'Université de Sidi Bel Abbès

Table des matières

Introduction générale

La statistique des durées de vie est un domaine actif de recherche stimulé par de nombreuses applications, notamment biomédicales, mais aussi en fiabilité, en actuariat ou en démographie ; ou en d'autres champs utilisant les modèles de durée.

Les modèles de survie permettent de modéliser la durée jusqu'à l'apparition d'un évènement d'intérêt. Les modèles de survie permettent aussi d'associer des facteurs d'exposition au risque de survenue de l'évènement dont l'explication est donnée grâce à un modèle de régression ; un des plus utilisés est le modèle de Cox dit aussi modèle à risques proportionnels. L'estimation des effets des facteurs peut être réalisée soit par des méthodes non paramétriques, des méthodes semi-paramétrique ou des méthodes paramétriques selon la problématique.

Dans cette thèse, nous nous intéressons aux modèles de durée dans une approche semi-markovienne. Ainsi le cas classique est considéré comme étant un processus à deux états dont l'un est transient et l'autre absorbant. Le temps influençant la transition est alors le temps depuis l'entrée dans l'état en cours (état transient). C'est le temps d'attente ou temps de séjour.

Les processus semi-markovien constituent alors une alternative intéressante puisqu'ils intègrent dans la définition du modèle les lois de temps de séjour dans un état. Ainsi on peut généraliser les modèles de durée à deux états ré-

currents ; voire plusieurs états de natures identiques ou différentes. Les modèles semi-markovien généralisent des modèles markoviens dans le sens où les lois de temps de séjour des états ne sont nécessairement exponentielles.

Les modèles multi-états sont considérés comme une généralisation des modèles de survie. Il sont caractérisés par un processus stochastique à espace d'états fini (ou infini dénombrable) pour décrire un phénomène. L'utilisation de ce modèle de processus permet de représenter un processus de sauts dont les trajectoires indiquent les différents états successivement occupés par un *individu* observé. 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 la dynamique du système de plusieurs manières. L'étude de ces modèles consiste à analyser les forces de passage (intensités de transition) entre les différents états.

Structure de la thèse

Notre travail est présenté comme suit :

Le premier chapitre consiste à présenter un historique sur les modèles de durées montrant ainsi la position du travail par rapport aux différentes contributions scientifiques dans ce domaine.

Le deuxième chapitre introduit les outils de probabilité nécessaires au travail qui s'ensuit. Des rappels de notions de base sur les modèles de durées y sont présentés ainsi que l'extension du modèle rendant ainsi les études classiques des modèles de durées comme étant un processus à 2 états dont un des états est absorbant.

Le troisième chapitre est consacré à la présentation de l'étude des modèles de durées dans un modèle semi-markovien

Dans le quatrième nous commençons, tout d'abord par présenter le modèle de Cox et sa généralisation dans le modèle semi-markovien. Par la suite, nous passons à l'étude des résultats existants pour l'estimation des paramètres de ce modèle.

Chapitre 1

Chapitre introductif

La statistique des durées de vie est un domaine actif de recherche stimulé par les résultats très satisfaisants dans plusieurs domaines d'application tels la biométrie (notamment en sciences biomédicales), en fiabilité des systèmes et en économie (notamment en actuariat). Le terme *durée de vie* désigne le temps écoulé jusqu'à la survenue de l'événement d'intérêt. L'inférence statistique pour ces données consiste à estimer leur loi de probabilité tout en tenant compte de l'hétérogénéité des données (cas de mélange de plusieurs groupes par exemple) et de leurs caractéristiques (interprétées par des variables explicatives). Les aspects restrictifs liés aux troncatures ou aux censures sont également étudiés.

Historiquement ce sont les démographes britanniques John Graunt et William Petty qui ont établi les premières statistiques sur les durées de survie de la population au milieu du XVII$^{\text{e}}$ siècle. Ce n'est qu'au XIX$^{\text{e}}$ siècle que des tables liées à des variables statistiques commencent à apparaître et la modélisation de la fonction de risque fut entamée. Citons le modèle de Benjamin Gompertz (1825) qui présente la fonction de risque comme suit

$$h(t; r, s) = rs^t \text{ où } t \in \mathbb{N},\ r,\ s \in \mathbb{R}_+^*.$$

Notons qu'à l'époque t était évalué en années. Cette formulation de la fonction de risque a été généralisée par William Makeham en 1860 sous la forme

$$h(t; r, s, u) = u + rs^t \text{ où } t \in \mathbb{N},\ r,\ s \in \mathbb{R}_+^* \text{ et } u \in \mathbb{R}$$

appelée formule de Gompertz-Makeham.

Dans le cas où la variable de durée T est continue, la fonction de risque, dite aussi fonction du hasard, exprimée en fonction de la loi de T s'écrit :

$$h(t) = \frac{f(t)}{1 - F(t)}$$

où f et F sont respectivement la densité et la fonction de répartition de T. Réciproquement, la densité s'écrit aussi sous la forme

$$f(t) = s(t)e^{\int_0^t -h(x)dx}.$$

Le modèle de Gompertz-Makeham a été exploité par les démographes pour étudier la dynamique des populations où la croissance (ou décroissance) s'exprime par l'équation différentielle :

$$\frac{dh(t; r)}{dt} = rh(t) \log\left(\frac{c}{h(t)}\right) \text{ avec } r = c^{te},\ h > 0 \text{ et } c \text{ la capacité limite du milieu.}$$

La solution générale s'écrit

$$h(t) = ce^{\left(se^{(-rt)}\right)} \text{ où } s = \log\left(\frac{h(t_0)}{c}\right).$$

Les problèmes soulevés dans ce contexte relèvent plutôt de la stabilité du système et de la recherche des points d'équilibre.

En utilisant le modèle de Gompertz-Makeham selon l'approche probabiliste, on a le modèle paramétrique suivant :

$$h(t;r,s) = \frac{rse^{-st}e^{-re^{-st}}}{1-e^{-re^{-st}}};\ t \in \mathbb{R}$$

ou plus généralement (avec dérive)

$$h(t;r,s) = \frac{(1+r(1-e^{-st}))se^{-st}e^{-re^{-st}}}{1-(1-e^{-st})e^{-re^{-st}}};\ t \in \mathbb{R}_+$$

Ce n'est que vers la deuxième moitié du XX$^{\text{e}}$ siècle que des résultats de grande importance sur les modèles de durées commencent à se développer. Les domaines d'application sont, notamment, la fiabilité des systèmes, l'économie et la biostatistique. En fiabilité des systèmes, les éléments observés ont les mêmes caractéristiques et peuvent être interchangés sans qu'il y ait un effet sur le modèle. Cependant, en biostatistique ou en économie, chaque élément a ses propres caractéristiques telles l'âge, le milieu social, ... qui sont considérées comme étant des variables pouvant intervenir pour expliquer le phénomène de durée.

En fiabilité, il est considéré que les travaux de Waloddi Weibull de 1951 [45] constituent la base sur laquelle se développent les modèles de fiabilité des systèmes. Il est à noter que ces modèles peuvent être exploités pour étudier des phénomènes de durées en biostatistique.

En biostatistique (notamment en médecine) et en économie (notamment en actuariat), le travail de Kaplan et Meier [21], en 1958, ainsi que celui de Cox [7], en 1972, sont souvent cités comme étant essentiels pour construire des modèles de durées utiles pour la description et la prévision des phénomènes étudiés dans ce contexte. Ces modèles sont exploités également en fiabilité.

Tous les modèles cités peuvent être utilisés aussi bien en fiabilité qu'en biostatistique qu'en économie ; il suffit seulement de mettre les hypothèses adéquates pour exploiter le modèle étudié.

Pour estimer la fonction de risque, trois grandes méthodes sont utilisées :

- La méthode paramétrique qui consiste à estimer les paramètres $r,\ s$ ... dans $h(t; r, s, ...)$ (*modèle de Gompertz-Makeham par exemple*).
- La méthode non paramétrique qui consiste à estimer le rapport $h(t) = \dfrac{f(t)}{1 - F(t)}$ dans un espace fonctionnel spécifique (*modèle de Kaplan-Meier par exemple*).
- La méthode semi-paramétrique qui consiste à estimer la fonction $h(t)$ dans un contexte paramétrique (*modèle de Cox par exemple*).

L'introduction de l'hétérogénéité dans le modèle de Cox (appelé fragilité en français ou *frailty* en anglais) a été introduit par Clayton (1978) en biométrie [6] et par Vaupel et al. (1979) en démographie [44].

Dans les années 80 des études paramétriques et non paramétriques ont fait ressortir l'intérêt d'utiliser une fragilité de loi gamma, nous en citons notamment les travaux de Hougaard [16]. En non paramétrique, la fragilité a été étudiée par plusieurs chercheurs, nous en citons Elberts et Heckman [9] et [15]. Ces dernières études se confinent dans le modèle MPH (Mixture of Proportional Hazards ou *mélange de hasards proportionnels*).

Des généralisations des modèles de survie ont été développées également en introduisant les modèles à risques compétitifs qui consistent à chercher le phénomène, parmi plusieurs, qui aurait été la cause première du changement d'état Chiang [4]. Cette approche fut mise en lien avec le renouvellement semi-markovien par Berman 1961 [3]. Toutefois l'exploitation de cette approche dans les modèles de durées ne fut mise en évidence qu'à partir des études sur les risques compétitifs dépendants dépassant ainsi la restriction au cas de risques compétitifs indépendants suggéré par le théorème de Tsiatis [43]. Ainsi l'étude des modèles de durées sur des systèmes à plus de deux états non tous transients a pris son importance, notamment l'étude des modèles semi-markoviens.

Depuis les année 90, on constate un progrès dans l'identification pour des modèles plus approchés que le modèle MPH à effet individuel. Ridder (1990) montre qu'il existe toujours deux modèles par observation équivalents, l'un avec une espérance finie et l'autre avec une espérance infinie.

Chapitre 2

Modèles de durées et extension à l'analyse multi-états

Les modèles de durées de vie sont des modèles statistiques dont l'objectif est de permettre l'analyse de variables aléatoires réelles (v.a.r.) positives pouvant être interprétées comme des durées. L'analyse peut être faite sur les variables elles-mêmes, comme elles peuvent être traitées en relation et plus particulièrement, d'étudier leurs variations en fonction d'autres variables. Il est à noter que toute v.a.r. peut être rendue positive (par une transformation positive bijective telle la transformation exponentielle) et donc, pour des durées, seule l'interprétation est spécifique.

Nous noterons, dans toute la suite, par T la variable de durées ; variable aléatoire réelle positive, de loi continue, admettant une espérance (ou éventuellement des moments d'ordre supérieur). Implicitement, nous supposerons que $\mathbb{P}(T = +\infty) = 0$

Au niveau des applications, cette variable représente la durée passée dans un état donné (la durée du chômage, la durée de l'activité, la durée de la carrière d'un individu, la durée passée dans un niveau hiérarchique précis, la durée d'attente

à un guichet, la durée de vie d'un leucémique ou d'un cancéreux, la durée de vie d'une machine,...), ou la durée séparant deux naissances, la durée séparant l'achat de certains biens de consommation, la durée séparant deux traitements pour un malade,...). Les applications sont donc très diverses, et sont présentes dans le domaine biomédical (la durée de survie d'un malade,...) en économie (chômage,...), en fiabilité (la durée de vie d'une machine ou la durée de son fonctionnement continu jusqu'à la prochaine panne, ...) en théorie des files d'attente (attente dans la file devant un guichet de la poste jusqu'au début de service, ...), et dans bien d'autres domaines encore.

2.1 Caractérisation de la loi d'une v.a.r. positive

Définition 2.1. La fonction de densité de probabilité de T, notée $f(t)$ est définie par :

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

Définition 2.2. La fonction de répartition $F(t)$ est la probabilité de décéder entre 0 et t :

$$F(t) = \mathbb{P}(T \leqslant t) = \int_0^t f(u)du$$

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

Définition 2.3. :On appelle fonction de survie S la probabilité que la durée de vie T soit supérieur à un temp t :

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

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

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

Il est clair que chacune des trois fonctions classiques précédentes caractérise entièrement la loi de T. D'autres fonctions peuvent également spécifier la loi de T.

Définition 2.4. La fonction de risque, notée h, est définie, pour tout t de $\mathbb{R}^+$, par :

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \frac{\mathbb{P}(t < T \leqslant t + \Delta t)}{\mathbb{P}(T > t)} \\ &= \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \frac{P[(t < T \leqslant t + \Delta t) \cap (T > t)]}{\mathbb{P}(T > t)} \\ &= \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \mathbb{P}(t < T \leqslant t + \Delta t) \end{aligned}$$

$h(t)$ est donc le taux instantané de sortie de l'état à la date t (même si $h(t)$ n'est pas nécessairement inférieur à 1.
Il existe d'autres terminologies classiques (suivant les domaines d'application) pour désigner h : taux de sortie du chômage, taux de perte d'emploi, taux de guérison, taux de mortalité, taux de panne, etc...

Théorème 2.1. *La fonction de risque caractérise la loi de T, et on a*

$$S(t) = \exp[-\int_0^t h(u)du] \qquad t \in \mathbb{R}^+$$

– **Démonstration :** Il suffit de prouver que h est en bijection avec S. Nous avons :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)}\frac{dS(t)}{dt} = -\frac{dlnS(t)}{dt}$$

En intégrant, grâce au fait que $S(0) = 1$, on a

$$lnS(t) = -\int_0^t h(u)du$$

ou encore

$$S(t) = \exp[-\int_0^t h(u)du]$$

Remarque 2.1.1. La démonstration du Théorème 1.1 permet d'introduire une autre fonction intéressante, le taux de risque cumulé ("Log-survival fonction") :

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

Remarque 2.1.2. La fonction h est positive, mais n'est pas une densité de probabilité puisque

$$h(+\infty) = -ln(S(+\infty)) = +\infty$$

.

Remarque 2.1.3. S est continue donc h est mesurable.
Comme

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0^+} \frac{\mathbb{P}(t < T \leqslant t + \Delta t)}{\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{f(t < T \leqslant t)\Delta t}{\Delta t} \end{aligned}$$

$h(t)$ s'interprète en tant que densité conditionnelle de la loi de T sachant que $(T > t)$, évaluée au point t. Si l'on s'intéresse à toute la loi conditionnelle, on est amené à définir la fonction de survie conditionnelle. Pour tout t_0 de $\mathbb{R}^+$, on examine toutes les lois conditionnelles de T sachant que $(T > t_0)$

Définition 2.5. La fonction de survie conditionnelle est définie par

$$S(t/t_0) = \mathbb{P}(T > t + t_0/T > t_0) \qquad t \in \mathbb{R}^+, t_0 \in \mathbb{R}^+$$

Remarque 2.1.4. $S(t/t_0)$ s'exprime à l'aide de la fonction de survie :

$$S(t/t_0) = \frac{\mathbb{P}(T > t + t_0)}{\mathbb{P}(t > t_0)} = \frac{S(t + t_0)}{S(t_0)}.$$

Remarque 2.1.5. $S(t/t_0)$ s'exprime à l'aide de la fonction de risque

$$\begin{aligned} S(t/t_0) &= \frac{S(t + t_0)}{S(t_0)} \\ &= \exp[-\int_0^{t+t_0} h(u)du + \int_0^{t_0} h(u)du] \\ &= \exp[-\int_{t_0}^{t+t_0} h(u)du] \end{aligned}$$

Théorème 2.2. *Les fonctions de survie conditionnelle caractérisent la loi de T.*

- **Démonstration :**

. Si on connaît $S(t)$, alors on connaît $S(t/t_0) = \frac{S(t+t_0)}{S(t_0)}$.

. Réciproquement, si on connaît les fonctions $S(t/t_0)$ pour toute date t_0 de $\mathbb{R}^+$ alors on connaît $S(t) = S(t/0)$. La fonction de survie s'interprète donc comme la fonction de survie conditionnelle à $t_0 = 0$, c'est à dire la fonction de survie à la naissance.

Définition 2.6. La durée de vie moyenne restante ("expected residual life") est la fonction définie par

$$r(t) = E[T - t/T > t]$$

Théorème 2.3. *La fonction r caractérise la loi de T.*

Démonstration. Il suffit de prouver qu'il existe une bijection entre les fonctions r et S.

Expression de r en fonction de S.

La loi conditionnelle de T, sachant $(T > t)$, a pour support l'intervalle $[t, +\infty[$; c'est une loi conditionnelle continue, de densité $\frac{f(t)}{S(t)}$.

On en déduit

$$\begin{aligned} r(t) &= \frac{\int_t^{+\infty} (u-t) f(t) du}{S(t)} \\ &= -\frac{\int_t^{+\infty} (u-t) dS(u)}{S(t)} \\ &= \frac{[-(u-t)S(u)]_t^{+\infty} + \int_t^{+\infty} S(u) du}{S(t)} \end{aligned}$$

Par une intégration par partie nous trouvons

$$r(t) = \frac{1}{S(t)} \int_t^{+\infty} S(u) du.$$

– Réciproquement

On a :

$$\frac{1}{r(t)} = \frac{S(t)}{\int_t^{+\infty} S(u) du}$$

D'où

$$\begin{aligned} -\frac{1}{r(t)} &= \frac{-S(t)}{\int_t^{+\infty} S(u)du} \\ &= \frac{\frac{d}{dt}\int_t^{+\infty} S(u)du}{\int_t^{+\infty} S(u)du} \\ &= \frac{d}{dt} ln[\int_t^{+\infty} S(u)du] \end{aligned}$$

En intégrant, il suit :

$$-\int_0^t \frac{1}{r(u)}du = ln\int_t^{+\infty} S(u)du - ln\int_0^{+\infty} S(u)du$$

C'est-à-dire :

$$\int_0^t \frac{1}{r(u)}du = -ln\int_t^{+\infty} S(u)du + lnr(0)$$

En remarquant que

$$\int_0^{+\infty} S(u)du = r(0) = E(T/T > 0) = E(T),$$

On a alors :

$$\int_0^{+\infty} S(u)du = r(0)\exp[-\int_0^t \frac{1}{r(u)}du].$$

Par dérivation, on obtient :

$$S(t) = \frac{r(0)}{r(t)}\exp[-\int_0^t \frac{1}{r(u)}du].$$

□

2.2 Modèle de Weibull

La loi exponentielle fut généralisée par Weibull en 1939. La fonction de densité d'une variable aléatoire de Weibull s'éscrit :

$$f(t) = \alpha\lambda t^{\alpha-1} \exp^{\{-(\lambda t^{\alpha})\}}$$

Où λ et α sont deux constantes strictement positives. Le paramètre λ est appelé paramètre d'échelle et α paramètre de forme, 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 :

Lorsque $\alpha > 1$, la fonction de risque est monotone croissante. Elle est monotone et croissante pour $\alpha < 1$ et constante si $\alpha = 1$. Les fonctions de survie et de risque associées sont respectivement :

$$S(t) = \exp -\lambda t^{\alpha}$$

$$h(t) = \alpha\lambda t^{\alpha-1}$$

La loi de Weibull est très largement utilisée dans les domaines industriel (fiabilité) et biomédical (analyse de durée de vie). Une caractéristique de cette loi, c'est qu'elle couvre à la fois le cas d'une fonction de risque croissante et décroissante.
Il existe d'autres lois avec des risques instantanés monotones. Citons notamment la loi Gamma et la loi de Gompertz.

2.3 Modèles multi-états

L'analyse de survie étudie le délai de survenue d'un évènement d'intérêt, ou le délai entre deux états successifs. Les modèles multi-états sont très utilisés ces derniers temps, permettent d'étudier des dynamiques plus complexes en utilisant

le notion de processus pour repésenter l'évolution d'un sujet à travers différents états successifs (Andersen et al, 1993 ; Hougaad, 1999). En épidémiologie, ils permettant de modéliser son évolution à travers les différents stades d'une maladie.

2.4 Modèle de Markov à temps continu

Définition 2.7. $\{X(t); 0 \leqslant t < \infty\}$ est une famille de variable alátoire où les valeurs prises par $X(t)$ sont des entiers$> 0 \in E = \{1, 2, ..., r\}$ pour $t_0 < t_1 < ..... < t_n < t_{n+1}$ nous avons la propriété markovienne suivante.

$$\mathbb{P}(X(t_{n+1}) = j/X(t_n), X(t_{n-1}), ..., X(t_0)) = \mathbb{P}(X(t_{n+1}) = j/X(t_n)) \tag{2.1}$$

pour simplifier l'écriture :

$$P_{ij}(t, t+s) = \mathbb{P}(X(t+s) = j/X(t) = i) \tag{2.2}$$

Classiquement la propriété suivante doit être respectée

$$\sum_j P_{ij}(t, t+s) = 1 \tag{2.3}$$

d'où

$$P_{ii}(t, t+s) = 1 - \sum_{j \neq i} P_{ij}(t, t+s) \tag{2.4}$$

Autrement dit, soit le processus reste dans le même état, soit il transite vers un autre état. Sous forme matricielle, ces probabilités de transition peuvent être notées :

$$P(t,t+s) = \begin{pmatrix} p_{11}(t,t+s) & p_{12}(t,t+s) & \dots & p_{1i}(t,t+s) \\ p_{21}(t,t+s) & p_{22}(t,t+s) & \dots & p_{2i}(t,t+s) \\ \vdots & \vdots & \vdots & \vdots \\ p_{i1}(t,t+s) & p_{i2}(t,t+s) & \dots & p_{ii}(t,t+s) \end{pmatrix} = (p_{ij}(t,t+s))_{i,j=1.....i}$$

A partir de la propriété markovienne nous pouvons écrire $\forall t, s > 0$

$$\begin{aligned} P_{ij}(0,t+s) &= \mathbb{P}(X(t+s) = j / X(0) = i) \\ &= \sum_k \mathbb{P}(X(t+s) = j / X(t) = k, X(0) = i) \\ &= \sum_k \mathbb{P}(X(t+s) = j / X(t) = k, X(0) = i).\mathbb{P}(X(t) = k / X(0) = i) \\ &= \sum_k P_{ik}(0,t) P_{kj}(t,t+s) \end{aligned}$$

Alors on aura la forme matricielle suivante :

$$P(0,t+s) = P(0,t)P(t,t+s) \tag{2.5}$$

Cette relation est appelée équation de *Chapman-Kolmogorov*.

Le paramètre d'intérêt en analyse de survie est **la force de transition** ou(**fct de risque instantané**), celle ci peut être définie pour $i \neq j$ comme suit :

$$\begin{aligned} S_{ij}(t) &= \lim_{\Delta t \to 0^+} \mathbb{P}(X(t+\Delta t) = j / X(t) = i)/\Delta t \\ &= \lim_{\Delta t \to 0^+} P_{ij}(t, t+\Delta t)/\Delta t \end{aligned} \tag{2.6}$$

De cette équation, on déduit que la quantité $S_{ij}(t) * \Delta t$ peut être considérée comme une approximation de la probabilité que le processus passe dans l'état j

entre t et $t + \Delta t$ conditionnellement au fait que ce processus soit dans l'état i en t ($i \neq j$). On remarque aussi que $S_{ij}(t)$ constitue la vitesse de transition de i vers j au temps t. Pour le cas où $i = j$, on définit $S_{ii}(t)$ à partir de la relation (2.4) :

$$\sum_{i \neq j} \mathbb{P}(X(t + \Delta t) = j / X(t) = i) = 1 - \mathbb{P}(X(t + \Delta t) = i / X(t) = i) \tag{2.7}$$

d'où

$$\sum_{i \neq j} \mathbb{P}(X(t + \Delta t) = j / X(t) = i)/\Delta t = 1 - \mathbb{P}(X(t + \Delta t) = i / X(t) = i)/\Delta t \tag{2.8}$$

On définit

$$\lim_{\Delta t \to 0^+} 1 - \mathbb{P}(X(t + \Delta t) = i / X(t) = i)/\Delta t = -S_{ii}(t) \tag{2.9}$$

On obteint alors

$$\sum_{i \neq j} S_{ij}(t) = -S_{ii}(t) \quad \text{et} \quad \sum_{j} S_{ij}(t) = 0 \tag{2.10}$$

2.5 Homogénéité et temps de séjour dans l'état

Dans les applications, le processus markovien est, le plus souvent considéré homogène. Les probabilités de transition sont alors définies par :

$$P_{ij}(t, t + s) = P_{ij}(0, s) = P_{ij}(s) \tag{2.11}$$

$P_{ij}(s)$ est indépendant de $t, \forall t \geqslant 0$ l'opérateur de *Chapman-Kolmogorov* peut s'écrire :

$$P(t + s) = P(t).P(s) \tag{2.12}$$

donc

$$\begin{aligned}
\frac{dP(s)}{ds} &= \frac{\lim_{\Delta s \to 0^+}(P(s+ds) - P(s))}{ds} \\
&= \frac{\lim_{\Delta s \to 0^+}(P(s)P(ds) - P(s))}{ds} \\
&= \frac{\lim_{\Delta s \to 0^+}(P(s)[P(ds) - I])}{ds} \\
&= P(s)\frac{\lim_{\Delta s \to 0^+}(P(ds) - I)}{ds} \\
&= P(s) * Q \qquad (2.13)
\end{aligned}$$

avec I, la matrice identité, et la matrice Q s'écrit :

$$Q = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1r} \\ q_{21} & q_{22} & \cdots & q_{2r} \\ \vdots & \vdots & \vdots & \vdots \\ q_{r1} & q_{r2} & \cdots & q_{rr} \end{pmatrix}$$

Définition 2.8. Générateur infinitésimal
La matrice Q, définie par :

$$Q(t) = \lim_{\Delta t \to 0^+} \frac{A(t, t+\Delta t) - I}{\Delta t}$$

est appelée Générateur infinitésimal, ou matrice de sauts, de la chaîne de Markov en temps continu.

Proposition 2.5.1. *Soit X un processus de Markov sur E. Il existe $Q \in \mathcal{Q}$ tel que pour tout $t \geqslant 0$*

$$A(t, t+\Delta t) = e^{\Delta t Q(t)} = \sum_{i=0}^{n} \frac{\Delta t^i}{i!} Q(t)^i$$

La matrice Q est le générateur infinitésimal du semi-groupe $A(t, t+\Delta t)$ c'est-à-dire que

$$\lim_{\Delta t \to 0^+} A(t, t+\Delta t) = I$$

est tout simplement la fonction exponentielle :

$$P(t) = \exp(Qt)$$

avec $P(0) = I$
Il est simple de définir l'exponentielle d'une matrice carrée.

Définition 2.9. Exponentielle de matrice
La série $\sum_{i=0}^{n} \frac{Q^i}{i!}$ est convergente dans l'espace des matrices. Sa somme est appelée exponentielle de la matrice Q :

$$\exp(Q) = e^Q = \sum_{i=0}^{\infty} \frac{Q^i}{i!}.$$

L'existence de l'exponentielle d'une matrice ne pose aucun problème lorsque celle-ci est de taille finie. On peut aussi la justifier lorsque Q est de taille infinie, mais sous certaines conditions qui seront vérifiées par la suite.

Propriété 1. *-Si Q est diagonale, à coefficients diagonaux, alors* $\exp^Q$ *est diagonale, à coefficients diagonaux*

avec

$$\Lambda(t) = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix}$$

-Si Q est diagonalisable, sous la forme $Q(t) = C\Lambda(t)C^{-1}$, alors $e^Q = C \exp^{\Lambda(t)} C^{-1}$.
- Si Q et M sont deux matrices telles que $QM = MQ$, alors : $e^{Q+M} = e^Q e^M = e^M e^Q$.

Démonstration. -Si Q est diagonale, à coefficients diagonaux $\lambda_0, \lambda_1, ...$, alors pour tout i, Q^i est diagonale, de coefficients diagonaux $\lambda_0^i, \lambda_1^i, ...,$. Ainsi pour tout $n \geqslant 0$, $\sum_{i=0}^n \frac{Q^i}{i!}$ est une matrice diagonale, de coefficients diagonaux $\sum_{i=0}^n \lambda_0^i, \sum_{i=0}^n \lambda_1^i, ...$, et on reconnaît des sommes partielles de séries exponentielles :

$$\forall p \in E \sum_{i=0}^{n} \frac{\lambda_p^i}{i!} \to_{n\to\infty} e^{\lambda(p)}.$$

Ainsi e^Q est diagonale, à coefficients diagonaux $e^{\lambda_0}, e^{\lambda_1,...,}$ -Si Q est diagonale, sous la forme $Q = C\Lambda^i(t)C^{-1}$, alors pour tout i, on a :

$$Q^i = C\Lambda^i(t)C^{-1}$$

d'où il vient :

$$\sum_{i=0}^{n} \frac{Q^i}{i!} = C\left(\sum_{i=0}^{n} \frac{\Lambda^i}{i!}\right)C^{-1} \to_{n\to\infty} Ce^{\Lambda}C^{-1},$$

c'est-à-dire :

$$e^Q = Ce^{\Lambda}C^{-1}$$

-Si Q et M commutent, le résultat se prouve de la même façon que pour l'exponentielle de nombres réels ou complexes :

$$\left(\sum_{i=0}^{n}\frac{Q^i}{i!}\right) * \left(\sum_{i=0}^{n}\frac{M^i}{i!}\right) = \left(\sum_{i=0}^{n}\frac{1}{i!}\right)\left(\sum_{k=0}^{j}C_j^k Q^k B^{j-k}\right).$$

Mais puisque Q et M commutent, on peut appliquer le binôme de Newton :

$$\sum_{k=0}^{j}C_j^k\frac{Q^k}{k!}\frac{M^{j-k}}{(j-k)!} = (Q+M)^i,$$

ce qui donne dans l'expression original :

$$\left(\sum_{i=0}^{n}\frac{Q^i}{i!}\right) * \left(\sum_{i=0}^{n}\frac{M^i}{i!}\right) = \sum_{i=0}^{n}\frac{(Q+B)^i}{i!}$$

Il reste alors à faire tendre n vers ∞ : le membre de gauche tend vers $e^Q e^M$ tandis que celui de droite tend vres e^{Q+M}, d'où l'égalité. Et puisque $Q+M = M+Q$, notre résultat est atteint. □

2.6 Censure et troncature

En général, pour une suite de données $t_1,\ t_2,\ ...\ t_n$, réalisation de la suite de v.a. $T_1,\ T_2,\ ...\ T_n$ (suites de durées dans un état du système), l'observateur n'accède pas à tous les T_i, $i = 1, ..., n$ à cause d'un empêchement, d'un arrêt des observations, d'une complication à l'accès à l'information, ... Le statisticien utilise un modèle pour chaque situation. Pour les modèles de durée, les plus utilisés sont généralement ceux de censure et ceux de troncatures.

La censure correspond à l'introduction d'une variable concurrente C qu'on appelle variable de troncature. Cette variable peut être fixe ou aléatoire. L'analyse des données censurées consiste à étudier la loi de la durée T et C en tenant

compte des données non observées. La troncature agit comme la censure, mais ne prend pas en considération les données non observées (c'est-à-dire ayant subi la troncature) ; la loi de la durée n'est, dans ce dernier cas, qu'une loi conditionnelle sur la partie non tronquée.

Une donnée est dite censurée si elle n'est pas observée au delà d'un certain seuil (prédéfini ou aléatoire), mais que l'information que le phénomène n'a pas été observé au delà du seuil est prise en considération. Si l'information que le phénomène n'a pas été observé au delà du seuil n'est pas prise en considération, on est devant le cas de données tronquées.

2.6.1 Types de censures

Il existe plusieurs types de censures ; nous en citons :

La censure à droite : Le modèle de censure à droite nécessite la construction d'une variable aléatoire $U_t = T_t \wedge C$

Définition 2.10. On appelle donnée censurée à droite, une réalisation de T_t pour laquelle $T_t > C$. Sinon, la donnée est dite non censurée.

En général U_t est notée T_t si $T_t \leqslant C$ et T_t^* si $T_t > C$.

La censure à gauche : Le modèle de censure à gauche nécessite la construction d'une variable aléatoire $V_t = T_t \vee C$

Définition 2.11. On appelle donnée censurée à gauche, une réalisation de T_t pour laquelle $T_t < C$. Sinon, la donnée est dite non censurée.

En général V_t est notée T_t si $T_t \geqslant C$ et T_t^* si $T_t < C$.

La censure à droite et à gauche : Le modèle de censure à droite (par C_d) et à gauche (par C_g) nécessite la construction d'une variable aléatoire $W_t = (T_t \wedge C_d) \wedge (T_t \vee C_g) =\equiv \wedge (T_t \wedge C_d), (T_t \vee C_g))$

Définition 2.12. On appelle donnée censurée à droite et à gauche, une réalisation de T_t pour laquelle $T_t < C_g$ ou $T_t > C_d$. Sinon, la donnée est dite non censurée.

En général W_t est notée T_t si $C_g \leqslant T_t \leqslant C_d$ et T_t^* sinon.

La censure par intervalles : Le modèle de censure par intervalles $\left(]C_g^{(i)}, C_d^{(i)}[\right)_{i=1,...,k}$ nécessite la construction d'une variable aléatoire $Z_t = \wedge_{i=1,...,k} \left((T_t \wedge C_g^{(i)}), (T_t \vee C_g^{(i)})\right)$. Quand k=1, nous retombons sur le cas qui précède.

Définition 2.13. On appelle donnée censurée par intervalles, une réalisation de T_t pour laquelle T_t est hors de tous les intervalles $\left(]C_g^{(i)}, C_d^{(i)}[\right)_{i=1,...,k}$. Sinon, la donnée est dite non censurée.

En général Z_t est notée T_t, si T_t est dans un des intervalles $]C_g^{(i)}, C_d^{(i)}[$, $i = 1, ..., k$, et T_t^* sinon.

Quand la censure est déterministe (fixe), on dit que la **censure de type I**; dans ce cas, tous les T_i^* sont égaux.

La **censure de type II** (ou censure avec attente) : est caractérisée par le nombre de fois où l'évènement d'intérêt s'est réalisé. Si r est le nombre choisi, en ordonnant les T_i sous la forme $T_{(1)} \leqslant T_{(2)} \leqslant ..... \leqslant T_{(n)}$, toutes les données $T_{(r+j)}$, $j = 1, ..., n - r$ sont censurées et l'on a $T^*_{(r+1)} = T^*_{(r+2)} = ... = T^*_{(n)}$.

La **censure de type III** (ou censure aléatoire) : considère que la censure (C ou $C_g^{(i)}$ ou $C_d^{(i)}$) est une v.a. ayant une loi (connue ou inconnue) et dépendante

ou indépendante de T. Cela permet de construire des modèles intéressants qui s'adaptent de manière pertinente aux problèmes posés. Dans le cas de censure de type III, on considère pour le cas de censure à droite qui peut être généralisé aux autres censures (à gauche et par intervalle(s) sans grande difficulté).

Au lieu d'étudier directement les T_i et les T_j^* $(i,\, j \in \{1,\, ...,\, n\})$, on étudie la suite de couples $((U_1,\, D_1), ..., (U_n\, D_n))$ où

$$U_i = T_i \wedge C_i \qquad \text{et} \qquad D_i = \begin{cases} 1 \text{ si } U_i \leqslant C_i \\ 0 \text{ sinon} \end{cases}$$

qui est plus simple à exploiter théoriquement. Ainsi, en statistique paramétrique, si les observations sont indépendantes et C indépendante de T, la vraisemblance s'écrit facilement, en conditionnant par rapport à la censure

$$L((U_1, D_1), ..., (U_n, D_n);\, \theta) = \prod_{i=1}^{n} [f_T(U_i;\theta) S_C(U_i;\theta)]^{D_i} [f_C(U_i;\theta) S_T(U_i;\theta)]^{1-D_i}$$

Dès que θ est bien estimé, la construction de la fonction de survie devient aisée.

En introduisant une covariable scalaire ou vectoriellle R pour expliquer le phénomène, nous considérons la suite de triplets $((U_i, D_i, R_i))_{i=1,...,n}$ pour construire la vraisemblance afin d'estimer θ. Dans certains cas particuliers de conditions sur R, sur C et sur θ, des résultats intéressants peuvent être obtenus. Mais ces résultats ont plus un impact théorique que pratique.

Cette approche peut être étendue à différents types et formes de censures avec leur lots de complexités du problème.

C'est pour cette raison que beaucoup d'auteurs s'intéressent à l'aspect non paramétrique ou semi-paramétrique qui permettent d'aborder le problème avec plus de facilité, mais convergeant lentement en général.

2.6.2 Troncature

On dit qu'il y a troncature, quand la variable d'intérêt n'est prise en considération que sur une partie de la durée des observations. Si la partie est connexe, nous avons trois possibilités qui se présentent :

- troncature à gauche : toutes les observations inférieures à une valeur r sont ignorées ;

- troncature à droite : toutes les observations supérieures à une valeur R sont ignorées ;

- troncature à gauche et à droite : toutes les observations inférieures à une valeur r ou supérieures à une valeur R sont ignorées ;

Si la partie est non connexe, on dit que la troncature est par intervalles.

La troncature est différente de la censure, car, dans ce cas, on perd complètement l'information sur les observations initiales, mais perdues le long de l'expérience.

En prenant l'exemple sur des données tronquées à droite par R et à gauche par r, la fonction de survie s'écrit

$$S(t \mid r \leqslant t < R) \begin{cases} 1 & \text{si } t > r \\ \dfrac{S(t) - S(R)}{S(r) - S(R)} & \text{si } r \leqslant t < R \\ 0 & \text{si } t \geqslant R \end{cases}$$

la fonction du hasard associée s'écrit facilement :

$$h(t \mid r \leqslant t < \ R) = h(t).\frac{S(t)}{S(t) - S(R)}$$

Il est à noter qu'il existe aussi des modèles où la troncature et la censure sont exploités simultanément pour étudier des cas pratiques

2.7 Estimation non paramétrique

La statistique non paramétrique concerne généralement les techniques statistiques qui ne dépendent pas de données appartenant à une distribution objet d'une hypothèse particulière. Celles-ci comprennent, notamment, les méthodes libres des distributions qui ne se reposent sur aucune hypothèse selon laquelle les données sont tirées d'une famille de distributions de probabilité donnée. Contrairement à la statistique paramétrique.

L'approche paramétrique induit des hypothèses sur la distribution des données mais elle a l'avantage de fournir des estimations en temps continu de n'importe quelle fonction caractérisant la distribution. L'approche non paramétrique fournit des estimations en temps discret et donc les fonctions estimées sont continues par morceaux. La fonction de risque étant la plus intéressante en terme d'interprétation, un lissage a posteriori de l'estimateur de Nelson-Aalen est envisageable en utilisant une méthode à noyau (Ramlau-Hansen, 1983). Une autre approche pour estimer des fonctions lisses sans faire d'hypothèse paramétrique est d'utiliser des fonctions splines (Rosenberg, 1995). Ce type d'approche et l'approche paramétrique, étant basés sur la vraisemblance, ont l'avantage de prendre en compte aisément des données censurées par intervalle.

Dans le cas de données i.i.d., l'estimateur de la f.r. de T est donné par l'expression

$$F_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1_{[0,\, t[}(T_i)$$

qui converge uniformément presque sûrement vers la f.r. de T : $n.F(t)$. Cet estimateur a de bonnes propriétés : il est sans biais et asymptotiquement gaussien. $n.F_n(t)$ est distribuée selon une loi binomiale de moyenne $n.F(t)$ et de variance $n.F(t).(1 - F(t))$. Plus généralement, on montre que $\sqrt{n}\,(F_n - F)$ converge vers un pont brownien. Ainsi, on peut utiliser asymptotiquement toutes les propriétés

du pont brownien sur F_n. Cette manière de faire est une approche non paramétrique. La recherche de la solution peut être faite dans des espaces fonctionnels de dimension infinie.

En général, moyennant des hypothèses supplémentaires, nous pouvons exploiter une autre approche de la f.r. empirique. Par exemple, si F est dérivable, F_n ne restitue cette propriété qu'asymptotiquement, aussi certain auteurs utilisent des noyaux régularisants pour obtenir des caractéristiques de régularité même pour de faibles échantillons. Toutefois, si les propriétés de régularité de la f.r. ne sont pas nécessaires pour l'étude, les techniques classiques peuvent donner des résultats exploitables et intéressants. Pour ce faire, nous présentons, à titre indicatif, l'approche de Kaplan-Meier sur la fonction de survie.

2.8 Estimateur de Kaplan-Meier

La fonction de survie S est inconnue et elle caractérise la loi de T. On va construire un estimateur en utilisant les observations

Kaplan et Meier (1958), ont proposé un estimateur de S nommé aussi estimateur produit-limite. Il repose sur l'idée suivante : un "individu" est en vie après l'instant t, c'est être en vie juste avant l'instant t et ne pas "mourir" en t. Cette idée se traduit par les relations suivantes

$$\begin{aligned} S(t) &= \mathbb{P}(T > t) \\ &= \mathbb{P}(T > t/T > t-1) * \mathbb{P}(T > t-1) \\ &= ... \\ &= \mathbb{P}(T > t/T > t-1) * ...\mathbb{P}(X > t'/X > t^{"}) * \mathbb{P}(T > 0) \end{aligned}$$

Si l'on choisit les instants de conditionnement au tepms de la production d'un évènement $t_{(i)}$ (mort, panne ou censure...) nous estimons des quantités de la forme :

$$\mathbb{P}(T > t_{(i)}/T > t_{(i-1)}) = p_i$$

où les $t_i < t_{(i-1)}$ et p_i est la probabilité de survivre pendant l'intervalle de temps $I_i =]t_{(i-1)}, t_{(i)}]$ sachant qu'on était était "vivant" au début de cet intervalle. Notons comme précédement, r_i le nombre des sujets qui sont "vivants" (donc à risque) juste avant l'instant $t_{(i)}$ et m_i le nombre d'évènements à l'instant $t_{(i)}$. Or $q_i = 1 - p_i$ est la probabilité de la survenue de l'évènement durant I_i. Un estimateur naturel de q_i est la fréquence

$$\hat{q}_i = \frac{M_i}{R_i} \tag{2.14}$$

Si on suppose qu'il n'y ait pas d'ex-aequo (plusieurs pannes en $t_{(i)}$) : si $D_i = 1$ c'est qu'il y a eu "évènement" en $t_{(i)}$ et donc $m_i = 1$, $D_i = 0$, c'est qu'il y a eu une censure en $t_{(i)}$ et donc $m_i = 0$. Par suite

$$\hat{p}_i = \begin{cases} 1 - \hat{q}_i = 1 - \frac{1}{r_i} & \text{si} D_i = 1 \\ 1 & \text{sinon} \end{cases}$$

On au $r_i = n - (i-1)$ (car il y'a eu $(i-1)$ "évènements" ou censures avant $t_{(i)}$ ou censure avant $t_{(i)}$ et il y'a n individus dans l'étude). L'estimateur de Kaplan-Meier dans ce cas est

$$\hat{S}_{KM}(t) = \prod_{X_{(i)} \leqslant t} \left(1 - \frac{1}{n-i+1}\right)^{D_{(i)}} \tag{2.15}$$

Remarque 2.8.1.

- L'estimateur de Kaplan-Meier est une fonction en escaliers qui fait des sauts à chaque instant t_i. La valeur du saut dépend du nombre d'évènements au temps t_i et aussi du nombre de censures à ce temps là.
- Traitement des ex-aequo : Dans le cas d'existence des ex-aequo on n'a plus m_i égale à 1 en $t_{(i)}$ mais au nombre d'évènements en $t_{(i)}$ et aussi on n'a plus $r_i = n - (i-1)$. Dans ce cas on doit garder r_i et m_i et l'estimateur de

Kaplan-Meier devient

$$\hat{S}_{KM}(t) = \prod_{X_{(i)} \leqslant t} \left(1 - \frac{M_i}{R_i}\right)^{D_{(i)}} \tag{2.16}$$

– L'estimateur de Kaplan-Meier est aussi l'estimateur du maximum de vraisemblance non paramétrique de la survie dans l'espace des fonctions de répartition (de dimension infinie)

Exemple On obseve les durées de vie de 10 appareils exprimées en mois : **1,3,4+,5,7+,8,9,10+,11,13+**

L'estimateur de Kaplan-Meier de la fonction de survie $S(t)$ calculé par la formule (1.15) vaut :

TABLE 2.1 – L'estimateur de Kaplan-Meier de la fonction de survie $S(t)$.

Temps	r_i	m_i	$\hat{S}_{KM}(t)$	Intervalle
0	10	0	1	[0,1[
1	10	1	(1-1/10)$\hat{S}(0)$ = 0,900	[1,3[
3	9	1	(1-1/9)$\hat{S}(1)$ = 0,800	[3,5[
5	7	1	(1-1/7)$\hat{S}(3)$ = 0,686	[5,8[
8	5	1	(1-1/5)$\hat{S}(5)$ = 0,589	[8,9[
9	4	1	(1-1/4)$\hat{S}(8)$ = 0,411	[9,11[
11	2	1	(1-1/2)$\hat{S}(9)$ = 0,206	[11,∞[

2.8.1 Propriétés de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier a des propriétés analoques à celles de la fonction de répartition empirique, en particulier il vérifie un théorème de normalité asymptotique globale. Mais il a aussi d'autres propriétés qui, elles sont typiques

de la présence de censure et ont l'intérêt de donner des idées lorsqu'on cherche à construire d'autres procédures d'estimation quand il y a de la censure.

2.8.2 Cohérence de l'estimateur de Kaplan-Meier

Définition 2.14. Un estimateur $\hat{SC}$ de la fonction de survie S est dit cohérent si pour t de R^+ il vérifie

$$\hat{SC}(t) = \frac{1}{n}\left[\sum_{i=1}^{n} I_{(T_i \geqslant t)} + \sum_{i=1}^{n} I_{(T_i<t, D_i=0)} \frac{\hat{SC}(t)}{\hat{SC}(T_i)}\right]$$

On remarque que cette équation est implicite car elle comporte $\hat{SC}$ dans les deux membres et pour différentes valeurs de l'argument. Elle est constituée de la manière suivante : les survivants au-delà de l'instant t sont ceux qui n'ont été avant cette date ni morts, ni censurés, et dont l'effectif constitue la première somme dans le crochet, et d'autre part ceux qui, ayant été censurés à l'instant T_i antérieur à t, survivent au-delà de t avec la probabilité conditionnelle

$$\frac{\hat{SC}(t)}{\hat{SC}(T_i)}$$

qui pondère chacun d'eux

Théorème 2.4. *L'estimateur de Kaplan-Meier est l'unique estimateur cohérent de la fonction de survie S*

Démonstration. Supposons qu'il n'y ait pas d'ex-aequo. La condition de cohérence s'écrit de manière suivante :

$$\hat{SC}(t)\left[n - \sum_{T_i<t}\left[\frac{(1-D_i)}{\hat{SC}(T_i)}\right]\right] = \sum_{i=1}^{n} I_{(T_i \geqslant t)} \qquad (*)$$

– Si $t \leqslant T_1, \hat{SC}(t) = 1 = \hat{S}_{KM}(t)$.

– $\hat{SC}$ est constante entre deux temps consécutifs, $]T_i, T_{j+1}]$, comme $\hat{S}_{KM}$.

– Il faut montrer que les sauts en T_j de $\hat{SC}$ et $\hat{S}_{KM}$ sont égaux.

1^{er} **Cas :** $D_j = 0$

Alors le membre de droite de (*) saute de -1 et le saut du membre de gauche est égal à :

$$-\hat{SC}(T_j)\frac{1}{\hat{SC}(T_j)} = -1$$

du fait que $\hat{SC}$ ne saute pas en T_j.

Il n'y a pas de saut de $\hat{SC}$ en un point de censure, c'est le cas de $\hat{S}_{KM}$

2^{me} **cas :** $D_j = 1$

Le membre de droite de saute de -1 une autre fois de

$$\hat{SC}(T_j^+)\left[n - \sum_{T_i \leqslant T_j}\left[\frac{(1-D_i)}{\hat{SC}(T_i)}\right]\right] = n - j$$

$$\hat{SC}(T_j)\left[n - \sum_{T_i \leqslant T_j}\left[\frac{(1-D_i)}{\hat{SC}(T_i)}\right]\right] = n - j + 1$$

On note que le facteur multipliant $\hat{SC}(T_j)$ dans la seconde expression est identique que T_i soit inférieur à T_j au sens large. Donc :

$$\frac{\hat{SC}(T_j^+)}{\hat{SC}(T_j)} = \frac{n-j}{n-j+1}$$

□

2.8.3 L'estimateur de Kaplan-Meier est GMLE pour S

Un estimateur GMLE est un estimateur du maximum de vraisemblance généralisé, défini ci-dessous.

Définition 2.15. Soit $\mathcal{P}$ une famille de probabilité sur $\mathbb{R}^n$ muni de la tribu borélienne $\mathbf{B}$, supposée non dominés. Etant donné un élément x de $\mathbb{R}^n$ et deux élés P_1 et P_2 de $\mathcal{P}$, soit $f(x; P_1, P_2) = \frac{dP_1}{d(P_1+P_2)}(x)$. On dit que $\hat{P}$ est GMLE pour la probabilité P qui gouverne X si

$$f(x; \hat{P}, P) \geqslant f(x; P, \hat{P}) \qquad \forall P \in \mathcal{P}$$

Théorème 2.5. *Si les lois de la survie Y et de la censure C sont diffusés et si $\mathcal{P}$ est n'importe quelle famille de probabilités qui contienne les probabilités chargeant l'observation $(T_i, D_i), i = 1, 2, ..., n$, alors $\hat{S}_{KM}$ est GMLE pour S.*

Démonstration.
– Il suffit de considérer dans $\mathcal{P}$ les probabilités qui donnent une probabilité strictement positive à l'observation $(T_1, D_1), ..., (T_n, D_n)$. En effet, si ce n'est pas le cas, le second membre de l'inégalité ci-dessus est nul.
– pour une telle P de $\mathcal{P}$ noterons :

$$p_i = P(Y \in]T_i, T_{i+1}]) \qquad i = 0, 1, 2, ..., n$$
$$T_0 = 0 et T_{n+1} = \infty$$

$$P((T_1, D_1), ..., (T_n, D_n)) = \prod_{i=1}^{n} [P(Y = T_i)]^{D_i} [P(Y \geqslant T_i)]^{1-D_i}$$

Si les p_i sont fixés, cette probabilité est maximisée en prenant :

$$\hat{P}(Y = T_i / T_i = 1) = p_i$$
$$\hat{P}(T_i < Y \leqslant T_{i+1} / D_i = 0) = p_i$$

$$V_{Max} = \prod_{i=1}^{n} \left[p_i^{D_i} (\sum_{j=1}^{n} p_j)^{1-D_i} \right]$$

– Quels sont les p_i qui maximisent V ?
Ce sont les :

$$\hat{p}_i = \left[\prod_{j=1}^{i-1} (1 - \frac{D_j}{n-j+1}) \right] \frac{D_i}{n-i+1}$$

Ce qui correspond à l'estimateur de Kaplan-Meier.
En effet, posons

$$\lambda_i = \frac{p_i}{\sum_{j=1}^{n} p_j}$$

(en fait λ_i) est le taux de mortalité au ième rang).
Alors

$$1 - \lambda_i = \frac{\sum_{j=i+1}^{n} p_j}{\sum_{j=1}^{n} p_j}$$

et par suite

$$\sum_{j=1}^{n} p_j = \prod_{j=1}^{i-1} (1 - \lambda_j)$$

Ecrivant la vraisemblance en fonction des λ_i on obtient :

$$\begin{aligned} V &= \prod_{i=1}^{n}[\lambda_i^{D_i}\prod_{j=1}^{i-1}(1-\lambda_j)] \\ &= \prod_{i=0}^{n-1}\lambda_i^{D_i}(1-\lambda_i^{n-j}). \end{aligned}$$

Chaque facteur est maximisé pour

$$\hat{\lambda}_i = \frac{D_i}{n-i+D_i} = \frac{D_i}{n-i+1}$$

ce qui donne

$$\begin{aligned} \hat{p}_i &= \hat{\lambda}_i(\sum_{j=i}\hat{p}_j) \\ &= \hat{\lambda}_i(\prod_{j=1}^{i-1}(1-\hat{\lambda}_j)) \\ &= \frac{D_j}{n-i+1}\prod_{j=1}^{i-1}(1-\frac{D_i}{n-j+1}). \end{aligned}$$

□

2.8.4 Consistance de l'estimateur de Kaplan-Meier :

Pour que l'estimateur de Kaplan-Meier $\hat{S}_{KM}$ de la fonction de survie, soit consistant la survie Y et la censure C n'aient aucune discontinuité commune. Notons F, G et ST respectivement les fonctions de répartition de la survie Y, de la censure C et de la durée observée $T = Y \wedge C$ (les Y_i et les C_i sont mutuellement indépendants). On adoptera aussi une notation particulière pour les deux

fonctions du temps suivante :

$$ST_0(t) = \mathbb{P}(T \geqslant t, D = 0)$$
$$ST_1(t) = \mathbb{P}(T \geqslant t, D = 1)$$

Alors

$$\begin{aligned} ST(t) &= ST_0(t) + ST_1(t) = \mathbb{P}(T \geqslant t) \\ &= (1 - F(t))(1 - G(t)) \end{aligned}$$

Théorème 2.6. *Si les lois de la survie Y et de la censure C n'ont aucune discontinuité commune, l'estimateur de Kaplan-Meier $\hat{S}_{KM}$ de la fonction de survie $S(t) = \mathbb{P}(T \geqslant t)$ est un estimateur consistant.*

Démonstration. Le principe de la démonstration est :

- On démontre que la fonction de survie S s'exprime en fonction de ST_0 et ST_1 : $S(t) = W(ST_0, ST_1, t)$.
- Or, l'etimateur de Kaplan-Meier de S s'exprime de la même façon en fonction des équivalents empiriques $\hat{ST_0}$ et $\hat{ST_1}$ de ST_0 et ST_1 :
  $\hat{S}_{KM}(t) = W(\hat{ST_0}, \hat{ST_1}, t)$ et $\hat{ST_0}$ et $\hat{ST_1}$ obéissent au théorème de Glivenko-Cantelli

Lemme 2.8.1. *On suppose que les f.r. F de Y (durée) et G de C (censure) n'ont aucune discontinuité commune. Alors*

$$S(t) = \exp\left\{ \int_0^t \frac{d(ST_1)}{ST_1 + ST_0} + \sum_{u \leqslant t} Log\left[\frac{ST_0(u^+) + ST_1(u^+)}{ST_0(u^-) + ST_1(u^-)} \right] \right\}$$

Le premier terme dans l'accolade correspond à une intégration sur les intervalles de continuité de ST_1. Le second terme est une sommation sur les points de discontinuité de ST_1.On note W cette fonctionnelle.

– On suppose pour commencer que ST_1 est continue en t :

$$\begin{aligned}\int_0^t \frac{dST_1(u)}{ST_1(u)+ST_0(u)} &= \int_0^t \frac{-(1-G(u))dF(u)}{(1-F(u))(1-G(u))} \\ &= \int_0^t \frac{-dF(u)}{1-F(u)} = [Log[1-F(u)]]_0^t \\ &= LogS(t)\end{aligned}$$

En effet : $ST_0(t) = \mathbb{P}(T \geqslant t, D = 0) = \mathbb{P}(Y \geqslant C \geqslant t) = \int_t^\infty (1-F)dG$

$ST_1(t) = \mathbb{P}(T \geqslant t, D = 1) = \mathbb{P}(C \geqslant Y \geqslant t) = \int_t^\infty (1-G)dF$ Par suite, en cas de continuité de ST_1 :

$$S(t) = \exp\left[\int_0^t \frac{dST_1}{ST_1+ST_0}\right].$$

Supposons que ST_1 saute en t, mais que ST_0 est continue en t :

$$\begin{aligned}Log\frac{ST_1(t^+)+ST_0(t^+)}{ST_1(t^-)+ST_0(t^-)} &= Log\frac{ST(t^+)}{ST(t^-)} \\ &= \frac{(1-F(t^+))(1-G(t^+))}{(1-F(t^-))(1-G(t^-))} \\ &= Log\frac{S(t^+)}{S(t^-)}\end{aligned}$$

Puisque $G(t^+) = G(t^-)$

Donc :

$$S(t^+) = S(t^-)\exp\left[Log\frac{ST_1(t^+)+ST_0(t^+)}{ST_1(t^-)+ST_0(t^-)}\right].$$

Lemme 2.8.2. *Si on note* $\hat{ST_0}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{T_(1\geqslant t, D_1=0)}$
$\hat{ST_1}(t) = \frac{1}{n}\sum I_{T_(1\geqslant t, D_1=1)}$ *Alors*

$$\hat{S}_{KM}(t) = W(\hat{ST_0}, \hat{ST_1}, t).$$

- On rappelle qu'on a toujours décidé lorsqu'il y avait , parmi des ex-aequo, à la fois des censures et des morts de faire précéder les censures par les morts.
- Comme $\hat{ST_1}$ est complètement discrète,il y a que la sommation qui intervient dans W,

Finalement on a le résultat énoncé dans le théorème qui est l'équivalent de Glivenko-Cantelli en présence de censure :

$$\hat{S}_{KM} \longrightarrow^{p.s} S \qquad \text{uniformément en t.}$$

En effet :
D'après le théorème de Glivenko-Cantelli :

$$\hat{ST_0} \longrightarrow^{p.s} ST_0 \qquad \text{uniformément en t.}$$

$$\hat{ST_1} \longrightarrow^{p.s} ST_1 \qquad \text{uniformément en t.}$$

Comme W est une fonction continue de ST_0 et ST_1 pour la norme sup :

$$\hat{S}_{KM}(t) = W(\hat{ST_0}, \hat{ST_0}, t) \longrightarrow^{p.s} \qquad W(ST_0, ST_1, t) = S(t) \qquad \text{uniformément en t.}$$

□

2.8.5 Normalité asymptotique

Le théorème de Donsker peut s'étendre au cas où une censure aléatoire droite est présente.

Théorème 2.7. *Si la survie Y de fonction de répartition $F = 1-S$, et la censure C, de fonction de répartition G, sont indépendantes, et si F et G n'ont aucune discontinuité commune*

$$\sqrt{n}(\hat{S}_{KM} - S) \quad \longrightarrow^{L} \quad Z$$

où Z est un processus gaussien centré, de la fonction de covariance

$$Cov(Z(t_1), Z(t_2)) = S(t_1)S(t_1) \int_0^{t_1 \wedge t_2} \frac{dF(u)}{[1 - F(u)]^2[1 - G(u)]}$$

Dans le cas non censuré, qui correspond à $1 - G(u) = 1$, redonne le résultat habituel. Le principe de la démonstration repose sur l'expression de Peterson de la survie. On fait un développement de la différence $\hat{S}_{KM} - S$, dont on ne garde que les termes principaux.

Si par exemple, la survie est exponentielle le paramètre λ et la censure fixe égale à c, la fonction de covariance vaut

$$e^{-\lambda(t_1+t_2)} \int_0^{t_1 \wedge t_2} \frac{\lambda e^{-\lambda u}}{e^{-2\lambda u} I_{u \leqslant c}} du = e^{-\lambda(t_1+t_2)}[e^{-\lambda(t_1 \wedge t_2 \wedge c)} - 1].$$

Si la survie et la censure sont toutes les deux exponentielle, de paramètre respectifs ν et μ

$$cov(Z(t_1), Z(t_2)) = e^{-\lambda(t_1+t_2)}[e^{(\lambda+\mu)(t_1 \wedge t_2)} - 1].$$

Corollaire 2.8.1.

$$L(\hat{S}_{KM}(t)) \equiv_{n\to\infty} N\left(S(t); \frac{S^2(t)}{n} \int_0^t \frac{dST_1(u)}{[ST(u)]^2}\right)$$

Du théorème précédent ; on a pour toute fonction f continue pour la norme sup, $f(Z_n)$ converge en loi vers $f(Z)$. En particulier

$$\hat{S}_{KM,n} = S + \frac{Z_n}{\sqrt{n}}.$$

Pour pouvoir utiliser ce corollaire, il faut avoir un estimateur de la variance asymptotique de $\hat{S}_{KM}$. Cette variance asymptotique fait intervenir trois fonctions, S, ST et ST_1 pour lesquelles nous avons déjà des estimateurs naturels :

$ST(t) = \mathbb{P}(T \geqslant t)$ $\qquad$ $\hat{ST}(t) = \frac{1}{n}\sum_{i=1}^{n} I_{T_i \geqslant t}$

$ST_1(t) = \mathbb{P}(T \geqslant t, D = 1)$ $\qquad$ $\hat{ST_1}(t) = \frac{1}{n}\sum_{i=1}^{n} D_i I_{T_i \geqslant t}$

$S(t) = \mathbb{P}(T \geqslant t)$ $\qquad$ $\hat{S}_{KM}(t) = \prod_{T_i<t}(1 - \frac{M(T_i)}{R(T_i)})$.

Comme $d\hat{ST_1}$ est nul, sauf aux instants de mort, on peut écrire, dans le cas où il n'y a pas d'ex-aequo :

$$\begin{aligned} d\hat{ST_1}(u) &= 0 \quad \text{si} \quad u \neq T_i \\ &= \frac{D_i}{n} \quad \text{si} \quad u = T_i \end{aligned}$$

$$1 - \hat{ST}(T_i^+) = \frac{1}{n}\sum_{j=1}^{n} I_{T_j \leqslant T_i}$$

$$1 - \hat{ST}(T_i^-) = \frac{1}{n}\sum_{j=1}^{n} I_{T_j < T_i}$$

Si on estime $[1 - ST(u)]^2$ par $[1 - ST(u^-)]\,[1 - ST(u^+)]$, on obteint pour estimateur de la variance asymptotique de $\hat{S}_{KM}$ l'estimateur de Greenwood.

Définition 2.16. Estimateur de Greenwood Si les observations censurées sont les $(T_i, D_i), i = 1, 2, ..., n$, avec éventuellement des ex-aequo, et $T_1^{'} < T_2^{'} < ... < T_k^{'}$ la suite des T_i réordonnés par ordre croissant, M_i le nombre des morts à l'instant $T_i^{'}$, R_i le nombre des sujets à risque à $T_i^{'}$ l'estimateur de Greenwood de la variance de $\hat{S}_{KM}$ est défini comme

$$\widehat{VarS_{KM}}(t) = \hat{S}_{KM}^2(t) \sum_{T_i^{'} \leqslant t} \frac{M_i}{R_i(R_i - M_i)}$$

En particulier, s'il n'y a pas d'ex-aequo :

$$\widehat{VarS_{KM}}(t) = \hat{S}_{KM}^2(t) \sum_{T_i \leqslant t} \frac{D_i}{R_i(n-i)(n-i+1)}$$

Remarque 2.8.2. On suppose toujours que, s'il y avais des ex-aequo, il étaient de même nature ; quand la mort et une censure sont ex-aequo, nous supposons que la mort précéde la censure.

Remarque 2.8.3. L'estimateur de Greenwood a d'abord été obtenu grâce aux considérations suivantes : $Log\hat{S}_{KM}(t) = \sum_{T_i \leqslant t} Log\hat{p}_i$. Les $\hat{p}_i$ ne sont pas indépendantes, mais si on en fait la conjecture, on a $Var(Log\hat{S}_{KM}(t)) = \sum_{T_i \leqslant t} Var(Log\hat{p}_i)$. Comme la loi de $R_i\hat{p}_i$ est la loi binomiale de paramètre R_i et p_i :

$$\begin{aligned} Var(Log\hat{p}_i) &\equiv Var(\hat{p}_i)\left(\frac{d}{dp_i}(Logp_i)\right)^2 \\ &= \frac{p_i q_i}{R_i}\frac{1}{p_i^2} = \frac{q_i}{R_i p_i}. \end{aligned}$$

On a finalement l'approximation suivante :

$$Var\left(Log\hat{S}_{KM}(t)\right) \equiv \sum \frac{q_i}{R_i p_i}$$

où q_i est estimé par M_i/R_i et p_i $(R_i - M_i)/R_i$, ce qui donne l'estimateur de Greenwood.

Remarque 2.8.4. L'estimateur de Greenwood est un estimateur consistant de la variance asymptotique de l'estimateur de Kaplan-Meier.

Chapitre 3

Risque proportionnels dans les modèles semi-markoviens

3.1 Introduction

Les modèles Markoviens homogènes ont été appliqués avec succès dans de nombreux domaines et sont utilisés de plus en plus fréquemment. Cependant, dans ces modèles, l'évolution du processus est indépendante du temps déjà passé dans l'état actuel. Dans le domaine clinique par exemple, cette hypothèse correspond rarement à la réalité. Les processus semi-Markoviens constituent alors une alternative intéressante puisqu'ils intègrent dans la définition du modèle les lois de 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 processus semi-markoviens ont été introduits simultanément par Lévy (1954) et Smith (1955). Les fondements de la théorie des processus semi-markoviens ont été introduits par Pyke (1961) dans deux travaux où le deuxième

est consacré aux processus d'espace d'état fini, la théorie de renouvellement classique à été généralisé aux processus semi-markoviens par Feller (1964). Des théorèmes limites des processus semi-markoviens ont été présenté par Yackel (1966), Grigorescu et Oprisan (1976), Athreya et al. (1978), Nummelin (1978) et Malinovskii (1987). Des modèles semi-markoviens sont décrits par Janssen (1986), Andersen et al. (1993), Janssen et Limnios (1999) et Csenki (2002). Ross (1970) et Puterman (1994) ont étudié les processus semi-markoviens de décision. Les Modèles semi-Markoviens ont été étudiés dans un cadre non-homogène par Vassiliou et Papadopoulou [1992] et Papadopoulou et Vassiliou [1994], alors que Sternberg et Satten [1999] se sont intéressés aux problèmes de données censurées par intervalles ou tronquées. Limnios et Oprisan (2001) donne une étude totale des processus semi-markoviens à temps continu et leurs applications en fiabilité. On peut ajouter qu'il est aussi possible d'obtenir des estimations non-paramétriques dans les modèles semi-Markovien en utilisant la théorie des processus de comptage Gill(1980), Andersen et al. (1993).

On considère un système aléatoire dans un espace d'états fini $E = \{1, ..., s\}$. On note par $\mathcal{M}_E$ l'ensemble des matrices réelles sur $E \times E$ et par $\mathcal{M}_E(\mathbb{N})$ l'ensemble des matrices fonctionnelle défini sur $\mathbb{N}$, à valeurs dans $\mathcal{M}_E$. Pour $\mathbf{A} \in \mathcal{M}_E(\mathbb{N})$, on écrit $\mathbf{A} = (\mathbf{A}(k); k \in \mathbb{N})$, où, pour $k \in \mathbb{N}$ fixé, $\mathbf{A}(k) = (A_{ij}(k); i, j \in E) \in \mathcal{M}_E$. On note que I_E la matrice identité.
On suppose que l'évolution du système dans le temps est décrite par :

- $\{X(t), t \in \mathbb{I} \subset \mathbb{R}\}$ un processus à temps continu.
- $E = \{1, 2, ..., N\}$ l'espace des états finis.
- $0 = T_0 < T_1 < ..... < T_n < ....$ les instants de changement d'état.
- $U_n = T_n - T_{n-1}$ le temps de séjour dans l'état après le n^{ime} instants de changement d'état.
- $(X_n)_{n\in\mathbb{N}}$ est une chaîne de Markov telle que $X_n = X(T_n)$

Définition 3.1. Une matrice à valeurs fonctionnelle $q = (q_{ij}(t)) \in \mathcal{M}_E(\mathbb{N})$ peut être dit noyau semi-Markovien à temps discret si elle satisfait les trois propriétés suivantes :

1. $0 \leqslant q_{ij}(t), i, j \in E, t \in \mathbb{N}$,
2. $q_{ij}(0) = 0, i, j \in E$,
3. $\sum_{t=0}^{\infty} \sum_{j \in E} q_{ij}(t) = 1, i \in E$.

Remarque 3.1.1. Au lieu de considérer les noyaux semi-Markov tel que défini précédemment, on peut être intéressé par des matrices fonctionnelle à valeurs non négatives, $q = (q_{ij}(t)) \in \mathcal{M}_E(\mathbb{N})$ qui satisfait

$$\sum_{t=0}^{\infty} \sum_{j \in E} q_{ij}(t) \leqslant 1$$

pour tout $i \in E$.

Définition 3.2. le processus (T, X) est dit semi-markovien si la distribution des temps de séjour $(T_{n+1} - T_n)$ satisfait la condition suivante :

$$\mathbb{P}(T_{n+1} - T_n \leqslant x, X_{n+1} = j | X_0, T_0, .., X_n, T_n) \quad = \quad \mathbb{P}(T_{n+1} - T_n \leqslant x, X_{n+1} = j | X_n) \tag{3.1}$$

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

Par ailleurs, si l'équation 3.1 est indépendante de n , et (X, T) est dite homogène, et le noyau semi-Markovien q à temps discret est défini par :

$$q_{ij}(t) = \mathbb{P}(X_{n+1} = j, U_{n+1} = t / X_n = i)$$

On note que, si (X, T) est une chaîne de renouvellement Markovien (homogène), on peut voir facilement que $(X_n) n \in \mathbb{N}$ est une chaîne de Markov homogène. On

note par $P = (p_{ij})_{i,j\in E} \in \mathcal{M}_E$ la matrice de transition de (X_n), elle est définie par $p_{ij} := \mathbb{P}(X_{n+1} = j/X_n = i), i, j \in E, n \in \mathbb{N}$. Notons aussi que, pour tout $i, j \in E, p_{ij}$ peut être exprimé dans le noyau semi-Markovien par $p_{ij} = \sum_{t=0}^{\infty} q_{ij}(t)$. On introduit le noyau semi-Markovien cumulé $Q = (Q(t); t \in \mathbb{N}) \in \mathcal{M}_E$ défini pour tout $i, j \in E$ et $t \in \mathbb{N}$ par :

$$Q_{ij}(t) := \mathcal{P}(X_{n+1} = j, U_{n+1} \leqslant t/X_n = i)$$

Lorsque l'on étudie l'évolution d'un renouvellement Markovien nous sommes intéressés à deux types de maintenir les distributions de temps : les distributions de temps de séjour dans un état donné et les distributions conditionnelles en fonction de l'état suivant à visiter.

Définition 3.3. Pour tout $i, j \in E$, définissons :

1. $f_{ij}(.)$, la distribution conditionnelle :

$$f_{ij}(t) = \mathbb{P}(U_{n+1} = t/X_n = i, X_{n+1} = j), t \in \mathbb{N}. \tag{3.2}$$

2. $F_{ij}(.)$, la distribution cumulative conditionnelle :

$$F_{ij}(t) = \mathbb{P}(U_{n+1} \leqslant t/X_n = i, X_{n+1} = j) = \sum_{l=0}^{t} f_{ij}(l)), t \in \mathbb{N}. \tag{3.3}$$

évidemment, pour tout $i, j \in E$ et pour tout $k \in \mathbb{N}$, on a

$$f_{ij}(t) = \begin{cases} \frac{q_{ij}(t)}{p_{ij}} & \text{si } p_{ij} \geqslant 0 \\ 0 & \text{si non} \end{cases}$$

Définition 3.4. $h_i(t)$ la distribution du temps de séjour dans l'état i :

1. $h_i(t) := \mathbb{P}(U_{n+1} = t/X_n = i) = \sum_{j\in E} q_{ij}(t), t \in \mathbb{N}$.
2. $H_i(t) := \mathbb{P}(U_{n+1} \leqslant t/X_n = i) = \sum_{j\in E}^{t} h_i(l), t \in \mathbb{N}$.

Pour la distribution cumulative de certain v.a, on note la fonction de survie par $S(t) = 1 - F(t) = \mathbb{P}(T > t)$.

$(X_n T_n)_{n\in\mathbb{N}}$ est une chaîne de renouvellement Markovien, avec le noyau semi-markovien $q_{ij=p_{ij}f_{ij}(k)}$ et la distribution du temps de séjour dans l'état i $h_i = \sum_{j\in E} q_{ij}(k)$. Notons que le processus $(X_n T_{n+1})_{n\in\mathbb{N}}$ est une chaîne de renouvellement Markovien avec le noyau semi-Markovien $\tilde{q}_{ij}(k) = p_{ij}h_j(k)$. De même, le processus $(X_{n+1}, T_n)_{n\in\mathbb{N}}$ est une chaîne de renouvellement Markovien son noyau $\check{q}_{ij}(k) = p_{ij}h_i(k)$. Une chaîne de Markov avec la matrice de transition $(p_{ij})_{ij} \in E$, c'est un cas particulier de la chaîne de renouvellement Markovien avec le noyau semi-Markovien

$$q_{ij}(k) = \begin{cases} p_{ij}(p_{ii})^{k-1} & \text{si} i \neq jetk \in \mathbb{N}^*, \\ 0 & \text{si non} \end{cases}$$

Quelques quantités importantes pour vérifier que l'évolution de la chaîne de renouvellement Markovien est la probabilité $\mathbb{P}(X_n = j, T_n = t/X_0), i, j \in E, n \in \mathbb{N}$.

Rappelons que, pour une chaîne de Markov finie $(X_n)_{n\in\mathbb{N}}$ de la matrice de transition $(p_{ij})_{i,j\in E}$ la n^{me} peut être écrite comme suit :

$$\mathbb{P}(X_n = j/X_0 = i) = p_{ij}^n$$

pour tous $n \in \mathbb{N}$,
où p_{ij}^n c'est l'élément (i, j) de la matrice produit de p n fois. Un résultat similaire va pour la probabilité $\mathbb{P}(X_n = j, T_n = t/X_0 = i)$ dans le contexte renouvellement Markovien :

Proposition 3.1.1. *(Voir [1]) Pour tous $i, j \in E$, pour tous n et $t \in \mathbb{N}$, on aura*

$$\mathbb{P}(X_n = j, T_n = t/X_0 = i) = q_{ij}^n(t) \tag{3.4}$$

Démonstration. On fait la démonstration pour $n = 0$, on a

$$\mathbb{P}(X_0 = j, T_0 = t/X_0 = i) = q_{ij}^0(t)$$

. Sur d'autre part, si $i = j$ et $t = 0$, Évidemment, pour $t \neq 0$ ou $i \neq j$, cette probabilité vaut 0

$$\begin{aligned}
\mathbb{P}(X_n = j, T_n = t/X_0 = i) &= \sum_{r\in E}\sum_{l=1}^{t-1} \mathbb{P}(X_n = j, T_n = t, X_1 = r, T_1 = l/X_0 = i) \\
&= \sum_{r\in E}\sum_{l=1}^{t-1} \mathbb{P}(X_n = j, T_n = t/X_1 = r, T_1 = l, X_0 = i)\mathbb{P}(X_1 = r, T_1 = l/X_0 = i) \\
&= \sum_{r\in E}\sum_{l=1}^{t-1} \mathbb{P}(X_{n-1} = j, T_{n-1} = t - l/X_0 = r)\mathbb{P}(X_1 = r, T_1 = l/X_0 = i) \\
&= \sum_{r\in E}\sum_{l=1}^{t-1} q_{rj}^{n-1}(t - l)q_{ir}(l) \\
&= q_{ij}^n(t)
\end{aligned}$$

et voila le résultat □

Remarque 3.1.2. Pour $(X_n, T_n)_{n\in\mathbb{N}}$ la chaîne de renouvellement Markovien dans l'espace des états E, on peut vérifier aussi que c'est une chaîne de Markov dans l'espace $E \times \mathbb{N}$. Notons par $P^{(n)}$ la n^{me} fonction de transition de cette chaîne de Markov, de la proposition précédente, pour tous $i, j \in E$, et pour tous n et $t \in \mathbb{N}$ on a

$$P_{(i,0),(j,k)}^{(n)} = q_{i,j}^{(n)}(k).$$

Comme application de la proposition précédente, on aura le lemme suivant, un résultat important pour les chaînes de renouvellement Markovien.

Lemme 3.1.1. *(Voir[1])* $(X,T) = (X_n, T_n)_{n\in\mathbb{N}}$ *est une chaîne de renouvellement Markovien, et* $q \in \mathcal{M}_E(\mathbb{N})$ *son noyau semi-markovien. Pour tous* $n, t \in \mathbb{N}$ *tels que* $n \geqslant t+1$ *on a* $q^{(n)}(t) = 0$.

Démonstration. Il est clair que le temps de saut $(T_n)_{n\in\mathbb{N}}$ du processus verifi la relation $T_n \geqslant n, n \in \mathbb{N}$. Ecrivant l'équation (3.4) pour n et $t \in \mathbb{N}$ tels que $n \geqslant t+1$, on obtient le résultat souhaité. □

La propriété du noyau semi-Markov temps discret mentionné est essentielle pour la simplicité et la précision numérique. Introduisons maintenant la chaîne semi-Markovienne, strictement liée aux celle de la chaîne de renouvellement Markovien.

Définition 3.5. Soit (X,T) est une chaîne de renouvellement Markovien. La chaîne $Z = (Z_t)_{t\in\mathbb{N}}$ est dite chaîne semi-markovienne associé à la chaîne de renouvellement Markovien (X,T) si

$$Z_t := X_{N(t),t\in\mathbb{N}},$$

où

$$N(t) := max\{n \in \mathbb{N}/T_n \leqslant t\} \tag{3.5}$$

c'est nombre total de transition dans l'intervalle $[0,t] \subset \mathbb{N}$ Ainsi que Z_t donne le système au temps t. On a aussi $X_n = Z_{T_n}$ et $T_n = min\{t > X_{n-1}/Z_t \neq Z_{t-1}\}, n \in \mathbb{N}$. Le vecteur ligne noté $\alpha = (\alpha_1, ..., \alpha_s)$ c'est la distribution initiale de la chaîne semi-Markovienne $Z = (Z_t)_{t\in\mathbb{N}}$ i.e., $\alpha_i := \mathbb{P}(Z_0 = i) = \mathbb{P}(X_0 = i), i \in E$.

Définition 3.6. La fonction de transition de la chaîne semi-Markovienne Z est la matrice à valeurs fonctionnelle $P = (P_{ij}(t); i,j \in E, t \in \mathbb{N}) \in \mathcal{M}_E(\mathbb{N})$. qui est définie par :

$$P_{ij}(t) := \mathbb{P}(Z_t = j/Z_0 = i), i,j \in E, t \in \mathbb{N}.$$

3.2 Estimation paramétrique des temps de séjour

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

$$h_{ij} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t < U_{n+1} \leqslant t + \Delta t / U_{n+1} > t, X_n = i, X_{n+1} = j)$$

Cette estimation va prendre que les fonctions de risque $h_{ij}(.)$ appartiennent à une classe de fonctions paramétriques. Les fonctions $S_{ij}(.)$ et $f_{ij}(.)$ correspondant respectivement aux fonctions de survie et de densité associées aux fonctions de risque $h_{ij}(.)$ peuvent s'écrire à partir de $h_{ij}(.)$.

$$\begin{aligned}
\frac{\partial S_{ij}(t)}{\partial t} &= -\lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t < U_{n+1} \leqslant t + \Delta t / X_n = i, X_{n+1} = j) \\
&= -\lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t < U_{n+1} \leqslant t + \Delta t / U_{n+1} > t, X_n = i, X_{n+1} = j) \\
&\times \; Pr(X_{n+1} > t / X_n = i, X_{n+1} = j) \\
&= -S_{ij}(t) \times -\lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t < U_{n+1} \leqslant t + \Delta t / X_n = i, X_{n+1} = j) \\
&= -S_{ij}(t) \times \alpha_{ij}(t).
\end{aligned}$$

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

$$S_{ij}(d) = \exp(-\int_0^d \alpha_{ij}(u)du). \tag{3.6}$$

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

$$f_{ij}(d) = -\frac{\partial S_{ij}(d)}{\partial d} = S_{ij}(d)\alpha_{ij}(d). \tag{3.7}$$

3.3 Modèle à risques proportionnels

Afin d'introduire des covariables on considère un modèle à risques proportionnels de "Cox".
On considère $(X_n, T_n)_{n>0}$ un processus semi-Markovien. Les covariables sont introduites dans les fonctions de risque des temps d'attente dans les états. La chaîne de Markov $(X_n)_{n>0}$ ne dépend pas du vecteur de covariables ainsi la chaîne conserve la probabilité de transition $p_{ij} = Pr(X_{n+1} = j/X_n = i)$. Les fonctions de risque du processus semi-Markovien ne dépendent pas des covariables.
Considérons $Z_{ij}(.) = (Z_{ij}^1(.), Z_{ij}^2(.), ..., Z_{ij}^{n_{ij}}(.))$, le vecteur de covariables associé à la transition de l'état i vers j, tel que n_{ij} le nombre de covariables pour cette transition. Les covariables peuvent être dépendantes du temps, il est nécessaire de supposer que la valeur des covariables ne change pas entre deux consultations. Pour simplifier les calculs et les écritures, on supposera par la suite que les covariables sont fixées au cours du temps d'attente : $Z_{ij}(t) = z_{ij}$. Les covariables vont modifier les fonctions d'intensité en suivant un modèle à risques proportionnels de Cox

$$h_{ij}(t - S_{N(t^-)}) \quad = \quad h_{ij,0}(t - S_{N(t^-)}) \exp(\beta_{ij}^T z_{ij}),$$

avec β_{ij} le vecteur des coefficients de régression associés à z_{ij} et $h_{ij,0}(.)$ est le risque de base. Ici la proportionnalité des risques est supposée au sein d'une même transition.
Nous intéressons maintenant aux fonctions de survie et de densité correspondant à des fonctions de risque de temps d'attente dépendantes de covariables. Considérons $\forall i, j \in E$, $h_{ij}(t, z) = h_{ij,0}(t) \exp \beta_{ij}^T z_{ij}$, alors d'après les équations 3.6 et 3.7 les fonctions de survie correspondantes sont données par :

$$\begin{aligned} S_{ij}(t,z) &= \exp(-\int_0^t h_{ij}(u)du) \\ &= \exp(-\int_0^t h_{ij,0}(u)e^{\beta_{ij}^T z_{ij}}du) \\ &= S_{ij,0}(t)^{e^{\beta_{ij}^T z_{ij}}} \end{aligned} \tag{3.8}$$

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

$$\begin{aligned} f_{ij}(d,z) &= S_{ij}(t,z)h_{ij}(t,z) \\ &= h_{ij,0}(t)e^{\beta_{ij}^T z_{ij}} S_{ij,0}(z)^{e^{\beta_{ij}^T z_{ij}}} \end{aligned} \tag{3.9}$$

3.4 Processus semi-markoviens à durée de transition bornée

3.4.1 Cas de durée continue

Définition 3.7. Une fonction matricielle $Q = (Q_{ij}(T_n, T_{n+1}, t))_{(i,j)\in E^2}$ est dite noyau cumulé semi-Markovien si elle satisfait la propriété suivante :
$Q_{ij}(T_n, T_{n+1}, t) = \mathbb{P}(X_{T_{n+1}} = j, U_{n+1} < t / X_{T_n} = i), i, j \in E, n \in \mathbb{N}, t \in \mathbb{I}$ où $\mathbb{I}$ est le support de U_{n+1}.
Par la suite $Q_{ij}(T_n, T_{n+1}, t)$ est notée abusivement $Q_{ij}(n, n+1, t)$. Dans le cas homogène, on a $Q_{ij}(n, n+1, t) = Q_{ij}(t)$.

Un processus semi-Markovien peut être entièrement déterminé par sa loi initiale et la fonction matricielle $Q = (Q_{ij}(T_n, T_{n+1}, t))_{(i,j)\in E^2}$.

Propriété 2. *Le noyau cumulé semi-Markovien Q vérifie :*

1. *Si Qij est dérivable par rapport à t, la fonction matricielle*

$$q_{ij}(n, n+1, t) = \frac{\partial Q_{ij}(n, n+1, t)}{\partial t}$$

caractérise le processus semi-Markovien $(X_t)_{t\in\mathbb{R}_+}$ *connaissant sa loi initiale.*

2. *En notant* $b = sup(\mathbb{I})$*, on a*
$Q_{ij}(n, n+1, b) = \mathbb{P}(X_{T_{n+1}} = j/X_{T_n} = i) = p_{ij}(n, n+1), i, j \in E, n \in \mathbb{N}.$

3. *La fonction de répartition de la durée en l'état i, conditionnellement à la transition à l'état j, s'écrit :*

$$\begin{aligned} F_{ij}(n, n+1, t) &= \mathbb{P}(U_{n+1} \leqslant t/X_{T_n} = i, X_{T_{n+1}} = j) \\ &= \frac{Q_{ij}(n, n+1, t)}{p_{ij}(n, n+1)} \ si\ p_{ij}(n, n+1) > 0 \ et\ t \in [0,\ b[\\ &= 1 \ si\ p_{ij}(n, n+1) > 0 \ et\ t \geqslant b \\ &= 0 \ si\ p_{ij}(n, n+1) = 0 \end{aligned}$$

4. *Si Qij est dérivable par rapport à t, la densité de la durée en l'état i, conditionnellement à la transition à l'état j, s'écrit :*

$$f_{ij}(n, n+1, t) = \frac{\partial F_{ij}(n, n+1, t)}{\partial t} = \frac{q_{ij}(n, n+1, t)}{p_{ij}(n, n+1)}.$$

5. *La fonction de répartition de la durée en l'état i (inconditionnelle), s'écrit :*

$$\begin{aligned} F_i(n,n+1,t) &= \mathbb{P}(U_{n+1} \leqslant t / X_{T_n} = i) \\ &= \sum_{j=1}^{s} \mathbb{P}(U_{n+1} \leqslant t, X_{T_{n+1}} = j / X_{T_n} = i) \\ &= \sum_{j=1}^{s} Q_{ij}(n,n+1,t) \\ &= \sum_{j=1}^{s} F_{ij}(n,n+1,t)\, p_{ij}(n,n+1) \end{aligned}$$

On en déduit alors la fonction de survie $S_i(n,n+1,t) = 1 - F_i(n,n+1,t)$ *la densité :*

$$f_i(n,n+1,t) \sum_{j=1}^{s} F_{ij}(n,n+1,t)\, p_{ij}(n,n+1)$$

la fonction de risque :

$$\begin{aligned} h_i(n,n+1,t) &= \frac{f_i(n,n+1,t)}{1 - F_i(n,n+1,t)} \text{ si } p_{ij}(n,n+1) > 0 \text{ et } F_i(n,n+1,t) < 1 \\ &= 0 \text{ sinon} \end{aligned}$$

et toutes les transformations de la fonction de survie permettant de bien la caractériser.

6. *Si l'on considère la fonction de risque conditionnement à un changement d'état vers j, on construit*

$$\begin{aligned} h_{ij}(n,n+1,t) &= \frac{f_{ij}(n,n+1,t)}{1 - F_i(n,n+1,t)} \text{ si } p_{ij}(n,n+1) > 0 \text{ et } F_i(n,n+1,t) < 1 \\ &= 0 \text{ sinon} \end{aligned}$$

qui seraient les intensités de transition instantanées du noyau semi-Markovien q.

Remarque 3.4.1. Si l'on note $N_{ij}(t)$ le processus de comptage du nombre de transitions observées de l'état i vers l'état j dans $[0,t]$, $t \in \mathbb{R}_+$, on vérifie que $N_{ij}(t)$, avec $N_{ij}(0) = 0$, est un processus *cadlag* avec sauts de 1. Il en de même des processus de comptage $N_i(t) = \sum_{j=1}^{s} N_{ij}(t)$ et $N(t) = \sum_{i=1}^{s} N_i(t)$. Le processus $N_{ij}(t)$ est construit comme suit :

$$N_{ij}(t) = \sum_{n \geqslant 0} 1_{\{T_n < t;\, X_{T_n}=i,\, X_{T_{n+1}}=j\}}(t)$$

Proposition 3.4.1. *Si le processus semi-markovien $(X_t)_{t\in\mathbb{R}_+}$ est asymptotiquement ergodique pour la classe des états récurrents (positifs) de E, alors $\frac{N_{ij}(t)}{N(t)} \to p_{ij}$ p.s.. La matrice de terme général p_{ij} est la matrice de transition asymptotique (qui existe toujours en cas d'ergodicité et est homogène).*

Proposition 3.4.2. *Si π_i est la $i^{ème}$ composante de la mesure stationnaire de la matrice de transition asymptotique du processus markovien sous-jacent, la $i^{ème}$ composante de la distribution du processus semi-markovien $(X_t)_{t\in\mathbb{R}_+}$ associé vérifie*

$$\Pi_i = \frac{\pi_i}{\mu_{ii}}$$

où μ_{ii} est le temps moyen de retour à l'état i partant de i.

Pour analyser de tels processus sur le plan opératoriel afin de les exploiter numériquement, différentes approches ont été exploitées (voir par exemple [19], [29], ...). Dans le cas de notre étude, comme la durée de séjour dans chaque état est majorée par une valeur finie $b = \sup\{t;\, t \in \mathbb{I}\}$, nous considérons une suite de morceaux de trajectoires de longueur b notés

$$\begin{array}{rcccl} {}^0X_n & : & (\Omega, \mathcal{A}, P) & \longrightarrow & (\mathcal{C}_{\mathbb{I}}^{ps}, \mathcal{X}) \\ & & \omega & \longmapsto & C_n^0 = \{(s, y_s); s \in]nb, (n+1)b]\} \end{array}$$

où $\mathcal{C}_{\mathbb{I}}^{ps}$ est l'espace vectoriel des fonctions scalaires continues p.s. définies dans l'intervalle $\mathbb{I}$.

De la même manière, pour tout $\tau \in \mathbb{I}$, on peut construire

$$\begin{array}{rcccl} {}^\tau X_n & : & (\Omega, \mathcal{A}, P) & \longrightarrow & (\mathcal{C}_{\mathbb{I}}^{ps}, \mathcal{X}) \\ & & \omega & \longmapsto & C_n^\tau = \{(s, y_s); s \in]\tau + nb, \tau + (n+1)b]\} \end{array}$$

qui est une transformation du processus $(X_t)_t$, semblable à la précédente, translatée de τ sur $\mathbb{R}_+$.

Sans perdre de généralité, nous restreignons notre étude au cas du processus $({}^0X_n)_n$.

Proposition 3.4.3. *Le processus $({}^0X_n)_n$ est un processus de Markov à temps discret et à valeurs dans $\mathcal{C}_{\mathbb{I}}^{ps}$.*

Tenant compte que le passé et le futur d'un tel processus sont indépendants connaissant le présent, nous montrons que $({}^0X_n)_n$ est un processus de Markov vérifiant

$\forall B,\ C_1\, C_2 \ldots C_n \in \mathcal{X}$

$P({}^0X_{n+1} \in B / {}^0X_n = C_n,\ \ldots, {}^0X_2 = C_2,\ {}^0X_1 = C_1) = P({}^0X_{n+1} \in B / {}^0X_n = C_n)$

L'existence de probabilité produit est assurée par le théorème de Ionescu Tulcea [32].

Commentaire : En utilisant une suite dénombrable de courbes de $\mathcal{C}_{\mathbb{I}}^{ps}$ issues d'une réalisation de $({}^0X_n)_n$, on obtient nécessairement un ou plusieurs points d'accumulation (qui sont ici des courbes). Pour chaque point d'accumulation, on peut construire une sous-suite qui converge uniformément presque sûrement vers

le point d'accumulation considéré. La courbe moyenne sera alors une moyenne pondérée des points d'accumulation trouvés. Nous pouvons également évaluer d'autres caractéristiques descriptives du processus telles que la dispersion autour de la courbe moyenne, la symétrie, ...

3.4.2 Cas de durée discrète

Dans les cas pratiques, souvent le temps est représenté par une durée (liée à une fréquence d'une horloge de référence : calendriers, heures, mn, seconde, fraction de seconde, ...). On considère, dans cette étude, que la durée de séjour dans chaque état est majorée par une valeur finie $v = \sup\{k;\, k \in I\}$. Ainsi les morceaux de trajectoires de longueur v notés $\widetilde{X}_n^v = (X_{vn}, X_{vn-1}, \ldots, X_{v(n-1)+1})$ vérifient

$$\begin{array}{rccc} \widetilde{X}_n^v : & (\Omega, \mathcal{A}, \mathrm{P}) & \longrightarrow & (E,\, \mathcal{E})^{\otimes v} \\ & \omega & \longmapsto & (X_{vn}(\omega), X_{vn-1}(\omega), \ldots, X_{v(n-1)+1}(\omega)) \end{array}$$

$\left(\widetilde{X}_n^v\right)_n$ est à vs états que l'on représente vectoriellement dans $\mathbb{R}^{vs}$ par :

$$\mathcal{X}_n = g \circ \widetilde{X}_n^v = \begin{pmatrix} 1_{(\widetilde{X}_n^v)^{-1}(1,\ldots 1)} \\ \vdots \\ 1_{(\widetilde{X}_n^v)^{-1}(i_1 \ldots i_v)} \\ \vdots \\ 1_{(\widetilde{X}_n^v)^{-1}(s \ldots s)} \end{pmatrix} \tag{3.10}$$

où g correspond à une fonction mesurable permettant de représenter les éléments de $(E,\, \mathcal{E})^{\otimes v}$ dans $\mathbb{R}^{vs}$. La $i^{\text{ème}}$ coordonnée de $\mathcal{X}_n$ vérifie

$$i = i_v + (i_{v-1} - 1)s + \ldots + (i_2 - 1)s^{v-2} + (i_1 - 1)s^{v-1}.$$

Ainsi l'état $\widetilde{X}_n^v(\omega) = (i_1 \dots i_v)$ peut être écrit $\widetilde{X}_n^v(\omega) = a_i$ avec $\mathcal{X}_n(\omega) = g(a_i) = e_i$ où e_i est le $i^{\text{ème}}$ vecteur de la base canonique de $\mathbb{R}^{vs}$.

Proposition 3.4.4. *Le processus $(\mathcal{X}_n)_n$ est markovien.*

Preuve : Notons T_1, T_2, T_3, ... les instants de changement d'état dans le processus semi-markovien $(X_n)_n$. Comme pour tout entier naturel $n > 0$, $T_{n+1} - T_n \leqslant v$, on montre aisément que
$P(\mathcal{X}_{n+1} = e_j / \mathcal{X}_n = e_i, \mathcal{X}_{n-1} = e_{i_1}, \mathcal{X}_{n-2} = e_{i_2}, ...) = P(\mathcal{X}_{n+1} = e_j / \mathcal{X}_n = e_i)$. □

Définition 3.8. Le processus $(X_n)_n$ générant le processus $(\mathcal{X}_n)_n$ ainsi construit est appelé processus v-markovien.

En exploitant la représentation du processus sous la forme (3.4.2), nous mettons facilement le lien entre l'étude du processus semi-markovien $(X_n)_n$ et celle du processus markovien $\mathcal{X}_n^v$.

Proposition 3.4.5. *Si $(X_n)_n$ est un processus v-markovien, alors $(X_n)_n$ est u-markovien pour tout $u \geqslant v$.*

Preuve : $(X_n)_n$ est un processus v-markovien si et seulement si connaissant $(X_n)_n$ sur v étapes successives son passé et son futur (par rapport à ces étapes) sont indépendants. Cette propriété reste vraie pour tout $u \geqslant v$. □

Définition 3.9. Le processus $(X_n)_n$ est dit processus v-markovien minimal, si $v = min\{u;\ (X_n)_n \text{ est } u - \text{markovien}\}$.

Proposition 3.4.6. *Si $(X_n)_n$ est un processus u-markovien pour tout $u \geqslant 1$, alors $(X_n)_n$ est markovien.*

Seul les états $1_{(\tilde{\mathcal{X}}_n^v)^{-1}(l,\dots l)}$; $l = 1, \dots, v$ coïncident avec les "trajectoires" du processus semi-markovien $(X_n)_n$.

Si $(X_n)_n$ est un processus v-markovien avec v connu, la construction de l'opérateur de transition est assez aisée. Pour v fixe mais inconnu, l'estimation de v nécessite des méthodes appropriées.

En décalant la trajectoire de i pas $i \in \{0, \ldots, v-1\}$, on construit de nouveaux morceaux de trajectoires de longueur v notés ${}^{(i)}\widetilde{X}_n^v = (X_{vn+i}, X_{vn-1+i}, \ldots, X_{v(n-1)+1+i})$ qui vérifient

$$\begin{array}{rccc} {}^{(i)}\widetilde{X}_n^v : & (\Omega, \mathcal{A}, \mathrm{P}) & \longrightarrow & (E,\, \mathcal{E})^{\otimes v} \\ & \omega & \longmapsto & (X_{vn+i}(\omega), X_{vn-1+i}(\omega), \ldots, X_{v(n-1)+1+i}(\omega)) \end{array}$$

$\left({}^{(i)}\widetilde{X}_n^v\right)_n$ est à vs états que l'on représente vectoriellement dans $\mathbb{R}^{vs}$ par :

$${}^{(i)}\mathcal{X}_n = g \circ^{(i)} \widetilde{X}_n^v = \begin{pmatrix} 1_{({}^{(i)}\widetilde{X}_n^v)^{-1}(1,\ldots 1)} \\ \vdots \\ 1_{({}^{(i)}\widetilde{X}_n^v)^{-1}(i_1 \ldots i_v)} \\ \vdots \\ 1_{({}^{(i)}\widetilde{X}_n^v)^{-1}(s \ldots s)} \end{pmatrix} \tag{3.11}$$

Le processus $\left({}^{(i)}\widetilde{X}_n^v\right)_n$ est markovien. Toutefois sa matrice de transition ne commute pas nécessairement avec celle de $\left({}^{(j)}\widetilde{X}_n^v\right)_n$ si $i \neq j$.

Ainsi par décalage de i ; $i \in \{0, \ldots, v-1\}$, nous pouvons construire v chaînes de Markov pour décrire le processus semi-markovien discret $(X_n)_{n\in\mathbb{N}}$. Nous exploitons ainsi v semi-groupes pour décrire $(X_n)_{n\in\mathbb{N}}$.

Proposition 3.4.7. *Si les v semi-groupes décrivant $(X_n)_{n\in\mathbb{N}}$ commutent entre eux le processus semi-markovien $(X_n)_n$ est markovien.*

Commentaire : En utilisant les morceaux de trajectoires (continues par morceaux ou discrètes), on transforme un processus semi-markovien en un processus

markovien dès qu'il a la propriété que la durée du temps de séjour dans chaque état est bornée. Ainsi, on peut exploiter les caractéristiques du processus semi-markovien, en utilisant des propriétés markoviennes du processus de "courbes" qui sont plus facilement manipulables techniquement.

Chapitre 4

Estimation semi-paramétriques dans le modèle de Cox généralisé

4.1 Les modèles à risques proportionnels

Ces modèles expriment un effet multiplicatif des diverses covariables sur la fonction de risque. On introduit une fonction de risque de base qui donne la forme générale du risque et qui est commune à tous les individus. Les modèles à risques proportionnels se caractérisent par la relation suivante, pour tous $t > 0$,

$$h(t|Z) = h_0(t)g(\beta, Z), \tag{4.1}$$

où Z est un vecteur de covariables, β le paramètre d'intérêt et g une fonction positive.
La fonction de risque est le produit d'une fonction qui ne dépend que du temps et d'une fonction qui n'en dépend pas. En générale on suppose que l'effet des covariables se résume à une quantité réelle $\beta^{'}Z$, c'est-à-dire $h(t|Z) = h_0 g(\beta^{'}Z)$.
Ce modèle est dit à risques proportionnels car, quels que soient deux individus i

et j qui ont pour covariables Z_i et Z_j, le rapport des fonctions de risque ne varie pas au cours du temps,

$$\frac{h(t|Z_i)}{h(t|Z_j)} = \frac{g(\beta^{'} Z_i)}{g(\beta^{'} Z_j)}. \tag{4.2}$$

Les fonctions de risque sont donc proportionnelles. C'est une conséquence du modèle mais c'est aussi une hypothèse qu'il faudra vérifier. Le rapport des fonctions de risque est par définition un risque relatif à l'instant t des sujets de caractéristiques Z_i par le rapport aux sujets de caractéristiques Z_j. Un cas particulier très important est le modèle de Cox, qui suppose que la fonction g est la fonction exponentielle, c'est-à-dire,

$$h(t|Z) = h_0 exp(\beta^{'} Z), \tag{4.3}$$

D'autre choix de fonctions h sont possibles, néanmoins la fonction exponentielle est très souvent utilisée dans la littérature car ses valeurs sont toujours positives et $exp(0) = 1$ Si h_0 et/ou h ont une forme inconnue, le modèle est dit semi-paramétrique.

4.2 Modèle de Cox

Le modèle de régression de Cox est l'un des plus utilisés pour l'analyse statistique des durées de vie. Cependant, l'inférence statistique pour ce modèle, basée sur la vraisemblance partielle de Cox.

$$h(t|Z) = h_0 exp(\beta^{'} Z), \tag{4.4}$$

où Z est un vecteur de covariables de dimension $p * 1$ et β un vecteur $(p * 1)$ de coefficient de régression.
Cosidérons,

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

4.2.1 Vraisemblance partielle de Cox

Le principe de la méthode est d'estimer uniquement le coefficient de régression β en considérant la fonction h_0 comme un paramètre de nuisance. Par conséquent, on ne cherche pas à estimer h_0. L'idée de Cox est qu'aucune information ne peut être donnée sur β par les intervalles pendant lesquels aucun événements n'a eu lieu, car on peut concevoir que h_0 soit nulle dans ces intervalles (On suppose que les moments où se produisent les censures n'apportent peu ou pas d'information sur β). On travaille alors conditionnellement à l'ensemble des instants où un décès a lieu. Supposons, dans un premier temps, qu'il n'y a qu'un seul décès à chaque temps d'événement (car le raisonnement provient du cas continu). La probabilité qu'il y ait un événement (décès) en T_i (dans l'intervalle $[T_i, T_{i+\Delta t}]$) est

$$\sum_{j \in R(T_i)} h_0(T_i) exp(\beta' Z_j), \tag{4.5}$$

La probabilité que l'individu i subisse l'événement en T_i sachant qu'un événement a eu lieu en T_i vaut

$$\frac{h0(T_i)exp(\beta' Z_{(i)})}{\sum_{j\in R(T_i)} h_0(T_i)exp(\beta' Z_j)} = \frac{exp(\beta' Z_{(i)})}{\sum_{j\in R(T_i)} \exp(\beta' Z_j)}$$

Le point important est que cette probabilité dépend uniquement du paramètre β. Comme il y a des contributions à la vraisemblance à chaque temps de décès, la vraisemblance partielle de Cox est définie comme le produit sur les temps de décs̀. La vraisemblance (partielle) totale est donc

$$L_{Cox}(\beta) = \prod_{i=1}^{D} \frac{exp(\beta' Z_{(i)})}{\sum_{i\in R(T_i)} exp(\beta' Z_j)}.$$

La vraisemblance partielle ne dépend pas de la fonction de risque de base $h_0(t)$. On peut donc estimer β, sans connaître la fonction de risque de base, par maximisation de la vraisemblance partielle de Cox. La vraisemblance partielle n'est pas une vraisemblance dans le sens statistique du terme, mais elle se comporte comme telle. Ainsi, on peut développer une théorie asymptotique similaire et l'utiliser pour estimer et tester les coefficients de régression β.

4.2.2 Evénements simultanés

Le raisonnement précédent suppose des temps d'évènements distincts. Dans le cas des données réelles, cette hypothèse n'est pas vérifiée (ex : mesure tous les mois ou trimestres). La probabilité que l'individu j décède en T_i est

$$p_i = \frac{\exp(\beta' Z_j)}{\sum_{k\in R(t_i)} \exp(\beta' Z_k)}$$

en présence de plusieurs événements, la méthode "exacte" consiste à admettre que les évènements se produisent les uns à la suite des autres. Cependant, on ne connaît pas l'ordre des événements, il faut donc considérer toutes les probabilités. Dans le cas de deux sujets s_1 et s_2 de caractéristiques Z_1 et Z_2 qui décèdent en T_i, la contribution exacte à la vraisemblance est

$$\frac{\exp(\beta' Z_1)\exp(\beta' Z_2)}{\sum_{j\in R(T_i)}\exp(\beta' Z_j)\times\sum_{j\in R(T_i)/s_1}\exp(\beta' Z_j)}+\frac{\exp(\beta' Z_1)\exp(\beta' Z_2)}{\sum_{j\in R(T_i)}\exp(\beta' Z_j)\times\sum_{j\in R(T_i)/s_2}\exp(\beta' Z_j)}.$$

Le problème de cette méthode est que le temps de calcul devient très long quand il y a beaucoup d'événements simultanés. Ainsi, on utilise le plus souvent l'approximation de Breslow qui consiste à supposer que la contribution des d_i événements en T_i est le produit des probabilités p_i pour les unités décédées en T_i i.e.

$$\sum_{j\in R(T_i)} exp(\beta' Z_j) \approx \sum_{j\in R(T_i)/k} \exp(\beta' Z_j)$$

$$L_B(T_i) = \prod_{\text{j :unitées décédées en}T_i} p_j = \frac{\exp\left(\beta'\left(\sum_{j:\text{unités décédées en}T_i} Z_j\right)\right)}{\left(\sum_{k\in R(T_i)}\exp(\beta' Z_j)\right)^{d_i}}$$

L'approximation de Breslow de la vraisemblance totale est :

$$\prod_{i=1}^{D} L_B(T_i)$$

où D est nombre de décès observés. La maximisation de cette vraisemblance est rapide. De plus, si le nombre d'événements simultanés n'est pas trop grand alors la méthode est assez précise.

4.2.3 Estimation

Estimation des coefficients de régression β

A partir de la vraisemblance partielle, on peut obtenir une estimation du vecteur de paramètre β de dimension $p \times 1$. Notons

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

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

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

L'estimateur de Cox β' du coefficient de régression est solution de l'équation :

$$U(\beta) = 0.$$

La solution exacte de ce problème se fait avec L'algorithme de Newton-Raphson, il est souvent utilisé par les logiciels pour obtenir une solution.
Un estimateur consistant de la matrice de variance-covariance de β peut se calculer à partir de l'inverse de la matrice d'information de Fisher,

$$\widehat{Var}\hat{\beta} = \{I(\hat{\beta})\}^{-1}$$

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

$$[I(\beta)]_{ij} = -\frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta_i \partial\beta_j}.$$

Algorithme de Newton-Raphson

Le but de l'algorithme est de trouver une racine de l'équation

$$\frac{\partial logL}{\partial\beta} = 0$$

Pour cela, on se donne une valeur initiale $\tilde{\beta}_0$ et on cherche le plan tangent en ce point à la fonction :

$$\beta \rightarrow d = \frac{\partial logL}{\partial\beta}$$

Ce plan, d'équation :

$$d = \frac{\partial logL}{\partial\beta}(\tilde{\beta}_0) + \frac{\partial^2 logL}{\partial\beta\partial^t\beta}(\tilde{\beta}_0)\left[\beta - \tilde{\beta}_0\right]$$

constitue une approximation de $d = \frac{\partial logL}{\partial\beta}$ d'où l'idée d'approcher la racine de :

$$\frac{\partial logL}{\partial\beta} = 0$$

par la racine de :

$$\frac{\partial logL}{\partial \beta}(\tilde{\beta}_0) + \frac{\partial^2 logL}{\partial \beta \partial^t \beta}(\tilde{\beta}_0)\left[\beta - \tilde{\beta}_0\right] = 0$$

c'est-à-dire par :

$$\tilde{\beta}_1 = \tilde{\beta}_0 - \left[\frac{\partial^2 logL}{\partial \beta \partial^t \beta}(\tilde{\beta}_0)\right]^{-1} \frac{\partial logL}{\partial \beta}(\tilde{\beta}_0)$$

on recommence alors la démarche en prenant $\tilde{\beta}_1$ comme valeur initiale et ainsi de suite. La formule de récurrence permettant de calculer $\tilde{\beta}_{h+1}$ en fonction de $\tilde{\beta}_h$ est :

$$\tilde{\beta}_{h+1} = \tilde{\beta}_h - \left[\frac{\partial^2 logL}{\partial \beta \partial^t \beta}(\tilde{\beta}_h)\right]^{-1} \frac{\partial logL}{\partial \beta}(\tilde{\beta}_h)$$

Si la suite des valeurs $\tilde{\beta}_h$ converge vers une limite $\tilde{\beta}$, celle-ci est forcément solution des équation de vraisemblance car :

$$\tilde{\beta} = \lim_{h\to\infty} \tilde{\beta}_{h+1} = \tilde{\beta} - \left[\frac{\partial^2 logL}{\partial \beta \partial^t \beta}(\tilde{\beta})\right]^{-1} \frac{\partial logL}{\partial \beta}(\tilde{\beta})$$

$$\Rightarrow \frac{\partial logL}{\partial \beta}(\tilde{\beta}) = 0$$

Estimation du risque cumulé de base H_0

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

$$\hat{H}_0(t) = \sum_{i:T_i \leqslant t} = \frac{d_i}{\sum_{j \in R(T_i)} \exp(\hat{\beta}' Z_j)}$$

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

$$S(t/Z) = \exp(-\int_0^t h(u/Z)du)$$

$$\Rightarrow \hat{S}(t/Z) = \exp(-\hat{H}_0(t)\exp(\hat{\beta}' Z)).$$

4.2.4 Interprétation des coefficients de régression

Par définition le risque relatif à l'instant t, pour deux vecteurs de covariables Z_i et Z_j, est égale à :

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

Dans le modèle de Cox, le risque relatif est constant au cours du temps :

$$RR(t) = RR = \exp(\beta'(Z_i - Z_j)).$$

Ainsi, dans un modèle de Cox avec une seule covariable Z, le risque relatif est

– pour une variable binaire codée 0 et 1 : $RR = \exp(\beta)$,

- pour une variable binaire codée a et b : $RR = \exp(\beta(b-a))$,
- pour une variable continue, $\exp(\beta)$ correspond au risque relatif pour une augmentation d'une unité de variable. Le risque relatif pour augmentation d'une unité de la variable quelle que soit la valeur de la covariable : c'est une hypothèse de log-linéarité.

Il y a donc deux hypothèses importantes à vérifier dans l'utilisation du modèle de Cox : l'hypothèse de risque proportionnels (risque relatif constant au cours du temps) et l'hypothèse de log-linéarité.

4.3 Modèle de cox généralisé

Considérons un modèle à risque proportionnel défini par l'expression suivante de la fonction de risque :

$$h(t/z) = h_0(t)exp(g({}^t\beta z_{ij})) \tag{4.6}$$

où $z(z_1, ..., z_p)$ est le vecteur ligne des variables exogène et β défini par ${}^t\beta = (\beta_1, ..., \beta_p)$ est le vecteur colonne des coefficient de régression.
Prenons le log de l'équation (4.6),

$$logh(t/z) = logh_0(t) + (g({}^t\beta z_{ij})$$

Les spécifications les plus courantes de la fonction g sont les suivantes :

- $exp({}^t\beta z)$ c'est le modèle original de Cox étudié dans la première partie.
- I l'identité
- $log(1 + exp({}^t\beta z))$ c'est le modèle logistique

alors par une dérivation par rapport à z_{ij} on aura :

$$\frac{\partial logh(t/z)}{\partial z_{ij}} = \beta g'({}^t\beta z_{ij})$$

et la moyenne s'écrit comme suit

$$E(\frac{\partial logh(t/z)}{\partial z_{ij}}) = \beta E(g'({}^t\beta z_{ij}))$$

on a $g(z) = E(y/z)$ alors elle peut être dérivable et la formule devient comme suit :

$$\frac{\partial E(y/z)}{\partial z_{ij}} = \beta G'(z_{ij}\beta)$$

En outre pour toute fonction de densité f

$$E[f(t)\frac{\partial g({}^t\beta z_{ij})}{\partial z_{ij}}] = \int g'({}^t\beta z_{ij})f(z)^2dz = -2\int g(z)f'f(z)dz = -2E(yf'(z))$$

$E(y/z)$ est proportionnel à β

Lemme 4.3.1.

$$E[f(t)\frac{\partial g({}^t\beta z_{ij})}{\partial z_{ij}}] = -2E(yf'(z))$$

Démonstration.

$$\int g'({}^t\beta z_{ij})f(z)^2dz = [gf(z)^2]_0^\infty - 2\int fgf'dz = -2E[yf']$$

Car :
$dv = dg$ $\quad v = g$
$u = f'$ $\quad dv = 2f'fdz$
où $[gf(t)^2]_0^\infty = 0$ □

On propose l'estimation de $-2E(yf'(z))$ Prenant δ est définit par $E(y\frac{\partial f(z)}{\partial z})$, on obtient l'estimateur efficace de δ en remplaçant f par son estimateur non paramétrique. L'estimateur de δ est donné par :

$$\delta_n = -\frac{2}{n}\sum_{i=1}^{n} y_i \frac{\partial f_n(z_i)}{\partial z}$$

où $\{y_i, z_i, i = 1, ..., n\}$ c'est les valeurs des observations et $f_n(z_i)$ c'est l'estimateur de la probabilité conjointe de la fonction de densité $f(x_i)$. Le fait que la probabilité conjointe de la fonction de densité de z est utilisée comme fonction de poids, le résultat de l'estimateur δ_n est appelé "densité moyenne pondérée de l'estimateur dérivé" en englais c'est "Density Weighted Average Derivative Estimator" (DWADE). Pour terminer l'estimation dans la relation précédente, l'estimateur de f doit être défini. Pour faciliter l'analyse l'estimateur utilisé est ceux du noyau. k représente la dimension de z, K le noyau et h_n c'est la fenêtre, alors l'estimateur est défini comme suit

$$f_n(z) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \Big(\frac{1}{h_n}\Big)^k K\Big(\frac{z - Z_j}{h_n}\Big)$$

On peut encore avoir que $\frac{\partial f_n(z)}{\partial z}$ l'estimateur de $\frac{\partial f(z)}{\partial z}$, donc sa formule sera la suivante :

$$\begin{aligned}\frac{\partial f_n(z)}{\partial z} &= \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \Big(\frac{1}{h_n}\Big)^k K'\Big(\frac{z - Z_j}{h_n}\Big)\Big(\frac{1}{h_n}\Big) \\ &= \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \Big(\frac{1}{h_n}\Big)^{k+1} K'\Big(\frac{z - Z_j}{h_n}\Big)\end{aligned}$$

où $K^{'}$ c'est la première dérivé de K. En remplaçant cette dernière formule dans l'estimateur de δ on aura :

$$\delta_n = -\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i} \Big(\frac{1}{h_n}\Big)^{k+1} K^{'}\Big(\frac{Z_i - Z_j}{h_n}\Big) Y_i$$

4.3.1 Cas où g est une fonction logistique

$$E(y/z) = G({}^t\beta z_{ij})$$

On va prendre

$$G({}^t\beta z_{ij}) = \frac{1}{1 + exp(-{}^t\beta z_{ij})}$$

Encore

$$\frac{\partial E(y/z)}{\partial z_{ij}} = \frac{{}^t\beta \exp(-{}^t\beta z_{ij})}{(1 + exp(-{}^t\beta z_{ij}))^2}$$

La vraisemblance s'écrit alors :

$$L(\beta) = \prod_{i=1}^{n} \left[\left(G({}^t\beta z_{ij}) \right)^{Y_i} \left(1 - G({}^t\beta z_{ij}) \right)^{1-Y_i} \right]$$

La log vraisemblance s'écrit :

$$\mathcal{L}(\beta)(logL(\beta)) = \sum_{i=1}^{n} Y_i log G({}^t\beta z_{ij}) + \sum_{i=1}^{n} (1 - Y_i)(1 - G({}^t\beta z_{ij})$$

Prenons la dérivée

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{Y_i - G({}^t\beta z_{ij})}{G({}^t\beta z_{ij}(1 - G({}^t\beta z_{ij}))} g({}^t\beta z_{ij}) z_{ij}$$

Cette dernière relation se simplifie car

$$g({}^t\beta z_{ij}) = G({}^t\beta z_{ij})\left[1 - G({}^t\beta z_{ij})\right] = \frac{exp(-{}^t\beta z_{ij})}{1 + exp(-{}^t\beta z_{ij})}$$

On obtient

$$\sum_{i=1}^{n}\left[Y_i - G({}^t\beta z_{ij})\right]({}^t z_{ij})$$

où g est la dérivée de G, et la matrice de dérivées secondes, ou Hessien :

$$\frac{\partial^2 l}{\partial\beta_i\beta_j} = -\sum_{i=1}^{n}\left[\frac{Y_i}{G^2({}^t\beta z_{ij})} + \frac{1-Y_i}{(1-G({}^t\beta z_{ij}))^2}\right]g^2({}^t\beta z_{ij})z_{ij}^t z_{ij} + \sum_{i=1}^{n}\frac{Y_i - G({}^t\beta z_{ij})}{G({}^t\beta z_{ij})(1-G({}^t\beta z_{ij}))}g'({}^t\beta z_{ij})z_{ij}^t z_{ij}$$

ainssi la matrice d'information de Fisher

$$I(\beta) = -E(\frac{\partial^2 l}{\partial\beta_i\beta_j})$$

Conclusion générale

Dans les modèles de durée multi-états, la fonction de risque (fonction du hasard) est identifiée à la "vitesse" de transition qui est donnée par la matrice d'intensité (générateur) dans le cas markovien. Ainsi, il devient possible d'étudier les différents risques et de les comparer dans une dynamique globale.

Dans le cas semi-markovien, nous avons montré que si le support de la loi de la durée pour chaque état est bornée, alors la dynamique semi-markovienne peut être rendue markovienne par une transformation appropriée. Comme les applications à un niveau fonctionnel sont possibles [27], [37], ... et on peut exploiter l'approche même quand l'ensemble des états a la puissance du continu ; cas des solution d'équations différentielles stochastique avec retard fini ou sans retard.

En termes d'applications, ces dernières années, l'intérêt des modèles multi-états ne cesse de croître, notamment en épidémiologie. Ils permettent, notamment, de représenter, de manière pertinente, l'évolution de l'état d'un patient à travers les différents stades d'une maladie, par exemple.

Bibliographie

[1] Barbu, V. S. and Limnios, N. *(2008). Semi-Markov Chains and Hidden Semi-Markov Models toward Applications (Springer-Verlag, New York).*

[2] Benchekor, A., Yousfate, A. *(2017). A strong Markov model for a discrete time inhomogeneous Semi-Markov Process with bounded phase staying. International Journal of Statistics and Economics. Vol. 18, Issue. 2, 1-9.*

[3] Berman, S.M. *(1963). Note on Extreme Values, Competing Risks and Semi-Markov Processes. The Annals of Mathematical Statistics, 34(3), 1104-1106.*

[4] Chiang, C.L. *(1961). On the Probability of Death from Specific Causes in the Presence of Competing Risks. Proceedings of 4th Berkeley Symposium on Mathematics, Statistics and Probability, 4,169-180. University of California Press.*

[5] Cinlar, E. *(1975). Introduction to Stochastic Processes. Norman J. Dover Publications, Inc. Mineola, New York.*

[6] Clayton, D.G. *(1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 65, 141-151.*

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

[8] Cox, D. R. *(1975). Partial likelihood. Biometrika, 62 : 269-276. 136.*

[9] Elbers, C. and Ridder, G. *(1982). True and spurious duration dependence : The identifiability of the proportional hazard model. Review of Economic Studies, 49, 403-411.*

[10] Franciszek, G. *(2014). Semi-Markov processes : application in system reliability and maintenance.Journal of Polish Safety and Reliability Association Summer Safety and Reliability Seminars, Volume 5.*

[11] Georgiadis, S. and Limnios, N. *(2014). Interval reliability for semi-Markov systems in discrete time. Journal de la Sociét Française de Statistique. Reliability Engineering & System Safety, 131, 282-290.*

[12] Gourieroux, C. *(1984). Econometrie des variables qualitatives. Collection "Economie et statistiques avancées". Paris.*

[13] Guo, S. *(2010). Survival Analysis. Published by Oxford University Press.*

[14] Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. *(2004). Nonparametric and Semiparametric Models (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA).*

[15] Heckman, J. and Singer B. *(1984). The Identifiability of the Proportional Hazard Model. The Review of Economic Studies, 51(2), 231-241.*

[16] Hougaard, P. *(1986). Survival Models for Heterogeneous Populations Derived from Stable Distributions. Biometrika, 73(2), 387-396.*

[17] Huber, C., Limnios, N., Mesbah, M. and Nikulin, M. (Edts) *(2008). Mathematical Methods in Survival Analysis, Reliability and Quality of Life. John Wiley & Sons, Inc. Hoboken, NJ 07030 USA.*

[18] Janssen, J. and Limnios, N. *(1999). Semi-Markov Models and Applications. Kluwer Academic Publishers Dordrecht/Boston/London.*

[19] Janssen, J. & Manca, R. *(2007).* Semi-Markov Risk Models for Finance, Insurance and Reliability. Springer.

[20] Jared, O and Søren, H. *(2011). Hidden Semi Markov Models for Multiple Observation Sequences : The mhsmm Package for R. Journal of Statistical Software, Volume 39, Issue 4.*

[21] Kaplan, E. L. and Meier, P. *(1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, vol. 53, 457-481.*

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

[23] Karlin, S. and Taylor H. M. A. *(1981). A second course in stochastic processes. Academic press, inc. edn, New York.*

[24] Korolyuk, V. S. and Turbin, A. F. *(1993). Decomposition of Large Scale Systems. Kluwer Academic, Singapore.*

[25] Korolyuk, V. S. and Swishchuk, A. *(1995). Random Evolution for Semi-Markov Systems. Kluwer Academic, Singapore.*

[26] Krol, A. and Saint-Pierre, P. *(2015). SemiMarkov : An R Package for Parametric Estimation in Multi-State Semi-Markov Models. Journal of Statistical Software, vol. 66, no. 6, 1-16.*

[27] Laksaci, A. and Yousfate, A. *(2002). Estimation fonctionnelle de la densité de l'opérateur de transition d'un processus de Markov à temps discret. C. R. Acad. Sci. Paris, Ser. I, 334, 1035–1038.*

[28] Lévy, P. *(1954). Processus semi-markoviens. Proceedings of the International Congress of Mathematics, 416-426.*

[29] Limnios, N. and Oprisan, G. *(2001). Semi-Markov Processes and Reliability. Statistics for Industry and Technology.*

[30] Limnios, N. and Ouhbi, B. & Sadek, A. *(2007). Empirical Estimator of Stationary Distribution for Semi-Markov Processes. Communications in Statistics - Theory and Methods, 34 :4, 987-995.*

[31] Listwon-Krol, A. and Saint-Pierre, P. *(2016). Package 'SemiMarkov' : Multi-States Semi-Markov Models. License GPL. Repository CRAN.*

[32] Neveu, J. *(1964).* Bases mathématiques du calcul de probabilités. Masson. Paris3.

[33] Nikulin, M. S., Balakrishnan, N., Mesbah, M. & Limnios, N. *(2004)* Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life. Statistics for Industry and Technology.

[34] Powell, J. L., Stock, J. H. and Stoker, T. M. *(1989). Semi-parametric of index coefficients.Econometrica, Vol. 57, No. 6, 1403-1430.*

[35] Proença, I. and Werwatz, A. *1994). Comparing Parametric and Semiparametric Binary Response Models, Sonderforschungsbereich 373 2000-20 (Humboldt Universitaet, Berlin).*

[36] Pyke, R. *(1961). Markov renewal processes with finitely maNew York states. Ann. of Math. Statist., 32, 1243-1259.*

[37] Roussas, G. *(1991). Recursive estimation of the transition distribution function of aMarkov process asymptotic normality, Statist. Probab. Lett., 11, 435–447.*

[38] Samuel, B., Roberto, D. & Pierre, D. *(2015). A mixture Cox-Logistic model for feature selection from survival and classification data. ICTEAM – Machine Learning Group.*

[39] Smith, W. L. *(1955). Regenerative stochastic processes. Royal Society of London Proceedings Series A, Vol 232, 6-31.*

[40] Takacs, L. *(1954). Some investigations concerning recurrent stochastic processes of a certain type. Magyar Tud. Akad. Mat. Kutato Int. Kzl., Vol 3, 115-128.*

[41] Takacs, L. *(1959). On a sojourn time problem in the theory of stochastic processes, 1rans. Amer. Math. Soc., 93, 531-540.*

[42] Themeau, R. M. and Grambsch, P. M. *(2000). Modeling Survival Data : Extending the Cox Modely. Statistics for Biology and Health.*

[43] Tsiatis, A. A. *(1975). A Nonidentifiability Aspect of the Problem of Competing. Proceedings of the National Academy of Sciences of the USA, 72(1), 20-22.*

[44] Vaupel, J.W., Manton, K.G. and Stallard, E. *(1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16, 439-454.*

[45] Weibull, W. *(1951). A Statistical Distribution Funtion of Wide Applicability. Journal of Applied Mechanics, 18, 293-297.*

[46] Ycart, B. *(2004). Processus markoviens de saut. Cahiers de Mathématiques Appliquées, CMA12.*

[47] Yousfate, A. *(1986). Décomposition canonique d'un processus qualitatif de type markovien stationnaire. Rev. Stat. & An. Données. Vol. 11, no. 1, 64-89.*