



N° Attribué par la bibliothèque

--	--	--	--	--	--	--	--	--	--



Année univ.: 2018/2019

Estimation de la fonction de risque pour les modèles Markoviens homogènes

Mémoire présenté en vue de l'obtention du diplôme de

Master Académique

Université de Saïda - Dr Moulay Tahar

Discipline : MATHÉMATIQUES

Spécialité : Analyse stochastique, statistique de processus et
applications

par

Louza Khadidja¹

Sous la direction de

Dr R. Rouane

Soutenue le 14/07/2019 devant le jury composé de

Dr. M. Kadi	Université Dr Tahar Moulay - Saïda	Président
Dr. R. Rouane	Université Dr Tahar Moulay - Saïda	Encadreur
Dr. F. Mokhtari	Université Dr Tahar Moulay - Saïda	Examinatrice
Dr. N. Hachemi	Université Dr Tahar Moulay - Saïda	Examinatrice

1. e-mail :joujoulaqueen@gmail.com

REMERCIEMENT

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail.

En second lieu, Je tiens à exprimer toute ma reconnaissance à ma directrice de mémoire, **Dr.Rachida Rouane**. Je la remercie de m'avoir encadré, orienté, aidé et conseillé.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leurs propositions.

Je remercie mes amis et camarades de promotion (2018/2019) pour ces cinq années passées ensemble, dans les meilleurs moments comme dans les pires.

Mes profonds remerciements vont à l'ensemble des enseignants qui ont participé à ma formation.

Enfin, je tiens également à remercier tous ceux qui ont contribué de près ou de loin à l'élaboration de ce mémoire.

Dédicace

Je dédis ce mémoire :

A mon père que dieu te garde pour moi.

A ma très chère mère, qui n'a pas cessé de m'encourager à aller de l'avant et de prier pour moi.

A ma très chère soeur Fouzia, et sa fille Fatima.

A ma petite soeur chérie Souad, je te souhaite le bon courage.

A mon cher frère Brahim et Abdelhafid vous êtes mon bonheur.

A toute ma famille et mes cousines.

A tout qui prend une place de mon coeur.

Table des matières

Introduction générale	5
1 Modèles de survie classiques	9
1.1 Concepts de base et notation	9
1.1.1 Distribution de la durée de survie	10
1.1.2 Espérance de durée de vie restante	12
1.1.3 Exemples :	12
1.2 Données censurées	13
1.2.1 Censure à droite	14
1.2.2 Censure à gauche	16
1.2.3 Censure par intervalle	16
1.3 Vraisemblance dans un modèle de survie censuré	17
1.3.1 Vraisemblance pour la censure à droite	17
1.3.2 Vraisemblance pour la censure par intervalle	17
1.4 Méthodes d'estimations	18
1.4.1 Estimation paramétrique	18
1.4.2 Estimation non-paramétrique	19
2 Extension du modèles de durées à l'analyse multi-états markoviens	25
2.1 Modèles multi-états	25
2.2 Définitions et propriétés	26
2.2.1 Processus Markovien	26
2.3 Modèle de Markov homogène	29
2.4 Vraisemblances d'un modèle multi-état	33
2.4.1 Données censurées par intervalle	33

2.4.2	Données exactes et censure à droite	36
3	Application	41
3.1	Application aux données AML	41
3.1.1	Méthode paramétrique	42
3.1.2	Méthode non paramétrique	44
3.2	Application aux données CAV	48
	Conclusion	53

Introduction générale

L'analyse des durées de vie est un domaine de la statistique qui étudie l'apparition d'un événement au cours du temps. Pour ce faire, il est nécessaire de disposer du temps de suivi de tous les individus, ainsi que du moment auquel l'événement est produit. De nombreuses études épidémiologiques s'intéressent en effet à l'analyse des temps d'événements. L'inférence statistique pour ces données consiste à estimer les fonctions associées à la distribution de la durée de survie et la fonction de hasard. L'estimation de cette fonction joue un rôle très important en statistique. Elle est utilisée dans l'analyse de risque ou pour l'étude des phénomènes de survie dans de nombreux domaines tels (médecine, géographique, fiabilité, ...). Des mécanismes de censure viennent très souvent affecter ces données et il est rare de pouvoir travailler à partir d'un échantillon complètement observé de durées de survie. Une littérature importante est consacrée au développement des méthodes pour traiter efficacement ces données incomplètes (Andersen et al., 1993; Kalbfleisch et Prentice, 2002). La censure à droite constitue sans doute l'exemple le plus populaire et a été très étudiée. La censure par intervalle décrit un phénomène où la durée de survie n'est jamais directement observée, c'est-à-dire la seule information dont on dispose est que la durée de survie appartient à un intervalle dont les bornes sont définies par des temps d'observations.

La fonction de survie intègre l'ensemble des événements observés avant t et décrit mal la dynamique instantanée du processus. La fonction de risque est la plus intéressante car elle décrit le futur immédiat du sujet et reflète des différences entre les modèles souvent moins explicites avec les fonctions de répartition et de survie. Cette fonction de risque peut être :

Paramétrable : exprimable sous forme d'une formule mathématique. C'est le cas du modèle exponentiel (qui suppose un risque instantané constant au

cours du temps) et du modèle de Weibull.

Non paramétrable : sans faire des hypothèses sur la distribution.

Dans ce contexte, les modèles multi-états sont une généralisation des modèles de survie, caractérisés par un processus stochastique à espace fini d'états, utilisé pour décrire l'évolution des sujets à travers différents états de santé dans le temps. Un modèle multi-états est un processus stochastique qui à tout temps peut occuper un état défini. Les états du processus sont souvent définis en fonction des symptômes cliniques présentés par les sujets au cours de leur suivi, ou en fonction des marqueurs biologiques. Un changement d'état est appelé transition. Lorsque le processus peut sortir d'un état pour ne plus jamais y revenir, cet état est dit transitoire. À l'inverse, un état à partir duquel un processus ne peut pas sortir est dit absorbant. Il existe une littérature abondante sur les modèles multi-états. On trouve les principales contributions dans le livre d'Andersen et al. (1993) et l'article de Hougaard (1999). Des modèles complexes avec un nombre important d'états et de transition variées sont envisageables. Toutefois, ils possèdent des structures qui ne sont que des extensions de structures standards (Hougaard, 1999). L'intérêt des modèles multi-états est d'étudier les taux instantanés d'échange au cours du temps entre les états, afin d'évaluer les effets des facteurs de risque. Les taux instantanés sont appelés intensités de transition ou fonctions de risque. Ces fonctions illustrent la force de circulation entre les états. L'hypothèse de modélisation de la fonction de risque est alors fondamentale. De nombreuses études utilisent les modèles markoviens homogènes. Ces derniers étant sans mémoire, l'évolution du processus est indépendante de la durée dans l'état actuel. Ce mémoire a pour objectif principal de présenter l'estimation paramétrique et non paramétrique de la fonction de risque pour les modèles multi-états markoviens, en développant la théorie et en l'appliquant à de données sous le logiciel R. Nous présentons les méthodes relative au modèle de Markov homogène, qu'il suppose que les intensités de transition entre les états sont constantes. Les méthodes d'estimation supposent que le mécanisme de censure (par intervalle, à droite) est indépendant de l'événement étudié.

Nous présentons notre travail dans trois chapitres. Dans le premier chapitre nous rappelons les concepts de base des modèles de survie qui seront

adoptés dans la suite de notre travail.

Le deuxième chapitre présente les liens qui existent entre la théorie des fonctions de risque et les processus stochastiques. Les processus stochastiques utilisés sont des chaînes de Markov homogènes. Le troisième chapitre est consacré à l'application de quelques procédures présentées dans les chapitres antérieurs. Nous avons pratiqué les résultats obtenus dans le cas d'une application sur des données médicales censurées à droite.

Chapitre 1

Modèles de survie classiques

Ce chapitre est consacré à une présentation générale sur les éléments de base des modèles de survie. Dans la première partie de ce chapitre, nous présentons le lien entre deux concepts centraux qui sont les fonctions de survie et de risque. Dans la deuxième partie, nous définissons la notion de censure, en particulier la censure à droite et la censure par intervalle. La troisième partie traite la construction des vraisemblances qui seront utilisées dans les chapitres suivants. Enfin, nous présentons les méthodes d'estimation et nous rappelons la définition de l'estimateur de Kaplan-Meier de la fonction de survie et de l'estimateur de Nelson-Aalen de la fonction de risque cumulée.

1.1 Concepts de base et notation

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement précis. L'événement étudié (communément appelé "décès") est le passage irréversible entre deux états (communément nommé "vivant" et "décès"). L'événement terminal n'est pas forcément la mort : il peut s'agir de l'apparition d'une maladie (par exemple, le temps avant une rechute ou un rejet de greffe), d'une guérison (temps entre le diagnostic et la guérison), la panne d'une machine (durée de fonctionnement d'une machine, en fiabilité).

Nous désignons par T une variable aléatoire continue positive définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ associé à une durée de vie (c'est à dire une v.a à valeurs dans $\mathbb{R}$) passée dans un certain état, l'origine des temps étant

prédéfinie.

Les durées de survie sont des variables aléatoires positives, de distribution, le plus souvent dissymétrique rendant difficile leur description par les distributions théoriques usuelles.

On cherche alors à estimer la distribution des temps de survie (fonction de survie), à comparer les fonctions de survie de plusieurs groupes ou à analyser la manière dont des variables explicatives modifient les fonctions de survie.

1.1.1 Distribution de la durée de survie

Supposons que la variable aléatoire T positive correspondant à la durée de survie absolument continue, alors sa loi de probabilité peut être définie par l'une des cinq fonctions équivalentes suivantes (chacune des fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions) :

Définition 1.1.1. (*Fonction de survie S*) La fonction de survie est, pour t fixé, la probabilité de survivre jusqu'à l'instant t , c'est-à-dire :

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

Définition 1.1.2. (*Fonction de répartition F*) La fonction de répartition représente, pour t fixé, la probabilité de mourir avant l'instant t , c'est-à-dire :

$$F(t) = \mathbb{P}(T \leq t) = 1 - S(t).$$

Définition 1.1.3. (*Densité de probabilité f*) C'est la fonction $f(t) > 0$ telle que pour tout $t > 0$, $F(t) = \int_0^t f(u) du$. Si la fonction de répartition F admet une dérivée au point t alors :

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t)}{\Delta t} = F'(t) = -S'(t).$$

Pour t fixé, la densité de probabilité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t .

Définition 1.1.4. (*Risque instantané $\lambda(t)$ (ou taux de hasard)*) Le risque instantané ou La fonction de risque (hasard), ou bien le risque instantané de changement d'état, pour t fixé caractérise la probabilité de mourir

dans un petit intervalle de temps après t , conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T > t)}{\Delta t}.$$

On montre facilement que :

$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$\lambda(t) = \frac{-d \log(S(t))}{d(t)}$$

On constate que la fonction de risque caractérise la loi de T (ou $S(t)$) du fait de la relation :

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

Cette formule conduit d'ailleurs à introduire le hasard cumulé.

Définition 1.1.5. (Taux de hasard cumulé Λ) *Le taux de hasard cumulé est l'intégrale du risque instantané λ :*

$$\Lambda(t) = \int_0^t \lambda(u) du = -\ln(S(t)).$$

Donc :

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

On en déduit que :

$$f(t) = \lambda(t) \exp \left(- \int_0^t \lambda(u) du \right).$$

La fonction de risque $\lambda(t)$ est la plus intéressante car elle décrit le futur immédiat du sujet et reflète des différences entre les modèles souvent moins explicites avec les fonctions de répartition et de survie.

Définition 1.1.6. (Fonction de survie conditionnelle)

$$\begin{aligned}
S(t \setminus t_0) &= \mathbb{P}(T > t + t_0 \setminus T > t_0) \\
&= \frac{S(t + t_0)}{S(t_0)} \\
&= \exp\left(-\int_{t_0}^t \lambda(u) du\right), \forall t > t_0.
\end{aligned}$$

1.1.2 Espérance de durée de vie restante

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

$$r(t) = \mathbb{E}(T - t \setminus T > t).$$

On montre que :

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

Preuve :

$$\begin{aligned}
r(t) &= \mathbb{E}(T - t \setminus T > t) \\
&= \frac{\mathbb{E}(T - t) \cdot \mathbb{1}(T > t)}{\mathbb{P}(T > t)} \\
&= \frac{1}{S(t)} \left\{ \int_t^{+\infty} u f(t) du - tS(t) \right\} \\
&= \frac{1}{S(t)} \left\{ [-uS(u)]_t^{+\infty} \int_t^{+\infty} S(u) du - tS(t) \right\}
\end{aligned}$$

$$\text{si } \lim_{u \rightarrow +\infty} uS(u) = 0, \text{ alors } r(t) = \frac{1}{S(t)} \int_t^{+\infty} S(u) du. \quad \square$$

1.1.3 Exemples :

La loi exponentiel :

Nous pouvons donc caractériser la loi de la durée T par une fonction de risque instantané constante $\forall t \in \mathbb{R}$:

$$\lambda(t) = \lambda.$$

Où, λ est une constante strictement positive. On obtient ainsi les fonctions définies ci-dessus :

$$\begin{aligned}
f(t) &= \lambda \exp(-\lambda t) \\
F(t) &= 1 - \exp(-\lambda t) \\
S(t) &= \exp(-\lambda t) \\
\Lambda(t) &= \lambda t \\
r(t) &= 1/\lambda = \mathbb{E}(t).
\end{aligned}$$

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

La loi de Weibull :

La généralisation de Weibull pour la loi exponentielle à la loi du même nom en introduisant un nouveau paramètre, de manière à ce que la fonction de risque soit la suivante : $\forall t \in \mathbb{R}^+$

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

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

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

$$\begin{aligned}
f(t) &= \lambda \alpha t^{\alpha-1} e^{-\lambda t^\alpha} \\
F(t) &= 1 - e^{-\lambda t^\alpha} \\
S(t) &= e^{-\lambda t^\alpha} \\
\Lambda(t) &= \lambda t^\alpha
\end{aligned}$$

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

1.2 Données censurées

Une des caractéristiques des données de survie est l'existence d'observation incomplète. En effet, les données sont souvent recueillies partiellement,

notamment, à cause des processus de censure. Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information au lieu d'observer des réalisations indépendantes et identiquement distribuées (i.i.d.) de durées T , on observe la réalisation de la variable T soumise à diverses perturbations, indépendantes ou non du phénomène étudié.

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Pour l'individu i , considérons :

- son temps de survie X_i
- son temps de censure C_i
- la durée réellement observée T_i .

1.2.1 Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'évènement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées, pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

1. La censure de type I :

Soit C une valeur fixée, au lieu d'observer les variables $X_1, \dots, X_n$ qui nous intéressent, on observe X_i uniquement lorsque $X_i \leq C$, sinon on sait uniquement que $X_i > C$. On utilise la notation suivante :

$$T_i = X_i \wedge C = \min(X_i, C).$$

Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles. Par exemple, on peut tester la durée de vie de n objet identique (ampoule) sur un intervalle d'observation fixé $[0, u]$. En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps u sont sacrifiées).

2. La censure de type II :

Elle est présente quand on décide d'observer les durées de survie de n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient X_i et T_i les statistiques d'ordre des variables

X_i et T_i . La date de censure est donc X_k et on observe les variables suivantes,

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

3. La censure de type III (ou censure aléatoire de type I) :

Soient $C_1, \dots, C_n$ des variables aléatoires i.i.d. On observe les variables

$$T_i = X_i \wedge C_i.$$

L'information disponible peut être résumée par :

- la durée réellement observée T_i
- un indicateur $\Delta t = \mathbb{1}_{X_t \leq C_t}$
- $\delta_i = 1$ si l'événement est observé (d'où $T_i = X_i$). On observe les "vraies" durées ou les durées complètes.
- $\delta_i = 0$ si l'individu est censuré (d'où $T_i = C_i$). On observe des durées incomplètes (censurées).

La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par :

- a** - La perte de vue : le patient quitte l'étude en cours et on ne le revoit plus (à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients "perdus de vue".
- b** - L'arrêt ou le changement du traitement : les effets secondaires ou l'inefficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.
- c** - La fin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'évènement). Ce sont des patients "exclus-vivants". Les "perdus de vue" (et les exclusions) et les "exclus-vivants" correspondent à des observations censurées mais les deux mécanismes sont de nature différentes (la censure peut être informative chez les "perdus de vue").

1.2.2 Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires (T, δ) :

$$T = X \vee C = \max(X, C),$$

$$\delta = \mathbb{1}_{\{T \geq C\}}.$$

Comme pour la censure à droite, on suppose que la censure C est indépendante à X .

Un des premiers exemples de censure à gauche rencontré dans la littérature considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger (les babouins passent la nuit dans les arbres).

Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).

1.2.3 Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'évènement, la seule information disponible est qu'il a eu lieu entre deux dates connues. Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (pas en continu), on sait alors uniquement que l'évènement s'est produit entre ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'évènement correspond au temps de la visite pour se ramener à la censure à droite.

1.3 Vraisemblance dans un modèle de survie censuré

1.3.1 Vraisemblance pour la censure à droite

Soit X une durée de vie aléatoire. On suppose que la loi $\mathbb{P}_X$ de X appartient à une famille de lois de probabilité $\{\mathbb{P}_\theta; \theta \in \Theta\}$ où $\Theta \subseteq \mathbb{R}^p$. La vraie loi de X est ainsi notée $\mathbb{P}_{\theta_0}$, où $\theta_0 \in \Theta$. On note $f_{X,\theta}(\cdot)$, $F_{X,\theta}(\cdot)$, $S_{X,\theta}(\cdot)$, $\lambda_{X,\theta}(\cdot)$, $\Lambda_{X,\theta}(\cdot)$: les densités, fonction de répartition, fonction de survie, fonction de risque instantané et fonction de risque cumulé de la variable durée de vie T , sous la loi $\mathbb{P}_\theta$, la variable de censure C est supposée indépendante de la variable T et on suppose que la distribution de T est entièrement définie par la connaissance d'un paramètre θ de dimension finie (à estimer par MV), on dit que la loi de la censure C est non informative. On note $f_C(\cdot)$, $F_C(\cdot)$, $S_C(\cdot)$ les densités, fonction de répartition et fonction de survie de la variable C .

Notons $(T_i, \delta_i)_{i \in \{1, \dots, n\}}$ un échantillon des variables (T, C) , où $T_i = \min(X_i, C_i)$ est la durée observée, et l'indicateur de censure $\delta = \mathbb{1}_{\{X < C\}}$, ainsi la vraisemblance associée à l'échantillon $(T_i, \delta_i)_{i \in \{1, \dots, n\}}$ s'écrit sous la forme :

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n (f_{X,\theta}(T_i) S_C(T_i))^{\delta_i} (S_{X,\theta}(T_i) f_C(T_i))^{1-\delta_i} \\ &= \prod_{i=1}^n \lambda_{X,\theta}(T_i)^{\delta_i} S_{X,\theta}(T_i) S_C(T_i)^{\delta_i} f_C(T_i)^{1-\delta_i}. \end{aligned}$$

Sous l'hypothèse (de censure non informative), le paramètre d'intérêt θ n'apparaît pas dans la loi de la censure on remarque qu'il est équivalent de chercher l'estimateur du maximum de vraisemblance de θ en maximisant l'expression :

$$\prod_{i=1}^n \lambda_{X,\theta}(T_i)^{\delta_i} S_{X,\theta}(T_i) \quad (1.1)$$

1.3.2 Vraisemblance pour la censure par intervalle

Soit S la fonction de survie de la variable d'intérêt T . La contribution individuelle d'un sujet i dont l'observation est censurée dans l'intervalle $[l_i, r_i]$

est $S(l_i; \theta) - S(r_i; \theta)$. La vraisemblance totale s'écrit donc :

$$L = \prod_{i=1}^n (S(l_i; \theta) - S(r_i; \theta)). \quad (1.2)$$

1.4 Méthodes d'estimations

Notre objectif est de modéliser la fonction de risque $\lambda(t)$. Plusieurs approches peuvent être envisagées entre les modèles non paramétriques et paramétriques.

1.4.1 Estimation paramétrique

En analyse de survie, l'approche paramétrique consiste à considérer que la distribution de la durée de survie étudiée appartient à une famille donnée de lois paramétriques. Kalbfleisch et Prentice (2002) ou encore Lawless (1984) décrivent un grand nombre des modèles paramétriques pouvant être utilisés en analyse de survie. Cox et Oakes (1984) discutent des critères permettant de choisir entre plusieurs modèles paramétriques. Les lois paramétriques considérées doivent être des lois adaptées pour des variables aléatoires positives, notamment pour cette raison que la loi normale ne convient pas dans ce contexte. Les modèles paramétriques les plus utilisés sont le modèle de Weibull et le modèle exponentiel. L'estimation des paramètres θ du modèle peut se faire par la méthode du maximum de vraisemblance. Nous commençons par des méthodes paramétriques concernant la fonction de survie.

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

La vraisemblance du modèle s'écrit alors :

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

Donc, la contribution à la vraisemblance peut s'écrire sous la forme :

$$\begin{aligned}
 L(\theta) &= \prod_{u=n-c} f(t_i, \theta) \prod_c S(t_i, \theta) \\
 &= \prod_u \theta \exp(-\theta t_i) \prod_c \exp(-\theta t_i) \\
 &= \theta^{n_u} \exp(-\theta \sum_u t_i) \exp(-\theta \sum_c t_i) \\
 &= \theta^{n_u} \exp(-\theta \sum_{i=1}^n t_i)
 \end{aligned}$$

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

$$\hat{\theta} = \frac{n_u}{\sum_{i=1}^n t_i}$$

donc on a : $S(t) = \exp(-\theta(t))$ alors,

$$\hat{S}(t) = \exp(-\hat{\theta}(t)).$$

1.4.2 Estimation non-paramétrique

L'approche paramétrique s'appuie sur des hypothèses concernant la distribution de données, lesquelles sont difficiles à vérifier avec des données incomplètes. Une solution de remplacement possible est l'utilisation de méthodes non paramétriques. L'estimateur non paramétrique le plus simple de la fonction de répartition est la distribution empirique. Il correspond à l'estimateur non paramétrique du maximum de vraisemblance pour des observations complètes. De nombreux estimateurs ont été développés afin de considérer les mécanismes de censure. Les plus connus sont l'estimateur de la fonction de survie de Kaplan et Meier (1958) et celui de la fonction de risque cumulé de Nelson-Aalen (Nelson, 1972 ; Aalen, 1978) pour traiter des données censurées à droite.

Estimateur de Kaplan-Meier de la fonction de survie

Nous présentons ici l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958) comme estimateur du maximum de vraisemblance non paramétrique de la fonction de survie $S(t)$. Considérons des durées de survie $T_i, i = 1, \dots, n$ soit n variables aléatoires indépendantes et identiquement distribuées, ayant la même fonction de survie $S(t)$. L'estimateur le plus naturel de $S(t)$ est la survie empirique :

$$\hat{S}_n(t) = \sum_1^n \mathbb{1}_{T_i > t}.$$

Malheureusement, dans le cas où les données sont censurées à droite, il est impossible d'utiliser la fonction de survie empirique c'est pourquoi il sert généralement de base à toute étude sur les durées. Pour dériver l'estimateur de Kaplan-Meier, partons de nouveau de l'échantillon avec censure à droite : $(x_1, \delta_1), (x_2, \delta_2), \dots, (x_n, \delta_n)$. Notons $t_1, \dots, t_k$ les temps de survie non censurés de cet échantillon qui sont distincts et que l'on énumère par ordre croissant. Notons de plus par d_j , le nombre de fois où t_j apparaît dans l'échantillon. Finalement, dans chaque intervalle $[t_j, t_{j+1}]$, supposons qu'il y a m_j observations censurées, notés $c_{j,1}, \dots, c_{j,m_j}$, également énumérées par ordre croissant. Avec ces notations, l'échantillon ordonné se présente comme suit :

$$c_{0,1}, \dots, c_{0,m_0}; t_1(d_1 \text{ fois}); c_{1,1}, \dots, c_{1,m_1}, \dots; t_k(d_k \text{ fois}); c_{k,1}, \dots, c_{1,m_k}.$$

L'estimateur de Kaplan-Meier est obtenu en maximisant la vraisemblance de ces données sous hypothèse qu'elles proviennent d'une loi discrète portée par les temps de survie non censures de l'échantillon.

Posons donc $t_0 = 0, t_{k+1} = \infty$ et soit

$$p(t_1), p(t_2), \dots, p(t_k), p(t_{k+1})$$

les probabilités que la loi discrète met sur les points $t_1, \dots, t_k, t_{k+1}(t_{k+1}$ sert à traiter les données censurées plus grandes que t_k). La vraisemblance est donnée par :

$$L = \prod_{i=1}^n p(x_i)^{\delta_i} S(x_i)^{1-\delta_i}.$$

S est la fonction de survie de la loi discrète que l'on peut récrire, avec les nouvelles notations, comme suit :

$$L = S(c_{0,1})S(c_{0,2}) \cdots S(c_{0,m_0}) \times p(t_1)^{d_1} \times S(c_{1,1})S(c_{1,2}) \cdots S(c_{1,m_1}) \times p(t_2)^{d_2} \\ \times S(c_{2,1})S(c_{2,2}) \cdots S(c_{2,m_2}) \times p(t_k)^{d_k} \times S(c_{k,1})S(c_{k,2}) \cdots S(c_{k,m_k})$$

On fait alors l'observation cruciale que, pour la loi discrète considérée, on a

$$S(c_{j,\ell}) = S(t_j) \quad \ell = 1, \dots, m_j, \quad j = 1, \dots, k$$

Ce qui permet de simplifier la vraisemblance comme suit :

$$L = \prod_{j=1}^k p(t_j)^{d_j} S(t_j)^{m_j} = \prod_{j=1}^k p_j^{d_j} S_j^{m_j} \quad (1.3)$$

(car $S(t_0) = S(0) = 1$). L'estimateur de Kaplan-Meier s'obtient en maximisant cette vraisemblance sur l'ensemble des vecteurs de probabilités $(p_1, \dots, p_k, p_{k+1})$. Une telle maximisation est possible mais laborieuse. On fait alors une seconde observation importante : il est équivalent de maximiser (1.3) sur $(S_1, \dots, S_k)$ plutôt que $(p_1, \dots, p_k)$ puisqu'il y a correspondance biunivoque entre les deux. Comme $p_j = S_{j-1} - S_j$, la vraisemblance (1.3) devient

$$L = \prod_{j=1}^k (S_{j-1} - S_j)^{d_j} S_j^{m_j} \quad (1.4)$$

qu'il s'agit maintenant de maximiser sur le cône : $1 \geq S_1 \geq S_2 \geq \dots \geq S_k \geq 0$. On peut se contenter de maximiser le logarithme de (1.4) :

$$\log L = \sum_{j=1}^k d_j \log(S_{j-1} - S_j) + \sum_{j=1}^k m_j \log(S_j).$$

Pour cela, il suffit de chercher un point critique car, comme on peut le démontrer, il n'y a de fait qu'un seul point critique, qui est un maximum et qui est, par ailleurs, à l'intérieur du cône. Calculons ce point critique. En dérivant par rapport à S_1 , on trouve :

$$\frac{\partial \log L}{\partial S_1} = \frac{-d_1}{S_0 - S_1} + \frac{m_1}{S_1} = \frac{m_1}{S_1} - \frac{d_1}{1 - S_1}$$

Si on prend cette dérivé égale à zéro et qu'on isole S_1 , on trouve l'estimateur :

$$\hat{S}_1 = \frac{m_1}{d_1 + m_1} = 1 - \frac{d_1}{d_1 + m_1} = 1 - \frac{d_1}{n_1}$$

(avec $n_1 = m_1 + d_1$). Pour S_2 , on a

$$\frac{\partial \log L}{\partial S_2} = \frac{-d_2}{S_1 - S_2} + \frac{m_2}{S_2}$$

d'où on déduit

$$\hat{S}_2 = \frac{\hat{S}_1 m_2}{m_2 + d_2} = \hat{S}_1 \left(1 - \frac{d_2}{n_2}\right) = \left(1 - \frac{d_1}{n_1}\right) \left(1 - \frac{d_2}{n_2}\right)$$

La formule générale se dégage :

$$\hat{S}_k = \prod_{j=1}^k \left(1 - \frac{d_j}{n_j}\right)$$

On voit bien que $(\hat{S}_1, \dots, \hat{S}_k)$ est dans le cône puisque $\hat{S}_1 < 1$ et chaque autre $\hat{S}_j$ est obtenu du précédent en multipliant ce dernier par un facteur plus petit que 1. Pour un temps t en général, nous avons donc

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right) \quad (1.5)$$

L'estimateur ainsi obtenu est l'estimateur de Kaplan-Meier de la fonction de survie $S(t)$ au temps t . On l'appelle également l'estimateur du produit limite. C'est une fonction constante par morceaux qui ne saute qu'aux temps t_j . Elle est continue à droite et possède des limites à gauche.

Estimateur de Nelson-Aalen du risque cumulé

L'estimateur de Nelson-Aalen (Nelson, 1972 ; Aalen, 1978) est donné par

$$\hat{A}(t) = \sum_{j:t_j \leq t} \left(\frac{d_j}{n_j}\right)$$

où n_j représente le nombre de sujets à risque juste avant le temps t_j et d_j représente le nombre de sujets qui ont subi l'évènement au temps t_j . L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut à chaque

temps de l'évènement t_1 . L'estimateur du risque cumulé a une interprétation moins immédiate que celui de la fonction de survie. Son intérêt réside surtout dans la pente de la courbe correspondante qui estime la fonction de risque $\lambda(t)$.

Chapitre 2

Extension du modèles de durées à l'analyse multi-états markoviens

Dans ce chapitre, nous nous intéressons aux processus markoviens homogènes à temps continu et à espace d'états fini, en donnant quelques notions et définitions sur les processus. Cette notion de processus est indispensable dans les modèles multi-états qui sont une généralisation naturelle des modèles de survie. Ensuite, nous présentons les vraisemblances d'un modèle muti-états décrites dans les articles Hougaard (1999) et Kalbfleisch et Prentice(2002).

2.1 Modèles multi-états

L'analyse de survie étudie le délai de survenue d'un évènement d'intérêt, ou autrement dit, le délai entre deux états successifs (communément état "vivant" et état "décédé"). Les modèles multi-états, aujourd'hui très populaires, permettent d'étudier des dynamiques plus complexes en utilisant la notion de processus pour représenter l'évolution d'un sujet à travers différents états successifs. L'étude de ces modèles consiste à analyser l'intensité de transition entre les différents états. En épidémiologie, ils permettent par exemple de modéliser son évolution à travers les différents stades d'une maladie.

2.2 Définitions et propriétés

2.2.1 Processus Markovien

Un processus de Markov $(X_{t \geq 0})$ à temps continu et à espace d'états fini $S = \{1, \dots, k\}$ est un processus dont l'évolution future $(X_t, t \geq s)$ dépend uniquement de la connaissance du temps présent s et de l'états en ce temps X_s , pour tout $t \geq s \geq 0$

$$\mathbb{P}(X_t = x_j | X_r = x_r, r \leq s) = \mathbb{P}(X_t = x_j | X_s = x_s). \quad (2.1)$$

Cette définition signifie que tout le passé du processus est résumé dans l'état précédent ou encore, le présent étant connu, le futur est indépendant du passé.

Définition 2.2.1. *Un processus de Markov $(X_{t \geq 0})$ à temps continu et à espace d'états fini $S = \{1, \dots, k\}$ est complètement défini par*

1. *Son vecteur des probabilités initiales, noté P_0 tel que*

$$P_0[j] = \mathbb{P}(X_0 = j), \quad j = 1, \dots, k, \quad (2.2)$$

$$\text{avec } \sum_{j=1}^k \mathbb{P}(X_0 = j) = 1,$$

2. *Ses matrices de probabilités de transition : $P(s, t) = (p_{ij}(s, t))$ telle que*

$$p_{ij}(s, t) = \mathbb{P}(X_t = j | X_s = i) \quad \forall, \quad 0 \leq s \leq t \text{ et } i, j \in S \quad (2.3)$$

avec

$$P(s, s) = I \quad \text{et} \quad \sum_{j=1}^k p_{ij}(s, t) = 1, \quad \forall 0 \leq s \leq t. \quad (2.4)$$

Théorème 2.2.1. (Équations de Chapman-Kolmogorov). *Soit $(X_t, t \geq 0)$ un processus de Markov à temps continu et à espace d'états fini $S = \{1, \dots, k\}$, et de probabilités de transition $p_{ij}(s, t), 0 < s < t, i, j \in S$. Alors*

$$p_{ij}(s, t) = \sum_{\ell=1}^k p_{i\ell}(s, u) p_{\ell j}(u, t),$$

soit en notation matricielle

$$P(s, t) = P(s, u)P(u, t) \quad \forall 0 \leq s \leq u \leq t, \quad (2.5)$$

où $P(s, t)$ est la matrice dont les éléments sont les probabilités de transition $p_{ij}(s, t)$, pour tout $i, j \in S$.

Preuve : Par la formule des probabilités totales, et la définition des probabilités conditionnelles, alors

$$\begin{aligned} P_{ij}(s, t) &= \mathbb{P}(X_t = j | X_s = i) \\ &= \frac{\mathbb{P}(X_t = j | X_s = i)}{\mathbb{P}(X_s = i)} \\ &= \sum_{\ell=1}^k \frac{\mathbb{P}(X_t = j, X_s = i, X_u = \ell)}{\mathbb{P}(X_s = i)} \quad 0 \leq s \leq u \leq t \\ &= \sum_{\ell=1}^k \frac{\mathbb{P}(X_t = j | X_s = i, X_u = \ell) \mathbb{P}(X_s = i, X_u = \ell)}{\mathbb{P}(X(s) = i)} \\ &= \sum_{\ell=1}^k \mathbb{P}(X_t = j | X_s = i, X(u) = \ell) \mathbb{P}(X_u = \ell | X(s) = i) \\ &= \sum_{\ell=1}^k p_{i\ell}(s, u) p_{\ell j}(u, t), \end{aligned}$$

où la dernière égalité provient de la propriété markovienne (2.1). $\square$

Le générateur infinitésimal (ou matrice de taux de transition) $(Q(t), t \geq 0)$ d'un processus de Markov est défini par

$$\begin{aligned} q_{ij}(t) &= \lim_{\Delta t \rightarrow 0^+} \frac{p_{ij}(t, t + \Delta t) - \delta_{ij}}{\Delta t}, \quad i \neq j \\ q_{ii}(t) &= \lim_{\Delta t \rightarrow 0^+} \frac{p_{ii}(t, t + \Delta t) - 1}{\Delta t} \end{aligned}$$

On peut démontrer que ces deux limites existent. On voit par ailleurs que

$$q_{ii}(t) = - \sum_{i \neq j} q_{ij}(t).$$

En effet, on a

$$\begin{aligned}
-\sum_{i \neq j} q_{ij}(t) &= -\sum_{i \neq j} \lim_{\Delta t \rightarrow 0^+} \frac{p_{ij}(t, t + \Delta t)}{\Delta t} \\
&= -\lim_{\Delta t \rightarrow 0^+} \sum_{i \neq j} \frac{p_{ij}(t, t + \Delta t)}{\Delta t} \\
&= -\lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} \sum_{i \neq j} p_{ij}(t, t + \Delta t) \\
&= -\lim_{\Delta t \rightarrow 0^+} \frac{1}{\Delta t} (1 - p_{ii}(t, t + \Delta t)) \\
&= \lim_{\Delta t \rightarrow 0^+} \frac{p_{ii}(t, t + \Delta t) - 1}{\Delta t} \\
&= q_{ii}(t)
\end{aligned}$$

Les modèles multi-états ne sont qu'une extension de modèles de survie classique. Le paramètre d'intérêt dans ces modèles est la fonction de risque instantané $t \mapsto (\lambda_1(t), \dots, \lambda_k(t))$, qui est définie par

$$\begin{aligned}
\lambda_i(t) &= \sum_{i \neq j} \lambda_{ij}(t) \\
\lambda_{ij}(t) &= \lim_{\Delta t \rightarrow 0^+} \frac{p_{ij}(t, t + \Delta t)}{\Delta t}, \quad i \neq j.
\end{aligned}$$

Le risque instantané λ est donc donné par le générateur infinitésimal : $\lambda_i(t) = -q_{ii}(t)$. Notons que $q_{ij}(t)\Delta t$ lorsque Δt est très petit représente la probabilité que le processus soit à l'état j au temps $t + \Delta t$ sachant que le processus est dans l'état i au temps t . La fonction $t \mapsto q_{ij}(t)$ représente donc l'intensité de transition de l'état i vers l'état j au temps t .

Exemple :

Le modèle de survie ne comportant que deux états avec une seule transition possible de l'état 0 à l'état 1. Ce modèle peut être considéré comme un modèle de Markov $(X_t, t \geq 0)$ à temps continu et à espace d'états fini $S = \{0, 1\}$. La matrice des probabilités de transition est définie par :

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

avec $p_{00}(0, t)$ et $p_{01}(0, t)$ qui correspondent aux fonctions $S(t)$, $F(t)$.

2.3 Modèle de Markov homogène

Un processus de Markov est homogène si la probabilité de transition de l'état i vers j satisfait la relation :

$$p_{ij}(s, t) = p_{ij}(0, t - s), \quad \forall s \leq t, i, j \in S.$$

Les probabilités de transition dépendent uniquement du temps entre deux transitions et non du moment où se produisent ces transitions. On peut alors simplement réécrire les probabilités en termes de l'écart : $P(s, t) = P(t - s)$ et l'équation de Chapman Kolmogorov (2.5) devient

$$P(t - s) = P(u - s)P(t - u).$$

Dans un processus de Markov homogène, les intensités de transition $q_{ij}(t)$ du processus sont indépendantes du temps, pour tout $i \neq j$,

$$q_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(\Delta t)}{\Delta t} = q_{ij} \quad (2.6)$$

Ainsi, la matrice des intensités de transition Q est de la forme

$$\begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1k} \\ q_{21} & q_{22} & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kk} \end{pmatrix}$$

Théorème 2.3.1. *Soit $(X_{t \geq 0})$ un processus de Markov homogène de matrice des intensités de transition $Q = (q_{ij})_{i,j \in S}$. Alors on a :*

1. Équation de Kolmogorov rétrograde

$$\frac{d}{dt}P(t) = QP(t), t \geq 0.$$

$$\text{Soit pour tout } i, j \in S : \frac{d}{dt}P_{ij}(t) = \sum_{\ell=1}^k q_{i\ell}p_{\ell j}(t), \quad t \geq 0.$$

2. *Équation de Kolmogorov progressive :*

$$\frac{d}{dt}P(t) = P(t)Q, \quad t \geq 0.$$

$$\text{Soit pour tout } i, j \in S : \quad \frac{d}{dt}P_{ij}(t) = \sum_{\ell=1}^k p_{i\ell}q_{\ell j}(t), \quad t \geq 0.$$

Preuve : Démontrons la première équation, soit $t, \Delta t \geq 0$ et $i, j \in S$.
Par la formule de Chapman-Kolmogorov (2.5), nous avons :

$$P_{ij}(t + \Delta t) = \sum_{\ell=1}^k p_{i\ell}(\Delta t)p_{\ell j}(t)$$

soit

$$\begin{aligned} P_{ij}(t + \Delta t) - p_{ij}(t) &= \sum_{\ell=1}^k p_{i\ell}(\Delta t)p_{\ell j}(t) - p_{ij}(t) \\ &= \sum_{\ell \neq i} p_{i\ell}(\Delta t)p_{\ell j}(t) + p_{ii}(\Delta t)p_{ij}(t) - p_{ij}(t) \\ &= \sum_{\ell \neq i} p_{i\ell}(\Delta t)p_{\ell j}(t) + (p_{ii}(\Delta t) - 1)p_{ij}(t). \end{aligned} \quad (2.7)$$

En divisant les deux membres de l'équation(2.7) par Δt , puis prenant la limite quand $\Delta t \mapsto 0$, nous obtenons :

$$\begin{aligned} \lim_{\Delta t \rightarrow 0} \frac{P_{ij}(t + \Delta t) - p_{ij}(t)}{\Delta t} &= \lim_{\Delta t \rightarrow 0} \sum_{\ell \neq i} \frac{p_{i\ell}(\Delta t)}{\Delta t} p_{\ell j}(t) + \lim_{\Delta t \rightarrow 0} \frac{(p_{ii}(\Delta t) - 1)}{\Delta t} p_{ij}(t) \\ &= \lim_{\Delta t \rightarrow 0} \sum_{\ell \neq i} \frac{p_{i\ell}(\Delta t)}{\Delta t} p_{\ell j}(t) + q_{ii}p_{ij}(t) \\ &= \sum_{\ell \neq i} \lim_{\Delta t \rightarrow 0} \frac{p_{i\ell}(\Delta t)}{\Delta t} p_{\ell j}(t) + q_{ii}p_{ij}(t) \\ &= \sum_{\ell \neq i} q_{i\ell}(t)p_{\ell j}(t) + q_{ii}p_{ij}(t) \\ &= \sum_{\ell=1}^k q_{i\ell}(t)p_{\ell j}(t) \end{aligned}$$

La deuxième équation de Kolmogorov se démontre de la même manière en

partant de :

$$P_{ij}(t + \Delta(t)) = \sum_{\ell=1}^k p_{i\ell}(t)p_{\ell j}(\Delta t)$$

Sachant que $p_{ii}(0) = 1$ et $p_{ij}(0) = 0$, la solution des équations ci-dessus est donnée par :

$$P(t) = \exp(tQ) = \sum_{s=0}^{\infty} \frac{t^s}{s!} Q^s. \quad (2.8)$$

En effet, dérivons l'équation (2.8) par rapport à t on trouve

$$\begin{aligned} \frac{d}{dt}P(t) &= \frac{d}{dt} \left\{ I + tQ + \frac{t^2 Q^2}{2!} + \frac{t^3 Q^3}{3!} + \dots \right\} \\ &= \left\{ 0 + Q + tQ^2 + 3\frac{t^2 Q^3}{3!} + \dots \right\} \\ &= \left\{ Q + tQ^2 + \frac{t^2 Q^3}{2!} + \dots \right\} \\ &= \left\{ I + Q + \frac{t^2 Q^2}{2!} + \dots \right\} Q \\ &= P(t)Q. \end{aligned} \quad (2.9)$$

la solution générale de l'équation (2.9) est sous la forme

$$P(t) = c \exp(tQ),$$

où c est une constante. Or, le processus ne peut pas changer d'état pendant un intervalle de temps nul. Donc il faut prendre en compte les contraintes $p_{ii}(0) = 1$ et $p_{ij}(0) = 0$, pour tout $i \neq j$. Elles sont satisfaites seulement si $c = 1$. $\square$

Remarque 2.3.1. Lorsque la matrice Q est diagonalisable, c'est-à-dire $Q = UDU^{-1}$, où D est la matrice diagonale des valeurs propres de la matrice des intensités de transition et U est la matrice des vecteurs propres correspondants. Alors

$$\exp(tQ) = U \exp(tD) U^{-1}$$

Exemple : (Équations de Chapman-Kolmogorov pour un processus de Markov homogène à deux état)

Le processus $(X_t, t \geq 0)$ associé est un processus dont les marques sont état 0 (vivant) et l'état 1 (décès). La matrice des intensités de transition est donnée par :

$$Q = \begin{pmatrix} -\lambda_{01} & \lambda_{01} \\ 0 & 0 \end{pmatrix}$$

avec $\lambda_{01} > 0$. On peut résoudre les équations de Kolmogorov en diagonalisant Q . Le polynôme caractéristique de Q est donné par $P_Q(\alpha) = \det(Q - \alpha I)$

$$\begin{aligned} \det(Q - \alpha I) &= \det \left[\begin{pmatrix} -\lambda_{01} & \lambda_{01} \\ 0 & 0 \end{pmatrix} - \alpha \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right] \\ &= \det \begin{pmatrix} -\lambda_{01} - \alpha & \lambda_{01} \\ 0 & -\alpha \end{pmatrix} \\ &= \alpha(\alpha + \lambda_{01}). \end{aligned}$$

La matrice Q admet deux valeurs propres distinctes 0 et $-\lambda_{01}$, donc elle est diagonalisable c'est-à-dire $Q = UDU^{-1}$ avec :

$$U = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, U^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}, D = \begin{pmatrix} 0 & 0 \\ 0 & -\lambda_{01} \end{pmatrix}$$

On utilise $P(t) = \exp(Qt) = U \exp(tD)U^{-1}$, on trouve

$$P(t) = \begin{pmatrix} e^{-\lambda_{01}t} & 1 - e^{-\lambda_{01}t} \\ 0 & 1 \end{pmatrix}.$$

En effet, $p_{00}(t) = e^{-\lambda_{01}t}$ et $p_{01}(t) = 1 - e^{-\lambda_{01}t}$ correspondent à la fonction de survie et la fonction de répartition d'une loi exponentielle présentée dans la section 1.2.

Remarque 2.3.2. Dans les modèles markoviens homogènes, les lois du temps de séjour dans un état sont définies de manière implicite et suivent toujours des lois exponentielles de paramètre $\sum_{i \neq j} \lambda_{ij}$. Ces lois sont dites sans mémoire car les fonctions de risque associées sont constantes au cours du temps. On peut aussi noter que le temps moyen passé dans l'état i avant de le quitter vaut $\mathbb{E}(T_i) = \frac{1}{\sum_{i \neq j} \lambda_{ij}}$.

2.4 Vraisemblances d'un modèle multi-état

Dans cette section, nous présentons les vraisemblances pour les données censurées par intervalles et les données censurées à droite.

2.4.1 Données censurées par intervalle

Dans ce cas, l'observation des différents états se fait en temps discret, à des instants successifs. Les temps de transition ne sont pas connus de façon exacte et le chemin pris pour aller d'un état à l'autre entre deux temps d'observations consécutifs est inconnu avec un nombre de chemins possibles pouvant être grand. Les modèles multi-états adaptés à de telles données sont souvent désignés dans la littérature sous le terme de "Markov models for panel data".

Dans ce type de modèles, décrits d'abord par Kalbfleisch et Lawless (1985), l'estimation est basée sur la résolution des équations différentielles de Kolmogorov qui permet d'obtenir les intensités de transition ainsi que les probabilités de transition. Soit $\{X_{th}, t \geq 0\}, h = 1, \dots, n$, des réalisations indépendantes d'un processus markovien homogène, que l'on observe aux temps $t_0, t_1, \dots, t_L (L > 1)$. On note $n_{ij}(\ell)$ le nombre de sujets qui étaient dans l'état i au temps $t_{\ell-1}$ et qui se trouvent dans l'état j au temps t_ℓ , $i, j \in S$. La fonction de vraisemblance conditionnelle à la loi initiale en t_0 s'écrit :

$$\begin{aligned} L &= \prod_{\ell=1}^L \left\{ \prod_{i,j=1}^k [p_{ij}(t_{\ell-1}, t_\ell)]^{n_{ij}(\ell)} \right\} \\ &= \prod_{\ell=1}^L \left\{ \prod_{i,j=1}^k p_{ij}(t_\ell - t_{\ell-1})^{n_{ij}(\ell)} \right\}. \end{aligned} \quad (2.10)$$

Dans l'exemple suivant, nous montrons que dans le cas d'un modèle de survie, la vraisemblance (2.10) est équivalente à la vraisemblance (1.2).

Exemple : Supposons qu'on observe aux temps $t_0, t_1, \dots, t_L$ la survenue de l'évènement d'intérêt pour un sujet $h, h = 1, \dots, n$. Supposons que le sujet h ait transité de l'état 0 vers l'état 1 entre t_2 et t_3 . Donc, sa contribution à la vraisemblance est donnée par :

$$L_h = p_{00}(t_0, t_1)p_{00}(t_1, t_2)p_{01}(t_2, t_3)p_{11}(t_3, t_4) \dots p_{11}(t_{L-1}, t_L), \quad (2.11)$$

or, $p_{11}(t_{\ell-1}, t_\ell) = 1, \ell \geq 4$ puisque l'état 1 est un état absorbant, donc l'équation (2.11) devient :

$$\begin{aligned} L_h &= \exp(-\lambda_{01}(t_1 - t_0)) \exp(-\lambda_{01}(t_2 - t_1))(1 - \exp(-\lambda_{01}(t_3 - t_2)))(1) \dots (1) \\ &= \exp(-\lambda_{01}(t_2 - t_0))(1 - \exp(-\lambda_{01}(t_3 - t_2))) \\ &= \exp(-\lambda_{01}(t_2 - t_0)) - \exp(-\lambda_{01}(t_3 - t_0)) \end{aligned} \quad (2.12)$$

$S(t) = \exp(-\lambda_{01}t)$, l'équation (2.12) s'écrit donc sous la forme

$$L_h = S(t_2 - t_0) - S(t_3 - t_0).$$

La vraisemblance totale pour n sujets est donnée par

$$L = \prod_{h=1}^n (S(\ell_h) - S(r_h)) .$$

où le sujet h a transité de l'état 0 vers l'état 1 dans l'intervalle $[\ell_h; r_h]$.

Procédure d'estimation de Kalbfleish et Lawess

Dans cette section, nous présentons l'algorithme proposé par Kalbfleish et Lawless (1985) pour maximiser la log-vraisemblance. En effet, puisque les probabilités de transition sont des fonctions compliquées à écrire, ces auteurs proposent de maximiser en Q . La vraisemblance (2.10), réécrite sous la forme

$$\mathcal{L}(Q) = \prod_{\ell=1}^L \left\{ \prod_{i,j=1}^k [\exp(Q(t_\ell - t_{\ell-1}))]_{i,j}^{n_{ij}(\ell)} \right\}, \quad (2.13)$$

par un algorithme du score appelé aussi algorithme de quasi-Newton pour estimer les intensités de transition. Dans l'équation (2.13), l'expression $(\exp(Q(t_\ell - t_{\ell-1})))_{i,j}$ représente l'élément (i, j) de la matrice $(\exp(Q(t_\ell - t_{\ell-1}))) = P(t_\ell - t_{\ell-1})$. Sous l'hypothèse que Q admet k valeurs propres distinctes $(\alpha_1, \dots, \alpha_k)$, les matrices Q et $P(t)$ se diagonalisent de la manière suivante :

$$Q = UDU^{-1} = U \text{diag}(\alpha_1, \dots, \alpha_k)U^{-1}, \quad (2.14)$$

et

$$\begin{aligned} P(t) &= \exp(Qt) = U \exp(Dt) U^{-1} \\ &= U \text{diag}(e^{\alpha_1 t}, \dots, e^{\alpha_k t}) U^{-1}. \end{aligned}$$

Cette décomposition permet de déduire une expression analytique du vecteur des scores pouvant être utilisée dans la procédure itérative. En effet, le vecteur des scores est défini par :

$$S(Q) = \left\{ \frac{\partial \log \mathcal{L}(Q)}{\partial \lambda_{uv}} \right\} = \left\{ \sum_{\ell=1}^L \sum_{i,j=1}^k n_{ij}(\ell) \frac{\partial \exp(Q(t_\ell - t_{\ell-1}))_{i,j} / \partial \lambda_{uv}}{\exp(Q(t_\ell - t_{\ell-1}))_{i,j}} \right\} \quad (2.15)$$

où

$$\frac{\partial P(t)}{\partial \lambda_{uv}} = \frac{\partial(\exp(Qt))}{\partial \lambda_{uv}} \quad (2.16)$$

$$\begin{aligned} &= \sum_{s=1}^{\infty} \frac{\partial}{\partial \lambda_{uv}} \left(\frac{Q^s t^s}{s!} \right) \\ &= \sum_{s=1}^{\infty} \left(\frac{\partial Q^s}{\partial \lambda_{uv}} \right) \frac{t^s}{s!} \\ &= \sum_{s=1}^{\infty} \sum_{r=0}^{s-1} Q^r \left(\frac{\partial Q}{\partial \lambda_{uv}} \right) Q^{s-1-r} \frac{t^s}{s!} \\ &= \sum_{s=1}^{\infty} \sum_{r=0}^{s-1} U D^r U^{-1} \left(\frac{\partial Q}{\partial \lambda_{uv}} \right) U D^{s-1-r} U^{-1} \frac{t^s}{s!} \end{aligned} \quad (2.17)$$

Posons $G_{uv} = U^{-1} \left(\frac{\partial Q}{\partial \lambda_{uv}} \right) U$, l'équation (2.16) devient

$$\begin{aligned} \frac{\partial P(t)}{\lambda_{uv}} &= \frac{\partial(\exp(Qt))}{\partial \lambda_{uv}} = \sum_{s=1}^{\infty} \sum_{r=0}^{s-1} U D^r G_{uv} D^{s-1-r} U^{-1} \frac{t^s}{s!} \\ &= U \left(\sum_{s=1}^{\infty} \sum_{r=0}^{s-1} D^r G_{uv} D^{s-1-r} \frac{t^s}{s!} \right) U^{-1} \\ &= U V_{uv} U^{-1}. \end{aligned}$$

La matrice

$$V_{uv} = \sum_{s=1}^{\infty} \sum_{r=0}^{s-1} D^r G_{uv} D^{s-1-r} \frac{t^s}{s!}$$

est une matrice $k \times k$ ayant pour éléments

$$\begin{aligned} v_{ij} &= (g_{uv})_{i,j} \frac{e^{\alpha_i t} - e^{\alpha_j t}}{\alpha_i - \alpha_j}, \quad i \neq j \\ v_{ii} &= (g_{uv})_{i,i} t e^{\alpha_i t} \end{aligned}$$

et $(g_{uv})_{i,j}$ est l'élément (i, j) de la matrice $G_{uv} = U^{-1} \left(\frac{\partial Q}{\partial \lambda_{kh}} \right) U$. Des dérivées secondes de la log-vraisemblance, on déduit la forme de matrice d'information

$$\mathbb{E} \left[- \frac{\partial^2 \log \mathcal{L}(Q)}{\partial \lambda_{uv} \partial \lambda_{u'v'}} \right] = \left\{ \sum_{\ell=1}^L \sum_{i,j=1}^k \frac{\mathbb{E}[N_i(t_{\ell-1})]}{p_{ij}(t)} \frac{\partial p_{ij}(t)}{\partial \lambda_{uv}} \frac{\partial p_{ij}(t)}{\partial \lambda_{u'v'}} \right\}$$

qui est estimée par

$$M(Q) = \left\{ \sum_{\ell=1}^L \sum_{i,j=1}^k \frac{N_i(t_{\ell-1})}{p_{ij}(t)} \frac{\partial p_{ij}(t)}{\partial \lambda_{uv}} \frac{\partial p_{ij}(t)}{\partial \lambda_{u'v'}} \right\}$$

où $N_i(\ell - l)$ représente le nombre de sujets dans l'état au temps $t_{\ell-1}$. La formule de récurrence de l'algorithme du score est donnée par

$$Q_{n+1} = Q_n + M(Q_n)^{-1} S(Q_n), \quad n \geq 0$$

avec $S(Q_n) = \left(\frac{\partial \log \mathcal{L}(Q)}{\partial \lambda_{11}}, \dots, \frac{\partial \log \mathcal{L}(Q)}{\partial \lambda_{kk}} \right)'$ et $M(Q_n)$ est une matrice de $k \times k$ d'éléments $\frac{\partial^2 \log \mathcal{L}(Q)}{\partial \lambda_{uv} \partial \lambda_{u'v'}}$.

Cette approche nous permet d'utiliser la matrice des intensités de transition, ses valeurs propres et vecteurs propres pour trouver la dérivée de la matrice des probabilités de transition afin d'éviter le calcul direct. (On va appliquer cette approche dans le chapitre suivant (exemple d'application 2).

2.4.2 Données exactes et censure à droite

Considérons un processus de Markov $(X_t, t \geq 0)$ observé en temps continu sur une période $[0, C]$ où le processus est dans l'état s_0 à l'instant $t_0 = 0$ et

notons t_m les temps exacts de transition dans les états $s_m, m = 1, \dots, E$. La contribution à la vraisemblance pour cette trajectoire, conditionnellement à X_{t_0} , est

$$\left(\prod_{m=1}^E \wp_{s_{m-1}}(t_{m-1}, t_m) \lambda_{s_{m-1}, s_m}(t_m) \right) \wp_{S_E}(t_E, C). \quad (2.18)$$

où $\wp_{s_{m-1}}(t_{m-1}, t_m)$ est la probabilité de ne pas transiter entre t_{m-1} et t_m , sachant que $X_{t_{m-1}} = s_{m-1}$. Cette probabilité est égale à

$$\wp_{s_{m-1}}(t_{m-1}, t_m) = \exp \left\{ - \int_{t_{m-1}}^{t_m} \lambda_{s_{m-1}}(u) du \right\}.$$

Si bien que la contribution à la vraisemblance (2.18) s'écrit

$$\left(\prod_{m=1}^E \exp \left\{ - \int_{t_{m-1}}^{t_m} \lambda_{s_{m-1}}(u) du \right\} \lambda_{s_{m-1}, s_m}(t_m) \right) \exp \left\{ - \int_{t_E}^C \lambda_{S_E}(u) du \right\}. \quad (2.19)$$

En regroupant les exponentielles, (2.19) s'écrit comme un produit de deux termes :

$$\left(\prod_{m=1}^E \lambda_{s_{m-1}, s_m}(t_m) \right) \exp \left\{ - \left(\int_0^{t_1} \lambda_{s_0}(u) du + \dots + \int_{t_E}^C \lambda_{S_E}(u) du \right) \right\}. \quad (2.20)$$

Comme Hougaard (1999), notons $K = E+1$, $t_K = C$, $s_K = -1$, et définissons deux indicatrices :

$$D_{ijm} = \mathbb{1}_{\{(s_{m-1}, s_m)\}}(i, j)$$

(valant 1 si la m^e transition correspond à la transition de l'état s_{m-1} vers l'état s_m) et

$$R_i(t) = \sum_{m=1}^K \mathbb{1}_{\{s_m = -1\} \times [t_{m-1}, t_m]}(i, t)$$

(valant 1 quand le processus est dans l'état s_{m-1} dans l'intervalle $[t_{m-1}, t_m[)$. Alors (2.20) se réécrit :

$$\begin{aligned} & \left(\prod_{m=1}^K \prod_{i \neq j} \lambda_{i,j}(t_m)^{D_{ijm}} \right) \exp \left\{ - \left(\sum_i \int_0^C R_i(u) \lambda_i(u) du \right) \right\} \\ &= \left(\prod_{i \neq j} \left\{ \prod_{m=1}^K \lambda_{i,j}(t_m)^{D_{ijm}} \right\} \right) \left(\prod_i \exp \left\{ - \int_0^C R_i(u) \lambda_i(u) du \right\} \right) \end{aligned}$$

Comme $\lambda_i(u) = \sum_{i \neq j} \lambda_{ij}(u)$, on peut remplacer dans le membre de droite et (2.19) devient simplement

$$\prod_{i \neq j} \left(\left\{ \prod_{m=1}^K \lambda_{ij}(t_m)^{D_{ijm}} \right\} \exp \left\{ - \int_0^C R_i(u) \lambda_{ij}(u) du \right\} \right).$$

Dans la partie précédente, nous avons présenté la vraisemblance pour un sujet observé sur une période $[0, C]$ et t_m représentent les temps exacts de transition dans les états $s_m, m = 1, \dots, E$. Supposons maintenant qu'on observe n sujets, pour chaque sujet $h = 1, \dots, n$ observé sur une période $[0, C_h]$, on note t_{m_h} les temps exacts de transition dans les états $s_{m_h}, m_h = 1, \dots, E_h$. Dans ce cas, nous avons n réalisations indépendantes $\{X_h(t), 0 \leq t \leq C_h\}, h = 1, \dots, n$, d'un processus markovien à temps continu, alors la vraisemblance totale est de la forme

$$\mathcal{L} = \prod_{h=1}^n \prod_{i \neq j} \left(\left\{ \prod_{m_h=1}^{K_{m_h}} \lambda_{ij}(t_{m_h})^{D_{hijm_h}} \right\} \exp \left\{ - \int_0^{C_h} R_{hi}(u) \lambda_{ij}(u) du \right\} \right). \quad (2.21)$$

Dans la cas d'un modèle de survie avec $i = 0$ et $j = 1$. soit $t_h, h = 1, \dots, n$, les temps exacts de transition de l'état 0 vers l'état 1 des n sujets, la vraisemblance (2.21) s'écrit

$$\mathcal{L} = \prod_{h=1}^n \left(\left\{ \lambda_{01}(t_h)^{D_{h01}} \right\} \exp \left\{ - \int_0^{C_h} R_{h0}(u) \lambda_{01}(u) du \right\} \right) \quad (2.22)$$

où $\lambda_{01}(t_h)^{D_{h01}}$ est la contribution à la vraisemblance pour les temps exacts de transition de l'état 0 vers l'état 1 et $\int_0^{C_h} R_{h0}(u)\lambda_{01}(u)du$ est la contribution à la vraisemblance pour les temps censurés à droite. En effet, nous remarquons que la vraisemblance (2.22) est équivalente à (1.1)

Dans le cas d'un modèle multi-états markovien homogène, l'estimateur de l'intensité de transition de l'état i vers l'état j est donné par :

$$\hat{\lambda}_{ij} = \frac{\sum_{h,m_h} D_{hijm_h}}{\sum_h \int_0^{C_h} R_{hi}(t)dt} = \frac{\sum_{h,m_h} \mathbb{1}_{\{(s_{m_h-1}, s_{m_h})\}}}{\sum_{h,m_h} \mathbb{1}_{\{s_{m_h-1}\}}(t_{m_h} - t_{m_h-1})}.$$

Nous détaillons maintenant le calcul de cet estimateur pour un exemple particulier.

Exemple : Supposons que nous avons n sujets qui étaient tous dans l'état 0 à l'instant $t_0 = 0$. Tous les sujets peuvent transiter vers deux états absorbants 1 ou 2. Donc, nous observons n réalisations indépendantes $(X_h(t), 0 \leq t \leq C_h, h = 1, \dots, n)$ d'un processus markovien à temps continu. Notons t_h les temps exacts de transition d'un sujet depuis l'état 0 vers un des deux états 1 ou 2. Dans ce cas, la vraisemblance totale (2.21) est de la forme :

$$\mathcal{L} = \prod_{h=1}^n \prod_{j=1,2} \left[\{\lambda_{0j}(t_h)^{D_{h0j}}\} \exp \left\{ - \int_0^{C_h} R_{h0}(u)\lambda_{0j}(u)du \right\} \right]. \quad (2.23)$$

En développant l'équation (2.23), on obtient

$$\begin{aligned} \mathcal{L} &= \prod_{h=1}^n \left[\{\lambda_{01}(t_h)^{D_{h01}} \lambda_{02}(t_h)^{D_{h02}}\} \times \exp \left\{ - \int_0^{C_h} R_{h0}(u)\lambda_{01}(u)du \right\} \right. \\ &\quad \left. \exp \left\{ - \int_0^{C_h} R_{h0}(u)\lambda_{02}(u)du \right\} \right] \\ \log(\mathcal{L}) &= \sum_{h=1}^n (D_{h01} \log(\lambda_{01}(t_h)) + D_{h02} \log(\lambda_{02}(t_h))) \\ &\quad - \int_0^{C_h} R_{h0}(u)\lambda_{01}(u)du - \int_0^{C_h} R_{h0}(u)\lambda_{02}(u)du. \end{aligned} \quad (2.24)$$

Lorsque les intensités du modèle sont constantes l'équation (2.24) devient

$$\begin{aligned} \log(\mathcal{L}) = \sum_{h=1}^n & \left(D_{h01} \log(\lambda_{01}) + D_{h02} \log(\lambda_{02}) - \lambda_{01} \int_0^{C_h} R_{h0}(u) du \right. \\ & \left. - \lambda_{02} \int_0^{C_h} R_{h0}(u) du \right). \end{aligned} \quad (2.25)$$

En dérivant l'équation (2.25) par rapport $\lambda_{0j}, j = 1, 2$, nous obtenons :

$$\frac{\partial \log(\mathcal{L})}{\partial \lambda_{0j}} = \sum_{h=1}^n \left(\frac{D_{h0j}}{\lambda_{0j}} - \int_0^{C_h} R_{h0}(u) du \right).$$

Ainsi, l'estimateur de l'intensité de transition est donné par

$$\hat{\lambda}_{0j} = \frac{\sum_{h=1}^n D_{h0j}}{\sum_{h=1}^n \int_0^{C_h} R_{h0}(t) dt}, \quad j = 1, 2$$

où D_{h0j} est égal à 1 si le sujet h transite vers l'état j et $R_{h0}(t)$ est le nombre de sujets encore dans l'état 0 au temps t .

Chapitre 3

Application

Nous présentons dans ce chapitre des exemples d'application simple pour illustrer notre travail. Cette application repose sur des données pour un test clinique (la leucémie myélogène aigue(AML)), des données de vasculopathie de l'allogreffe cardiaque (CAV). Nous devons modéliser les fonctions de survie et de risque instantané de changement d'état pendant le temps à l'évènement.

3.1 Application aux données AML

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

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

Tableau 3.1 : Les données de l'étude d'entretien AML.

Les nombres avec + correspondent à des patients perdus de vue à la date considérée (+ indique une valeur censuré). Dans l'analyse de survie on tient compte de toutes les observations censurées ou non.

En effet dans les problèmes d'estimations statistiques si on élimine les observations censurées on perd de l'information puisque on ne tient pas compte des patients ayant des durées de rémission plus longues.

3.1.1 Méthode paramétrique

Nous commençons par des méthodes paramétriques concernant la fonction de survie $S(t)$ et la fonction de risque cumulatif $\Lambda(t)$.

Nous considérons ici les données d'AML.

Sous le logiciel R, pour exécuter une analyse de survie, il est nécessaire d'installer la bibliothèque d'analyse de survie et des données (**survival**).

Ajustement exponentiel

Les commandes de R sont :

```
>library(survival)
>Data(aml)
>weib.ajst <- survreg(Surv(aml$time, aml$status)~1,
+, dist="weib", scale=1)
>weib.ajst
Call:
survreg(formula=Surv(aml$time, aml$status)~1,
+ dist = "weib", scale = 1)
Coefficients:
(Intercept)
3.628776
Scale fixed at 1
Loglik(model)= -83.3    Loglik(intercept only)= -83.3
n= 23
```

l'intercept = 3.628776, ce qui égale $\hat{\mu} = -\log(\hat{\lambda}) = \log(\hat{\theta})$.

Les commandes suivantes produisent un intervalle de confiance à 95% pour la moyenne θ et l'estimation de $S(t)$

```
> coeff <- weib.ajst$coeff # muhat
> var <- weib.ajst$var
> thetahat <- exp(coeff) # exp(muhat)
```

```
> thetahat
> C.I.mean1 <- c(thetahat,exp(coeff-1.96*sqrt(var)),
+ exp(coeff+1.96*sqrt(var)))
> names(C.I.mean1) <- c("mean1","LCL","UCL")
> C.I.mean1
  mean1      LCL      UCL
37.66667 23.73140 59.78483
> muhat <- weib.ajst$coeff
> time.u <- time[status == 1]
> nu <- length(time.u)
> scalehat <- rep(exp(muhat),nu)
> Shat <- 1 - pweibull(time.u,1,scalehat)
> LCL <- exp(log(Shat)*exp(1.96/sqrt(nu)))
> UCL <- exp(log(Shat)*exp(-1.96/sqrt(nu)))
> C.I.Shat <- data.frame(time.u,Shat,LCL,UCL)
> round(C.I.Shat,5)
```

	time.u	Shat	LCL	UCL
1	9	0.78746	0.68438	0.86024
2	13	0.70813	0.57822	0.80457
3	18	0.62010	0.46837	0.74002
4	23	0.54301	0.37939	0.68065
5	31	0.43911	0.27082	0.59540
6	34	0.40549	0.23866	0.56626
7	48	0.27962	0.13231	0.44804
8	5	0.87569	0.81002	0.91977
9	5	0.87569	0.81002	0.91977
10	8	0.80865	0.71383	0.87475
11	8	0.80865	0.71383	0.87475
12	12	0.72718	0.60311	0.81814
13	23	0.54301	0.37939	0.68065
14	27	0.48831	0.32054	0.63660
15	30	0.45092	0.28248	0.60544
16	33	0.41640	0.24894	0.57581
17	43	0.31931	0.16334	0.48712
18	45	0.30280	0.15013	0.47109

Tableau 3.1.1 : Résultats de l'estimation de la fonction de survie.

3.1.2 Méthode non paramétrique

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

$$\hat{S}(t) = \prod_{T_i \leq t} p_i = \prod_{T_i \leq t} \left(\frac{n_i - d_i}{n_i} \right) = \prod_{T_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

Or p_i est la probabilité de survivre au-delà de l'intervalle de temps $I_i = [X_{(i-1)}, X_{(i)}[$ sachant qu'était vivant au début de l'intervalle. Notons :

- n_i le nombre des sujets qui sont vivants (donc à risque de mourir) juste avant l'instant T_i
- d_i le nombre des morts à l'instant $T_{(i)}$.

- On pose $q_i = 1 - p_i$. q_i est la probabilité de mourir durant l'intervalle I_i sachant que l'individu était vivant au début de cet intervalle.

Alors l'estimateur naturel de q_i est : $\hat{q}_i = \frac{d_i}{n_i}$

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

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

L'estimations de la fonction de risque cumulatif construit avec K-M :

$$\hat{\Lambda}(t) = -\log \prod_{t(i) \leq t} \left(\frac{n_i - d_i}{n_i} \right)$$

les résultats obtenus d'après l'étude des données d'AML sont présentés comme suit :

```
> library(survival)
> attach(aml)
> Surv(aml$time, aml$status) # Objet de surv
> km.fit <- survfit(Surv(time, status)~1,
+type="kaplan-meier", data=aml)
> km.fit
> summary(km.fit) # survival est l'estimation de S(t)
> H.hat <- -log(km.fit$surv)
> op <- par(mfrow=c(2,1))
> plot(km.fit, xlab=temps_jusqu'à_la_rechute(en_semaines),
+_ylab=proportion_sans_rechute, col='red')
> mtext(Courbe_de_survie_de_K-M, 3, line=-1, cex=0)
> plot(km.fit, lty=2:3, fun="cumhaz",
+_xlab="Temps", ylab=Hazard_cumulative)
> par(op)
> abline(h=0)
```

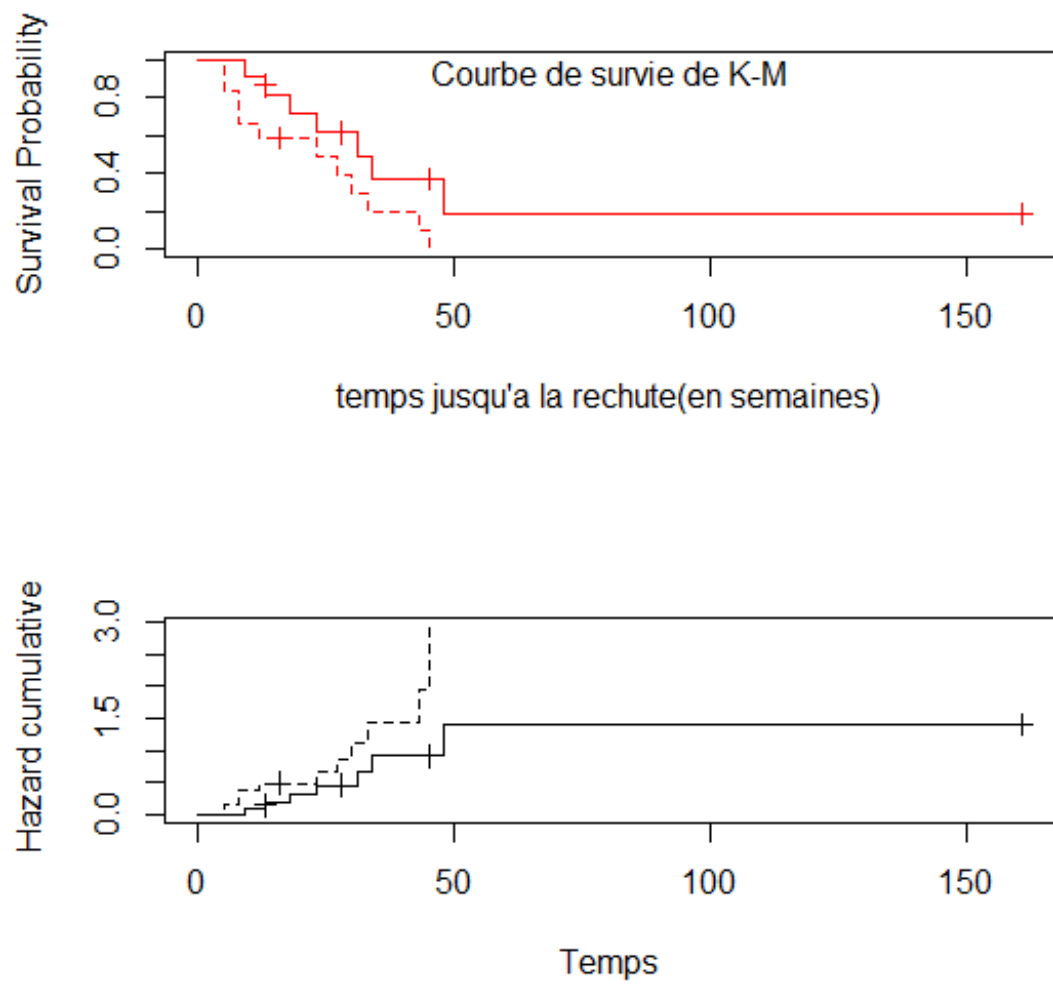


FIGURE 3.1 – La courbe des estimateurs de K-M

Discussion :

La figure 3.1 représente l'estimateur de Kaplan-Meier de la fonction de survie et la fonction de hasard cumulatif. La ligne continue correspond au groupe qui a reçu le traitement (groupe A), et la ligne intermittente correspond au groupe qui n'ont pas reçu le traitement (groupe B)

La courbe KM est une fonction en escalier droit continue, chaque fois qu'un patient est suivi moins de 161 semaines, il est dit censuré et représenté par un trait vertical sur la courbe. Plus qu'il y a de patients censurés plus la marche d'escalier à chaque décès est haute. À 48 semaines la proba de survie est de 40% pour le groupe A et de 20% pour le groupe B. Alors cette différence est significative. Les personnes à risque au début de l'intervalle t_i sont les personnes qui ont survécu (pas morte, perdu ou retiré) de l'intervalle précédent t_{i-1} . Les commandes suivantes fournissent les valeurs des estimateurs de KM :

```
>kpm <-survfit (Surv(time,status)~1,type="kaplan-meier",
+data=aml)
> kpm
Call: survfit(formula=Surv(time, status)~1,data=aml,
+type="kaplan-meier")
> summary(kpm )#survival est l'estimation de S(t)
Call: survfit(formula = Surv(time, status)~1,
+data=aml,type="kaplan-meier")
```


time	n.risk	n.event	survival	std.err	lower 95%CI	upper 95%CI
5	23	2	0.9130	0.0588	0.8049	1.000
8	21	2	0.8261	0.0790	0.6848	0.996
9	19	1	0.7826	0.0860	0.6310	0.971
12	18	1	0.7391	0.0916	0.5798	0.942
13	17	1	0.6957	0.0959	0.5309	0.912
18	14	1	0.6460	0.1011	0.4753	0.878
23	13	2	0.5466	0.1073	0.3721	0.803
27	11	1	0.4969	0.1084	0.3240	0.762
30	9	1	0.4417	0.1095	0.2717	0.718
31	8	1	0.3865	0.1089	0.2225	0.671
33	7	1	0.3313	0.1064	0.1765	0.622
34	6	1	0.2761	0.1020	0.1338	0.569
43	5	1	0.2208	0.0954	0.0947	0.515
45	4	1	0.1656	0.0860	0.0598	0.458
48	2	1	0.0828	0.0727	0.0148	0.462

Tableau 3.1.2 : Résultats des estimateurs de K-M.

3.2 Application aux données CAV

L'ensemble de données CAV provient d'une base de données de receveurs de greffe du coeur. Sharples et al. (2003) ont étudié la progression des vasculopathies d'allogreffe coronaire (CAV). Ces vasculopathies représentent l'une des causes de décès les plus fréquentes chez les survivants après une transplantation cardiaque. Cet ensemble de données est disponible dans le package `msm` (Jackson, 2011) du logiciel R. .

L'objectif de cette section est d'estimer les intensités de transition entre les différents états pour un modèle de markov homogène lorsque les données sont censurées par intervalle, en utilisant le logiciel R..

```
> library(msm)
> data(cav)
> cav
```

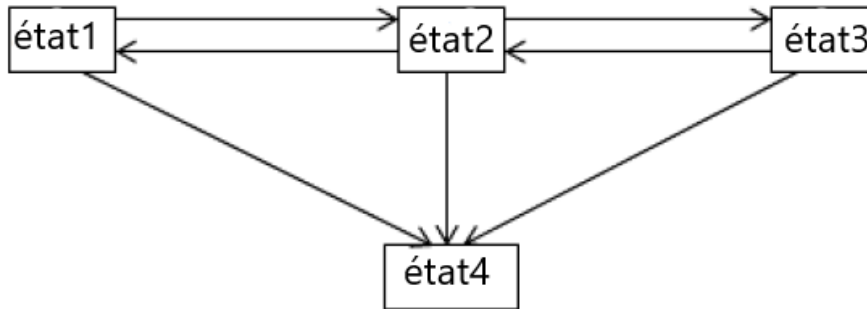


FIGURE 3.2 – Modèle multi-état pour les sujets de cav.

	PTNUM	age	years	dage	sex	pdiag	cumrej	state
1	100002	52.4958	0	21	0	IHD	0	1
2	100002	53.4986	1.00273	21	0	IHD	2	1
3	100002	54.4986	2.00273	21	0	IHD	2	2
4	100002	55.5890	3.09315	21	0	IHD	2	2
5	100002	56.4958	4.0000	21	0	IHD	3	2
6	100002	57.4931	4.99726	21	0	IHD	3	3
7	100002	58.3506	5.85479	21	0	IHD	3	4
8	100003	29.5068	0.00000	17	0	IHD	0	1
9	100003	30.6958	1.18904	17	0	IHD	1	1
10	100003	31.5150	2.00821	17	0	IHD	1	3

Tableau 3.2 : Un extrait des données cav

La base de données est constituée de 614 individus, ce qui représente 2816 observations. Approximativement, tous les ans après une greffe cardiaque, chaque patient subit une angiographie, pour diagnostiquer la CAV. Le résultat de l'essai est dans l'ensemble $\{1, 2, 3, 4\}$ représentant respectivement cav-Absente, cav-Moyenne, cav-Grave et mort donne la période de l'essai(test) depuis la greffe du coeur.

D'autres variables incluent l'âge (âge à l'écran), le dage (âge de distributeur), le sexe (0=male,1=femelle), le pdiag (diagnostic primaire, ou la raison de la

greffe-IHD représente la maladie cardiaque ischémique, l'IDC représente la cardiomyopathie dilatée idiopathique), le cumrej (nombre cumulatif des épisodes de rejet), et firstobs, un indicateur qui égale 1 quand l'observation correspond à la greffe du patient (première observation), et 0 quand l'observation correspond à une angiographie postérieure.

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

Le modèle est défini par la figure(3.2). La structure étudiée sera ainsi définie par quatre états transitives.

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

$$q_{rs} = \lim_{\Delta t \rightarrow 0^+} \mathbb{P}[X_{t+\Delta t} = s | X_t = r] / \Delta t.$$

L'objectif est de d'estimer la matrice Q des intensités de transition donnée par :

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

Les résultats de l'estimation du maximum de vraisemblance de la matrice des intensités de transition et les intervalles de confiance à 95% sont donnés comme suit :

```
cav.msm<-msm(state~years, subject=PTNUM, data=cav,
+qmatrix=Q, death=4)
cav.msm
```

Intensité	Estimés	IC à 95%
état 1 - état 1	-0.17037	(-0.19027,-0.15255)
état 1 - état 2	0.12787	(0.11135, 0.14684)
état 1 - état 4	0.04250	(0.03412, 0.05294)
état 2 - état 1	0.22512	(0.16755, 0.30247)
état 2 - état 2	-0.60794	(-0.70880,-0.52143)
état 2 - état 3	0.34261	(0.27317, 0.42970)
état 2 - état 4	0.04021	(0.01129, 0.14324)
état 3 - état 2	0.13062	(0.07952, 0.21457)
état 3 - état 3	-0.43710	(-0.55292,-0.34554)
état 3 - état 4	0.30648	(0.23822, 0.39429)

Tableau 3.3 : Estimation des intensités de transition.

Discussion :

À partir de la matrice des intensités de transition estimée, nous remarquons que la vitesse de transition de l'état cav-absent (état 1) vers cav-moyenne (état 2) est trois fois supérieure à celle vers le décès (état 4). Cela signifie que les sujets qui transitent de l'état cav-absent (état 1) vers cav-moyenne (état 2) ont trois fois plus de chance de développer des symptômes que de mourir. Après le début de la maladie (état 2), la progression vers les symptômes graves (état 3) est plus rapide que le rétablissement (état 1) ou le décès (état 4).

D'après la commande suivante :

```
>cav.p<-pmatrix.msm(cav.msm,t=10,ci="normal")
```

Après 10 ans la matrice d'estimation des probabilités de transition est :

	état 1	IC	état 2	IC
état 1	0.3094	(0.26766,0.3451)	0.0975	(0.0775,0.1129)
état 2	0.1716	(0.12855,0.2147)	0.0655	(0.0456,0.0830)
état 3	0.0589	(0.03520,0.0906)	0.0297	(0.0166,0.0459)
état 4	0		0	
	état 3	IC	état 4	IC
état 1	0.0878	(0.0666,0.1057)	0.5052	(0.4681,0.5654)
état 2	0.0779	(0.0511,0.1009)	0.6848	(0.6301,0.7611)
état 3	0.0466	(0.0257,0.0726)	0.8646	(0.8059,0.9163)
état 4	0		1.0000	(1.0000,1.0000)

Tableau 3.4 : Probabilités de transition estimées après $t = 10$.

Le tableau 3.4 est montré que les probabilités de survie après dix ans avec cav-grave (état 3) et cav-moyenne (état 2) sont faibles par rapport à celle de cav-absent (état 1) parce que les sujets qui se trouvent encore avec cav-absent (état 1) n'ont pas encore développé les symptômes.

Conclusion

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 dans le cas markovien. Ainsi, il devient possible d'étudier les différents risques et de les comparer dans une dynamique globale.

Dans ce mémoire, nous avons étudié les méthodes d'estimations paramétriques et non paramétriques de la fonction de risque relatif au modèle de Markov homogène. Ce modèle est le moins complexe, il suppose que les intensités de transition entre les états sont constantes dans le temps. En épidémiologie, cette contrainte est souvent trop forte. Les modèles markoviens non homogènes permettent de se soustraire à cette hypothèse en modélisant une matrice de probabilités dépendante du temps.

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, par exemple l'évolution de l'état d'un patient à travers les différents stades d'une maladie.

Bibliographie

- [1] Aalen, O. O. (1975). Statistical inference for a family of counting processes. (Thèse de doctorat). University of California, Berkeley.
- [2] Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4), 701-726.
- [3] Andersen, P. K. et Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2), 91-115.
- [4] Antoniadis Anestis, Sapatinas Theofanis. (2007) Estimation and inference in functional mixed-effects models tom 51. 4793-4813. Elsevier B.V. Computational Statistics
- [5] Avram Florin. (2010). Les Probabilités du Bonheur, et les Processus de Markov et de Levy dans l'Actuariat, Files d'attente et Fiabilité. Florida State University. Mémoire de Master.
- [6] Beyersmann, J., Allignol, A. et Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. New York : Springer. 5(4), 315-327.
- [7] Blum J., Susarla V. (1980) Maximal Deviation Theory of Density and Failure Rate Function Estimates Based on Censored Data, *Multivariate Analysis V*, Ed. by P. Krishnaiah, 213-222, North-Holland, Amsterdam.
- [8] Boudreau, C. and Lawless, J. F. (2010) Survival Analysis Based on the Proportional Hazards Model and Survey Data. 506- 516. *Canadian Journal of Statistics*.
- [9] Bouzebda Salim et Elhattab Issam (2009) A Strong Consistency of a Nonparametric Estimate of Entropy under Random Censorship.
- [10] Breslow, N. and J. Crowley. (1974) A large sample study of the life table and product limit estimates under random censorship. 2, 437-453. *Ann. Statist.*

- [11] Cases.C, Lollivier.S. (1993)L'économétrie des modèles de durée avec SAS.Document de travail CREST,9344BIS.43-84.Actes des journées de méthodologie statistique.
- [12] Commenges, D. (1999). « Multi-state models in epidemiology ». Lifetime Data Analysis,
- [13] Embury, S.H., Elias, L., Heller, P.H., Hood, C.E. Greenberg, P.L.,et Schrier, S.L. (1977). Remission maintenance therapy in acute myelogenous leukemia. 126,267-272.Western Journal of Medicine.
- [14] Gray, B. (2014). cmprsk (version 2.2.7). [Package R]. Vienne : The R Foundation for Statistical Computing.
- [15] Gurler Ulkue Wang Jane-Ling . ((1993))Nonparametric estimation of hazard functions and their derivatives under truncation model. Vol.45, No. 2, 249-264.Ann.Inst. Statist. Math.
- [16] Horová Ivana , Pospisil Zdenek et Zelinka Jiri . (2007)Semiparametric Estimation of Hazard Function for Cancer Patients.Volume 69, Part 3, pp. 494-513.The Indian Journal of Statistics.
- [17] Hougaard, P. (1999). Multi-state models : a review. Lifetime Data Analysis, 5(3), 239- 264.
- [18] Kalbfleisch, J., et Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. Journal of the American Statistical Association, 80(392) , 863-871.
- [19] Kalbfleisch, J. D. et Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data. Chichester : John Wiley & Sons.
- [20] Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observation. Journal of the American Statistical Association, 53(282), 457-481.
- [21] Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. Biometrics, 42(4), 855-865.
- [22] Lawless, J. F. (1984). Statistical Models and Methods for Lifetime Data. Chichester : John Wiley & Sons.
- [23] Nelson, W.B. (1972). A short life test for comparing a sample with previous accelerated test results.14, 175-185. Technometrics

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