UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES MULTI-ÉTATS MARKOVIENS EN ANALYSE DE SURVIE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

LEILA BOURMOUCHE

JUILLET 2016

REMERCIEMENTS

Tout d'abord, je remercie chaleureusement ma directrice de recherche, Madame Juli Atherton, pour ses précieux conseils, ses encouragements constants et ses observations critiques, dont l'attention et les remarques ont été toujours fort stimulantes et sans lesquels ce travail n'aurait pu aboutir. Je tiens à souligner ses magnifiques qualités humaines, sa gentillesse et son humanisme. Je ne pourrai jamais assez la remercier pour tout ce qu'elle m'a apporté. Également, il me fait vraiment plaisir de remercier bien sincèrement mon codirecteur de recherche Monsieur René Ferland pour sa vaste expertise en mathématiques et sa rigueur scientifique. De plus, je le remercie pour sa lecture critique et attentionnée de mon mémoire. J'apprécie énormémemt le temps qu'il a passé à corriger aussi bien les mathématiques que la présentation générale. Cela a grandement contribué à une meilleure version de mon mémoire.

Ces années passées au Département de mathématiques de l'UQÀM ont été pour moi particulièrement enrichissantes, c'est la raison pour laquelle je tiens à exprimer mes reconnaissances à tou(te)s mes professeur(e)s.

Je suis également très reconnaissante envers Madame Gisèle Legault, pour son soutien technique en LaTeX, sa disponibilité et sa gentillesse.

Je tiens à remercier Monsieur Michel Adès, toujours disponible et accueillant, ses conseils, ses échanges et encouragements furent très utiles.

J'adresse mes sincères remerciements à Mélissa, Chantal, Stéphanie, Jill, pour m'avoir soutenue au cours de ces années de maîtrise.

Je tiens à remercier mes collègues pour les bons moments passés ensemble au labo-

ratoire. J'adresse mes sincères remerciements à Fabio, tu es vraiment une bonne personne, à Nadia pour sa sympathie et son sens de l'humour, à Fils, merci pour les discussions partagées, ta gentillesse. C'est un réel plaisir pour moi de vous avoir connus, vous avez fait de ces années de maîtrise des moments inoubliables.

Je voudrais exprimer ma profonde reconnaissance à ma mère, mes soeurs et mes frères qui m'ont toujours encouragée malgré la distance qui nous sépare.

Le plus fort de mes remerciements est pour Noureddine. Tu as toujours été un soutien important et tu m'as toujours suivi dans mes choix. Avec tout mon amour, merci.

DÉDICACE

À la mémoire de mon père

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

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. Ce mémoire a pour objectif principal de présenter les modèles multi-états markoviens, en développant la théorie et en l'appliquant à deux ensembles de données : `sir.adm` et `cav`. Dans un premier temps, nous présentons la méthode relative 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. Dans un second temps, nous présentons la théorie des processus de comptage afin d'introduire des méthodes d'estimation non paramétriques dans le cadre d'un modèle de Markov non homogène. Dans ce modèle, les intensités de transition dépendent du temps. Les méthodes d'estimation supposent que le mécanisme de censure est indépendant de l'événement étudié. Les applications ont été réalisées avec les bibliothèques `msm`, `mstate` et `etm` du logiciel R.

Mots-clés : Modèles multi-états, processus de Markov homogène et non homogène, processus de comptage, estimateur de Nelson-Aalen, estimateur de Aalen-Johansen.

INTRODUCTION

L'analyse des historiques d'événements est un domaine important de la statistique appliquée à la santé. 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, à comparer les fonctions de survie de plusieurs groupes ou analyser l'effet des facteurs de risque. Des mécanisme 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 définie par des temps d'observations. Bien souvent, dans l'analyse des durées de survie, on fait l'hypothèse d'indépendance de la durée de survenue de l'événement d'intérêt et de la durée de censure. L'étude d'un événement unique constitue l'une des limites de ce type d'approche. Certaines extensions plus récentes peuvent néanmoins prendre en compte la répétition ou le renouvellement de l'événement étudié lorsqu'il n'est pas unique.

Dans ce contexte, les modèles multi-états connaissent un intérêt grandissant, notamment avec de nombreuses applications en épidémiologie et en recherche clinique. Ils sont considérés comme une généralisation des modèles classiques de survie dans le cas où plusieurs événements se produisent successivements au fil du temps. Ils offrent la

possibilité d'étudier l'évolution des sujets à travers les différents états et de se focaliser sur les risques instantanés au cours du temps de survenue des événements associés à ces états. 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 abandante 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 des formes de transition variées sont envisageables. Toutefois, ils possédent des structures qui ne sont que des extensions de structures standards (Hougaard, 1999).

Les figures 0.1, 0.2 et 0.3 représentent trois exemples simples de modèles multi-états à trois états :

- *Modèle à états progressifs.* Le modèle à états progressifs, présenté à la figure 0.1, permet de prendre en compte des états de transition se succédant avant l'état terminal (Hougaard, 1999). Le principal atout de cette approche par rapport à l'analyse de survie classique est de pouvoir caractériser la progression de la maladie en plus de l'étude de la durée de survenue de l'événement terminal. Ce modèle permet également d'étudier l'impact des covariables sur la vitesse de progression de la maladie.
- *Modèle illness-death.* Le modèle *illness-death* permet d'étudier l'évolution d'un sujet atteint d'une maladie irréversible, particulièrement lorsque cette maladie augmente le risque de décès. Le modèle présenté à la figure 0.2, comporte un état absorbant et deux états transitoires, l'état 0 représente l'absence de maladie, l'état 1 correspond à l'état de maladie et l'état 2 représente le décès. Il permet

Figure 0.1 Structure du modèle multi-états progressifs.

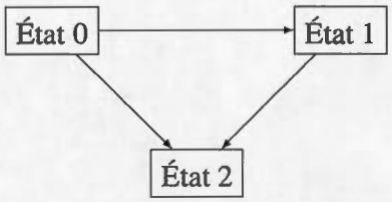

Figure 0.2 Structure du modèle *illness-death*.

de comparer le taux de mortalité chez des sujets sains à celui des sujets malades.

- *Modèle des risques concurrents*. Le modèle des risques concurrents peut être appliqué lorsque plusieurs événements terminaux peuvent se produire (Andersen *et al.*, 1993). Ce modèle, présenté en figure 0.3, permet de prendre en compte plusieurs événements d'intérêt exclusifs comme, par exemple, les différentes causes de décès. Les événements sont dits en concurrence. Ce modèle possède un état transitoire et plusieurs états absorbants, lorsqu'un événement a eu lieu, les autres ne peuvent pas se produire. Dans ce type de modèle, deux approches sont proposées : modélisation en variables latentes ou en causes spécifiques (Crowder, 2001; Pintilie, 2006).

Ces structures, même si elles couvrent beaucoup de problématiques recontrées en pratique, peuvent être adaptées à des situations particulières. Cependant, le choix de la structure représente donc un point majeur.

Tout 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 fonc-

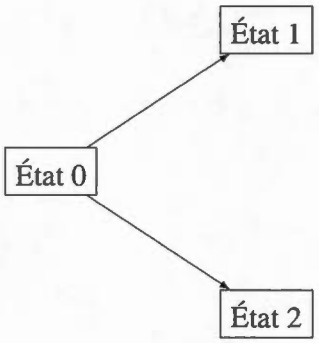

Figure 0.3 Structure du modèle de risques concurrents.

tion 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. 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 dépendantes du temps.

Aalen et Johansen (1978) font partie des premiers à avoir introduit les modèles de Markov pour analyser des données multi-états. Depuis, ces modèles ont été appliqués à de nombreuses problématiques, telle que celle du VIH (Gentleman *et al.*, 1994) ou du cancer (Kay, 1986; Hsieh *et al.*, 2002). Plusieurs modèles statistiques sont possibles, on distingue les approches paramétrique, non paramétrique et semi-paramétrique.

L'approche paramétrique suppose que les intensités de transition appartiennent à une famille particulière de lois, qui dépendent d'un nombre fini de paramètres. L'approche non paramétrique ne fait aucune hypothèse sur la distribution des intensités de transition. Elle a été beaucoup développée à partir de la théorie des processus de comptage et des intégrales stochastiques. L'estimation des intensités de transition est réalisée en utilisant le produit intégral appelé estimateur de Nelson-Aalen (Aalen et Johansen, 1978).

L'approche semi-paramétrique est une solution de remplacement à l'approche para-

métrique ou non paramétrique. On la trouve généralement dans le modèle de survie classique pour évaluer l'effet des facteurs sur l'événement étudié à travers le modèle des risques proportionnels de Cox (Cox, 1972).

De manière similaire à un modèle de survie, le mécanisme de censure à droite est toujours présent dans l'utilisation des modèles multi-états. Par exemple, dans les études épidémiologiques, si le processus qui modélise les événements prend fin avant un état absorbant, cela conduit à des temps d'observation censurés à droite. Par contre, le mécanisme de censure par intervalles intervient lorsque les temps exacts de transition ne sont pas connus (on sait seulement que les transitions se sont produites pendant un intervalle de temps). La présence de ces mécanismes crée des problèmes particuliers dans l'analyse de données et doit être soigneusement examinée lors de la construction des fonctions de vraisemblance des modèles multi-états. Dans la plupart des cas, les mécanismes de censure sont supposés être indépendants du processus d'événement.

Dans ce mémoire, on s'intéresse à l'application de ces approches pour deux ensembles de données nommés `sir.adm` et `cav` :

- `sir.adm` est extrait d'une étude de cohorte qui a été menée dans cinq unités de soins intensifs (médicale, chirurgicale, neurochirurgie et deux interdisciplinaires) dans un hôpital universitaire allemand de février 2000 à juillet 2001 (une période d'étude de 18 mois). Le but était d'étudier l'effet des infections nosocomiales dans les unités de soins intensifs (Wolkewitz *et al.*, 2008). Nous nous intéressons aux risques concurrents (décès ou sortie). Après l'admission à l'unité de soins intensifs (état 0), le patient peut soit sortir vivant (état 1) ou mourir au cours de son séjour (état 2). La pneumonie nosocomiale représente un facteur de risque de décès dans les unités de soins intensifs. C'est une infection pulmonaire qui est contractée à l'hôpital dans les 48 heures qui suivent l'admission. Pour chaque sujet, le temps de séjour dans l'unité de soins intensifs est connu exactement.

- `cav` est extrait d'une base de données de receveurs de greffe de coeur. Sharples *et al.* (2003) ont étudié la progression des vasculopathies d'allogreffe cardiaque (CAV). Ces vasculopathies représentent l'une des causes de décès les plus fréquentes chez les survivants après une transplantation cardiaque. Chaque sujet subit à chaque année une angiographie. Le résultat du test peut être cav-absent (état 1), cav-moyenne (état 2), cav-grave (état 3). Le sujet peut aussi être décédé pendant l'année (état 4). L'état d'un sujet est connu seulement à des dates de diagnostic. Si par exemple, le sujet est diagnostiqué cav-moyenne une année après la greffe de coeur, on sait seulement que cav-moyenne est survenue durant cette année. Cela donne lieu à des données censurées par intervalle.

Les objectifs que nous poursuivons dans ce mémoire sont multiples :

- expliciter la vraisemblance d'un modèle multi-états makovien en présence de censure à droite ou de censure par intervalle ;
- adapter ces vraisemblances à un modèle de survie classique ;
- présenter les méthodes d'estimation basées sur la théorie des processus de comptage, particulièrement les estimateurs non paramétriques de Nelson-Aalen (Aalen et Johansen, 1978) des intensités de transition cumulées et de Aalen-Johanson des probabilités de transition (Aalen et Johansen, 1978) ;
- illustrer ces concepts sur les données de `cav` en utilisant un modèle à quatre états ;
- estimer les intensités et les probabilités de transition entre les états en utilisant la bibliothèque `msm` (Jackson, 2011) du logiciel R.
- mener une analyse semblable pour les données de `sir.adm` avec un modèle de risques concurrents en utilisant les bibliothèques `mstate` (de Wreede *et al.*, 2011) et `etm` (Allignol *et al.*, 2011) du logiciel R.

Le présent mémoire est composé de quatre chapitres ainsi que deux annexes contenant respectivement des notions mathématiques relatives à la théorie des processus de

comptage et les codes du logiciel R. Le premier chapitre de ce mémoire comporte une revue de littérature des modèles multi-états et une description des ensembles de données utilisés dans notre travail. Il se termine par un résumé des bibliothèques R utilisés pour l'analyse des modèles multi-états. Dans le deuxième chapitre, nous rappelons les concepts de base des modèles de survie qui seront adoptés dans la suite de notre travail. Le troisième chapitre présente l'inférence des modèles markoviens homogènes tout en donnant quelques notions et définitions sur les processus. Ce modèle est appliqué à l'ensemble de données `cav`. Le quatrième chapitre présente les modèles markoviens non homogènes. La théorie des processus de comptage est utilisée pour obtenir des estimations non paramétriques. Ces méthodes sont appliquées à l'ensemble de données `sir.adm`.

CHAPITRE I

SURVOL DES MODÈLES MULTI-ÉTATS

Dans la première partie de ce chapitre, nous proposons une revue de littérature des modèles multi-états. Ensuite, nous décrivons les ensembles de données qui seront utilisés tout au long de notre travail. Enfin, nous présentons des bibliothèques R pour l'analyse des modèles multi-états.

1.1 Revue de littérature

L'analyse des historiques d'événements sont étudiées dès le XVII^{e} siècle par divers personnages fondateurs de cette théorie. John Graunt (1620-1674), marchand londonien et fondateur de la statistique démographique, conçoit les bases théoriques de l'élaboration des tables de mortalité grâce à l'étude des dénombrements de population et des bulletins de mortalité de Londres. L'économiste William Petty (1623-1687) systématise et théorise les études démographiques sur les naissances, décès, nombre de personnes par famille, etc. L'astronome anglais Edmund Halley (1662-1742) établit la première table de mortalité véritable en s'appuyant sur 5 ans d'état civil de la ville polonaise de Breslau. Suite à ces travaux d'experts, liés à des préoccupations pratiques, se développe au XIX^{e} siècle l'utilisation financière des données de survie, à une large échelle, par les actuaires des compagnies d'assurance. Durant ce siècle, apparaîssent également les premières modélisations concernant la probabilité de mourir à un certain âge, probabilité qui sera par la suite désignée sous le terme de fonction de risque. Au XX^{e} siècle,

beaucoup de disciplines sont susceptibles d'avoir recours à de tels types de données.

Jusqu'en 1950, la communauté des statisticiens s'intéresse peu à l'analyse des données de survie, la principale contribution étant celle de Greenwood (1926), qui propose une formule pour l'erreur standard d'une table de survie. En 1951, Weibull propose un modèle paramétrique dans le domaine de fiabilité, il fournit une nouvelle distribution de probabilité qui sera par la suite fréquemment utilisée en analyse de survie, la loi de Weibull. En 1958, Kaplan et Meier présentent d'importants résultats concernant l'estimation non paramétrique de la fonction de survie.

L'année 1972 se révèle une date fondamentale, en effet, un modèle semi-paramétrique voit le jour, grâce aux travaux de Cox. Ce modèle est le plus utilisé en analyse des données de survie. Il ne fait aucune hypothèse sur la loi du temps de survie.

L'analyse des données de survie a pour première particularité de ne concerner que des variables aléatoires positives. Une deuxième particularité est qu'elle fait appel à des données incomplètes, censurées ou tronquées. Cependant, une des limites de ce type d'approche est l'étude d'apparition d'un événement unique.

Dans les études longitudinales comme dans les enquêtes de cohorte, les sujets sont observés au cours du temps et les informations les concernant sont recueillies à plusieurs reprises. Le modèle de survie peut être résumé comme un processus à deux états où une seule transition est possible, de l'état vivant à l'état mort (par exemple).

Dans certaines études, l'état vivant peut être divisé en deux ou plusieurs états intermédiaires, dont chacun correspond à une étape particulière du phénomène observé. Dans une telle étude, le modèle de survie classique est insuffisant, puisque le phénomène observé est composé finalement de plusieurs états, d'où la naissance des modèles multi-états. Pour simplifier, les modèles multi-états ne sont qu'une extension du modèle de survie, introduite pour faire face aux études des phénomènes qui se décomposent en

plusieurs états. Les états peuvent être transitoires ou absorbants. Un état est dit absorbant s'il est impossible d'en sortir, il est dit transitoire s'il mène à un état absorbant.

Les modèles multi-états les plus rencontrés dans la littérature sont donnés dans l'article de Hougaard (1999). Deux de ces modèles ont en particulier connu un grand nombre d'applications ces dernières années : le modèle de survie et le modèle des risques concurrents. Les modèles multi-états ne cessent de connaître un intérêt croissant. Ils sont caractérisés par un processus stochastique ayant un espace fini d'états. La notion de processus est utilisée pour représenter les différents états successivement occupés à chaque temps d'observation. En épidémiologie, ils permettent, par exemple, de représenter l'évolution d'un sujet à travers les différents stades d'une maladie. Après définition de différents stades (états), les modèles multi-états permettent d'étudier de nombreuses dynamiques complexes. L'étude de ces modèles consiste à analyser les forces de passage (intensités de transition) entre les différents états.

Cependant, dès que le modèle comprend des états réversibles, il est nécessaire de faire une hypothèse sur l'histoire du sujet. Les modèles de type markovien sont très utiles car ils supposent que l'information sur les états précédents est résumée par l'état présent. Le nombre de publications sur le sujet est très important. On pourra se référer, par exemple, aux travaux de Andersen et Keiding (2002), Hougaard (1999), Andersen *et al.* (1993) et Commenges (1999).

Dans ces modèles de type markovien, les intensités de transition entre les états peuvent être constantes dans le temps ou non ; on parle d'un modèle homogène dans le premier cas et de modèle non homogène dans le second.

Les modèles de Markov homogènes ont été appliqués avec succès dans de nombreux domaines, en particulier en épidémiologie, dans la modélisation des stades du cancer (Kay, 1986; Hsieh *et al.*, 2002), des stades du diabète (Marshall et Jones, 1995) ou encore des stade de l'infection par VIH (Gentleman *et al.*, 1994; Longini *et al.*, 1989).

Si le modèle de Markov homogène est régulièrement utilisé, il impose cependant des contraintes fortes sur le comportement de l'évolution de la maladie. En effet, les intensités de transition sont supposées constantes sur une longue période ce qui est restrictif dans de nombreuses maladies. Un modèle de Markov non homogène permet souvent de mieux modéliser l'évolution du processus mais rend souvent l'inférence plus compliquée.

En 1975, Odd Aalen expose dans sa thèse des méthodes d'inférence statistique pour une famille de processus de comptage (Aalen, 1978). Depuis, la théorie des processus de comptage a été largement étudiée et a permis le développement de plusieurs estimateurs (Andersen *et al.*, 1993). Elle fournit, par l'intermédiaire de la théorie des martingales, un cadre rigoureux à de nombreuses problématiques. Le nombre de publications liées au sujet est très important. On pourra noter les nombreuses contributions de O. Aalen, P. K. Andersen, R. D. Gill, N. Keiding et P. Hougaard.

La théorie des processus de comptage permet d'obtenir des estimateurs non paramétriques dans les modèles de Markov non homogènes. En particulier, l'estimateur de Aalen-Johansen (Aalen et Johansen, 1978) permet d'obtenir une estimation de la matrice des probabilités de transition dans un modèle de Markov. Cet estimateur peut être également adapté à un modèle de régression semi-paramétrique afin d'étudier les effets des covariables (Andersen *et al.*, 1991). Les estimateurs ainsi obtenus par la théorie des processus de comptage constituent une généralisation, pour les modèles markoviens, des estimateurs de Kaplan-Meier et de la vraisemblance partielle de Cox.

Ces méthodes consituent une solution de remplacement intéressante quand l'hypothèse d'homogénéité est trop forte. Cependant, l'utilisation de ces méthodes reste limitée car la théorie des processus de comptage fait intervenir des notions mathématiques plus complexes et la programmation des estimateurs pourrait devenir très complexe.

1.2 Présentation des ensembles de données

Dans cette section, nous présentons les deux ensembles de données qui seront utilisés dans la suite de ce document. Dans le premier ensemble étudié, `sir.adm`, les temps de transition entre les états sont connus exactement (sauf s'il y a censure à droite) alors que pour le second ensemble `cav`, ces temps ne sont pas exactement connus (dans tous les cas). Toutefois, l'ensemble `cav` inclut des temps censurés par intervalle.

1.2.1 L'ensemble de données `sir.adm`

Les données de `sir.adm` concernent des patients admis dans les unités de soins intensifs d'un hôpital, et dont certains d'entre eux développent une pneumonie nosocomiale. Les données font partie de la bibliothèque `mvna` du logiciel R. L'échantillon d'intérêt provient d'une étude de cohorte qui a été menée dans un hopital universitaire allemand de février 2000 à juillet 2001, dans le but d'étudier l'effet des infections nosocomiales dans les unités de soins intensifs (Wolkewitz *et al.*, 2008). Seulement les patients avec une durée de séjour de plus de deux jours ont été inclus. Un extrait de cet ensemble de données est présenté dans le tableau 1.1 et le tableau 1.2 donne une description des caractéristiques des sujets.

Tableau 1.1 Quelques sujets de `sir.adm`.

id	pneu	status	time	age	sex
41	0	1	4	75.34153	F
395	0	1	24	19.17380	M
710	1	1	37	61.56568	M
3138	0	1	8	57.88038	F
3154	0	1	3	39.00639	M
3178	0	1	24	70.27762	M

L'histogramme 1.1 montre que l'âge des sujets se situent majoritairement entre 50 ans

Tableau 1.2 Caractéristiques des sujets de `sir.adm`.

Caractéristiques de l'échantillon	$n = 747$
Covariable continue	Moyenne (écart type)
Âge	57.48 (18.76)
Covariable binaire	(%)
Sexe Feminin	40.96
Pneumonie à l'admission	12.98
Statut	(%)
Sortie de l'USI	87.95
Décès à l'USI	10.17
Censure	1.87

et 75 ans. Donc, en majorité, les sujets sont agés.

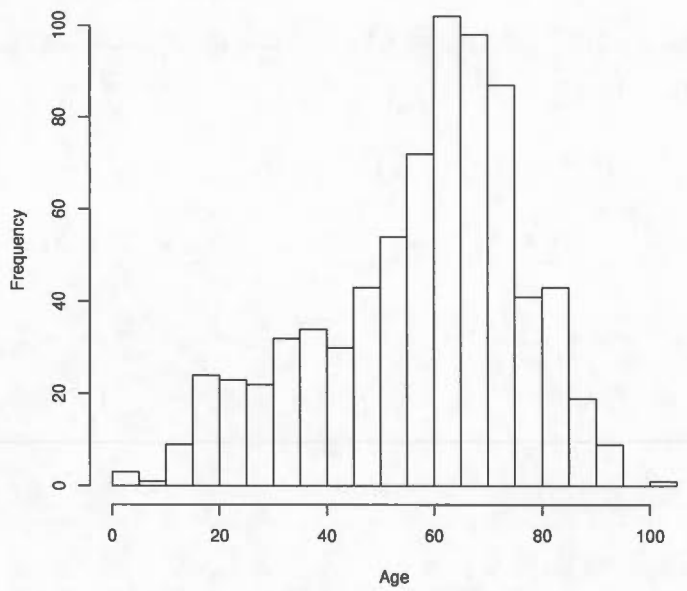

Figure 1.1 Distribution de l'âge pour les sujets de `sir.adm`.

La figure 1.2 montre que les femmes restent moins longtemps à l'unité des soins intensifs et la majorité des sujets quittent les unités des soins intensifs durant les 50 premiers jours.

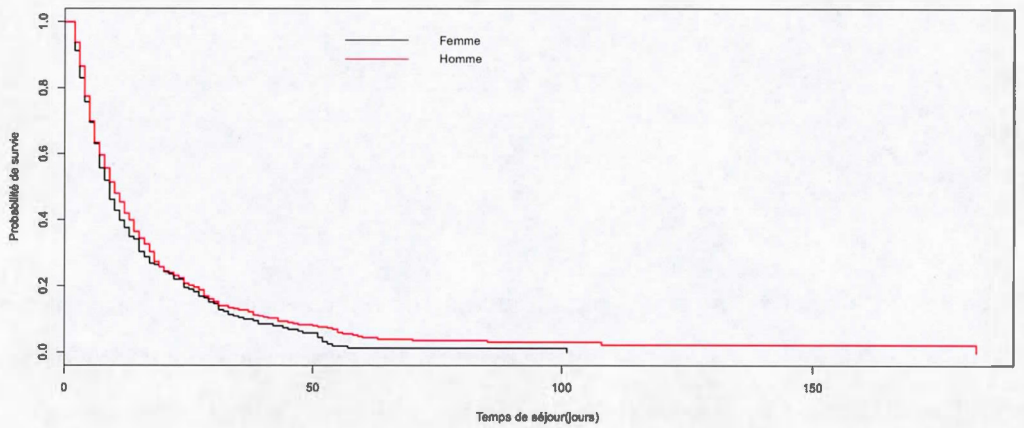

Figure 1.2 Probabilités de sortie selon le sexe.

La figure 1.3 montre que les sujets sans pneumonie restent moins longtemps aux unités des soins intensifs que ceux avec la pneumonie.

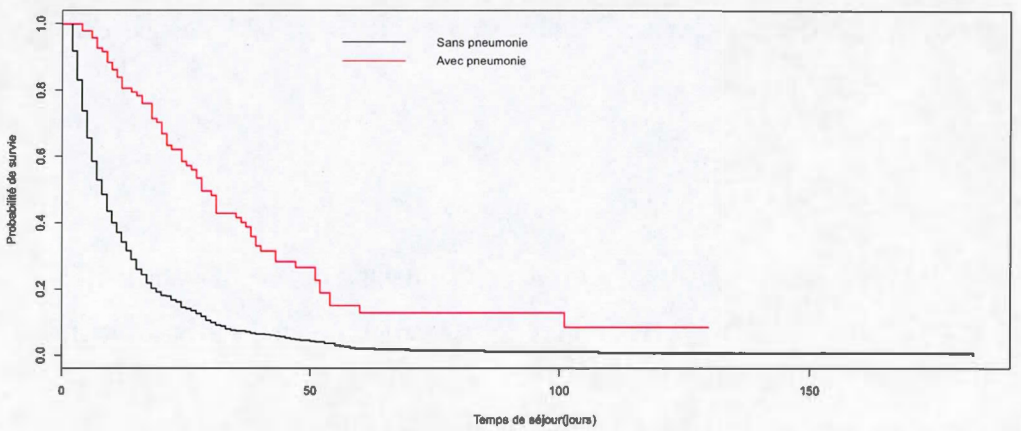

Figure 1.3 Probabilités de sortie selon la pneumonie.

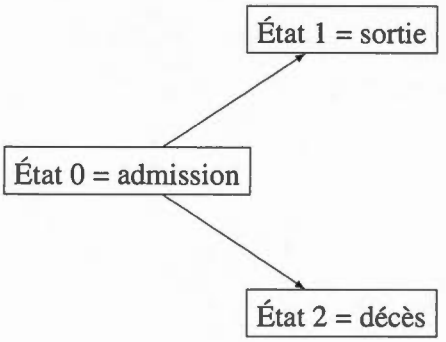

Figure 1.4 Modèle multi-états pour les sujets de `sir.adm`.

L'état de santé du sujet est modélisé par un modèle de risques concurrents (la figure 1.4). Dans ce modèle, les sujets sont initialement dans l'état 0 et peuvent décéder ou sortir de l'unité des soins intensifs. Dans l'ensemble de données `sir.adm`, le temps de séjour correspond au temps de transition de l'état 0 vers l'état 1 ou de l'état 0 vers l'état 2. Ils sont observés en temps continu et connus exactement.

Nous utiliserons cet ensemble de données pour illustrer le calcul de la vraisemblance dans le cas d'un modèle makovien homogène avec des temps exacts dans la section 3.3.2. Nous l'utiliserons également, à la section 4.3, pour effectuer l'analyse non paramétrique d'un modèle markovien non homogène à l'aide de la bibliothèque `mstate` (de Wreede *et al.*, 2011).

1.2.2 L'ensemble de 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). Cet ensemble de données est disponible dans la bibliothèque `msm` du logiciel R.

L'ensemble de données est constitué de 622 sujets, ce qui représente 2846 observa-

tions. Approximativement tous les ans après une greffe de coeur, chaque sujet subit une angiographie pour diagnostiquer la CAV. Le résultat du test est dans l'ensemble $\{1,2,3,4\}$ représentant respectivement cav-absent, cav-moyenne, cav-grave et décès. Donc, on peut modéliser l'état du sujet par un modèle multi-états (à 4 états) qui est illustré par la figure 1.5.

Les données sont spécifiées comme une série d'observations par sujet. Le tableau 1.3 donne un extrait de l'ensemble de données `cav`.

Tableau 1.3 Quelques données de `cav`.

PTNUM	age	years	dage	sex	pdiag	cumrej	state	firstobs	statemax
100002	52.4958	0.0000	21	0	IHD	0	1	1	1
100002	53.4986	1.0027	21	0	IHD	2	1	0	1
100002	54.4986	2.0027	21	0	IHD	2	2	0	2
100002	55.5890	3.0931	21	0	IHD	2	2	0	2
100002	56.4958	4.0000	21	0	IHD	3	2	0	2
100002	57.4931	4.9972	21	0	IHD	3	3	0	3
100002	58.3506	5.8547	21	0	IHD	3	4	0	4
100003	29.5068	0.0000	17	0	IHD	0	1	1	1
100003	30.6958	1.1890	17	0	IHD	1	1	0	1
100003	31.5150	2.0082	17	0	IHD	1	3	0	3

`PTNUM` est l'indicateur de sujet ; `age` (âge à l'examen) et `dage` (âge de distributeur) sont des covariables continues ; `sexe` et `firstobs`, des indicateurs correspondant à la greffe du sujet ou bien à une angiographie postérieure, sont des covariables binaires ; `pdiag` donne le premier diagnostic du sujet et `cumrej` correspond au nombre cumulatif d'épisodes de rejets.

Le tableau 1.4 donne quelques statistiques pour certaines covariables.

Les transitions observées sont décrites dans le tableau 1.5. Ainsi il y avait 148 décès

Tableau 1.4 Caractéristiques des sujets de cav.

Caractéristiques de l'échantillon	$n = 622$
Covariable continue	Moyenne (écart type)
age	48.94 (11.9)
dage	28.77 (11.35)
Covariable binaire	(%)
Sexe masculin	86.01

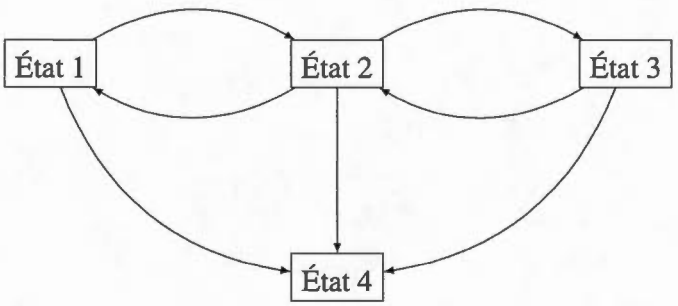

Figure 1.5 Modèle multi-état pour les sujets de cav.

depuis l'état 1, 48 depuis l'état 2 et 55 depuis l'état 3. Par contre, il y a seulement 4 occasions où une observation de cav-grave fut suivie d'une observation cav-absent.

Dans l'ensemble de données cav, l'observation des différents états se fait à intervalles réguliers. Donc les temps de transition exacts ne sont pas connus. De plus, le chemin pour aller d'un état à l'autre entre deux temps d'observation consécutifs est lui aussi inconnu et le nombre de transitions possibles dans l'intervalle peut être élevé. Les modèles multi-états adaptés à de telles données sont souvent désignés dans la litérature sous le terme *Markov models for panel data.*

À la section 3.3.1, nous utiliserons cet ensemble de données pour illustrer la vraisem-

Tableau 1.5 Transitions observées dans `cav`.

Transition	Effectif	Fréquence (%)
$1 \rightarrow 1$	1367	61.5
$1 \rightarrow 2$	204	9.2
$1 \rightarrow 3$	44	2
$1 \rightarrow 4$	148	6.7
$2 \rightarrow 1$	46	3
$2 \rightarrow 2$	134	6
$2 \rightarrow 3$	54	2.4
$2 \rightarrow 4$	48	2.1
$3 \rightarrow 1$	4	0.1
$3 \rightarrow 2$	13	0.6
$3 \rightarrow 3$	107	4.8
$3 \rightarrow 4$	55	2.5

blance dans le cas d'un modèle makovien homogène avec des temps censurés par intervalles.

1.3 Résumé des bibliothèques R pour les modèles multi-états

Des bibliothèques R existent pour l'analyse des modèles multi-états (Beyersmann *et al.*, 2011). La bibliothèque `mstate` (de Wreede *et al.*, 2011) peut être utilisée pour des données censurées à droite ou tronquées à gauche. Elle fournit des estimations non paramétriques et semi-paramétriques pour les intensités de transition et les probabilités de transition. La bibliothèque `msm` (Jackson, 2011) peut être utilisée lorsque les données sont censurées par intervalle. Elle fournit des estimations des intensités de transition et des probabilités de transition. Les intensités de transition sont supposées constantes ou constantes par morceaux (entre deux temps d'observation consécutifs) et sont es-

timées par résolution des équations différentielles de Kolmogorov. Les bibliothèques `changeLOS` (Wrangler *et al.*, 2006), `mvna` (Allignol *et al.*, 2008) et `etm` (Allignol *et al.*, 2011) fournissent des estimateurs non paramétriques pour les modèles multi-états. La bibliothèque la plus spécialisée est `changeLOS` ; elle est basée sur des méthodes décrites dans Schulgen et Schumacher (1996). La bibliothèque `mvna` calcule les estimés des intensités de transition pour des données censurées à droite ou tronquées à gauche. Par contre, elle ne calcule pas les estimés des probabilités de transition. La bibliothèque `etm` calcule les probabilités de transition pour des données censurées à droite et tronquées à gauche. La bibliothèque `cmprsk` (Gray, 2015) fournit des estimateurs des probabilités de transition dans un modèle de risques concurrents. La bibliothèque `SmoothHazard` (Touraine *et al.*, 2014) permet d'estimer les intensités de transition et les probabilités de transition de façon paramétrique ou semi-paramétrique pour des données censurées par intervalle dans un modèle *illness-death*.

Il y a beaucoup d'autres bibliothèques développées par des utilisateurs de R pour l'analyse de survie en général et les modèles multi-états disponibles sur le site du CRAN (*Comprehensive R Archive Network*). On poura aussi se référer au *Journal of Statistical Software*. Dans notre travail, nous avons utilisé la bibliothèque `msm` pour le traitement de l'ensemble de données 1.2.2 via un modèle markovien homogène, et les bibliothèques `etm` et `mstate` pour l'ensemble de données 1.2.1 via un modèle markovien non homogène.

CHAPITRE II

ANALYSE DE SURVIE CLASSIQUE

L'objectif de ce chapitre est de familiariser le lecteur avec 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 de 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.

2.1 Les éléments de base

On suppose que la durée de survie T est une variable aléatoire non négative dont la loi est absolument continue. La distribution de T est caractérisée par l'une des cinq fonctions suivantes définies pour $t \geq 0$, chacune pouvant être obtenue à partir de l'une des quatre autres.

Définition 2.1.1. *La* fonction de survie *S au temps t est la probabilité de survivre jusqu'à l'instant t :*

$$S(t) = P[T > t], \quad t \geq 0.$$

Définition 2.1.2. *La fonction de répartition (cumulative distribution function) au temps*

t est la probabilité pour que l'évènement survienne avant t :

$$F(t) = P[T \leq t] = 1 - S(t).$$

Définition 2.1.3. *La densité de probabilité f au temps t est la probabilité pour que l'évènement survienne dans l'intervalle de temps* $[t, t+dt[$ *(où dt est infinitésimal) :*

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}.$$

Définition 2.1.4. *La* fonction de risque instantané *au temps t est définie par*

$$\begin{aligned} \alpha(t) &= \lim_{\Delta t \to 0} \frac{P\big(t \leq T < t+\Delta t \mid T \geq t\big)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t)}{\Delta t P(T \geq t)} \\ &= \frac{f(t)}{S(t)} = -\frac{d}{dt}\big(\ln S(t)\big), \end{aligned} \tag{2.1}$$

et donne la probabilité que l'évènement survienne dans l'intervalle de temps $[t, t+dt[$ *sachant qu'il n'a pas eu lieu avant l'instant t.*

Définition 2.1.5. *On définit la* fonction de risque cumulé *par*

$$A(t) = \int_0^t \alpha(u)du.$$

Proposition 2.1.1. *La fonction de survie peut aussi s'exprimer en fonction du risque cumulé :*

$$\begin{aligned} S(t) &= P(T > t) \\ &= \exp\left(-\int_0^t \alpha(u)du\right) \\ &= \exp(-A(t)). \end{aligned}$$

2.2 Censure

Une des caractéristique des données de survie est l'existence d'observations incomplètes. En effet, dans les enquêtes épidémiologiques, les données sont souvent recueillies de façon incomplète. La censure fait partie des processus générant ce type

de données (Andersen *et al.*, 1993; Kalbfleisch et Prentice, 2002). Elle doit être prise en compte dans l'écriture de la vraisemblance. Il existe trois types de censure : censure à droite, censure à gauche et censure par intervalle. Le phénomène de censure le plus courant est la censure à droite. Il correspond au cas où le sujet n'aurait pas subi l'événement à sa dernière observation. Un autre phénomène de censure souvent rencontré dans les études épidémiologiques est la censure par intervalle. La censure par intervalle correspond au cas où l'événement d'intérêt se serait produit entre deux temps d'observation.

2.2.1 Censure à droite

Dans tout ce qui suit, nous utiliserons les notations suivantes. Il y a n durées de survie T_i, $i = 1,\ldots,n$, indépendantes et identiquement distribuées. Il y a aussi n temps de censure C_i, $i = 1,\ldots,n$ qui ne sont pas nécessairement indépendants. Supposons que C_i est la valeur réalisée du temps de censure pour le sujet i et T_i est la valeur réalisée de son temps de survie. Si $C_i < T_i$, nous dirons que le temps de survie pour le sujet i est *censuré à droite*. Si le temps de censure arrive avant le temps de décès, nous savons seulement que le temps de décès pour ce sujet est plus grand que son temps censuré. Nous ne pouvons pas ignorer les informations sur ces sujets censurés. Si nous ignorons les sujets censurés dans nos analyses, nous aurions tendance à sous-estimer le temps de survie.

Généralement, dans des données de cohorte, une durée T_i est censurée à droite si le sujet i est

- *perdu de vue* : sa surveillance est interrompue alors qu 'il n'a pas encore subi l'événement (pour cause de déménagement par exemple),
- *exclu vivant* : à la date de fin d'étude le sujet n'a pas encore subi l'événement.

Il peut être très difficile de faire l'inférence pour un temps de survie en présence de censure à droite. La définition 2.2.1 ci-dessous donne un cas de censure à droite pour

lequel l'inférence sur les temps de survie est encore facile. Mais avant, nous présentons un exemple de censure à droite aléatoire.

Exemple 2.2.1. *(Données de Freireich) Freireich, en 1963, a réalisé un essai thérapeutique ayant pour objectif de comparer les durées de rémission avant rechute (en semaines) de 21 sujets atteints de leucémie selon qu'ils ont reçu ou non un médicament appelé la 6-mercaptopurine (6 M-P) ; le groupe témoin a recu un placebo et l'essai a été fait à double insu.*

Tableau 2.1 Durée de rémission (en semaine) selon le traitement

6 M-P	6	6	6	6^+	7	9^+	10	10^+	11^+	13	16	17^+	19^+
	20^+	22	23	25^+	32^+	32^+	34^+	35^+					
Placebo	1	1	2	2	3	4	4	5	5	8	8	8	8
	11	11	12	12	15	17	22	23					

Les nombres suivis du signe + correspondent à des sujets qui ont été perdus de vue à la date considérée. Ils sont donc exclus vivants de l'étude et on sait donc seulement d'eux que leur durée de survie est supérieure à celle indiquée. Par exemple, le quatrième sujet traité par 6 M-P a eu une durée de rémission supérieure à 6 *semaines, alors que les trois premiers ont eu une durée de rémission égale à* 6 *semaines. On dit que les perdus de vue ont été censurés à droite et ce problème de la censure demande un traitement particulier. En effet, si l'on se contentait d'éliminer les observations incomplètes, c'est-à-dire les* 12 *sujets censurés du groupe traité, on perdrait beaucoup d'information car on ne tiendrait pas compte des sujets qui ont justement les durées de rémission les plus longues.*

Définition 2.2.1. *La censure à droite est dite* aléatoire *si, pour chaque sujet i, la variable aléatoire C_i est indépendante de la variable aléatoire T_i.*

Définition 2.2.2. *La censure à droite est dite* indépendante *si pour chaque sujet i,*

$$\begin{aligned}\alpha(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T_i < t + \Delta t \mid T_i \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T_i < t + \Delta t \mid T_i \geq t, C_i \geq t)}{\Delta t}.\end{aligned}$$

Remarque 2.2.1. *La censure aléatoire est une censure indépendante, par contre le contraire n'est pas toujours vrai.*

Dans la suite de ce chapitre, on suppose que les durées de survie T_i et les durées de censure C_i satisfont la définition 2.2.2. Cette hypothèse ne serait pas vérifiée si, par exemple, des sujets cessaient d'être suivis à cause d'une aggravation de leur état. Dans un tel cas, une différence intrinsèque dans les temps de survie est masquée par la censure si bien que l'égalité dans la définition 2.2.2 cesse d'être vérifiée pour des raisons liées aux T_i et non aux C_i. Dans leur article, Andersen et Keiding (2002) expliquent plus en détails comment on doit intuitivement comprendre la notion de censure indépendante.

Nous verrons, dans la section 2.3, que l'hypothèse d'indépendance des variables T_i et C_i est fondamentale afin d'obtenir une vraisemblance simple. Une autre hypothèse qui simplifie la vraisemblance est que la censure est non informative. La censure est supposée *non informative* si sa contribution à la vraisemblance ne dépend pas des paramètres qui interviennent dans la distribution de la durée de survie.

En théorie, l'hypothèse de censure à droite indépendante ne permet pas de dire si la censure est informative ou non. Kalbfleisch et Prentice (2002) parviennent à donner un exemple de censure indépendante et informative. Mais cet exemple n'est pas vraiment réaliste et, en pratique, la censure à droite est habituellement supposée indépendante et non informative.

Tout au long de notre travail, ces hypothèses seront utilisées. Ces hypothèses sont raisonnables si les données censurées proviennent de sujets qui n'ont pas connu l'évé-

nement avant la fin de l'étude, c'est-à-dire les sujets exclus-vivants. Elles sont moins clairement vérifiées quand les données censurées correspondent à des sujets perdus de vue.

2.2.2 Censure par intervalle

La durée de survie T est dite *censurée par intervalle* si au lieu de l'observer de façon exacte, la seule information dont on dispose est qu'elle est comprise entre deux dates connues. La censure par intervalle se rencontre généralement dans les études de cohorte où les sujets ne sont pas observés en temps continu mais plutôt à des visites successives. Par exemple, si l'on s'intéresse à l'âge où survient une maladie et que le sujet i est diagnostiqué malade au cours d'une visite, on sait seulement que $T_i \in [L_i, R_i]$ où R_i est l'âge à la visite de diagnostic et L_i est l'âge à la visite précédente.

Comme avec la censure à droite, il est possible de définir, pour la censure par intervalle, les notions de censure indépendante ou non informative. On pourra consulter la page 79 du livre de Kalbfleisch et Prentice (2002) et Sun (2007) pour plus de détails.

2.3 Vraisemblances

Comme nous l'avons vu précédemment, les études d'analyse de survie donnent lieu à des données censurées, dont il faut tenir compte dans l'écriture des fonctions de vraisemblance. Dans cette section, nous allons présenter le calcul de la vraisemblance d'un modèle de survie pour deux cas : lorsque les données sont censurées à droite pour les observations exactes, ou lorsque les données sont censurées par intervalle. Nous ferons les calculs sous l'hypothèse que la censure est aléatoire et non informative, quoique les résultats sont encore vrais si la censure est indépendante et non informative.

Soit θ le vecteur des paramètres du modèle. Supposons que l'on étudie n sujets. Les estimateurs des paramètres sont obtenus en maximisant la vraisemblance détaillée ci-

dessous. La vraisemblance représente la probabilité d'observer l'échantillon d'après le modèle et elle est le produit des n contributions individuelles.

2.3.1 Vraisemblance pour la censure à droite

Pour chaque sujet i, considérons le cas d'une censure aléatoire à droite C_i indépendante de la durée d'intérêt T_i comme nous l'avons définie 2.2.1 dans la section précédente. Supposons que les variables T_i et C_i ont f et g pour densités respectives, et S et G pour fonctions de survie. La distribution de T_i est définie par un paramètre θ de dimension finie. Toute l'information est contenue dans le couple (X_i, Δ_i), où $X_i = \min(T_i, C_i)$ est la durée observée, et $\Delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$ est l'indicatrice (d'absence) de censure.

Supposons que l'on dispose de n durées observées avec leur indicatrice :

$$(x_1, \delta_1),\ (x_2, \delta_2),\ \ldots,\ (x_n, \delta_n).$$

Donc (x_i, δ_i) est une réalisation de (X_i, Δ_i) et ces couples sont supposés indépendants. Sous cette hypothèse d'indépendance, la vraisemblance L est le produit des contributions L_i de chaque observation :

$$L\big((x_1, \delta_i), (x_2, \delta_2) \ldots, (x_n, \delta_n); \theta\big) = \prod_{i=1}^{n} L_i\big(x_i, \delta_i; \theta\big).$$

L'écriture des contributions individuelles dépend du type d'observation :

- si le temps d'observation considéré est un temps exact d'événement ($\delta_i = 1$), la contribution à la vraisemblance est $f(x_i; \theta)$;
- si le temps d'observation considéré est un temps de censure à droite ($\delta_i = 0$), on sait seulement que le temps d'événement est plus grand que le temps de censure, la contribution à la vriasemblance sera $S(x_i; \theta)$.

Ainsi, si le sujet i a subi l'événement, sa contribution à la vraisemblance est

$$\begin{aligned} &P(X_i \in [x_i, x_i + dx[\,, \delta_i = 1\,; \theta) \\ &\qquad = \ P(T_i \in [x_i, x_i + dx[\,, C_i \geq T_i\,; \theta) = \big[f(x_i; \theta) G(x_i)\big]^{\delta_i} \end{aligned}$$

et si le sujet i est censuré à droite, la contribution est plutôt

$$P(X_i \in [x_i, x_i + dx[\,, \delta_i = 0; \theta)$$
$$= \; P(C_i \in [x_i, x_i + dx[\,, C_i < T_i; \theta) = \left[g(x_i) S(x_i; \theta) \right]^{1-\delta_i}.$$

Donc, la contribution à la vraisemblance pour le sujet i peut s'écrire sous la forme :

$$L_i(x_i; \delta_i; \theta) = \left[f(x_i; \theta) G(x_i) \right]^{\delta_i} \times \left[g(x_i) S(x_i; \theta) \right]^{1-\delta_i}.$$

Mais puisque la censure est non informative, le paramètre θ n'apparaît pas dans la loi de la censure. On peut donc se limiter, pour les fins de la maximisation, aux termes relatifs à la loi de survie, c'est-à-dire considérer que la vraisemblance est

$$L\big((x_1, \delta_i), (x_2, \delta_2), \ldots, (x_n, \delta_n); \theta\big) = \prod_{i=1}^{n} f(x_i; \theta)^{\delta_i} S(x_i; \theta)^{1-\delta_i}. \tag{2.2}$$

En utilisant la définition 2.1.4, la vraisemblance totale (2.2) peut aussi s'écrire

$$L\big((x_1, \delta_i), (x_2, \delta_2), \ldots, (x_n, \delta_n); \theta\big) = \prod_{i=1}^{n} S(x_i; \theta) \alpha(x_i; \theta)^{\delta_i}. \tag{2.3}$$

2.3.2 Vraisemblance pour la censure par intervalle

Considérons le cas d'une censure par intervalle. Soit S la fonction de survie de la variable d'interêt 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

$$L = \prod_{i=1}^{n} \Big(S(l_i; \theta) - S(r_i; \theta) \Big). \tag{2.4}$$

À noter que ces fonctions de vraisemblance seront également abordées sous l'approche markovienne dans la section 3.3.1 pour des données en panel, ainsi que dans la section 3.3.2 pour des données censurées à droite.

2.4 Les méthodes d'estimation

Le modèle de survie classique peut être vu comme un modèle multi-états particulier, défini par deux états et une transition. C'est ce que montre la figure 2.1 ci-dessous. L'objectif est alors de modéliser la fonction de risque $\alpha(t)$ présentée dans la définition 2.1.4. Plusieurs approches peuvent être envisagées entre les modèles non paramétriques, semi-paramétriques et paramétriques.

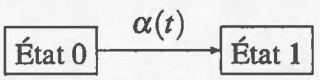

Figure 2.1 Modèle de survie classique.

2.4.1 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 de 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, c'est 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.

2.4.1.1 Loi exponentielle

La distribution exponentielle, qui ne dépend que d'un paramètre α, est la seule distribution continue qui admet un risque instantané constant. Pour $t \geq 0$ et avec $\alpha > 0$,

$$\begin{aligned} \alpha(t) &= \alpha, \\ S(t) &= \exp(-\alpha t), \\ f(t) &= \alpha \exp(-\alpha t), \\ F(t) &= 1 - \exp(-\alpha t), \\ A(t) &= \alpha t. \end{aligned}$$

C'est la distribution à risque constant ou sans mémoire.

2.4.1.2 Loi de Weibull

La distribution de Weibull est une généralisation de la loi exponentielle. Elle est caractérisée par deux paramètres α et γ, qui sont des paramètres de forme et d'échelle respectivement. Pour $t \geq 0$ et avec $\alpha > 0$ et $\gamma > 0$,

$$\begin{aligned} \alpha(t,\alpha,\gamma) &= \alpha\gamma t^{\gamma-1}, \\ S(t,\alpha,\gamma) &= \exp(-\alpha t^{\gamma}), \\ f(t,\alpha,\gamma) &= \alpha\gamma t^{\gamma-1}\exp(-\alpha t^{\gamma}), \\ F(t,\alpha,\gamma) &= 1 - \exp(-\alpha t^{\gamma}), \\ A(t,\alpha,\gamma) &= \alpha t^{\gamma}. \end{aligned}$$

Par ailleurs, si $\gamma = 1$, la fonction de risque devient constante et donc nous retrouvons la loi exponentielle.

2.4.2 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.

2.4.2.1 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, $i = 1,\ldots,n$, soit n variables aléatoires indépendantes et identiquement distibuées, ayant la même fonction de survie $S(t)$. L'estimateur le plus naturel de $S(t)$ est la survie empirique :

$$\hat{S}_n(t) = \sum_{i=1}^{n} \mathbf{1}_{\{T_i > t\}}.$$

Malheureusement, dans le cas où les données sont censurées à droite, il est impossible d'utiliser la fonction de survie empirique puisqu'elle fait intervenir des données non observées. En effet, en présence de censure, T_i n'est pas connue.

Pour dériver l'estimateur de Kaplan-Meier, partons de nouveau de l'échantillon avec censure de la section 2.3.1 : (x_1, δ_1), (x_2, δ_2), ..., (x_n, δ_n). Notons $t_1,\ldots,t_k$, 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'échan-

tillon. Finalement, dans chaque intervalle $[t_j, t_{j+1})$, supposons qu'il y a m_j observations censurées, notés $c_{j,1}, \ldots, c_{j,m_j}$, également énumérées par ordre croissant. Avec ces notations, l'échantillon ordonné se présente comme suit :

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

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

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

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

les probabilités que la loi discrète met sur les points $t_1, \ldots, t_k, t_{k+1}$ (t_{k+1} sert à traiter les données censurées plus grandes que t_k). En raisonnant comme dans la section 2.3.1 (censure aléatoire et non informative), 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 :

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

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

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

ce qui permet de simplifier la vraisemblance comme suit :

$$L = \prod_{j=1}^{k} p(t_j)^{d_j} S(t_j)^{m_j} = \prod_{j=1}^{k} p_j^{d_j} S_j^{m_j} \tag{2.5}$$

(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 $(p_1, \ldots, p_k, p_{k+1})$. Une telle maximisation est possible mais laborieuse. On fait alors une seconde observation importante : *il est équivalent de maximiser (2.5) sur* $(S_1, \ldots, S_k)$ *plutôt que* $(p_1, \ldots, p_k)$ *puisqu'il y a correspondance biunivoque entre les deux.* Comme $p_j = S_{j-1} - S_j$, la vraisemblance (2.5) devient

$$L = \prod_{j=1}^{k} (S_{j-1} - S_j)^{d_j} S_j^{m_j} \tag{2.6}$$

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

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

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 dérivée 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 $(\hat{S}_1, \ldots, \hat{S}_k)$ est dans le cône puisque $\hat{S}_1 \leq 1$ 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

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{2.7}$$

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 (càdlàg).

2.4.2.2 Estimateur de Nelson-Aalen du risque cumulé

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

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

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'événement au temps t_j. L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut à chaque temps de l'événement t_j.

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 $\alpha(t)$ (voir équation (2.1))

Exemple 2.4.1. *Calculons l'estimateur de Kaplan-Meier sur les données de l'exemple 2.2.1 pour le groupe traité par la 6-mercaptopurnie (6 M-P). Alors la fonction de survie* $\hat{S}(t)$, *estimée par la méthode de Kaplan-Meier, est une fonction en escalier, non nulle*

pour tout $t \geq 0$, *telle que*

$$\begin{aligned}
\hat{S}(t) &= 1 \quad si\ 0 \leq t < 6 \\
&= 0.857 \quad si\ 6 \leq t < 7 \\
&= 0.807 \quad si\ 7 \leq t < 10 \\
&= 0.753 \quad si\ 10 \leq t < 13 \\
&= 0.690 \quad si\ 13 \leq t < 16 \\
&= 0.627 \quad si\ 16 \leq t < 22 \\
&= 0.538 \quad si\ 22 \leq t < 23 \\
&= 0.448 \quad si\ 23 \leq t.
\end{aligned}$$

Sa courbe est tracée à la figure 2.2. On remarque en particulier dans cet exemple que

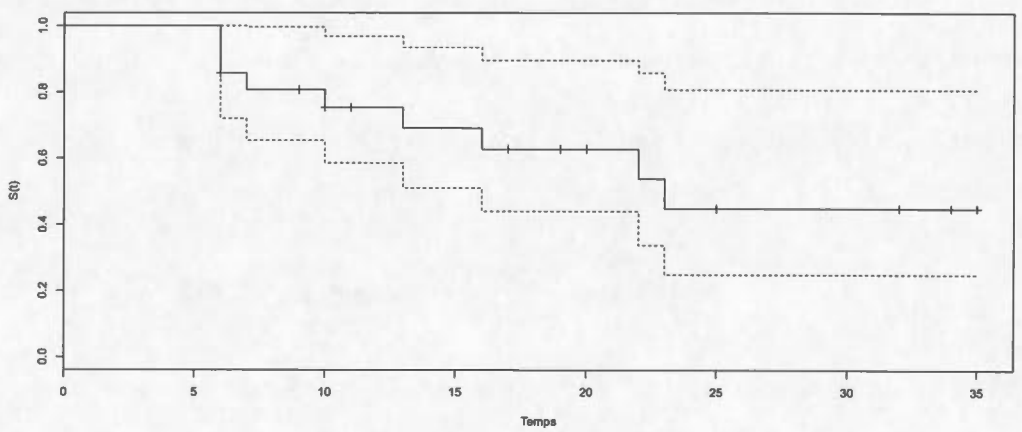

Figure 2.2 Courbe de Kaplan-Meier de la fonction de survie.

l'estimateur de Kaplan-Meier n'atteint jamais 0. Ceci est dû au fait que les dernières observations sont censurées.

2.4.3 Les modèles de risques proportionnels

Nous nous somme concentrés jusqu'ici sur le délai jusqu'à la survenue de l'événement d'intérêt mais l'un des principaux objectifs de l'analyse de survie et de l'épidémiologie est d'évaluer l'effet, sur ce délai, des covariables auxquelles sont exposés les sujets.

Dans un modèle de risques proportionnels, l'effet des diverses covariables sur la fonction de risque est de type multiplicatif. On introduit une fonction de risque de base qui donne la forme générale du risque et qui est commune à tous les sujets et on ajoute l'effet des covariables en multipliant par un facteur qui dépend de celles-ci. Le modèle de risques proportionnels le plus largement utilisé est le modèle de Cox (1972) :

$$\alpha(t \mid \vec{Z}_i) = \alpha_0(t)\exp(\vec{\beta}'\vec{Z}_i),$$

où i est l'indice du sujet, $\vec{Z}_i$ est le vecteur des covariables, $\vec{\beta}$ est le vecteur des coefficients de régression et $\alpha_0(t)$ est la fonction de risque de base, c'est-à-dire le risque instantané des sujets pour lesquels toutes les covariables sont nulles.

Considérons un sujet qui a pour caractéristiques $\vec{Z}_i$ par rapport à un autre sujet qui a pour caractéristiques $\vec{Z}_j$. Le rapport des risques instantanés est donné par

$$\frac{\alpha(t \mid \vec{Z}_i)}{\alpha(t \mid \vec{Z}_j)} = \frac{\alpha_0(t)\exp(\vec{\beta}'\vec{Z}_i)}{\alpha_0(t)\exp(\vec{\beta}'\vec{Z}_j)} = \frac{\exp(\vec{\beta}'\vec{Z}_i)}{\exp(\vec{\beta}'\vec{Z}_j)} = \exp(\vec{\beta}'(\vec{Z}_i - \vec{Z}_j))$$

et on voit qu'il est constant au cours du temps. C'est *l'hypothèse des risques proportionnels* caractéristique du modèle de Cox.

L'intérêt de ce modèle semi-paramétrique réside dans la séparation de la fonction du risque $\alpha_0(t)$ du coefficient de régression $\vec{\beta}$. Ainsi, en maximisant la vraisemblance partielle de Cox (Cox, 1975), l'estimation des paramètres $\vec{\beta}$ peut se faire sans spécifier ni estimer la fonction de risque $\alpha_0(t)$. L'estimation du risque de base cumulé peut s'effectuer notamment avec l'estimateur de Nelson-Aalen. Pour plus de détails, on peut

consulter Nelson (1972) et Aalen (1978). Par ailleurs, on peut aussi construire l'estimateur de Nelson-Aalen à partir de l'estimateur de Kaplan-Meier (voir Kalbfleisch et Prentice (2002)).

CHAPITRE III

MODÈLES DE MARKOV HOMOGÈNES

Dans ce chapitre, nous nous intéressons aux processus markoviens homogènes à temps continu et à espace d'états fini ; on se base sur l'ouvrage de référence de Andersen *et al.* (1993). Nous commençons par un rappel des notions sur les processus qui seront nécessaires par la suite. Cette notion de processus est indispensable dans les modèles multi-états. Ensuite, nous présentons les vraisemblances d'un modèle muti-états principalement décrites dans les articles Hougaard (1999) et Kalbfleisch et Lawless (1985). Enfin, nous présentons un modèle de Markov homogène qui permet d'observer l'impact de certains facteurs de risque et de pouvoir le quantifier par l'intermédiaire du modèle des risques proportionnels.

3.1 Définitions et propriétés

Cette section est consacrée aux définitions et propriétés des processus markoviens.

3.1.1 Processus stochastique

Soit $(\Omega, \mathscr{A}, \mathbb{P})$ un espace de probabilité, $\mathscr{T}$ l'espace des temps et S l'espace d'états. Un processus stochastique $\{X(t), t \in \mathscr{T}\}$ est une application définie par

$$\begin{aligned} X : \mathscr{T} \times \Omega &\rightarrow S \\ (t, \omega) &\mapsto X(t, \omega) \end{aligned}$$

telle que, pour $t \in \mathscr{T}$, la fonction $\omega \mapsto X(t,\omega)$ est une variable aléatoire sur $(\Omega, \mathscr{A}, \mathbb{P})$ à valeur dans S.

Pour un ω donné, la fonction $t \mapsto X(t,\omega)$ est la trajectoire du processus en ω.

3.1.2 Propriété de Markov

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

$$P(X(t) = x_j \mid X(r) = x_r, r \leq s) = P(X(t) = x_j \mid X(s) = x_s). \tag{3.1}$$

Définition 3.1.1. *Un processus de Markov $\{X(t), t \geq 0\}$ à temps continu et à espace d'états fini $S = \{1,\ldots,k\}$ est complètement défini par*

1. *son vecteur des probabilités initiales, noté $\boldsymbol{P}_0$ tel que*

$$\boldsymbol{P}_0[j] = P(X(0) = j), \quad j = 1,\ldots,k, \tag{3.2}$$

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

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

$$p_{ij}(s,t) = P(X(t) = j \mid X(s) = i), \ \forall \ 0 \leq s \leq t \ \textit{et} \ i,j \in S \tag{3.3}$$

   avec

$$\boldsymbol{P}(s,s) = \mathbf{I} \ \textit{et} \ \sum_{j=1}^{k} p_{ij}(s,t) = 1, \ \forall \ 0 \leq s \leq t. \tag{3.4}$$

Théorème 3.1.1 (Équations de Chapman-Kolmogorov). *Soit $\{X(t), t \geq 0\}$ un processus de Markov à temps continu et à espace d'états fini $S = \{1,\ldots,k\}$, et de probabilités de transition $p_{ij}(s,t)$, $0 \leq s \leq t$, $i,j \in S$. Alors*

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

soit en notation matricielle

$$\boldsymbol{P}(s,t) = \boldsymbol{P}(s,u)\boldsymbol{P}(u,t), \ \forall \ 0 \leq s \leq u \leq t, \tag{3.5}$$

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

DÉMONSTRATION. Par la formule des probabilités totales, et la définition des probabilités conditionnelles, alors

$$\begin{aligned}
p_{ij}(s,t) &= P(X(t) = j \mid X(s) = i) \\
&= \frac{P(X(t) = j, X(s) = i)}{P(X(s) = i)} \\
&= \sum_{\ell=1}^{k} \frac{P(X(t) = j, X(s) = i, X(u) = \ell)}{P(X(s) = i)}, \ 0 \leq s \leq u \leq t \\
&= \sum_{\ell=1}^{k} \frac{P(X(t) = j \mid X(s) = i, X(u) = \ell) P(X(s) = i, X(u) = \ell)}{P(X(s) = i)} \\
&= \sum_{\ell=1}^{k} P(X(t) = j \mid X(s) = i, X(u) = \ell) P(X(u) = \ell \mid X(s) = i) \\
&= \sum_{\ell=1}^{k} p_{i\ell}(s,u) p_{\ell j}(u,t),
\end{aligned}$$

où la dernière égalité provient de la propriété markovienne (3.1). □

Le *générateur infinitésimal* $\{\mathbf{Q}(t), t \geq 0\}$ d'un processus de Markov est défini par

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

On peut démontrer (sous de faibles hypothèses) que ces deux limites existent. On voit facilement par ailleurs que

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

En effet, on a

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

En analyse de survie, les modèles multi-états sont des cas particuliers de processus markoviens. Le paramètre d'intérêt dans ces modèles est la *fonction de risque instantané* $t \mapsto (\alpha_1(t),\ldots,\alpha_k(t))$, qui est définie par

$$\alpha_i(t) = \sum_{j\neq i} \alpha_{ij}(t) \tag{3.6}$$

$$\alpha_{ij}(t) = \lim_{\Delta t\to 0} \frac{p_{ij}(t,t+\Delta t)}{\Delta t}, \quad i\neq j. \tag{3.7}$$

Le risque instantané α est donc donné par le générateur infinitésimal : $\alpha_i(t) = -q_{ii}(t)$. 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 $t+\triangle t$ 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 3.1.1. *Le modèle de survie (Figure 2.1) 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*

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

Avec $p_{00}(0,t)$ et $p_{01}(0,t)$ qui correspondent aux fonctions $S(t)$, $F(t)$ respectivement présentées dans la section 2.1.

3.2 Homogénéité et temps de séjour dans l'état

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 récrire les probabilités en termes de l'écart : $\mathbf{P}(s,t) = \mathbf{P}(t-s)$ et l'équation de Chapman-Kolmogorov (3.5) devient

$$\mathbf{P}(t-s) = \mathbf{P}(u-s)\mathbf{P}(t-u).$$

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

$$q_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t,t+\Delta t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{p_{ij}(\Delta t)}{\Delta t} = q_{ij}. \tag{3.8}$$

Ainsi, la matrice des intensités de transition $\mathbf{Q}$ est de la forme

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

Théorème 3.2.1. *Soit $\{X(t), t \geq 0\}$ un processus de Markov homogène de matrice des intensités de transition $\boldsymbol{Q} = (q_{ij})_{i,j \in S}$. Alors on a :*

1. *(équation de Kolmogorov rétrograde)*

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

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

2. *(équation de Kolmogorov progressive)*

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

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

DÉMONSTRATION. Démontrons la première équation, soit $t, \Delta t \geq 0$ et $i, j \in S$. Par la formule de Chapman-Kolmogorov (3.5), nous avons

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

soit

$$\begin{aligned} p_{ij}(t+\Delta t) - p_{ij}(t) &= \sum_{\ell=1}^{k} p_{i\ell}(\Delta t)p_{\ell j}(t) - p_{ij}(t) \\ &= \sum_{\ell \neq i} p_{i\ell}(\Delta t)p_{\ell j}(t) + p_{ii}(\Delta t)p_{ij}(t) - p_{ij}(t) \\ &= \sum_{\ell \neq i} p_{i\ell}(\Delta t)p_{\ell j}(t) + (p_{ii}(\Delta t) - 1)p_{ij}(t). \end{aligned} \tag{3.9}$$

En divisant les deux membres de l'équation (3.9) par Δt, puis prenant la limite quand $\Delta t \to 0$, nous obtenons

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

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 de l'une ou de l'autre des équations ci-dessus est donnée par

$$\mathbf{P}(t) = \exp(t\mathbf{Q}) = \sum_{s=0}^{\infty} \frac{t^s}{s!} \mathbf{Q}^s. \tag{3.10}$$

En effet, dérivons l'équation (3.10) par rapport à t

$$\begin{aligned}
\frac{d}{dt}\mathbf{P}(t) &= \frac{d}{dt}\left\{\mathbf{I} + t\mathbf{Q} + \frac{t^2\mathbf{Q}^2}{2!} + \frac{t^3\mathbf{Q}^3}{3!} + \cdots\right\} \\
&= \left\{0 + \mathbf{Q} + t\mathbf{Q}^2 + 3\frac{t^2\mathbf{Q}^3}{3!} + \cdots\right\} \\
&= \left\{\mathbf{Q} + t\mathbf{Q}^2 + \frac{t^2\mathbf{Q}^3}{2!} + \cdots\right\} \\
&= \left\{\mathbf{I} + \mathbf{Q}t + \frac{t^2\mathbf{Q}^2}{2!} + \cdots\right\}\mathbf{Q} \\
&= \mathbf{P}(t)\mathbf{Q}.
\end{aligned} \tag{3.11}$$

Il est facile de remarquer que la solution générale de l'équation (3.11) est sous la forme

$$\mathbf{P}(t) = c\exp(t\mathbf{Q}),$$

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

Remarque 3.2.1. *Étant donné le lien précédemment évoqué entre $\boldsymbol{Q}$ et les fonctions de risque instantanné, nous écrirons indifféremment :*

$$\boldsymbol{Q} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1k} \\ q_{21} & q_{22} & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kk} \end{pmatrix} = \begin{pmatrix} -\alpha_1 & \alpha_{12} & \cdots & \alpha_{1k} \\ \alpha_{21} & -\alpha_2 & \cdots & \alpha_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & -\alpha_k \end{pmatrix}.$$

Remarque 3.2.2. *Lorsque la matrice $\boldsymbol{Q}$ est diagonalisable, c'est-à-dire $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$, où $\boldsymbol{D}$ est la matrice diagonale des valeurs propres de la matrice des intensités de transition et $\boldsymbol{U}$ est la matrice des vecteurs propres correspondants. Alors*

$$\exp(t\boldsymbol{Q}) = \boldsymbol{U}\exp(t\boldsymbol{D})\boldsymbol{U}^{-1}.$$

Exemple 3.2.1. *Résolution des équations de Chapman-Kolmogorov pour un processus de Markov homogène à deux états.*

Le processus $\{X(t), t \geq 0\}$ *associé est un processus dont les marques sont état* 0 *(vivant) et état* 1 *(décès). La matrice des intensités de transition est donnée par*

$$\boldsymbol{Q} = \begin{pmatrix} -\alpha_{01} & \alpha_{01} \\ 0 & 0 \end{pmatrix},$$

avec $\alpha_{01} > 0$. *On peut résoudre les équations de Kolmogorov en diagonalisant* $\boldsymbol{Q}$. *Le polynôme caractéristique de* $\boldsymbol{Q}$ *est donné par* $P_{\boldsymbol{Q}}(\lambda) = det(\boldsymbol{Q} - \lambda \mathbf{I})$

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

La matrice $\boldsymbol{Q}$ *admet deux valeurs propres distinctes* 0 *et* $-\alpha_{01}$, *donc elle est diagonalisable c'est-à-dire* $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1}$ *avec*

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

On utilise $\boldsymbol{P}(t) = \exp(\boldsymbol{Q}t) = \boldsymbol{U}\exp(t\boldsymbol{D})\boldsymbol{U}^{-1}$, *on trouve*

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

En effet, $p_{00}(t) = e^{-\alpha_{01}t}$ *et* $p_{01}(t) = 1 - e^{-\alpha_{01}t}$ *correspondent à la fonction de survie et la fonction de répartition d'une loi exponentielle présentée dans la section 2.4.1.1.*

Remarque 3.2.3. *En général, la variable aléatoire* T_i *(le temps passé dans l'état i avant de le quitter) suit une loi exponentielle de paramètre* $\sum_{i \neq j} \alpha_{ij}$. *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. 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 $E(T_i) = \frac{1}{\sum_{i \neq j} \alpha_{ij}}$.

3.3 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.

3.3.1 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*, comme l'ensemble de données présenté dans la section 1.2.2.

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 $\{X_h(t), t \geq 0\}$, $h = 1, \ldots, n$, des réalisations indépendantes d'un processus markovien homogène, que l'on observe aux temps $t_0, t_1, \ldots, t_L$ ($L > 1$). 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

$$\begin{aligned} \mathscr{L} &= \prod_{\ell=1}^{L} \left\{ \prod_{i,j=1}^{k} \left[p_{ij}(t_{\ell-1}, t_\ell) \right]^{n_{ij}(\ell)} \right\} \\ &= \prod_{\ell=1}^{L} \left\{ \prod_{i,j=1}^{k} p_{ij}(t_\ell - t_{\ell-1})^{n_{ij}(\ell)} \right\}. \end{aligned} \tag{3.12}$$

Dans l'exemple suivant, nous montrons que dans le cas d'un modèle de survie, la vraisemblance (3.12) est équivalente à la vraisemblance (2.4) donnée dans la section 2.3.2.

Exemple 3.3.1. *Supposons qu'on observe aux temps $t_0, t_1, \ldots, t_L$ la survenue de l'événement d'intérêt pour un sujet h, $h = 1, \ldots, n$, 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*

$$L_h = p_{00}(t_0,t_1)p_{00}(t_1,t_2)p_{01}(t_2,t_3)p_{11}(t_3,t_4)\cdots p_{11}(t_{L-1},t_L), \qquad (3.13)$$

or, $p_{11}(t_{\ell-1},t_\ell) = 1$, $\ell \geq 4$ puisque l'état 1 est un état absorbant, donc l'équation (3.13) *devient*

$$\begin{aligned} L_h &= \exp(-\alpha_{01}(t_1-t_0))\exp(-\alpha_{01}(t_2-t_1))(1-\exp(-\alpha_{01}(t_3-t_2)))(1)\cdots(1) \\ &= \exp(-\alpha_{01}(t_2-t_0))(1-\exp(-\alpha_{01}(t_3-t_2))) \\ &= \exp(-\alpha_{01}(t_2-t_0)) - \exp(-\alpha_{01}(t_3-t_0)). \qquad (3.14) \end{aligned}$$

Comme nous l'avons montré dans l'exemple 3.2.1, $S(t) = \exp(-\alpha_{01}t)$, l'équation (3.14) *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

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

où le sujet h a transité de l'état 0 vers l'état 1 dans l'intervalle $[\ell_h; r_h]$. Ainsi, nous remarquons que la vraisemblance (3.15) *est équivalente à la vraisemblance donnée dans la section 2.3.2.*

3.3.1.1 La procédure d'estimation de Kalbfleish et Lawess

Dans cette section, nous présentons l'algorithme proposé par Kalbfleisch 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 (3.12), récrite sous la forme

$$\mathscr{L}(\mathbf{Q}) = \prod_{\ell=1}^{L}\left\{\prod_{i,j=1}^{k}\left[\exp(\mathbf{Q}(t_\ell - t_{\ell-1}))\right]_{i,j}^{n_{ij}(\ell)}\right\}, \tag{3.15}$$

par un algorithme du score appelé aussi algorithme de quasi-Newton pour estimer les intensités de transition. Dans l'équation (3.15), l'expression $(\exp(\mathbf{Q}(t_\ell - t_{\ell-1})))_{i,j}$ représente l'élément (i,j) de la matrice $\exp(\mathbf{Q}(t_\ell - t_{\ell-1})) = \mathbf{P}(t_\ell - t_{\ell-1})$. Sous l'hypothèse que $\mathbf{Q}$ admet k valeurs propres distinctes $(\lambda_1,\ldots,\lambda_k)$, les matrices $\mathbf{Q}$ et $\mathbf{P}(t)$ se diagonalisent de la manière suivante :

$$\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^{-1} = \mathbf{U}\,\mathrm{diag}(\lambda_1,\ldots,\lambda_k)\,\mathbf{U}^{-1}, \tag{3.16}$$

et

$$\begin{aligned}\mathbf{P}(t) = \exp(\mathbf{Q}t) &= \mathbf{U}\exp(\mathbf{D}t)\mathbf{U}^{-1}\\ &= \mathbf{U}\,\mathrm{diag}(e^{\lambda_1 t},\ldots,e^{\lambda_k t})\,\mathbf{U}^{-1}.\end{aligned}$$

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

$$S(\mathbf{Q}) = \left\{\frac{\partial \log \mathscr{L}(\mathbf{Q})}{\partial \alpha_{uv}}\right\} = \left\{\sum_{\ell=1}^{L}\sum_{i,j=1}^{k} n_{ij}(\ell)\,\frac{\partial \exp(\mathbf{Q}(t_\ell - t_{\ell-1}))_{i,j}/\partial \alpha_{uv}}{\exp(\mathbf{Q}(t_\ell - t_{\ell-1}))_{i,j}}\right\} \tag{3.17}$$

où

$$\begin{aligned}\frac{\partial \mathbf{P}(t)}{\partial \alpha_{uv}} &= \frac{\partial(\exp(\mathbf{Q}t))}{\partial \alpha_{uv}} \qquad (3.18)\\ &= \sum_{s=1}^{\infty}\frac{\partial}{\partial \alpha_{uv}}\left(\frac{\mathbf{Q}^s t^s}{s!}\right)\\ &= \sum_{s=1}^{\infty}\left(\frac{\partial \mathbf{Q}^s}{\partial \alpha_{uv}}\right)\frac{t^s}{s!}\\ &= \sum_{s=1}^{\infty}\sum_{r=0}^{s-1}\mathbf{Q}^r\left(\frac{\partial \mathbf{Q}}{\partial \alpha_{uv}}\right)\mathbf{Q}^{s-1-r}\,\frac{t^s}{s!}\\ &= \sum_{s=1}^{\infty}\sum_{r=0}^{s-1}\mathbf{U}\mathbf{D}^r\mathbf{U}^{-1}\left(\frac{\partial \mathbf{Q}}{\partial \alpha_{uv}}\right)\mathbf{U}\mathbf{D}^{s-1-r}\mathbf{U}^{-1}\,\frac{t^s}{s!}. \qquad (3.19)\end{aligned}$$

Posons $\mathbf{G}_{uv} = \mathbf{U}^{-1}\left(\frac{\partial \mathbf{Q}}{\partial \alpha_{uv}}\right)\mathbf{U}$, l'équation (3.19) devient

$$\begin{aligned}\frac{\partial \mathbf{P}(t)}{\alpha_{uv}} = \frac{\partial(\exp(\mathbf{Q}t))}{\partial \alpha_{uv}} &= \sum_{s=1}^{\infty}\sum_{r=0}^{s-1}\mathbf{U}\mathbf{D}^r\mathbf{G}_{uv}\mathbf{D}^{s-1-r}\mathbf{U}^{-1}\frac{t^s}{s!} \\ &= \mathbf{U}\left(\sum_{s=1}^{\infty}\sum_{r=0}^{s-1}\mathbf{D}^r\mathbf{G}_{uv}\mathbf{D}^{s-1-r}\frac{t^s}{s!}\right)\mathbf{U}^{-1} \\ &= \mathbf{U}\mathbf{V}_{uv}\mathbf{U}^{-1}.\end{aligned}$$

La matrice

$$\mathbf{V}_{uv} = \sum_{s=1}^{\infty}\sum_{r=0}^{s-1}\mathbf{D}^r\mathbf{G}_{uv}\mathbf{D}^{s-1-r}\frac{t^s}{s!}$$

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

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

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

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

qui est estimée par

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

où $N_i(\ell-1)$ 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

$$\mathbf{Q}_{n+1} = \mathbf{Q}_n + M(\mathbf{Q}_n)^{-1}S(\mathbf{Q}_n),\ n \geq 0$$

avec $S(\mathbf{Q}_n) = \left(\frac{\partial \log \mathscr{L}(\mathbf{Q})}{\partial \alpha_{11}}, \ldots, \frac{\partial \log \mathscr{L}(\mathbf{Q})}{\partial \alpha_{kk}}\right)'$ et $M(\mathbf{Q}_n)$ est une matrice $k \times k$ d'éléments $\frac{\partial^2 Log\mathscr{L}(\mathbf{Q})}{\partial \alpha_{uv}\partial \alpha_{u'v'}}$.

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. Dans l'exemple suivant, nous allons utiliser la bibliothèque `msm` (Jackson, 2011) du logiciel R pour appliquer cette approche en utilisant l'ensemble de données présenté dans la section 1.2.2.

Exemple 3.3.2. *En pratique, le package* `msm` *(Jackson, 2011) du logiciel* R *peut être utilisé pour estimer les intensités de transition ainsi que les probabilités de transition pour un modèle de Markov homogène, lorsque les données sont censurées par intervalle. En effet, nous avons utilisé l'ensemble de données présenté dans la section 1.2.2. Soit* $X_h(t)$ *l'état occupé par un sujet h au temps t, les sujets se déplaçant parmi les états* 1 *à* 4 *selon un processus de Markov homogène à temps continu* $\{X(t), t \geq 0\}$ *qui a* $S = \{1,2,3,4\}$ *comme ensemble d'états possibles. Le modèle est défini par la figure 1.5 dans la section 1.2.2. L'objectif est d'estimer la matrice* $\boldsymbol{Q}$ *des intensités de transition donnée par*

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

Nous considèrons un modèle sans covariables. Le code R est fourni en annexe dans la section B.2. Les résultats de l'estimation du maximum de vraisemblance de la matrice des intensités de transition et les intervalles de confiance à 95% *sont présentés dans le tableau 3.1.*

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

Tableau 3.1 Estimation des intensités de transition.

Intensité	Estimés	IC à 95%
α_{11}	−0.17037	[−0.19027; −0.15255]
α_{12}	0.12787	[0.11135; 0.14684]
α_{14}	0.04250	[0.03412; 0.05294]
α_{21}	0.22512	[0.16755; 0.30247]
α_{22}	−0.60794	[−0.70880; −0.52143]
α_{23}	0.34261	[0.27317; 0.42970]
α_{24}	0.04021	[0.01129; 0.14324]
α_{32}	0.13062	[0.07952; 0.21457]
α_{33}	−0.43710	[−0.55292; −0.34554]
α_{34}	0.30648	[0.23822; 0.39429]

progression vers les symptômes graves (état 3) est plus rapide que le rétablissement (état 1) ou le décès (état 4).

La matrice des estimés des probabilités de transition lorsque $t = 10$ *ans est présentée dans le tableau 3.2. En effet, nous remarquons 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.*

3.3.2 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'intant $t_0 = 0$ et notons t_m les temps exacts de transition dans les états s_m, $m = 1, \ldots, E$. La contribution à la vraisemblance

Tableau 3.2 Probabilités de transition estimées pour un laps de temps de $t = 10$.

	État 1	État 2	État 3	État 4
État 1	0.30940656	0.09750021	0.08787255	0.5052207
État 2	0.17165172	0.06552639	0.07794394	0.6848780
État 3	0.05898093	0.02971653	0.04665485	0.8646477
État 4	0.00000000	0.00000000	0.00000000	1.0000000

pour cette trajectoire, conditionnellement à $X(t_0)$, est

$$\left(\prod_{m=1}^{E} \wp_{s_{m-1}}(t_{m-1}, t_m) \alpha_{s_{m-1}, s_m}(t_m) \right) \wp_{s_E}(t_E, C). \tag{3.20}$$

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

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

si bien que la contribution à la vraisemblance (3.20) s'écrit

$$\left(\prod_{m=1}^{E} \exp\left\{ - \int_{t_{m-1}}^{t_m} \alpha_{s_{m-1}}(u) du \right\} \alpha_{s_{m-1}, s_m}(t_m) \right) \exp\left\{ - \int_{t_E}^{C} \alpha_{s_E}(u) du \right\}. \tag{3.21}$$

En regroupant les exponentielles, (3.21) s'écrit comme un produit de deux termes :

$$\left(\prod_{m=1}^{E} \alpha_{s_{m-1}, s_m}(t_m) \right) \exp\left\{ - \left(\int_0^{t_1} \alpha_{s_0}(u) du + \cdots + \int_{t_E}^{C} \alpha_{s_E}(u) du \right) \right\} \tag{3.22}$$

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

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

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

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

(valant 1 quand le processus est dans l'état s_{m-1} dans l'intervalle $[t_{m-1}, t_m[$). Alors (3.22) se récrit :

$$\left(\prod_{m=1}^{K}\prod_{i\neq j}\alpha_{i,j}(t_m)^{D_{ijm}}\right)\exp\left\{-\left(\sum_i\int_0^C R_i(u)\alpha_i(u)du\right)\right\}$$
$$= \left(\prod_{i\neq j}\left\{\prod_{m=1}^{K}\alpha_{i,j}(t_m)^{D_{ijm}}\right\}\right)\left(\prod_i\exp\left\{-\int_0^C R_i(u)\alpha_i(u)du\right\}\right)$$

Comme $\alpha_i(u) = \sum_{i\neq j}\alpha_{ij}(u)$, on peut remplacer dans le membre de droite et (3.21) devient simplement

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

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

$$\mathscr{L} = \prod_{h=1}^{n}\prod_{i\neq j}\left(\left\{\prod_{m_h=1}^{K_{m_h}}\alpha_{ij}(t_{m_h})^{D_{hijm_h}}\right\}\exp\left\{-\int_0^{C_h}R_{hi}(u)\alpha_{ij}(u)du\right\}\right). \qquad (3.23)$$

Dans la cas d'un modèle de survie (Figure 2.1) avec $i = 0$ et $j = 1$, soit t_h, $h = 1,\ldots,n$, les temps exacts de transition de l'état 0 vers l'état 1 des n sujets, la vraisemblance (3.23) s'écrit

$$\mathscr{L} = \prod_{h=1}^{n}\left(\left\{\alpha_{01}(t_h)^{D_{h01}}\right\}\exp\left\{-\int_0^{C_h}R_{h0}(u)\alpha_{01}(u)du\right\}\right) \qquad (3.24)$$

où $\alpha_{01}(t_h)^{D_{h01}}$ est la contribution à la vraisemblance pour les temps exacts de transition de l'état 0 vers l'état 1 et $\int_0^{C_h}R_{h0}(u)\alpha_{01}(u)du$ est la contribution à la vraisemblance pour les temps censurés à droite. En effet, nous remarquons que la vraisemblance (3.24) est équivalente à (2.2), celle que nous avons présentée dans la section 2.3.1.

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

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

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

Exemple 3.3.3. *Considérons le modèle de risques concurrents représenté dans la figure 1.4 de la section 1.2.1. 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* $\{X_h(t), 0 \leq t \leq C_h\}$, $h = 1, \ldots, n$ *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* (3.23) *est de la forme*

$$\mathscr{L} = \prod_{h=1}^{n} \prod_{j=1,2} \left[\{\alpha_{0j}(t_h)^{D_{h0j}}\} \exp\left\{ -\int_0^{C_h} R_{h0}(u)\alpha_{0j}(u)du \right\} \right]. \tag{3.25}$$

En développant l'équation (3.25), *on obtient*

$$\begin{aligned} \mathscr{L} &= \prod_{h=1}^{n} \left[\{\alpha_{01}(t_h)^{D_{h01}} \alpha_{02}(t_h)^{D_{h02}}\} \right. \\ &\qquad \left. \times \exp\left\{ -\int_0^{C_h} R_{h0}(u)\alpha_{01}(u)du \right\} \exp\left\{ -\int_0^{C_h} R_{h0}(u)\alpha_{02}(u)du \right\} \right] \\ \log(\mathscr{L}) &= \sum_{h=1}^{n} \left(D_{h01}\log(\alpha_{01}(t_h)) + D_{h02}\log(\alpha_{02}(t_h)) \right. \\ &\qquad \left. - \int_0^{C_h} R_{h0}(u)\alpha_{01}(u)du - \int_0^{C_h} R_{h0}(u)\alpha_{02}(u)du \right). \end{aligned} \tag{3.26}$$

Lorsque les intensités du modèle sont constantes l'équation (3.26) *devient*

$$\begin{aligned} \log(\mathscr{L}) &= \sum_{h=1}^{n} \left(D_{h01}\log(\alpha_{01}) + D_{h02}\log(\alpha_{02}) \right. \\ &\qquad \left. - \alpha_{01}\int_0^{C_h} R_{h0}(u)du - \alpha_{02}\int_0^{C_h} R_{h0}(u)du \right). \end{aligned} \tag{3.27}$$

En dérivant l'équation (3.27) *par rapport* α_{0j}, $j = 1, 2$, *nous obtenons*

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

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

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

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. Pour illustrer cette approche, nous avons utilisé l'ensemble de données présenté dans la section 1.2.1. Dans cet ensemble de données, tous les sujets étaient observés sur la même période $[0,C]$, donc dans ce cas $C_h = C$, d'où

$$\begin{aligned}
\hat{\alpha}_{01} &= \frac{\sum_{h=1}^{n} D_{h01}}{\sum_{h=1}^{n} \int_0^{C} R_{h0}(t)dt} \\
&= \frac{D_{101} + D_{201} + \cdots + D_{n01}}{\int_0^C R_{10}(t)dt + \int_0^C R_{20}(t)dt + \cdots + \int_0^C R_{n0}(t)dt} \\
&= \frac{657}{3017} \\
&= 0.2177.
\end{aligned}$$

De la même manière, nous trouvons $\hat{\alpha}_{02} = 0.0251$.

3.4 Prise en compte de covariables

Dans la plus part des applications, on dispose de covariables sur chaque sujet. Il est intéressant d'étudier l'impact des ces covariables sur les paramètres du modèle. Le modèle peut être généralisé de manière simple en considérant un modèle de régression pour lequel les intensités sont proportionnelles (Cox, 1972). Les intensités de transition peuvent s'écrire

$$\alpha_{ij}(\vec{Z}) = \alpha_{ij0} \exp(\vec{\beta}'\vec{Z}), \quad i \neq j$$

avec $\vec{Z}$ un vecteur de covariables indépendantes du temps de dimension m, $\vec{\beta}$ un vecteur de m coefficients de régression et α_{ij0} est l'intensité de transition de base associée à la transition de l'état i vers l'état j. Ce modèle est souvent utilisé dans la littérature

(Andersen *et al.*, 1991; Marshall et Jones, 1995). En effet, les estimateurs des intensités de transition sont toujours positifs quelles que soient les valeurs de $\vec{Z}$ et $\vec{\beta}$. De plus, ce modèle fournit des résultats en terme de risques relatifs qui sont facilement interprétables, comme dans le modèle des risques proportionnels de Cox présenté dans la section 2.4. Le modèle peut être adapté afin de prendre en compte des covariables dépendantes du temps. En effet, il est nécessaire de supposer que les valeurs des covariables dépendantes du temps restent constantes entre les deux temps d'observations consécutifs. Cette hypothèse peut être forte dans certaines situations, notamment, si le temps entre deux observations est grand.

CHAPITRE IV

MODÈLES DE MARKOV NON HOMOGÈNES

Les modèles de Markov homogènes sont de plus en plus appliqués en épidémiologie. Cependant, dans de nombreuses applications, l'hypothèse d'homogénéité (intensités de transition constantes au cours du temps) n'est pas vérifiée et s'avère trop restrictive. Les modèles de Markov non homogènes sont une solution de remplacement permettant de considérer des intensités de transition qui dépendent du temps. Grâce au travail fondamental d'Aalen (1975, 1978), la théorie des processus de comptage a été largement étudiée et a permis le développement de plusieurs estimateurs (Andersen *et al.*, 1993). Elle permet aussi d'obtenir des estimateurs non paramétriques dans les modèles de Markov non homogènes.

Dans ce chapitre, nous abordons l'estimateur de Aalen-Johanson (Aalen et Johansen, 1978) qui donne une estimation de la matrice des probabilités de transition dans un modèle de Markov non homogène. Cet estimateur peut également être adapté à un modèle de régression semi-paramétrique afin d'étudier les effets des covariables (Andersen *et al.*, 1991). Les estimateurs ainsi obtenus par la théorie des processus de comptage constituent une généralisation, aux modèles markoviens, de l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958). La première partie de ce chapitre présente les résultats essentiels liés aux processus de comptage et leur relation avec le processus de Markov. La prise en compte de la censure est également considérée. La deuxième partie présente

les estimateurs non paramétriques de Nelson-Aalen et de Aalen-Johansen. Dans la dernière partie de ce chapitre, nous présentons l'application de ces méthodes à l'ensemble de données `sir.adm`.

4.1 Processus de Markov et processus de comptage

La notion de processus de comptage constitue la base de la théorie des processus stochastiques. À partir de ces processus ponctuels, différentes classes de processus ayant des trajectoires à sauts peuvent être envisagées, par exemple, les processus de Markov.

4.1.1 Processus de Markov non homogène

Dans le chapitre précédent, nous avons introduit la notion du processus de Markov à temps continu et à espace fini d'états, particulièrement le processus de Markov homogène. Nous nous sommes intéressés au paramètre qui permet de définir un processus de Markov homogène, spécifiquement à l'intensité de transition.

Maintenant, considérons $(X(t), t \geq 0)$ un processus de Markov à temps continu, ayant un espace fini d'états $S = \{1, \ldots, k\}$. La mesure d'intensité cumulée est un autre paramètre qui permet de définir un processus de Markov. C'est une matrice $k \times k$ de fonctions, notée $\mathbf{A} = (A_{ij})$, telle que

$$A_{ij}(t) = \int_0^t \alpha_{ij}(u)du, \ \forall \ i \neq j, \tag{4.1}$$

$$A_{ii}(t) = -\sum_{i \neq j} A_{ij}(t), \ \forall \ t \geq 0 \tag{4.2}$$

et telle que, pour $i \neq j$, $A_{ij}(.)$ soit une fonction cadlag (continue à droite avec une limite à gauche) croissante, nulle en zéro. La fonction $A_{ij}(.)$ est l'intensité cumulée pour les transitions de l'état i vers l'état j, alors que $A_{ii}(.)$ est (l'oppposé de) la fonction d'intensité cumulée pour les transitions qui quittent l'état i. Les fonctions $\alpha_{ij}(.)$ sont les intensités de transition définies par l'équation (3.7) dans la section 3.1. En effet, il est facile de démontrer l'équation (4.2) en utilisant les équations (3.6) et (3.7). Dans

la section 3.2, nous avons présenté les équations différentielles de Kolmogorov pour un processus de Markov homogène. Dans la cas d'un processus de Markov non homogène, c'est-à-dire quand les intensités de transition dépendent du temps, les équations différentielles de Kolmogorov sont définies par

1. équation de Kolmogorov progressive :

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\mathbf{Q}(t) \tag{4.3}$$

2. équation de Kolmogorov rétrograde :

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = \mathbf{Q}(s)\mathbf{P}(s,t) \tag{4.4}$$

avec $\mathbf{P}(s,t)$ est la matrice des probabilités de transition et $\mathbf{Q}(t)$ est la matrice des intensités de transition, c'est-à-dire la dérivée de $\mathbf{A}(t)$.

La notion de produit intégral est introduite pour écrire la vraisemblance du processus et pour obtenir une relation entre la matrice des intensités de transition et la matrice des probabilités de transition. Dans ce qui suit, nous utilisons la notation $\mathscr{P}_{]s,t]}$ pour désigner un produit intégral (la définition est présentée dans l'appendice A).

Théorème 4.1.1. *Soit* $\mathbf{A}$ *une matrice de fonctions, de dimension* $k \times k$*, donnant des intensités cumulées. Soit* $\mathbf{I}$ *la matrice indentité. Alors la matrice*

$$\mathbf{P}(s,t) = \mathscr{P}_{]s,t]}\left(\mathbf{I} + d\mathbf{A}(u)\right), \ 0 \leq s \leq t, \tag{4.5}$$

est la matrice des probabilités de transition d'un processus de Markov à espace d'états fini $S = \{1,\ldots,k\}$*. Lorsque* $\mathbf{A}$ *est une fonction étagée et cadlag, le produit intégral* (4.5) *devient un produit fini sur les temps de sauts de* $\mathbf{A}$*, ainsi*

$$\mathbf{P}(s,t) = \prod_{\ell=1}^{K} \left(\mathbf{I} + \Delta\mathbf{A}(t_\ell)\right), \ 0 \leq s \leq t,$$

où $s < t_1 < \cdots < t_K < t$ *sont les temps de sauts et* $\Delta\mathbf{A}(t_\ell) = \mathbf{A}(t_\ell) - \mathbf{A}(t_{\ell-1})$.

4.1.2 Processus de comptage

Pour simplifier la présentation, nous donnons d'abord, les résultats sans censure et dans la section 4.1.4 nous ajouterons la censure à droite. Dans tout ce qui suit, nous étudierons des processus indexés par $[0,\tau]$, ou τ est une constante fixée, et on supposera l'existence d'une filtration $\left(\mathscr{F}_t, t \in [0,\tau]\right)$. Dans l'appendice A, nous présentons également des définitions et propriétés liées aux processus de comptage.

Définition 4.1.1. *Soit* $(\Omega, \mathscr{A}, \mathbb{P})$ *un espace de probabilité où* $\mathbb{P}$ *est la mesure de probabilité sur* $(\Omega, \mathscr{A})$. *Un processus aléatoire de comptage* $N(.)$ *sur* $[0,\tau]$ *est une fonction aléatoire, à valeurs dans les entiers naturels,*

$$\begin{aligned} N : [0,\tau] \times \Omega &\rightarrow \mathbb{N} \\ (t,\omega) &\mapsto N(t,\omega) \end{aligned}$$

qui est càdlàg, nulle en zéro, croissante et ayant des sauts d'amplitude 1.

Le concept de filtration permet de définir l'ensemble des événements observés à l'instant t, c'est-à-dire toute l'information disponible à l'instant t. La filtration naturelle $\left(\mathscr{F}_t^N, t \in [0,\tau]\right)$ associée au processus $N(.)$ est définie par la σ-algèbre

$$\mathscr{F}_t^N = \sigma\left(N(u), 0 \leq u \leq t\right).$$

Définition 4.1.2. *Un processus de comptage* $\mathbf{N}(.) = (N_{ij}(.), i \neq j, i, j \in S)$ *est appelé processus de comptage* multivarié *si chacune de ses composantes est un processus de comptage univarié.*

Théorème 4.1.2. *Soit* $\mathbf{A}$ *la matrice des intensités de transition cumulées d'un processus de Markov* $\left(X(t), t \in [0,\tau]\right)$. *Définissons*

$$Y_i(t) = \mathbf{1}_{\{X(t^-)=i\}} \tag{4.6}$$

$$N_{ij}(t) = \#\{s \leq t : X(s^-) = i, X(s) = j\}, \ \forall \ i \neq j. \tag{4.7}$$

Alors $\mathbf{N}(.) = (N_{ij}(.), i \neq j, i, j \in S)$ *est un processus de comptage multivarié. Le processus d'intensité cumulée (ou le compensateur de* $\mathbf{N}(.)$*) par rapport à* $\left(\mathscr{F}_t^X, t \in [0,\tau]\right)$ *est* $\Lambda(.) = (\Lambda_{ij}(.), i \neq j)$ *donné par*

$$\Lambda_{ij}(t) = \int_0^t Y_i(u) dA_{ij}(u). \tag{4.8}$$

De plus, les processus $M_{ij}(.)$ *définies par*

$$M_{ij} = N_{ij} - \Lambda_{ij}$$

sont des martingales.

Le processus de comptage $N_{ij}(t)$ compte le nombre de transitions obervées du processus $X(t)$ de l'état i vers l'état j dans l'intervalle $[0,t]$. La quantité $Y_i(t)$ nous renseigne sur l'état du processus $X(t)$: si le processus $X(t)$ est dans l'état i juste avant l'instant t alors $Y_i(t) = 1$; $Y_i(t) = 0$ sinon. De plus, d'après l'équation (4.2), le compensateur $\Lambda(.) = (\Lambda_{ij}(.), i \neq j)$ de $\mathbf{N}(.)$ vérifie

$$\Lambda_{ij}(t) = \int_0^t Y_i(u) \alpha_{ij}(u) du.$$

Dans ce cas, le théorème précédent implique que le processus $\mathbf{N}(.)$ a un processus d'intensité multiplicative $\lambda(.) = (\lambda_{ij}(.), i \neq j)$ par rapport à $\left(\mathscr{F}_t^X, t \in [0,\tau]\right)$ qui vérifie

$$\lambda_{ij}(t) = Y_i(t) \alpha_{ij}(t), \quad i \neq j.$$

Notons que $\lambda_{ij}(.)$ sont des variables aléatoires puisque $Y_i(.)$ est aléatoire et $\alpha_{ij}(.)$ est déterministe.

Remarque 4.1.1. *La connaissance du processus de comptage* $\mathbf{N}(.) = (N_{ij}(.), i \neq j)$ *sur* $[0,t]$ *apporte la même information que l'obervation du processus* $\left(X(u),\ 0 \leq u \leq t\right)$.

4.1.3 Vraisemblance

Dans cette section, nous présentons la vraisemblance d'un sujet. On considère le processus de comptage $\mathbf{N}(t) = (N_{ij}(t), i \neq j, i, j \in S)$ par rapport à la filtration naturelle

$\mathscr{F}_t = \sigma(\mathbf{N}(u), 0 \leq u \leq t)$ et $\Lambda(t) = (\Lambda_{ij}(t), i \neq j)$ son compensateur par rapport à $\mathscr{F}_t$. La vraisemblance basée sur l'observation de $\mathbf{N}(t)$ conditionnellement à l'état initial (Andersen *et al.*, 1993) est donnée par

$$\mathscr{L} = \mathscr{P}_{t\in[0,\tau]} \left\{ \left(\prod_{i\neq j} (d\Lambda_{ij}(t))^{\Delta N_{ij}(t)} \right) \left[1 - \sum_{i\neq j} \Lambda_{ij}(t) \right]^{1-\sum_{i\neq j} \Delta N_{ij}(t)} \right\}, \tag{4.9}$$

où $\Delta N_{ij}(t) = N_{ij}(t) - N_{ij}(t^-)$. D'après les équations (4.2) et (4.8), la vraisemblance s'écrit

$$\mathscr{L} = \mathscr{P}_{t\in[0,\tau]} \left\{ \left(\prod_{i\neq j} (\alpha_{ij}(t) Y_i(t))^{\Delta N_{ij}(t)} \right) \left[1 - \sum_{i\neq j} \alpha_{ij}(t) Y_i(t) \right]^{1-\sum_{i\neq j} \Delta N_{ij}(t)} \right\}. \tag{4.10}$$

4.1.4 Processus de comptage et censure à droite

Le phénomène de censure à droite peut arrêter l'observation du processus de Markov $\{X(t), t \geq 0\}$ et donc aussi du processus de comptage $\mathbf{N}(.)$.

Pour chaque sujet, on considère $\mathscr{C}(t)$ le processus de censure à droite, définie par

$$\mathscr{C}(t) = \mathbf{1}_{\{t \leq C\}}$$

où C est la variable aléatoire de censure à droite.

Suite à la censure à droite, le processus réellement observé est $\mathbf{N}^c(.)$ le processus de comptage censuré à droite qui représente la partie observable de $\mathbf{N}(.)$:

$$\mathbf{N}^c(t) = \left(N_{ij}^c(t), i \neq j, i, j \in S \right)$$

avec

$$\mathbf{N}_{ij}^c(t) = \int_0^t \mathscr{C}(u) dN_{ij}(u).$$

4.1.5 Censure à droite indépendante

Un problème fréquent avec la censure à droite est qu'elle peut modifier les intensités des événements d'intérêt. Par exemple, dans un essai clinique, si les sujets particulièment malades sont enlevés de l'étude, les sujets qui restent, qui sont encore à risque, ne sont plus représentatifs de l'échantillon total. La notion de censure à droite indépendante permet d'éviter ce problème (définition 2.2.2).

Sous l'hypothèse de censure à droite indépendante, la mesure d'intensité (le compensateur) $\Lambda_{ij}^c(.)$ est

$$\Lambda_{ij}^c(t) = \int_0^t \mathscr{C}(u) d\Lambda_{ij}(u) \tag{4.11}$$

et le processus censuré à droite $Y_i^c(.)$ (qui représente la partie observable du processus $Y_i(.)$) est tel que

$$Y_i^c(t) = \mathscr{C}(t) Y_i(t), \quad i = 1, \ldots, k. \tag{4.12}$$

D'après (4.8) et (4.12), l'équation (4.11) peut s'écrire

$$\Lambda_{ij}^c(t) = \int_0^t \mathscr{C}(u) Y_i(u) \alpha_{ij}(u) du = \int_0^t \alpha_{ij}(u) Y_i^c(u) du.$$

Proposition 4.1.1. *Soit* $\mathbf{N}^c(t) = \left(N_{ij}^c(t), i \neq j, i, j \in S\right)$ *un processus de comptage censuré. Si la censure à droite est indépendante alors la vraisemblance s'écrit*

$$\mathscr{L} = \mathscr{P}_{t\in[0,\tau]} \left(\prod_{i\neq j} (d\Lambda_{ij}^c(t))^{\Delta N_{ij}^c(t)} \left[1 - \sum_{i\neq j} d\Lambda_{ij}^c(t) \right]^{1-\sum_{i\neq j} \Delta N_{ij}^c(t)} \right). \tag{4.13}$$

4.2 Estimation non paramétrique

Dans cette section, nous nous plaçons dans le cadre d'un mécanisme de censure indépendante. Les processus considérés sont des processus censurés. Nous pésentons l'estimateur de Nelson-Aalen des fonctions d'intensité de transition cumulée et l'estimateur de Aalen-Johansen de la matrice des probabilités de transition.

4.2.1 Observations et notations

Considérons un échantillon $X_1(.),\ldots,X_n(.)$ des processus de Markov non homogènes indépendants, de mêmes lois et à espace fini d'états $S=\{1,\ldots,k\}$. Le processus $X_h(.)$ associé au sujet h représente l'état du sujet au temps t. Pour tout $h=1,\ldots,n$, nous définissons :

- le processus $N_{ijh}(t)$ qui compte le nombre de transitions de l'état i vers l'état j dans $[0,t]$ pour le sujet h :

$$N_{ijh}(t)=\#\{s\leq t : X_h(s^-)=i, X_h(s)=j\},\ \forall\ i\neq j\,;$$

- $Y_{ih}(.)$ un indicateur pour savoir si $X_h(.)$ est dans l'état i juste avant le temps t :

$$Y_{ih}(t)=\mathbf{1}_{\{X(t^-)=i\}},\ i=1,\ldots,k\,;$$

- le processus $N_{ij}(t)$ qui compte le nombre total de transitions de l'état i vers l'état j dans $[0,t]$ (dans tout l'échantillon) :

$$N_{ij}(t)=\sum_{h=1}^{n} N_{ijh}(t),\ \forall\ i\neq j\,;$$

- $Y_i(t)$ qui donne le nombre total de sujets à risque dans l'état i juste avant l'instant t, c'est-à-dire le nombre de sujets succeptibles de subir une transition à partir de l'état i :

$$Y_i(t)=\sum_{h=1}^{n} Y_{ih}(t),\ \forall\ i=1,\ldots,k.$$

La vraisemblance totale associée aux n processus de comptage $N_h(t)=(N_{ijh}(t), i,j\in S, i\neq j)$, $h=1,\ldots,n$, conditionnellement à l'état initial, est donnée par

$$\mathscr{L}=\mathscr{P}_{t\in[0,\tau]}\left(\prod_{h=1}^{n}\prod_{i\neq j}(\alpha_{ij}(t)Y_{ih}(t))^{\Delta N_{ijh}(t)}\left(1-\sum_{h=1}^{n}\sum_{i\neq j}\alpha_{ij}(t)Y_{ih}(t)\right)^{1-\sum_{h=1}^{n}\sum_{i\neq j}\Delta N_{ijh}(t)}\right) \quad (4.14)$$

où $\Delta N_{ij}(t)=N_{ij}(t)-N_{ij}(t^-)$.

4.2.2 Estimation des intensités de transition cumulées

Pour les démonstrations des propriétés de cette section, nous renvoyons le lecteur au livre de Andersen *et al.* (1993). Supposons que les intensités de transition vérifient $\int_0^t \alpha_{ij}(u)du < \infty$ pour tout $i \neq j$ et pour tout $t \geq 0$. Un estimateur des intensités de transition cumulées $A_{ij}(t) = \int_0^t \alpha_{ij}(u)du$ a été introduit par Aalen et Johansen (1978) pour généraliser celui de Nelson (1972) aux processus de comptage.

Définition 4.2.1. *L'estimateur de Nelson-Aalen des fonctions d'intensité de transition cumulée est défini par*

$$\hat{A}_{ij}(t) = \int_0^t \frac{J_i(u)}{Y_i(u)} dN_{ij}(u), \ \forall \ i \neq j \tag{4.15}$$

où $J_i(t) = \mathbf{1}_{\{Y_i(t)>0\}}$. *(On pose* $\frac{J_i(t)}{Y_i(t)} = 0$ *lorsque* $Y_i(t) = 0$.*)*

Remarque 4.2.1. *Nous pouvons écrire* $\hat{A}_{ij}(t)$ *comme une simple somme*

$$\hat{A}_{ij}(t) = \sum_{\{t_K \leq t\}} \frac{N_{ij}(t_K)}{Y_i(t_K)}$$

où les t_K *sont les temps de sauts,* $N_{ij}(t_K)$ *(équation (4.7)) le processus qui compte le nombre de transitions de l'état i vers l'état j au temps* t_K *et* $Y_i(t_K)$ *(équation (4.6)) le processus relatif au nombre de sujets à risque juste avant l'instant* t_K *pour la transition de l'état i vers l'état j.*

4.2.3 Estimation des probabilités de transition

Rappelons que la matrice des probabilités de transition est définie en fonction de la matrice des intensités de transition cumulées par le produit intégral (équation (4.5)) :

$$\mathbf{P}(s,t) = \mathscr{P}_{u\in]s,t]}(\mathbf{I} + d\mathbf{A}(u)), \ \ 0 < s \leq t, \ \ s,t \in [0,\tau]$$

où $\mathbf{P}(s,t) = (p_{ij}(s,t))_{i,j}$ et $\mathbf{A}(t) = (A_{ij}(t))_{i,j}$. L'estimateur introduit par Aalen et Johansen (1978) utilise cette relation et l'estimateur de Nelson-Aalen pour obtenir une estimation de la matrice des probabilités de transition.

Définition 4.2.2. *L'estimateur de Aalen-Johansen de la matrice des probabilités de transition est défini par*

$$\hat{\mathbf{P}}(s,t) = \mathscr{P}_{u\in]s,t]}(\mathbf{I}+d\hat{\mathbf{A}}(u)),\ \forall\ i\neq j, \tag{4.16}$$

où $\hat{\mathbf{P}}(s,t) = (\hat{p}_{ij}(s,t))_{i,j}$ *et* $\hat{\mathbf{A}}(t) = (\hat{A}_{ij}(t))_{i,j}$ *est l'estimateur de Nelson-Aalen.*

Remarque 4.2.2. *En pratique, il y a un nombre fini de transitions. Soient* $s < t_1 < t_2 < \cdots < t_K < t$*, les temps de transition entre deux états, l'estimateur de Aalen-Johansen devient un produit fini de matrice tel que*

$$\hat{\mathbf{P}}(s,t) = \prod_{l=1}^{K}(\mathbf{I}+\Delta\hat{\mathbf{A}}(t_l)). \tag{4.17}$$

Exemple 4.2.1. *Dans le cas d'un modèle de risques concurrents à 3 trois états (figure 0.3), supposons* $s = t_0 = 0$*, alors*

$$\mathbf{I}+\Delta\hat{\mathbf{A}}(t_\ell) = \begin{pmatrix} 1-\dfrac{\Delta N_{01}(t_\ell)+\Delta N_{02}(t_\ell)}{Y_0(t_\ell)} & \dfrac{\Delta N_{01}(t_\ell)}{Y_0(t_\ell)} & \dfrac{\Delta N_{02}(t_\ell)}{Y_0(t_\ell)} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

avec $\Delta\mathbf{A}(t_\ell) = \mathbf{A}(t_\ell) - \mathbf{A}(t_{\ell-1})$*, pour* $\ell = 2,\ldots,k$*, et* $\Delta\mathbf{A}(t_1) = \mathbf{A}(t_1)$*.*

4.2.4 Cas particulier : Modèle de survie

Le modèle de survie peut être considéré comme un modèle de Markov non homogène particulier $\{X(t), t \geq 0\}$ à temps continu et à espace d'états fini $S = \{0,1\}$ (figure 2.1). Le processus de comptage associé est un processus dont les marques sont : état 0 (vivant), état 1 (décès). La matrice des probabilités de transition et la matrice des intensités de transitions sont

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

Nous appliquons l'équation différentielle de Kolmogorov (4.3) et nous obtenons l'équation suivante

$$\frac{\partial p_{00}(s,t)}{\partial t} = p_{00}(s,t^-)\alpha_{00}(t). \tag{4.18}$$

En intégrant l'équation (4.18) par rapport à la seconde variable, nous obtenons alors la probabilité de rester dans l'état 0 entre l'instant s et t sous la forme

$$\begin{aligned} p_{00}(s,t) &= \int_s^t p_{00}(s,u)\alpha_{00}(u)du \\ &= \int_s^t p_{00}(s,u^-)dA_{00}(u). \end{aligned} \tag{4.19}$$

Puisque **A** est une matrice des intensités de transition cumulées d'un processus markovien, d'après l'équation (4.2), nous avons $dA_{00}(u) = -dA_{01}(u)$, l'équation (4.19) est de la forme

$$Z(s,t) = W(s,t) + \int_s^t Z(s,u^-)dB(u),$$

avec $Z(s,t) = p_{00}(s,t)$, $W(s,t) = 0$ et $B(u) = -A_{01}(u)$. Donc, d'après le théorème A.0.1 présenté dans l'appendice A), il existe une unique solution de la forme

$$p_{00}(s,t) = \mathscr{P}_{]s,t]}(1 - dA_{01}(u)). \tag{4.20}$$

En prenant $s = 0$, $p_{00}(0,t)$ est la probabilité d'être vivant au temps t sachant qu'on est vivant au temps 0, c'est-à-dire que $p_{00}(0,t)$ représente la fonction de survie $S(t)$ présentée dans la section 2.1. Ainsi, nous pouvons déduire, de (4.20) et (4.15), l'estimateur de Kaplan-Meier présenté dans la section 2.4.

En effet, supposons qu'on observe n sujets et considérons l'échantillon $X_1(.),\ldots,X_n(.)$ des processus de Markov non homogènes indépendants et identiquement distribués à espace d'états fini $S = \{0,1\}$. Le processus $X_h(.)$ associé au sujet h représente l'état du sujet au temps t. Pour $h = 1,\ldots,n$, nous définissons :

- $Y_{0h}(t) = \mathbf{1}_{\{X_h(t^-)=0\}}$ qui vaut 1 si le sujet h est dans l'état 0 (vivant) juste avant le temps t, 0 sinon ;

- $N_{01h}(t) = \mathbf{1}_{\{X_h(t)=1\}}$ qui vaut 1 si le sujet h a transité vers l'état 1 (décédé) au temps t, 0 sinon ;
- $Y_0(t) = \sum_{h=1}^{n} Y_{0h}(t)$, le nombre de sujets à risque juste avant le temps t ;
- $N_{01}(t) = \sum_{h=1}^{n} N_{01h}(t)$, le nombre total de sujets qui ont transité vers l'état 1 au temps t.

En effet, d'après l'équation (4.16), l'estimateur Aalen-Johansen de la probabilité $p_{00}(0,t)$ est donné par

$$\hat{p}_{00}(0,t) = \mathscr{P}_{]0,t]}(1 - d\hat{A}_{01}(u)) \tag{4.21}$$

où $\hat{A}_{01}(u)$ est l'estimateur de Nelson-Aalen (équation (4.15)) :

$$d\hat{A}_{01}(u) = \frac{J_0(u)}{Y_0(u)} dN_{01}(u).$$

En pratique, le nombre d'évéments est fini. Soient $s < t_1 < \cdots < t_K < t$, les temps de transition entre deux états, l'estimateur (4.21) devient

$$\hat{p}_{00}(0,t) = \prod_{t_\ell \leq t} \left(1 - \frac{J_0(t_\ell)}{Y_0(t_\ell)} \Delta N_{01}(t_\ell)\right).$$

Nous retrouvons donc l'estimateur de Kaplan-Meier présenté dans la section 2.4.2.1 du chapitre 2.

4.3 Analyse de données (`sir.adm`)

Dans cette section, nous effectuons l'estimation non paramétrique d'un modèle à deux risques concurrents en utilisant l'ensemble de données présenté dans la section 1.2.1. Dans cet ensemble de données, les sujets ont été observés de l'admission à l'unité de soins intensifs jusqu'à la fin de l'étude. Dans tout ce qui suit, nous faisons l'hypothèse que les patients sont observés de façon continue. Le temps de chaque patient est initialisé à $t = 0$ dès l'entrée dans l'unité de soins intensifs. Pour chaque sujet, les instants de changement d'état sont connus.

Nous considérons le modèle de risques concurrents de la figure 0.3. Notons $X_h(t)$ l'état du sujet h au temps t. Alors $\left(X_h(t), t \geq 0\right)$ est un processus de Markov non homogène à valeurs dans $S = \{0,1,2\}$, l'ensemble des états représentant respectivement l'admission en unité de soins intensifs, la sortie du service de réanimation et le décès. La sortie et le décès sont des états absorbants. Tous les sujets sont exposés à deux risques concurrents (la sortie et le décès).

Dans cette section, on se propose :

- d'estimer les intensités de transition cumulées et les probabilités de transition entre les états par l'approche non paramétrique ;
- d'utiliser des modèles semi-paramétriques similaires à ceux présentés dans la section 2.4.3 pour étudier l'effet de la pneumonie.

L'application a été réalisée en utilisant les bibliothèques `mstate` (de Wreede *et al.*, 2011) et `etm` (Allignol *et al.*, 2011) du logiciel R. Le code R est fourni dans l'appendice B.

4.3.1 Estimation non paramétrique des intensités de transition sans covariables

Dans cette section, nous effectuons l'estimation du modèle sans tenir compte des covariables. La figure 4.1 présente les estimateurs de Nelson-Aalen (équation (4.15)) des intensités de transition cumulées. L'intensité cumulée de l'état 0 vers l'état 1 est plus élevée que l'intensité cumulée de l'état 0 vers l'état 2. Ce résultat est en partie expliqué par le taux élevé de sujets sortant du service de réanimation 87.95% (tableau 1.2). Nous remarquons aussi que la pente est beaucoup plus élevée pour la courbe de l'intensité de transition de l'état 0 vers l'état 1 entre le début de l'étude et le 75^{e} jour. Ce phénomène est dû au fait qu'un grand nombre de transitions de ce type se produit pendant cette période.

Les estimateurs de Aalen-Johanson des probabilités de transition (équation (4.16)) sont

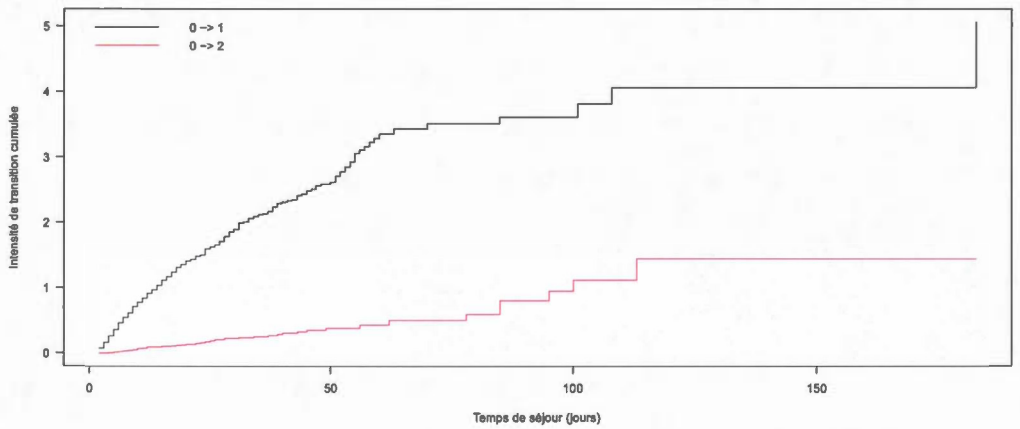

Figure 4.1 Estimation non paramétrique (Nelson-Aalen) des intensités cumulées.

représentés dans la figure 4.2. Comme pour les intensités de transition cumulées, les estimés des probabilités de transition montrent de manière graphique que la probabilité de transition de l'état 0 vers l'état 1 est plus élevée que la probabilité de transition de l'état 0 vers l'état 2.

4.3.2 Estimation non paramétrique des intensités de transition avec covariables

Dans cette section, nous allons stratifier l'ensemble de données pour observer l'effet de la covariable pneumonie. Ainsi, dans tout ce qui suit, la pneumonie sera considérée comme une covariable indépendante du temps et sa valeur sera fixée à l'admission à l'unité de soins intensifs. La covariable pneu sera codée 0 si le sujet n'a pas la pneumonie à l'admission et elle sera codée 1 si le sujet a la pneumonie à l'admission. L'objectif est d'observer l'effet de la pneumonie sur le risque de mourir à l'unité de soins intensifs.

Les figures 4.3 et 4.4 présentent les estimateurs de Nelson-Aalen (équation (4.15)) des intensités de transition dans les deux groupes. L'intensité associée à la transition de l'état 0 vers l'état 1 est plus élevée dans les deux strates.

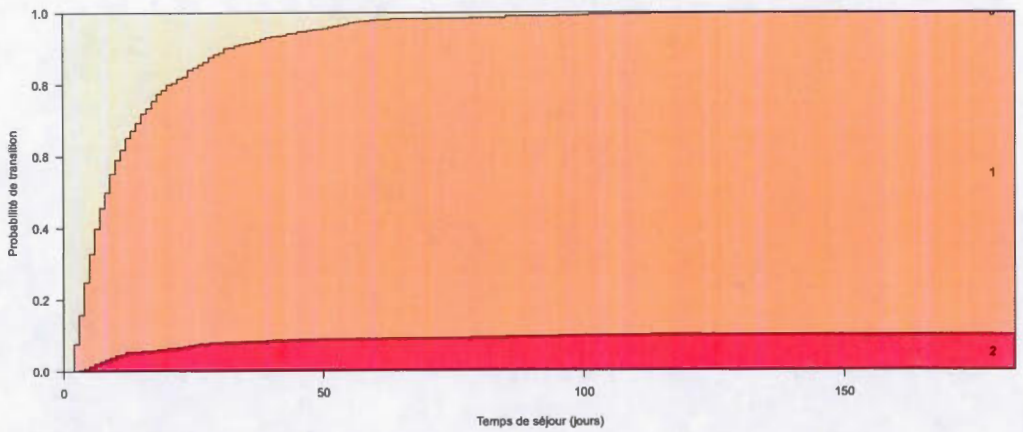

Figure 4.2 Estimation de Aalen-Johanson des probabilités de transition $\hat{P}(0,t)$.

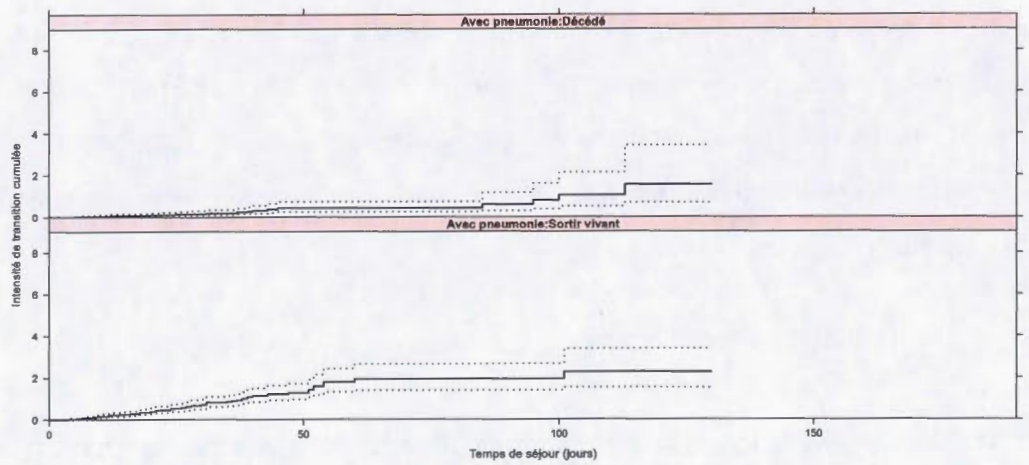

Figure 4.3 Estimation de Nelson-Aalen des intensités cumulées avec pneumonie.

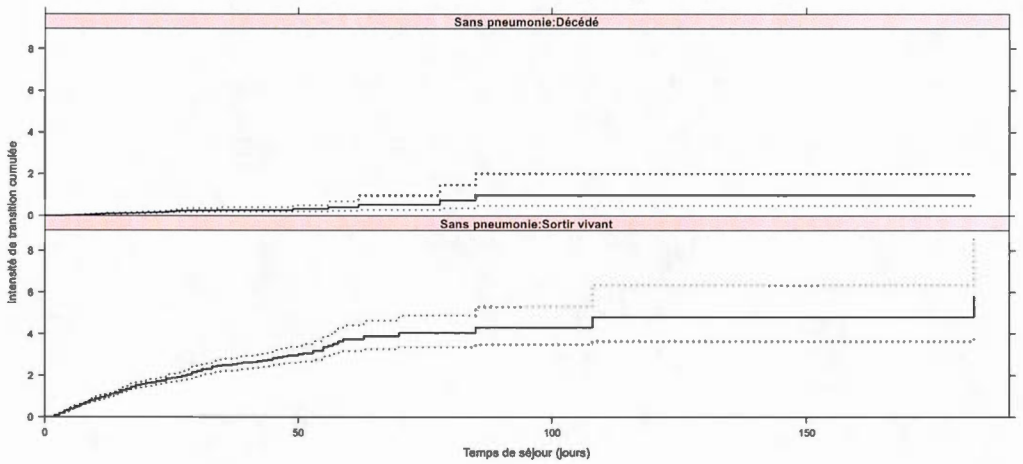

Figure 4.4 Estimation de Nelson-Aalen des intensités cumulées sans peumonie.

Les estimateurs de Aalen-Johansen des probabilités de transition (équation (4.16)) dans chacun des groupes sont représentés dans les figures 4.5 et 4.6. Les deux courbes en rouge représentent les probabilités de transitions de l'état 0 vers l'état 2. Les deux courbes montrent que la probabilité de transition de l'état 0 vers l'état 2 est plus élevée dans les deux groupes. On conclut que la pneumonie semble réduire le risque de toutes causes à la fin de séjour dans l'unité de soins intensifs, et les sujets atteints de pneumonie à l'admission restent plus longtemps dans l'unité de soins intensifs et le risque de mourir est plus important.

4.3.3 Estimation semi-paramétrique

Un modèle semi-paramétrique de risques proportionnels est ajusté pour prendre en compte l'effet des covariables. Nous considérons les deux modèles suivants :

$$\alpha_{0j}(t) = \alpha_{0j0}(t)\exp(\beta_{0j}Z_1), \quad j = 1,2 \tag{4.22}$$

$$\alpha_{0j}(t) = \alpha_{0j0}(t)\exp(\beta_{0j1}Z_1 + \beta_{0j2}Z_2 + \beta_{0j3}Z_3), \quad j = 1,2. \tag{4.23}$$

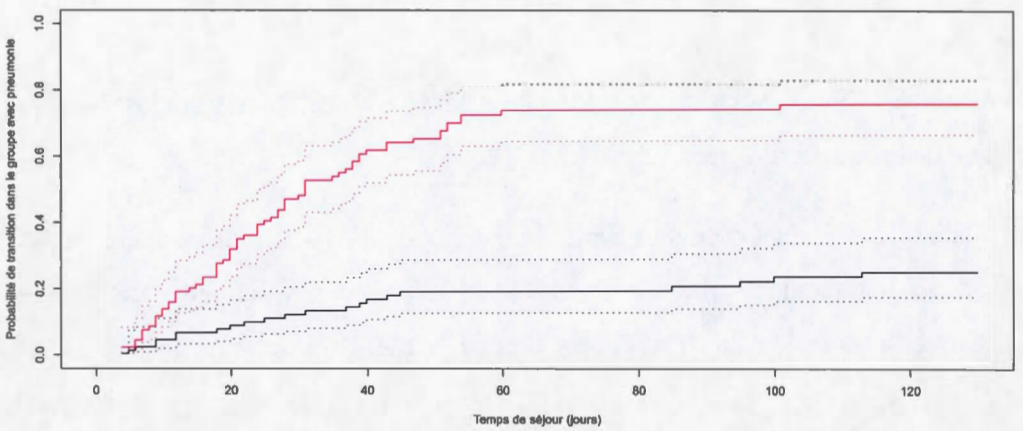

Figure 4.5 Estimation de Aalen-Johansen des probabilités de transition (pneu = 1).

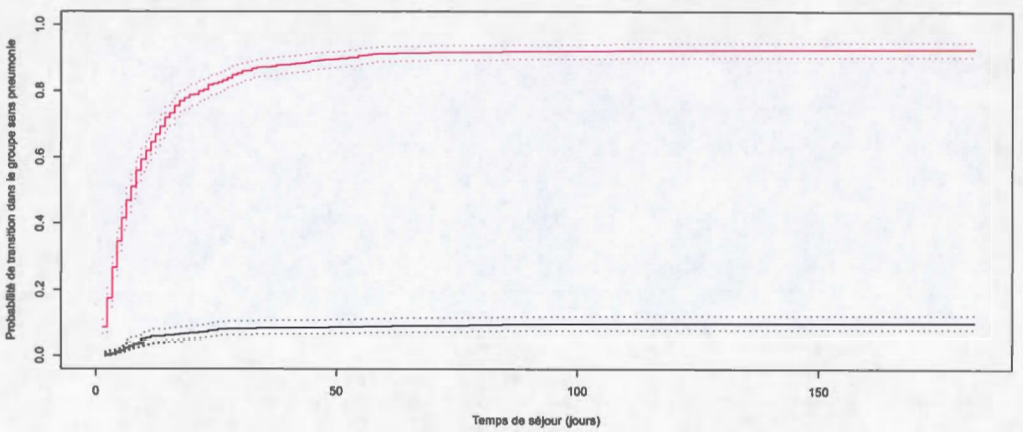

Figure 4.6 Estimation de Aalen-Johansen des probabilités de transition (pneu = 0).

La covariable Z_1 est associée à la pneumonie :

$$Z_1 = \begin{cases} 0 & \text{si le sujet avait la pneumonie à l'admission,} \\ 1 & \text{sinon.} \end{cases}$$

Les covariables Z_2 et Z_3 sont associées au sexe et à l'âge respectivement, et les fonctions de risque de base sont différentes.

Le modèle (4.22) permet d'étudier l'effet de la pneumonie par l'intermédiaire de β_{0j}. Si β_{0j} est statistiquement différent de zéro alors la pneumonie influence de manière significative la transition de l'état 0 vers l'état $j = 1,2$. Par contre, le modèle (4.23) comprend les trois covariables. Les résultats sont rapportés ci-dessous dans les tableaux 4.1 et 4.2 :

Tableau 4.1 Estimation semi-paramétrique des coefficients dans le modèle (4.22).

Transition : $0 \to j$	$\hat{\beta}_{0j}$	$\exp(\hat{\beta}_{0j})$	$\mathrm{SE}(\hat{\beta}_{0j})$	p
$0 \to 1$	-1.064	0.345	0.130	3.3e-16
$0 \to 2$	-0.161	0.852	0.268	5.5e-01

Tableau 4.2 Estimation semi-paramétrique des coefficients dans le modèle (4.23).

Transition : $0 \to j$	Covariable : Z_k	$\hat{\beta}_{0jk}$	$\exp(\hat{\beta}_{0jk})$	$\mathrm{SE}(\hat{\beta}_{0jk})$	p
$0 \to 1$	Z_1	-1.07850	0.340	0.13052	1.1e-16
$0 \to 2$	Z_1	-0.08168	0.922	0.27093	7.6e-01
$0 \to 1$	Z_2	-0.11829	0.888	0.08000	1.4e-01
$0 \to 2$	Z_2	-0.40427	0.667	0.23570	8.6e-02
$0 \to 1$	Z_3	-0.00316	0.997	0.00217	1.4e-01
$0 \to 2$	Z_3	0.01411	1.014	0.00741	5.7e-02

Dans les deux modèles, l'effet de la covariable Z_1 sur la transition $0 \to 1$ est significatif (les probabilités critiques sont $p = 3.3 \times 10^{-16}$ et $p = 1.1 \times 10^{-16}$ respectivement).

Dans le modèle (4.23), on ne peut rejeter l'hypothèse selon laquelle le coefficient des covariables sexe et âge est nul pour les transitions $0 \rightarrow 1$ et $0 \rightarrow 1$. Cela signifie qu'elles n'ont aucun effet sur les transitions.

CONCLUSION

Dans de nombreuses études épidémiologiques, les sujets peuvent subir plusieurs événements. Dans ce contexte, les modèles multi-états qui fournissent une vision complète et détaillée de l'évolution des sujets dans le temps sont des méthodes intéressantes. Ils sont considérés comme une généralisation du modèle de survie classique. La notion de processus est utilisée dans ces modèles afin de présenter les différents états successivement occupés à chaque temps d'observation. L'étude de ces modèles consiste à analyser les forces de passage entre les différents états appelées intensités de transition. Cependant, dans l'utilisation des modèle multi-états l'hypothèse markovienne est généralement considérée. Cette hypothèse suppose que l'évolution future du processus ne dépend pas de son passé, mais uniquement de l'état présent.

Dans ce mémoire, nous nous sommes intéressés à expliquer l'approche markovienne pour des données multi-états, en développant la théorie et en l'appliquant à deux ensembles de données : `sir.adm` et `cav`. Dans un premier temps, nous avons rappelé les concepts de base et les principales méthodes d'estimation du modèle de survie classique. Dans un second temps, nous avons abordé la méthode relative au modèle de Markov homogène. Ce modèle est le plus simple des modèles de type markovien car il suppose que les intensités de transition sont constantes au cours du temps. Cette hypothèse d'homogénéité simplifie les méthodes d'estimation. Cependant, elle impose une contrainte qui est souvent trop forte dans de nombreuses applications. Dans ce cadre, nous avons présenté les vraisemblances lorsque les temps de transition entre les états sont, soit possiblement censurés à droite, soit censurés par intervalle. Ces vraisemblances étaient adaptées aux modèles particuliers des modèles muli-états markoviens

homogènes : le modèle de survie classique et le modèle des risques concurrents pour des données censurés à droite. A l'aide de la bibliothèque `msm` du logiciel R, nous avons estimé les intensités de transition ainsi que les probabilités de transition pour l'ensemble de données `cav` où les temps de transition entre les états sont censurés par intervalle.

La théorie des processus de comptage est ensuite présentée afin d'introduire des méthodes d'estimation non paramétrique dans le cadre d'un modèle de Markov non homogène. Dans ce modèle, les intensités de transition dépendent du temps. La méthodologie des processus de comptage fournit un cadre rigoureux qui permet notamment de généraliser, aux modèles markoviens, les estimateurs des modèles de survie classiques. Les méthodes d'estimation présentées dans le cadre du modèle de Markov non homogène supposent que le mécanisme de censure n'apporte aucune information sur la survenue des événements. Pour illustrer cette approche, nous avons appliqué différents modèles : le modèle sans covariable, le modèle stratifié selon la covariable pneumonie, le modèle des risques proportionnels. Chaque méthode d'estimation est appliquée à l'ensemble de données `sir.adm`. L'application est réalisée en utilisant les bibliothèques `mstate` et `etm` du logiciel R.

Il serait intéressant d'aborder aussi les modèles semi-markoviens qui n'ont pas été présentés dans le présent travail. Ces modèles généralisent les modèles markoviens dans le sens où ils permettent d'intégrer la notion du temps de séjour dans le calcul des intensité de transition. L'intérêt de ce type de modèle est contenu dans le choix de la distribution du temps de séjour dans l'état. La probabilité de rester dans un état peut alors dépendre de la durée passée dans ce état. Un modèle semi-markovien dont les temps de séjour suivent des lois exponentielles devient un modèle markovien homogène. Nous pourrions appliquer un modèle semi-markovien à l'ensemble de données `cav` puisque les temps de séjour dans les états sont inconnus. Par contre, cette approche ne peut pas être appliquée à l'ensemble de données `sir.adm` puisque les temps de séjour dans les unités de soins intensifs sont connus exactement.

Les méthodes d'estimation dans les modèles multi-états varient en fonction du type de données incomplètes. Dans ce travail, nous avons abordé seulement le mécanisme de censure à droite et le mécanisme de censure par intervalle. Le mécanisme de troncature à gauche peut être présent dans l'analyse de données des modèles multi-états. La troncature à gauche peut être présentée comme un processus qui ne respecte pas l'origine choisie, mais seulement le fait que l'événement survient plus tard dans le temps.

APPENDICE A

RAPPELS SUR LE PRODUIT INTÉGRAL

Nous rappelons ici un résultat complémentaire concernant le produit intégral.

Définition A.0.1. *Soit* $(\Omega, \mathscr{A}, \mathbb{P})$ *un espace de probabilité où* $\mathbb{P}$ *est la mesure de probabilité sur* $(\Omega, \mathscr{A})$. *Un* processus stochastique *est une fonction à deux variables*

$$\begin{aligned} X : [0, +\infty[\times \Omega &\to \mathbb{R} \\ (t, \omega) &\mapsto X(t, \omega), \end{aligned}$$

où t *représente le temps et* ω *le hasard. Pour tout* $t \geq 0$, *la fonction* $X_t : \omega \mapsto X(t, \omega)$ *est une variable aléatoire réelle appelée coordonnée à l'instant* t. *Pour tout* $\omega \in \Omega$, *la fonction* $t \mapsto X(t, \omega)$ *est la trajectoire de* ω.

Définition A.0.2. *Soit* $(\Omega, \mathscr{A}, \mathbb{P})$ *un espace de probabilité. Une* filtration *est une famille de tribus* $\{\mathscr{F}_t\}$, $t \geq 0$, *telle que*

$$\mathscr{F}_s \subset \mathscr{F}_t \subset \mathscr{A}, \ \forall \ s \leq t.$$

Définition A.0.3. *Soient un espace de probabilité* $(\Omega, \mathscr{A}, \mathbb{P})$ *et* $\{\mathscr{F}_t\}$ *une filtration. Un processus* X *est dit* $\{\mathscr{F}_t\}$*-adapté si* X_t *est* $\mathscr{F}_t$*-mesurable pour chaque* $t \geq 0$.

Définition A.0.4. *Soit* $\{\mathbf{X}(t), t \geq 0\}$ *une matrice* $k \times k$ *de processus cadlag, nul en 0 et à variation bornée. Soit* $t_0 = s < t_1 < \cdots < t_K = t$ *une partition de* $]s,t]$. *On appelle* produit intégral

$$\mathscr{P}_{]s,t]}(\mathbf{I} + d\mathbf{X}) = \lim_{\max |t_i - t_{i-1}| \to 0} \prod \{\mathbf{I} + \mathbf{X}(]t_{h-1} - t_h])\},$$

où $\mathbf{I}$ *représente la matrice indentité* $k \times k$.

Théorème A.0.1. *Soient* $\mathbf{Z}$ *et* $\mathbf{W}$ *des matrices* $k \times p$ *de fonctions cadlag. Pour* $\mathbf{W}$ *fixée, l'unique solution* $\mathbf{Z}$ *de l'équation*

$$\mathbf{Z}(t) = \mathbf{W}(t) + \int_0^t \mathbf{Z}(s^-)\mathbf{X}(ds),$$

est

$$\begin{aligned} \mathbf{Z}(t) &= \mathbf{W}(t) + \int_0^t \mathbf{W}(s^-)X(du)\mathscr{P}_{]s,t]}(\mathbf{I}+d\mathbf{X}) \\ &= \mathbf{W}(0)\mathscr{P}_{[0,t]}(\mathbf{I}+d\mathbf{X}) + \int_0^t \mathbf{W}(ds)\mathscr{P}_{]s,t]}(\mathbf{I}+d\mathbf{X}). \end{aligned}$$

APPENDICE B

CODES R

B.1 Code pour l'ensemble de données sir.adm

```
##### Ensemble de données sir.adm
library(mvna)
data(sir.adm)

##### Premières lignes de l'ensemble de données
head(sir.adm)

##### Nombre de sujets
n<-nrow(sir.adm)

##### Nombre de sujets avec et sans pneumonie
table(sir.adm$pneu)

##### Nombre de sujets décédés et ceux qui sont sortie vivant
des unitées de soins intensifs
à la fin de l'étude
table(sir.adm$status)

##### Nombre de femmes et d'hommes
table(sir.adm$sex)
library(mstate)

#### Matrice des transitions possibles ####
tmat <- trans.comprisk(2, names = c("0", "1", "2"))
sir.adm$stat1<-as.numeric(sir.adm$status==1)
sir.adm$stat2<-as.numeric(sir.adm$status==2)
```

```
#### Transformer les données  ####
sir.admlong<-msprep(time=c(NA,"time","time"),
status=c(NA, "stat1", "stat2"), data=sir.adm,
keep=c("sex","age"), trans=tmat)
events(sir.admlong)

######## Estimation des intensités de transition cumulées
library(survival)
IT<-coxph(Surv(Tstart,Tstop,status)~strata(trans),
data=sir.admlong)
msfa <- msfit(object = IT, vartype = "aalen", trans = tmat,
variance=TRUE)
head(msfa$Haz)
tail(msfa$Haz)
head(msfa$varHaz)
tail(msfa$varHaz)

###### La courbe des intensités de transition cumulées
plot(msfa, las=1,xlab = "Temps de séjour (jours)",
ylab="Intensité de transition cumulée")

##### Estimation des probabilités de transition
pt<- probtrans(msfa, predt = 0, method = "aalen",
variance=TRUE)
summary(pt, from = 1)

###### La courbes des probabilités de transitions
library(colorspace)
statecols <- heat_hcl(3, power = c(1,2))[c(1, 2, 3)]
ord <- c(3,2,1)
plot(pt, ord = ord, xlab = "Temps de séjour (jours)",
ylab="Probabilité de transition", las = 1,
type = "filled", col = statecols[ord])

###### Stratifier selon la covariable pneumonie

### Transformer l'ensemble de données
to<-ifelse(sir.adm$status==0,"cens",
ifelse(sir.adm$status==1,2,1))
my.sir.data<-data.frame(id=sir.adm$id, from=0,to,
```

```
time=sir.adm$time, pneu=sir.adm$pneu)
head(my.sir.data)
table(my.sir.data$to)
# Matrice des transitions possibles
tra<-matrix(ncol=3,nrow=3,FALSE)
dimnames(tra)<-list(c("0","1","2"),c("0","1","2"))
tra[1,2:3]<-TRUE
tra

##### Estimateur de Nelson-Aalen de l'intensité de transition cumulée

## sans pneumonie
my.nelaal.nop<-mvna(my.sir.data[my.sir.data$pneu==0, ],
c("0","1","2"),tra,"cens")

## avec pneumonie
my.nelaal.p<-mvna(my.sir.data[my.sir.data$pneu==1, ],
c("0","1","2"),tra,"cens")

##### La courbe de l'estimateur de Nelson-Aalen

## sans pneumonie
library(lattice)
dessin.nop<-xyplot(my.nelaal.nop,tr.choice=c("0 2","0 1"),
lwd=2,layout=c(1,2),
strip=strip.custom(factor.levels=c("Sans pneumonie:Sortir vivant",
"Sans pneumonie:Décédé"),par.strip.text=list(font=2)),
ylim=c(0,9),xlim=c(0,190),xlab = "Temps de séjour (jours)",
 ylab="Intensité de transition cumulée",
scales=list(alternating=1,x=list(at=seq(0,150,50)),
y=list(at=seq(0,8,2))))

## avec pneumonie
dessin.p<-xyplot(my.nelaal.p,tr.choice=c("0 2","0 1"), lwd=2
,layout=c(1,2),
strip=strip.custom(factor.levels=c("Avec pneumonie:Sortir vivant",
"Avec pneumonie:Décédé"),par.strip.text=list(font=2)),
ylim=c(0,9),xlim=c(0,190),xlab = "Temps de séjour (jours)",
ylab="Intensité de transition cumulée",
scales=list(alternating=1,x=list(at=seq(0,150,50)),
y=list(at=seq(0,8,2))))
```

```
##### Estimateur Aalen-Johansen des probabilités de transition

## sans pneumonie
library(etm)
my.sir.etm.nop <- etm(my.sir.data[my.sir.data$pneu == 0, ],
c("0", "1", "2"), tra, "cens", s = 0)

## avec pneumonie
my.sir.etm.p <- etm(my.sir.data[my.sir.data$pneu == 1, ],
c("0", "1", "2"), tra, "cens", s = 0)

##### La courbe de l'estimateur de Aalen-Johansen

## sans pneumonie
plot(my.sir.etm.nop, tr.choice = '0 1', col = 1, lwd = 2,
conf.int = TRUE, ci.fun = "cloglog", legend = FALSE,
ylab="Probabilité de transition dans le groupe sans pneumonie",
xlab = "Temps de séjour (jours)")
lines(my.sir.etm.nop, tr.choice = '0 2', col = "red",
lwd = 2, conf.int = TRUE, ci.fun = "cloglog")

## avec pneumonie
plot(my.sir.etm.p, tr.choice = '0 1', col = 1, lwd = 2,
conf.int = TRUE, ci.fun = "cloglog", legend = FALSE,
ylab="Probabilité de transition dans le groupe avec pneumonie",
 xlab = "Temps de séjour (jours)")
lines(my.sir.etm.p, tr.choice = '0 2', col = "red",
lwd = 2, conf.int = TRUE, ci.fun = "cloglog")
```

B.2 Code pour l'ensemble de données cav

```
##### Ensemble de données cav
library(msm)
head(cav)

#### matrice de transition
statetable.msm(state, PTNUM, data=cav)

### matrice initiale
Q<- rbind(c(0, 0.25, 0, 0.25), c(0.166, 0, 0.166, 0.166)
```

```
, c(0, 0.25, 0, 0.25), c(0, 0, 0, 0))

### Estimation des intensités de transition
cav.msm<- msm(state ~ years, subject=PTNUM, data=cav,
qmatrix = Q, death=4)

### Estimation des probabilités de transition au temps t=10 ans
cav.p<-pmatrix.msm(cav.msm, t=10, ci="normal")
```

RÉFÉRENCES

Aalen, O. O. (1975). *Statistical inference for a family of counting processes.* (Thèse de doctorat). University of California, Berkeley.

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, *6*(4), 701–726.

Aalen, O. O. et Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, *5*(3), 141–150.

Allignol, A., Beyersmann, J. et Schumacher, M. (2008). `mvna` : An R Package for the Nelson-Aalen Estimator in Multistate Models. *R News*, *8*(2), 48–50.

Allignol, A., Schumacher, M. et Beyersmann, J. (2011). Empirical transition matrix of multi-state models : the `etm` package. *Journal of Statistical Software*, *38*(4), 1–15.

Andersen, P. K., Borgan, Ø., Gill, R. D. et Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York : Springer.

Andersen, P. K., Hansen, L. S. et Keiding, N. (1991). Non- and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous markov process. *Scandinavian Journal of Statistics*, *18*(2), 153–167.

Andersen, P. K. et Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*(2), 91–115.

Beyersmann, J., Allignol, A. et Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. New York : Springer.

Commenges, D. (1999). « Multi-state models in epidemiology ». *Lifetime Data Analysis*, *5*(4), 315–327.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B. Methodological*, *34*, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Cox, D. R. et Oakes, D. (1984). *Analysis of Survival Data*. London : Chapman & Hall.

Crowder, M. J. (2001). *Classical Competing Risks*. London : Chapman & Hall / CRC.

de Wreede, L. C., Fiocco, M. et Putter, H. (2011). mstate : An R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software*, *38*(7), 1–30.

Gentleman, R., Lawless, J., Lindsey, J. et Yan, P. (1994). Multi-state markov models for analysing incomplete disease history data with illustrations for HIV disease. *Statistics in Medicine*, *13*(8), 805–821.

Gray, B. (2014). *cmprsk* (version 2.2.7). [Package R]. Vienne : The R Foundation for Statistical Computing.

Hougaard, P. (1999). Multi-state models : a review. *Lifetime Data Analysis*, *5*(3), 239–264.

Hsieh, H.-J., Chen, T. H. H. et Chang, S.-H. (2002). Assessing chronic disease progression using non-homogeneous exponential regression Markov models : an illustration using a selective breast cancer screening in Taiwan. *Statistics in Medicine*, *21*(22), 3369–3382.

Jackson, C. (2011). Multi-state models for panel data : the `msm` package for R. *Journal of Statistical Software*, *38*(8), 1–29.

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.

Kalbfleisch, J. D. et Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Chichester : John Wiley & Sons.

Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, *53*(282), 457–481.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, *42*(4), 855–865.

Lawless, J. F. (1984). *Statistical Models and Methods for Lifetime Data*. Chichester : John Wiley & Sons.

Longini, I. M., Clark, W. S., Byers, R. H., Ward, J. W., Darrow, W. W., Lemp, G. F. et Hethcote, H. W. (1989). Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, *8*(7), 831–843.

Marshall, G. et Jones, R. H. (1995). Multi-state models and diabetic retinopathy. *Statistics in Medicine*, *14*(18), 1975–1983.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, *14*(4), 945–966.

Pintilie, M. (2006). *Competing Risks : A Practical Perspective*. Chichester : John Wiley & Sons.

Schulgen, G. et Schumacher, M. (1996). Estimation of prolongation of hospital stay attributable to nosocomial infections : new approaches based on multistate models. *Lifetime Data Analysis*, *2*(3), 219–240.

Sharples, L. D., Jackson, C. H., Parameshwar, J., Wallwork, J et Large, S. R. (2003). Diagnostic accuracy of coronary angiography and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation*, *76*(4), 679–682.

Sun, J. (2007). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York : Springer.

Touraine, C., Joly, P. et Gerds, T. A. (2014). *SmoothHazard* (version 1.2.3). [Package R]. Vienne : The R Foundation for Statistical Computing.

Wolkewitz, M., Vonberg, R. P., Grundmann, H., Beyersmann, J., Gastmeier, P., Bärwolff, S., Geffers, C., Behnke, M, Rüden, H. et Schumacher, M. (2008). Risk factors for the development of nosocomial pneumonia and mortality on intensive care units : application of competing risks models. *Critical Care*, *12*(2/R44), 1–9.

Wrangler, M., Beyersmann, J. et Schumacher, M. (2006). changeLOS : An R package for change in length of hospital stay based on the Aalen-Johansen estimator. *R News*, *6*(2), 1–35.