Modèles multi-états de type Markovien et application à l'asthme

Philippe Saint-Pierre

UNIVERSITE MONTPELLIER I
U.F.R. de MEDECINE

Année 2005 N° attribué par la bibliothèque

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER I

Formation Doctorale : Biostatistique

Ecole Doctorale : Information, Structures, Systèmes
Discipline : Epidémiologie, Economie de la santé, Prévention

présentée et soutenue publiquement
par

Philippe SAINT PIERRE

Le 29 Avril 2005

Titre :

MODELES MULTI-ETATS DE TYPE MARKOVIEN ET APPLICATION A L'ASTHME

Directeur de thèse : Professeur Jean-Pierre DAURES
Co-encadrant : Professeur Philippe GODARD

JURY

M. Michel AUBIER	Professeur à l'Université Paris VII	Rapporteur
M. Jean-Pierre DAURES	Professeur à l'Université Montpellier I	Directeur de thèse
M. Gilles DUCHARME	Professeur à l'Université Montpellier II	Examinateur
M. Abdelkader EL HASNAOUI	Biométricien à GlaxoSmithKline	Examinateur
M. Philippe GODARD	Professeur à l'Université Montpellier I	Directeur de thèse
Mme. Catherine HUBER	Professeur à l'Université Paris V	Rapporteur
M. Gérard TAP	M. de Conférences à l'Université Toulouse III	Examinateur

REMERCIEMENTS

En premier lieu, je tiens à exprimer mes sincères remerciements à Jean-Pierre Daurès pour m'avoir encadré pendant cette thèse. Il a su me faire confiance en me laissant une grande liberté d'initiative tout en me faisant profiter de sa rigueur et de ses compétences mathématiques. Au delà du contexte professionnel, je lui exprime toute mon estime.

Je voudrais aussi remercier Philippe Godard qui a co-encadré ce travail. Le médecin qu'il est a su faire confiance au statisticien que je suis. Il m'a beaucoup apporté par sa vision clinique du problème. Ce fut pour moi un très grand honneur de collaborer avec un éminent spécialiste de l'asthme.

Je suis également très reconnaissant envers Catherine Huber et Michel Aubier pour l'intérêt qu'ils ont bien voulu accorder à ma thèse en acceptant d'en être les rapporteurs.

J'adresse mes remerciements sincères à Gilles Ducharme, Gérard Tap et Abdelkader El Hasnaoui pour avoir accepté de participer à ce jury.

Je tiens à remercier toute l'équipe de l'IURC pour m'avoir supporté au cours de ces années de thèse. Vous m'avez tous aidé, de près ou de loin, à réaliser cette thèse. Merci à Nicolas pour m'avoir fait partager son expérience.

Je remercie tout particulièrement Yohan D., Christophe & Julie, Yohann F. et Christel pour tous les bons moments passés au boulot et en dehors. Vous êtes pour beaucoup dans ce travail, autant sur le plan humain que scientifique.

Merci à Marie, Sandy, Nathalie, Françoise, Séverine, Christine, Pascale pour leur bonne humeur quotidienne et leurs nombreux coups de main, Christophe C. pour ses conseils éclairés, Faiza, Pierre et Eve mes collègues de thèse. Je pense aux anciens de l'IURC : Yannick, Claude & Karine, Christophe B. pour m'avoir supporté dans le bureau, Delphine, Remy et Florian pour les foots, Claudine pour la collaboration scientifique et les moments partagés aux congrès.

Un merci particulier à Sylvie pour sa gentillesse, son efficacité et son aide dans l'organisation des congrès et de la thèse.

Je décerne une mention spéciale à Benoît pour tous les moments partagés. Du parc de Balma au Laos en passant par le Burkina, tu as toujours été là !

"Spéciale dédicace" à Charly & Amandine pour leur enthousiasme et leur motivation débordantes. Je vous souhaite sincèrement le meilleur pour le "BiblioBrousse". Je suis fier de vous avoir comme amis.

Je remercie évidemment Cathy & Guillaume pour leur amitié fidèle malgré le temps qui passe ainsi qu'Olivier pour toutes les soirées "play" et les week-ends "snow".

J'ai des remerciements tous particuliers à adresser aux anciens Balmanais pour leurs amitiés et les nombreuses soirées aussi bien survoltées que mémorables. Je pense à Benoît, PJ, Cathy, Olivier C., Laurent, Sophie, Bruno, François J., François C., Bertrand, ainsi que Johan A. et sa petite famille.

Comment pourrai-je oublier mes collègues toulousains de licence et de maîtrise avec qui j'ai partagé une bonne partie de ma vie universitaire. Les soirées enflammées, le séjour en Espagne entre "playboys", mais aussi les révisions de partiels resteront pour moi, de

grands et de bons souvenirs. Cruz, Cécile, Aurélie, Ernest, Nico S., Géraldine & Patrick, Hélène et Franck : merci ! Je suis heureux de continuer à vous voir malgré la distance.

Je remercie également Amélie & Alex et Laurie & Benoît pour nous avoir accueillis à Montpellier.

Je ne peux oublier Kaz dit Alexis pour son hospitalité et pour tous les bons moments passés de Ouaga à Songo. J'ai une pensée particulière pour sa famille qui m'a accueilli et m'a permis de partager une tranche de vie dans un monde loin de nos valeurs et de nos considérations occidentales.

Je pense également à Cédric Bonnet, aux familles Helson, Buxant, Ravel et Mangenot, les vieux amis de Bouaké. La cour de l'école Bambi restera pour moi un terrain de jeu inoubliable.

Un grand merci à la famille de Céline qui m'a toujours apporté son soutien : Jo & Bernard, Pierre & Aurèle, Fabienne & Bruno.

Je voudrais remercier spécialement ma famille qui m'a toujours soutenu. En particulier, mes parents qui ont toujours cru en moi, mon frère qui, fort de son expérience, a su me conseiller, ma soeur et Olivier pour leurs encouragements. Aude, maintenant que je suis initié, je pourrai te conseiller ! "Spéciale dédicace" à Mamie qui malheureusement n'a pas pu participer à cet événement mais à qui je pense.

Ces remerciements ne seraient pas complets si je n'évoquais la présence et le soutien inconditionnel de Céline. Elle m'a suivi à Montpellier (et à Delhi !), elle m'a supporté au quotidien et elle m'a toujours encouragé dans les moments difficiles. Elle m'a également aidé dans toutes les étapes de la rédaction : des problèmes mathématiques à la relecture en passant par la rédaction. Je ne te remercierai jamais assez pour tout ce que tu m'as apporté.

La rédaction d'une thèse donne la chance de pouvoir remercier toutes les personnes qui nous sont chères et qui ont rendu ce travail possible. Famille, collègues ou amis, si j'en suis arrivé là, c'est bien grâce à vous. Même si les mots et la place manquent, puissent ces quelques lignes exprimer l'ampleur de ma reconnaissance.

Philippe.

Table des matières

Liste des figures

Liste des tableaux

Introduction

1 Contexte

Dans de nombreux domaines, décrire l'évolution des phénomènes dans le temps est d'un intérêt capital, en particulier pour aborder les problématiques de la prédiction et de la recherche de facteurs causaux. Les observations correspondent souvent à des mesures d'une même caractéristique faites à plusieurs instants. Ces données, appelées **mesures répétées**, se distinguent de celles présentes dans les modélisations statistiques traditionnelles. En effet,

- la même variable est mesurée plusieurs fois sur un même individu en des temps différents : les réponses ne peuvent plus être considérées comme étant indépendantes comme c'est le cas dans l'analyse de régression usuelle,
- plusieurs individus sont inclus dans l'échantillon : les réponses ne forment pas une série chronologique simple.

En épidémiologie, les données de cohorte constituent la source majeure de mesures répétées. Ces données sont obtenues quand on observe un groupe de patients au cours du temps afin d'identifier, décrire ou quantifier un phénomène. La variable qu'on cherche à expliquer peut être de différentes formes.

- Dans le cas du VIH, la mesure de la charge virale reflète l'avancée du virus dans l'organisme et la mesure de lymphocytes CD4 reflète le réservoir immunologique. Ce sont des marqueurs fondamentaux pour définir l'état de santé du patient. Ces marqueurs sont des variables quantitatives.
- Dans le cas du cancer, l'état de la maladie peut être renseigné par le nombre de métastases. Il y a un nombre fini de modalités pour la variable étudiée.
- Dans le suivi d'une transplantation, le phénomène d'intérêt est par exemple le rejet de la greffe. La variable étudiée est binaire.
- Dans de nombreuses maladies, la variable prenant les valeurs « sain », « malade » et « décès » peut être considérée comme la variable à expliquer.

Une autre spécificité des données de cohorte réside dans le fait que souvent, les mesures ne sont pas effectuées à des intervalles de temps fixés. Ainsi, les temps d'observation ne sont pas les mêmes pour tous les individus et le temps écoulé entre deux observations n'est généralement pas fixé (et varie d'un individu à l'autre). L'espace des temps de mesure est dit continu par opposition au cas discret où les temps d'observation sont fixés pour tous les individus.

Dans la plupart des maladies chroniques, des données de cohorte sont mises en place afin de mieux comprendre l'évolution de la maladie. Dans le cas de l'asthme et des allergies, en

réponse à l'augmentation rapide de la prévalence, de nombreuses cohortes ont été mises en oeuvre dans le but de mieux connaître les différents facteurs de risque de ces affections. En 1996, plusieurs CHU français ont collecté des données sur des patients asthmatiques suivis à l'hôpital. La base obtenue est une cohorte observationnelle reflétant l'activité réelle d'un hôpital avec des patients venant consulter à des instants quelconques. L'espace des temps d'observation étant continu, certains modèles statistiques usuels pour données longitudinales ne conviennent plus. Par exemple, l'analyse de la variance pour mesures répétées, qui nécessite des temps d'observation fixés, n'est pas adaptée.

Les modèles multi-états constituent une alternative intéressante pour modéliser des données de type mesures répétées. D'un point de vue théorique, l'objectif de ce document est d'étudier et de développer des méthodes statistiques pour les modèles multi-états. D'un point de vue clinique, l'objectif est d'analyser une cohorte sur l'asthme afin de fournir aux cliniciens des outils pour comprendre l'évolution des patients asthmatiques.

2 Modélisation

2.1 Alternatives aux modèles multi-états

Les modèles à effets aléatoires (où modèles mixtes) sont fréquemment utilisés pour étudier la corrélation entre des observations. Par l'intermédiaire des effets aléatoires, ils permettent d'expliciter les différences entre les individus sans observer les déterminants de cette variabilité inter-individuelle. Ils prennent ainsi en compte une hétérogénéité entre individus ou entre groupes d'individus. Ces modèles peuvent être généralisés et de nature linéaire ou non (Pinheiro et Bates [2000]). La corrélation entre les observations d'un même individu peut être modélisée par l'intermédiaire des modèles marginaux (Diggle et al. [1994]). La matrice de covariance des observations est définie de manière à prendre en compte les corrélations intra-individus. La théorie des équations d'estimation généralisées (Liang et Zeger [1986]) permet par exemple d'estimer ce type de modèles.

Les modèles de régression mettant en jeu des variables fonctionnelles peuvent également être utilisés. De manière plus précise, ces modèles considèrent le cas d'un problème de régression où la variable à expliquer est réelle et la variable explicative est une fonction. Pour chaque individu, les données mesurées au cours du temps sont considérées comme une courbe. Dans le contexte de la régression, ces modèles ont été introduits pour prendre en compte la corrélation entre les observations d'un même individu et le fait que le nombre de variables explicatives (nombre de mesures pour un même individu) est souvent supérieure à la taille de l'échantillon. Le modèle linéaire et le modèle linéaire généralisé traditionnels peuvent être adaptés afin de tenir compte de la nature fonctionnelle des variables explicatives (Saint-Pierre [2001], Cardot et al. [1999]).

2.2 Les modèles multi-états

Depuis une trentaine d'années, les modèles multi-états ne cessent de connaître un intérêt croissant. Ces modèles utilisent la notion d'« état » et de processus pour décrire un

phénomène. La notion de processus est utilisée pour représenter les différents états successivement occupés à chaque temps d'observation. En épidémiologie, ils permettent par exemple, de représenter l'évolution d'un patient à travers les différents stades d'une maladie. Après définition des différents stades, 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.

Un nombre important de publications statistiques concerne les modèles multi-états. Cependant, l'application de ces modèles dépasse rarement le cadre des revues spécialisées. Cette situation s'explique en partie, par l'absence de logiciels adaptés et la méconnaissance des méthodes statistiques. La popularité et la richesse des modèles de survie, en particulier du modèle de Cox, dessert l'utilisation de ces modèles dans le domaine appliqué. Il est pourtant des situations où l'étude d'un délai d'apparition d'un événement ne peut apporter qu'une réponse partielle au problème posé.

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

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

- la durée du suivi (temps depuis l'inclusion dans l'étude),
- le temps depuis la dernière transition (durée dans l'état présent).

Il existe plusieurs possibilités pour définir les intensités de transition $\alpha(t, d)$, où t représente la durée du suivi et d la durée passée dans l'état. Lorsque $\alpha(t, d) = \alpha$, le modèle est dit homogène par rapport au temps t. Lorsque $\alpha(t, d) = \alpha(t)$ le modèle est dit non-homogène. Dans le cas où les intensités de transition dépendent de la durée du suivi, $\alpha(t, d) = \alpha(d)$, le modèle est semi-Markovien homogène par rapport au temps t. Enfin, lorsque $\alpha(t, d)$ dépend des deux échelles de temps, le modèle est semi-Markovien non-homogène.

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

Plusieurs modèles statistiques sont possibles, on distingue les approches paramétrique, non-paramétrique et semi-paramétrique.

L'approche paramétrique stipule que les intensités de transition appartiennent à une classe particulière de fonctions, qui dépendent d'un nombre fini de paramètres. L'avantage de cette approche est la facilitation attendue de la phase d'estimation des paramètres. L'inconvénient est l'inadéquation pouvant exister entre le modèle retenu et le phénomène étudié.

L'approche non-paramétrique ne nécessite aucune hypothèse sur la forme des intensités de transition et c'est là son principal avantage. L'inconvénient d'une telle approche est la nécessité de disposer d'un nombre important d'observations. En effet, le problème de l'estimation d'un paramètre fonctionnel est délicat puisqu'il appartient à un espace de dimension infinie.

L'approche semi-paramétrique est une sorte de compromis entre les deux approches précédentes. Les intensités de transition appartiennent à une classe de fonctions pour partie dépendant de paramètres et pour partie s'écrivant sous forme de fonctions non-paramétriques. Cette approche est très répandue en analyse de survie au travers du modèle de régression de Cox (Therneau et Grambsch [2000]).

Le modèle peut également faire intervenir un effet aléatoire qui agit de manière multiplicative sur les intensités de transition. Dans les études de survie, ces modèles permettent de tenir compte de la dépendance entre les temps d'événement sont appelés modèle de fragilité (frailty model) (Therneau et Grambsch [2000], Hougaard [1995]). Plus généralement, ces modèles permettent de prendre en compte des variables omises dans la modélisation (par exemple, les variables non observées, celles dont les effets sont déjà bien connus ou celles dont il n'est pas certain qu'elles influencent les intensités) (Huber-Carol et Vonta [2004], Hougaard [2000], Nielsen et al. [1992], Andersen et al. [1993]). Ces modèles sont particulièrement intéressants quand on peut distinguer des groupes d'individus : par exemple, les personnes d'une même famille auront des caractéristiques génétiques communes. Les caractéristiques génétiques étant différentes d'une famille à l'autre, il est intéressant d'avoir un effet aléatoire spécifique à chaque famille.

Une particularité des données de cohorte réside dans le fait qu'elles ne sont que partiellement observées à cause des différents phénomènes de censure et troncature (à droite, à gauche, par intervalles). Par exemple, le mécanisme de censure à droite est toujours présent car on n'observe pas un phénomène sur un temps infini ; le mécanisme de censure par intervalles intervient quand les temps exacts de transition ne sont pas connus (on sait uniquement que les transitions se sont produites pendant un intervalle de temps) (Commenges [2002]). Les méthodes d'estimation varient en fonction du type de données incomplètes.

Les modèles statistiques faisant intervenir des données censurées considèrent le plus souvent que le processus de censure est indépendant du processus d'événement. Dans les études de survie par exemple, cela suppose que le fait qu'un patient soit censuré n'apporte aucune information sur la survenue de l'événement. Dans le cas du VIH par exemple, les patients qui arrêtent le suivi sont souvent ceux dont l'état est le plus grave et dont le moral est affecté. Il est alors nécessaire de proposer des alternatives afin de tenir compte de l'information comprise dans la censure (Robins et Finkelstein [2000], Minini et Chavance [2004], Little [1995]).

3 Structure du document

Le chapitre I décrit la base de données sur l'asthme qui sera analysée tout au long de cette thèse. Les notions cliniques essentielles, la définition des états de contrôle de l'asthme, la structure des différents modèles étudiés et les covariables sont décrites.

Le chapitre II présente les problématiques classiques de l'inférence dans le modèle de Markov homogène tout en donnant quelques notions et définitions sur les processus. Ce modèle est appliqué au cas de l'asthme afin d'obtenir des résultats cliniques.

Le chapitre III présente les modèles de semi-Markov et deux méthodes d'estimation des paramètres du modèle : une méthode paramétrique permettant d'estimer la distribution des temps de séjour et une méthode non-paramétrique permettant d'estimer les intensités du processus par des fonctions constantes par morceaux. L'application de ces méthodes à la base de données sur l'asthme est également discutée.

Le chapitre IV s'attache, quant à lui, au modèle de Markov non-homogène. La théorie des processus de comptage est utilisée pour obtenir des estimations non-paramétriques et semi-paramétriques des intensités de transitions. Les résultats obtenus dans le cas de l'asthme sont présentés pour chacune des méthodes.

Le chapitre V décrit une méthode d'estimation permettant de prendre en compte un mécanisme de censure informative. La méthode IPCW (Robins et Finkelstein [2000]) est tout d'abord présentée dans le cadre de données de survie ; elle est ensuite généralisée à certains modèles multi-états, aux modèles progressifs et au modèle avec deux états réversibles. Ces méthodes sont appliquées au cas de l'asthme où il semble que les patients qui se portent bien arrêtent le suivi.

Une partie résume ensuite le travail entrepris dans les chapitres précédents.

Enfin, les différentes annexes fournissent des compléments concernant la théorie statistique des processus de comptage et un « guide » visant à faciliter la programmation des méthodes d'estimation dans le modèle de Markov non-homogène.

Chapitre I

Présentation de la base de données

1 Introduction

L'asthme est une maladie chronique réversible provoquant une obstruction bronchique plus ou moins grave. L'asthme mortel était rare et méconnu avant 1960, année où l'on a enregistré une épidémie d'asthmes mortels dans les pays anglo-saxons. Une meilleure compréhension du traitement de l'asthmatique a permis de réduire le nombre annuel de décès mais, depuis 1977, une seconde épidémie est apparue dans la plupart des pays développés touchant de plus en plus d'enfants. L'OMS a modifié en 1979 la classification des maladies, les conditions de diagnostic ont été améliorées, la consommation de médicaments anti-asthmatiques a augmenté mais cela reste insuffisant puisque la mortalité ne diminue pas. Chaque année en France, on dénombre environ 2000 décès dont une grande partie pourrait être évitée par une meilleure connaissance de la maladie, en particulier sa gravité, son évolution, le suivi et le traitement. En France, il y a 3.5 millions de personnes touchées par l'asthme, plus particulièrement des enfants et des jeunes adultes. Aux Etats-Unis, il y a plus de 17 millions d'asthmatiques et il y en aurait plus de 300 millions dans le monde (Masoli et al. [2004]). Cette prévalence est en constante augmentation dans les pays en voie de développement et commence à se stabiliser dans les pays industrialisés.

Au sujet de la prévalence importante de l'asthme, il est important de noter qu'il semble de plus en plus probable qu'il y ait une relation entre l'indice de masse corporelle (IMC $= poids/taille^2$) et les symptômes de l'asthme. Depuis quelques années déjà, l'obésité est un problème majeur de santé publique dans les pays industrialisés. On note une augmentation de la prévalence de l'obésité pour les femmes et les hommes et pour tous les âges (adulte et enfant). Par exemple, une étude récente de la population américaine a montré que la prévalence du surpoids (IMC $\geqslant 25$) est de 65% et que celle de l'obésité (IMC $\geqslant 30$) est de 31% (Flegal et al. [2002]). Dans de nombreux pays européens (Angleterre, Allemagne, Pologne), le pourcentage de personnes obèses dépasse les 15%. Ces cinq dernières années, de nombreux travaux ont été publiés afin de montrer que le surpoids et l'obésité étaient des facteurs de risque pour l'asthme à la fois pour les adultes et pour les enfants (Weiss et Shore [2004]). De plus, l'IMC serait lié à la sévérité de l'asthme et une perte de poids améliorerait le fonctionnement pulmonaire, les symptômes de morbidité et le niveau général de santé (Akerman et al. [2004]). L'étude la plus récente suggère que l'obésité pourrait être en relation avec de nombreuses inflammations (Weisberg et al. [2003]).

L'asthme est une maladie capricieuse, complexe et d'évolution souvent imprévisible. Notre compréhension de l'évolution de l'asthme est encore insuffisante et bien des asthmes mortels pourraient être prévenus par une amélioration du suivi thérapeutique et une meilleure éducation du patient. Certaines études ont montré qu'en France, l'asthme insuffisamment pris en charge aboutirait trop fréquemment à des formes graves nécessitant une hospitalisation en urgence. De plus, on sait que près de 50% des personnes asthmatiques ne suivent pas le traitement prescrit. Il parait donc indispensable de mieux comprendre l'évolution de cette maladie. En effet, une meilleure connaissance de l'évolution des patients asthmatiques peut permettre (i) d'éduquer les malades pour qu'ils disposent des compétences nécessaires à la bonne prise en charge de leur maladie au quotidien et pour qu'ils adhérent en toute confiance aux prescriptions et aux conseils du médecin, (ii) aux cliniciens de mieux adapter les thérapeutiques et les protocoles de suivi afin d'éviter les changements d'états conduisant aux états de la maladie les plus sévères.

Cependant, dans la littérature, peu d'études se sont intéressées à la modélisation longitudinale de l'évolution de l'asthme dans un cadre Markovien. Jain [1986] a ajusté un modèle de Markov à temps discret, Redline et al. [1989] ont utilisé un modèle autorégressif de type Markovien, alors que Korn et Whittemore [1979] et Ware et al. [1988] présentent des méthodes d'analyse de données longitudinales appliquées à l'asthme. Plus récemment, Duchateau et al. [2003] ont utilisé un modèle de fragilité pour étudier les crises d'asthme ; Boudemaghe et Daures [2000] et Combescure et al. [2003] ont appliqué un modèle de Markov homogène sans covariable pour modéliser l'évolution du contrôle de l'asthme. On note aussi que certaines études (Chouaid et al. [2004], Price et Briggs [2002], Paltiel et al. [2001]) se sont intéressées à des analyses coût-efficacité du traitement de l'asthmatique.

Dans ce qui suit, nous présentons la base de données longitudinales (données répétées dans le temps) sur l'asthme. Nous définissons le contrôle de l'asthme, les structures des modèles utilisés et les différentes covariables étudiées. Enfin, nous décrivons succinctement les caractéristiques des patients lors de leur inclusion dans l'étude.

2 Présentation de la base de données

Afin d'étudier l'évolution de l'asthme et les facteurs qui conditionnent son évolution, l'ARIA (Association de Recherche en Intelligence Artificielle dans le cadre de l'asthme et des maladies respiratoires), coordonnée par le professeur P. GODARD, a mis au point en 1994 une cohorte de patients asthmatiques. Cette base est constituée de diverses informations sur des patients suivis entre 1994 et 2002. Elle est alimentée par un réseau d'experts en asthmologie issue des CHU français. Les dossiers proviennent essentiellement de patients suivis dans des centres hospitaliers de Bordeaux, Grenoble, Marseille, Montpellier, Paris, Strasbourg et Tarbes.

La base est constituée de patients adultes avec un asthme persistant diagnostiqué depuis au moins un an. Le diagnostic, le suivi et le traitement des patients sont effectués en accord avec les recommandations internationales du National Institutes of Health [1997]. La base de données reflète l'activité d'un hôpital avec des patients venant consulter à des intervalles de temps variables. Il a été demandé aux patients de venir consulter tous les trois mois ou de venir en fonction de leurs besoins personnels. La base comprend 871 patients, soit

au total 2386 consultations. A chaque visite, le médecin complète un formulaire (Juniper et al. [1999]) renseignant plus de 1700 variables liées à l'asthme. Dans toute cette thèse, on s'intéresse à l'étude de l'évolution de l'asthme et à l'analyse des données répétées dans le temps. Ainsi, nous utilisons pour l'analyse une sous base contenant uniquement les patients ayant au moins deux consultations. L'échantillon comprend 406 patients avec un total de 1639 consultations. Le temps de suivi chronologique est l'échelle de temps de référence. Ce temps est mesuré à partir de la première consultation du patient. Ainsi, quelque soit le moment d'entrée dans l'étude, le temps associé à la première consultation sera 0.

Les méthodes présentées aux chapitres III, IV et V prennent en compte certaines durées dans l'écriture des modèles. Le modèle semi-Markovien (*cf.* chapitre III) nécessite de définir la durée dans le dernier état visité. Les modèles de Markov non-homogènes (*cf.* chapitre IV et V) font intervenir la durée de suivi. La définition de ces quantités passe par la désignation d'une date de fin d'étude à partir de laquelle il n'est plus tenu compte de l'information. Cette date commune à tous les individus est appelée **date de point**. Dans notre cas, cette date est définie comme la dernière consultation renseignée dans la base de données, ici, le 4 février 2002.

2.1 Définition des états de contrôle

Afin d'étudier l'évolution de l'asthme par des modèles multi-états, il a d'abord été nécessaire de définir des états représentant globalement les différents états de santé dans lesquels peut se retrouver le patient lors de son suivi. Dans le cas de l'asthme, l'état de santé du patient est souvent établi à l'aide de la notion de **contrôle** (Godard et al. [1998], Boulet et al. [1999], Cockcroft et Swystum [1996]). Le contrôle de la maladie renvoie à l'appréciation des événements (cliniques, fonctionnels et thérapeutiques) sur une période plus courte (7 à 30 jours). Le terme contrôle est synonyme de « maîtrise » de la maladie et reflète d'un certain point de vue « l'activité » de la maladie sur quelques semaines. Le contrôle est devenu depuis quelques années déjà un marqueur fondamental dans le suivi de l'asthmatique. D'après les recommandations canadiennes (Boulet et al. [1999]), les critères pris en compte sont les symptômes (fréquence et intensité), le degré d'obstruction bronchique, sa variabilité et la pression thérapeutique (Tableau I.1).

Définissons les notions intervenant dans la définition du contrôle.

- Le débit expiratoire de pointe (DEP) qui mesure (en litre par minute) le souffle au moment où l'air sort le plus vite des bronches.
- Une exacerbation (ou attaque d'asthme) est un épisode de dégradation progressive, sur quelques jours, d'un ou plusieurs signes cliniques, ainsi que des paramètres fonctionnels d'obstruction bronchique. Une exacerbation est dite grave quand elle nécessite une corticothérapie per os (par voie orale) ou si le débit expiratoire de pointe (DEP) a chuté de plus de 30% au-dessous des valeurs initiales pendant 2 jours successifs.

- Le **contrôle optimal** est atteint lorsque tous les critères du tableau I.1 sont satisfaits (c'est-à-dire que pour tout $i = 1, ..., 7$, $S_i = 0$) et que le médecin obtient le meilleur compromis pour le patient entre le degré de contrôle, l'acceptation du traitement et la survenue éventuelle d'effets (définition de l'Agence National d'Accréditation et d'Evaluation en Santé (ANAES [2004])).

Critères	Valeurs
S_1 : Symptômes diurnes	$S_1 = 0$, si < 4 jours/semaine $S_1 = 1$, si $\geqslant 4$ jours/semaine
S_2 : Symptômes nocturnes	$S_2 = 0$, si < 1 nuit/semaine $S_2 = 1$, si $\geqslant 1$ nuit/semaine
S_3 : Activité physique	$S_3 = 0$, si absence de limitation $S_3 = 1$, si présence de limitation
S_4 : Exacerbations	$S_4 = 0$, si légères ou peu fréquentes $S_4 = 1$, si fréquentes
S_5 : Absentéisme professionnel ou scolaire	$S_5 = 0$, si pas d'absentéisme $S_5 = 1$, si absentéisme
S_6 : Utilisation de bêta-2 mimétiques	$S_6 = 0$, si < 4 doses/semaine $S_6 = 1$, si $\geqslant 4$ doses/semaine
S_7 : DEP (Débit Expiratoire de Pointe)	$S_7 = 0$, si > 85 % de la meilleure valeur personnelle $S_7 = 1$, si $\leqslant 85$ % de la meilleure valeur personnelle

Tab. I.1 – Critères pour le contrôle de l'asthme.

- Le **contrôle sous-optimal** est le minimum à rechercher chez tous les patients. Il est atteint lorsque tous les critères du tableau sont satisfaits ($\forall\ i = 1, ..., 7,\ S_i = 0$) mais le meilleur compromis pour le patient entre le degré de contrôle, l'acceptation du traitement et la survenue éventuelle d'effets n'est pas obtenu (ANAES [2004]).
- Le **contrôle inacceptable** est défini par la non-satisfaction d'un ou plusieurs critères de contrôle (c'est-à-dire qu'il existe i tel que $S_i = 1$, $i = 1, ..., 7$). Il représente un état insatisfaisant qui nécessite une adaptation de la prise en charge (ANAES [2004]).

Le stade sous-optimal est introduit car un contrôle optimal (ou excellent) est parfois impossible à obtenir. Les trois niveaux de contrôle permettent ainsi de définir l'état de santé du patient à chaque consultation.

Remarque 1 *Plusieurs références bibliographiques sur l'asthme sont disponibles dans Saint-Pierre et al. [2005a] et Combescure et al. [2003].*

Remarque 2 *La définition des états de contrôle à partir des variables de la base de données est détaillée en annexe page 187.*

2.2 Définitions des modèles

Au cours de ce document, les méthodes statistiques présentées sont appliquées à la base de données sur l'asthme. Trois types de modèles seront considérés : un modèle à trois états, un modèle à deux états et un modèle de survie.

Modèle à trois états

Les trois états de contrôle étant réversibles, le modèle suppose que toutes les transitions entre les états sont possibles. Le modèle est représenté par la figure I.1.

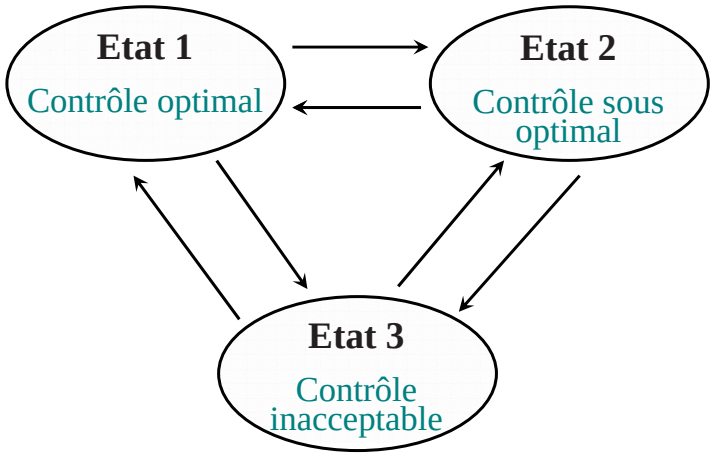


FIG. I.1 – Modèle à trois états de contrôle pour l'asthme.

Le tableau I.2 décrit la représentativité des 1233 transitions observées et renseignées dans la base de données. On comptabilise une transition $i \rightarrow j$ quand on observe un patient dans l'état i à une consultation et qu'il est observé dans l'état j à la consultation suivante $(i, j = 1, ..., 3)$. Dans de nombreux cas, on n'observe pas de changement d'état de contrôle d'une consultation à l'autre (transition $i \rightarrow i$). On dénombre un plus grand nombre de passage de $3 \rightarrow 1$ et $3 \rightarrow 2$ que de $1 \rightarrow 3$ et $2 \rightarrow 3$, ce qui est cohérent avec l'effort du médecin pour obtenir une amélioration de l'état de contrôle. Notons qu'entre deux consultations consécutives, il peut y avoir des changements d'état qui ne sont pas observés par les médecins et qui ne sont pas renseignés dans la base de données.

Transition	Effectif	Pourcentage
$1 \rightarrow 1$	130	10.6%
$1 \rightarrow 2$	95	7.7%
$1 \rightarrow 3$	44	3.6%
$2 \rightarrow 1$	112	9.1%
$2 \rightarrow 2$	169	13.7%
$2 \rightarrow 3$	71	5.8%
$3 \rightarrow 1$	114	9.2%
$3 \rightarrow 2$	120	9.7%
$3 \rightarrow 3$	378	30.6%

TAB. I.2 – Représentativité des transitions.

Ce modèle à trois états sera utilisé dans tous les chapitres sauf dans le chapitre V où la méthode ne permet pas d'étudier ce type de modèle.

Modèle à deux états

Devant l'impossibilité d'obtenir ce contrôle optimal chez tous les patients asthmatiques, un rapport de l'ANAES (Agence National d'Accréditation et d'Evaluation en Santé) datant de septembre 2004 (ANAES [2004]) préconise de fonder le contrôle de l'asthme sur des critères acceptables. Le contrôle de l'asthme est ainsi défini par deux états :

- un état de contrôle **acceptable** regroupant l'état **optimal** et l'état **sous-optimal**,
- un état de contrôle **inacceptable**.

Le modèle est représenté par la figure I.2.

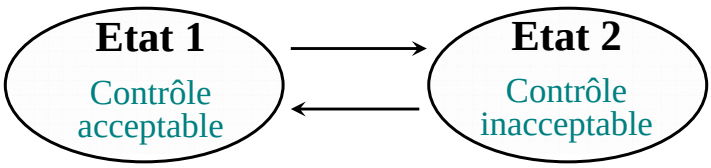


FIG. I.2 – Modèle à deux états de contrôle pour l'asthme.

Le tableau I.3 décrit la représentativité des 1233 transitions. On dénombre 115 transitions de l'état de contrôle **acceptable**) vers l'état de contrôle **inacceptable** ($1 \rightarrow 2$) et plus du double de transitions de l'état **inacceptable** vers l'état **acceptable** ($2 \rightarrow 1$).

Transition	Effectif	Pourcentage
$1 \rightarrow 1$	506	41%
$1 \rightarrow 2$	115	9.3%
$2 \rightarrow 1$	234	19%
$2 \rightarrow 2$	378	30.7%

TAB. I.3 – Représentativité des transitions.

Cette considération clinique permet de bien différencier le stade le plus grave de la maladie, mais permet aussi de considérer un modèle plus simple avec moins de paramètres. Ce modèle sera étudié dans le cadre du modèle de Markov homogène (chapitre II) pour approfondir les résultats cliniques et être en accord avec les récentes recommandations. Il sera également utilisé dans l'application des méthodes prenant en compte la censure informative dans un modèle à deux états réversibles (chapitre V).

Modèle de survie

Le chapitre V présente une méthode d'estimation pour données de survie sous hypothèse de censure informative. Dans le même temps, il est dans l'intérêt des pneumologues de bien comprendre le passage dans un état de contrôle inacceptable. Des données de survie sont obtenues en sélectionnant pour chaque patient, une séquence d'au moins deux consultations

consécutives commençant par un état acceptable et qui se termine soit par la fin du suivi soit par un état inacceptable. Le modèle de survie est représenté par la figure I.3.

Après cette sélection, la base comprend 334 patients et un total de 777 consultations. Deux types de suivi sont rencontrés :

- Les patients qui subissent l'événement, c'est-à-dire des patients qui passent dans l'état de contrôle inacceptable ($24,8\%$; $n = 83$).
- Les patients qui ne subissent par l'événement avant la fin de l'étude ($75,2\%$; $n = 251$).

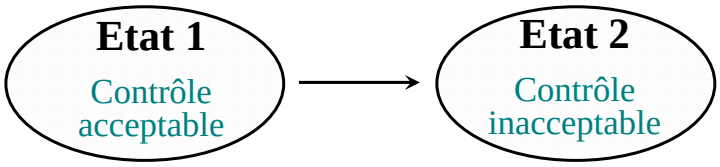


FIG. I.3 – Modèle de survie pour l'asthme.

2.3 Covariables

Les résultats cliniques présentés concernent essentiellement l'impact du surpoids sur l'évolution de l'asthme. Les covariables étudiées sont le traitement par corticoïdes oraux et inhalés, l'indice de masse corporelle, la présence d'exacerbations, la dose cumulée de corticoïdes oraux pendant l'année avant l'inclusion et la sévérité. Cependant, de nombreuses autres covariables ont été étudiées comme par exemple, la présence de rhinite, la durée de l'asthme, le tabagisme, l'âge ou le sexe, l'atopie (test positif à au moins un allergène). Les résultats concernant ces covariables ne sont pas présentés dans ce travail car ils sont moins intéressants et ont parfois une influence mineure sur la maladie. Par exemple, dans notre cas, le tabagisme ne modifie pas de manière significative l'évolution de la maladie (Ce résultat peut s'expliquer par le fait que les patients suivis ont des asthmes persistants ; ainsi, il y a peu de fumeurs dans la base de données et ceux qui fument semblent avoir les asthmes les moins sévères (Tableau I.6)). De nombreuses autres covariables n'ont pas été prises en compte car elles sont trop liées à la définition des états de contrôle (par exemple, le débit d'air expiré). Les covariables étudiées seront codées de manière binaire (codage (0,1)). Ce codage facilite l'interprétation clinique des résultats en terme de risque relatif, permet de mieux différencier certains groupes de patients et permet d'avoir des effectifs suffisants dans chacune des classes ce qui n'est pas toujours le cas avec un codage en plusieurs classes. Définissons un peu plus en détail les covariables et les codages considérés.

- La dose journalière de corticoïdes oraux : les patients n'ayant pas de traitement par corticoïdes oraux seront codés 0. La variable vaudra 1 sinon.
- La dose journalière de corticoïdes inhalés : les patients ayant une posologie inférieure ou égale à 500 μg seront codés 0. La variable vaudra 1 sinon.
- L'indice de masse corporelle : l'IMC est défini comme le poids (en kilogrammes) divisé par la taille (en mètres) au carré et sert d'indicateur du surpoids et de l'obésité. Un IMC < 20 correspond à un poids insuffisant, un IMC entre 20 et 25 correspond à un poids normal, un IMC entre 25 et 30 correspond au surpoids, un IMC entre 30

et 35 correspond à l'obésité et un IMC $\geqslant 35$ correspond à une obésité morbide. Afin d'étudier l'effet du surpoids, nous considèrerons le codage binaire suivant : les patients avec un IMC < 25 seront codés 0 et les patients avec un IMC $\geqslant 25$ seront codés 1. Ce codage permet d'avoir des effectifs suffisants dans chaque classe : en effet, on dénombre trop peu de patients obèses et de patients avec un poids trop faible dans la base.

- La sévérité de l'asthme : la sévérité tient compte de l'histoire de la maladie sur une période plus ou moins longue : 6 à 12 mois. Le niveau de sévérité est fondé sur l'importance des symptômes, le niveau de perturbation des paramètres fonctionnels et sur les paramètres thérapeutiques. La sévérité peut se définir par le niveau de pression thérapeutique nécessaire pour obtenir un contrôle durable de la maladie. Cette notion est différente de celle du contrôle de l'asthme (Combescure et al. [2003], Cockcroft et Swystum [1996]). La sévérité est codée en 5 catégories numérotées de 1 à 5 et de sévérité croissante. En suivant les recommandations des cliniciens, les modalités seront regroupées en deux classes : les patients sévères codés 1 (sévérité 4-5) et les patients non sévères (sévérité 1-2-3) codés 0. Cette manipulation permet de simplifier l'interprétation clinique en différenciant bien les asthmes les plus graves et permet d'avoir des effectifs plus conséquents dans chacune des classes.
- Le nombre d'exacerbations : cette variable renseigne sur la fréquence des crises ou attaques entre deux consultations. Elle sera codé 0 si le nombre d'exacerbations entre deux consultations est nul et 1 si elles sont au nombre de 1 ou plus. Ceci nous permettra d'une part de comparer les patients qui n'ont pas d'exacerbations à ceux qui en ont et d'autre part d'éviter le problème du très faible effectif dans certaines classes.
- La dose cumulée de corticoïdes oraux pendant l'année avant l'inclusion : cette variable renseigne sur les antécédents de traitement par corticoïdes oraux. Elle est particulièrement intéressante dans l'étude de l'impact du surpoids sur l'évolution de l'asthme. En effet, les traitements par corticoïdes oraux représentent un des biais les plus courants dans l'étude du surpoids car un de leurs effets secondaires est une prise de poids. Les patients ayant une dose totale inférieure ou égale à 2 grammes sont codés par 0 et patients ayant une dose supérieure à 2 grammes sont codés par 1. Ce seuil est fixé par des spécialistes car apparemment une dose cumulée sur l'année inférieure à 2 grammes n'entraînerait pas de prise de poids significative.

3 Description de la base à l'inclusion

Nous présentons ici certaines caractéristiques des patients à l'inclusion dans l'étude. Les variables présentées sont liées au traitement de l'asthme, l'état de la maladie mais aussi le tabagisme, le sexe, l'age et l'IMC. La table I.4, décrit les variables en fonction de l'état de contrôle à l'inclusion : état optimal et sous optimal d'un côté et état inacceptable de l'autre. Les tableaux I.5 et I.6 donnent une description des variables en fonction, respectivement, de l'IMC (IMC < 25 et IMC $\geqslant 25$) et de la sévérité (non sévère et sévère). Chaque tableau fournit les p-values obtenues avec les tests de Kruskall-Wallis et du Chi-2 pour repérer des liens significatifs entre les groupes pour chacune des variables.

Le tableau I.4 fournit des résultats prévisibles à savoir que la sévérité, la présence d'exacerbations et la présence d'une corticothérapie orale sont corrélées avec l'état de contrôle : ces

caractéristiques sont associées avec un contrôle inacceptable. De même, le débit expiratoire de pointe décroît significativement avec un état de contrôle inacceptable.

D'après le tableau I.5, plusieurs variables sont liées avec les groupes associés à l'indice de masse corporelle. En particulier, les patients en surpoids ont plus tendance à avoir un asthme sévère, une corticothérapie orale ou inhalée, ou des antécédents de corticoïdes oraux. Le surpoids semble lié à la sévérité et à la corticothérapie, en particulier par voie orale.

La description des patients à l'inclusion en fonction de la sévérité est présentée dans le tableau I.6. Le débit expiratoire de pointe, le traitement et l'IMC sont liés avec la sévérité. Comme précédemment, le surpoids est associé à la sévérité : près de 50% des patients sévères sont en surpoids. De plus, le tabagisme est associé avec la sévérité : en effet, il y a plus d'anciens fumeurs parmi les patients avec un asthme sévère. Ces patients semblent plus réceptifs aux recommandations médicales.

Variables	Etat de contrôle		p-value[#]
	Contrôle acceptable ($n = 163$)	Contrôle inacceptable ($n = 243$)	
Age	41.8 ± 17.8	42.4 ± 15.3	0.84
Femme	84 (51.5)	148 (60.9)	0.06
IMC < 25	109 (66.9)	151 (62.1)	0.33
Asthme sévère	12 (7.4)	79 (32.5)	< 0.01
Au moins une exacerbation	31 (19)	94 (38.7)	< 0.01
Rhinite	107 (69.9)	184 (76.7)	0.14
Atopie	78 (47.8)	122 (50.2)	0.64
Durée de l'asthme	17 ± 14.5	18.3 ± 14.5	0.36
Debit Expiratoire de Pointe	86.6 ± 17.8	71.2 ± 22.8	< 0.01
Antécédents corticoïdes oraux > 2g	14 (8.6)	40 (16.5)	0.03
Traitement avec corticoïdes oraux	19 (11.7)	54 (22.2)	< 0.01
Dose corticoïdes inhalés $> 500\ \mu g$	83 (50.9)	138 (56.8)	0.29
Tabagisme			
Non fumeur	124 (76.1)	166 (68.3)	0.21
Ancien fumeur	22 (13.5)	47 (19.3)	
Fumeur	17 (10.4)	30 (12.4)	

Les résultats sont présentés avec la moyenne ± écart-type ou le nombre (%).
[#] p-values calculées avec les tests de Kruskall-Wallis ou du Chi-2.

Tab. I.4 – Caractéristiques des patients lors de l'inclusion dans l'étude en fonction de l'état de contrôle (acceptable et inacceptable).

Variables	Indice de Masse Corporelle		p-value#
	IMC < 25 ($n = 260$)	IMC $\geqslant 25$ ($n = 146$)	
Age	38.1 ± 16.4	49.3 ± 13.7	< 0.01
Femme	160 (61.5)	72 (49.3)	0.02
Asthme sévère	48 (18.5)	43 (29.5)	0.01
Au moins une exacerbation	74 (28.5)	51 (34.9)	0.17
Rhinite	196 (75.4)	95 (65.1)	0.02
Atopie	145 (55.9)	55 (37.7)	< 0.01
Durée de l'asthme	16.5 ± 12.8	20.1 ± 16.9	0.16
Debit Expiratoire de Pointe	79.3 ± 21.9	73.8 ± 22.5	0.03
Antécédents corticoïdes oraux > 2g	27 (10.4)	27 (18.5)	0.03
Traitement avec corticoïdes oraux	36 (13.8)	37 (25.3)	< 0.01
Dose corticoïdes inhalés $> 500\ \mu g$	128 (49.2)	93 (63.7)	< 0.01
Tabagisme			
Non fumeur	183 (70.4)	107 (73.3)	0.78
Ancien fumeur	45 (17.3)	24 (16.4)	
Fumeur	32 (12.3)	15 (10.3)	

Les résultats sont présentés avec la moyenne ± écart-type ou le nombre (%).
p-values calculées avec les tests de Kruskall-Wallis ou du Chi-2.

TAB. I.5 – Caractéristiques des patients lors de l'inclusion dans l'étude en fonction de l'Indice de Masse Corporelle (IMC < 25 et IMC $\geqslant 25$).

Variables	Sévérité de l'asthme		p-value#
	Non sévère ($n = 315$)	Sévère ($n = 91$)	
Age	41.4 ± 16.5	44.7 ± 15.8	0.12
Femme	188 (59.7)	44 (48.3)	0.05
IMC < 25	212 (67.3)	48 (52.8)	0.01
Au moins une exacerbation	95 (30.2)	30 (33)	0.61
Rhinite	223 (73.8)	68 (74.7)	0.86
Atopie	162 (51.4)	38 (41.8)	0.1
Durée de l'asthme	17.6 ± 14.5	18.4 ± 14.9	0.78
Debit Expiratoire de Pointe	83.8 ± 17.3	55.1 ± 23.1	< 0.01
Antécédents corticoïdes oraux > 2g	20 (6.3)	34 (37.4)	< 0.01
Traitement avec corticoïdes oraux	37 (11.7)	36 (39.6)	< 0.01
Dose corticoïdes inhalés > 500 μg	153 (48.6)	68 (74.7)	< 0.01
Tabagisme			
Non fumeur	233 (74)	57 (62.6)	< 0.01
Ancien fumeur	43 (13.6)	26 (28.6)	
Fumeur	39 (12.4)	8 (8.8)	

Les résultats sont présentés avec la moyenne ± écart-type ou le nombre (%).
p-values calculées avec les tests de Kruskall-Wallis ou du Chi-2.

TAB. I.6 – Caractéristiques des patients lors de l'inclusion dans l'étude en fonction de la sévérité de l'asthme (non sévère et sévère).

Chapitre II

Modèle de Markov homogène et extensions

1 Introduction

Dans l'analyse de données longitudinales et particulièrement en épidémiologie, les sujets sont souvent suivis par intermittence et l'information recueillie se présente sous la forme de mesures ou d'états de santé en plusieurs temps discrets. Il est alors utile de modéliser le passage des individus entre les différents stades de la maladie. Dans ce contexte, les modèles multi-états qui fournissent une vision complète et détaillée de l'évolution de la maladie sont des méthodes intéressantes. Les modèles multi-états à temps continus sont particulièrement utiles quand les temps de consultation varient d'un individu à l'autre et quand les temps entre consultations sont variables.

Dans les modèles multi-états, l'hypothèse de Markov est couramment considérée. Cette hypothèse suppose que l'évolution future du processus dépend uniquement de l'état du processus au temps t, autrement dit, l'histoire du processus est résumée par l'état au temps t. Dans les modèles de Markov, on peut faire différentes hypothèses sur l'évolution du processus. En particulier, les paramètres du modèle peuvent dépendre ou non de la durée totale du suivi : on parle alors de modèle non-homogène et de modèle homogène. L'hypothèse d'homogénéité qui rend les forces de transition constantes au cours du temps simplifie la modélisation et donne des résultats facilement interprétables d'un point de vue clinique.

Les modèles de Markov homogènes ont été appliqués avec succès dans de nombreux domaines, en particulier en épidémiologie : par exemple, dans la modélisation des stades du cancer (Kay [1986], Hsieh et al. [2002]), des stades de l'infection par VIH (Gentleman et al. [1994], Longini et al. [1989]) ou encore des stades du diabète (Marshall et Jones [1995]). De nombreuses publications sont consacrées au développement de la méthodologie. Par exemple, Cook [1999] utilise des effets aléatoires dans le cas particulier d'un modèle à deux états, Satten [1999] utilise des effets aléatoires dans un modèle progressif, Aguirre-Hernandez et Farewell [2002] développent un test d'ajustement pour des modèles de Markov stationnaires, Kousignian et al. [2003] utilisent la méthode Monte Carlo Markov Chain pour estimer les paramètres. Récemment, plusieurs travaux traitent de l'estimation dans le cas

de données censurées par intervalles (Chen et Cook [2003], Satten [1999], Joly et al. [2002]). Une vue d'ensemble sur ces modèles est donnée par Hougaard [1999] et Commenges [1999].

Le modèle de Markov homogène est régulièrement utilisé, cependant il impose 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. Il est important d'envisager d'autres hypothèses moins restrictives par l'intermédiaire des modèles non-homogènes. Par exemple, l'ajustement d'un modèle à intensités constantes par périodes (Lindsay et Ryan [1993], Alioum et Commenges [2001]) permet de considérer un modèle non-homogène tout en conservant l'hypothèse d'homogénéité au sein d'une même période. Ce modèle sera utilisé pour tester la validité de l'hypothèse d'homogénéité. Une approche par les processus de comptage des modèles non-homogènes sera présentée au chapitre IV.

L'hypothèse d'homogénéité de la population est aussi contraignante et peut être plus raisonnable si elle est appliquée dans des sous-groupes. L'incorporation de covariables dans le modèle permet ainsi d'obtenir des estimations des probabilités de transition ajustées à chaque groupe de patients. De plus, en considérant un modèle à risques proportionnels, on peut étudier et quantifier l'impact des covariables sur l'évolution de la maladie. L'introduction de covariables dépendantes du temps permet d'être plus précis dans l'utilisation de certaines variables.

L'objectif de ce chapitre est (i) de décrire la méthodologie des modèles de Markov homogènes à temps continu et d'en donner quelques extensions possibles, (ii) d'appliquer ces méthodes à la modélisation de l'évolution du contrôle de l'asthme. Il reprend principalement des résultats faisant l'objet de deux publications liées aux modèles de Markov homogènes.

- Un premier article (Saint-Pierre et al. [2003]) publié dans la revue *Statistics in Medicine*. Ce travail reprend une grande partie des problématiques présentes dans ce type de modèle : estimation, modélisation avec covariables dépendantes du temps, modèle non-homogène par périodes, test des paramètres du modèle et des tests d'adéquation. Ce travail décrit également l'application de ces méthodes à un modèle à trois états de contrôle de l'asthme.
- Un deuxième article (Saint-Pierre et al. [2005a]) soumis dans la revue internationale de pneumologie *Chest*. Ce travail approche l'application des modèles Markoviens d'un point de vue clinique. Il traite de l'impact négatif du surpoids sur l'évolution du contrôle de l'asthme et a pour objet d'informer les médecins sur les conséquences du surpoids sur le contrôle du patient.

Ce chapitre se décompose en trois parties. La première partie est consacrée aux définitions et propriétés des processus Markoviens homogènes. La deuxième partie décrit le modèle Markovien homogène, l'écriture de la vraisemblance, l'incorporation de covariables et l'extension à un modèle homogène par périodes. La troisième partie discute l'application de ces méthodes à un modèle à trois états et présente des résultats sur l'impact négatif du surpoids dans un modèle à deux états de contrôle.

2 Définitions et notations

2.1 Processus

Soient $\mathcal{F}$ l'espace des temps et $(S, \mathcal{S})$ l'espace des états (un espace mesurable).

Définition 1 Un processus stochastique $(\Omega, \mathcal{A}, P, \{\xi(t), t \in \mathcal{F}\})$ est la donnée d'une fonction aléatoire dependante du temps $(t \in \mathcal{F})$ et du hasard $(\omega \in \Omega)$

$$\begin{aligned} \xi : \mathcal{F} \times \Omega &\to S \\ : (t, \omega) &\longmapsto \xi(t, \omega) \end{aligned}$$

telle que, pour $t \in \mathcal{F}$, la fonction $\omega \longmapsto \xi(t, \omega)$ est une variable aléatoire sur $(\Omega, \mathcal{A}, P)$ à valeurs dans $(S, \mathcal{S})$.

Pour un ω donné, la fonction $t \longmapsto \xi(t, \omega)$ est la trajectoire du processus.

La notion de processus élargit la notion de variable aléatoire. Un processus sera noté $\{\xi(t), t \in \mathcal{F}\}$.

Si S est fini ou dénombrable alors le processus $\{\xi(t), t \in \mathcal{F}\}$ est à espace d'états discret. Dans le cas contraire, $\{\xi(t), t \in \mathcal{F}\}$ est un processus à espace d'états continu.

Si $\mathcal{F} = \{t_n\}$ alors le processus est dit à temps discret. Si $\mathcal{F}$ est un intervalle de temps de l'axe $\mathbb{R}$ ou si l'ensemble des valeurs de $\mathcal{F}$ est continu, alors $\{\xi(t), t \in \mathcal{F}\}$ est un processus à temps continu (Tableau II.1). Nous parlerons de processus quand l'ensemble des valeurs de $\mathcal{F}$ est continu et de chaîne dans le cas contraire. Notons que la variable t représente en général le temps mais t peut être de dimension multiple. Les résultats présentés dans cette thèse concernent des processus à temps continu et à espace d'états fini.

		Espace des temps F	
		Discret	Continu
Espace d'état K	Discret	Chaîne stochastique à espace d'état discret	Processus stochastique à espace d'état discret
	Continu	Chaîne stochastique à espace d'état continu	Processus stochastique à espace d'état continu

Tab. II.1 – Processus stochastiques.

Exemple : Prenons l'exemple d'un service de chirurgie.

– Le nombre d'opérations en attente au temps de la $t^{ème}$ opération est un processus à espace de temps discret et à espace d'états discret (« nombre d'opérations »)

$$\mathcal{F} = \{1, 2, ...\} \text{ et } S = \{0, 1, 2, ...\}.$$

– Le nombre d'opérations en cours au temps t, forme un processus à temps continu et à espace d'états discret

$$\mathcal{F} = \{t \mid 0 \leqslant t < \infty\} \text{ et } S = \{0, 1, 2, ...\}.$$

– Le temps d'attente pour la $t^{ème}$ opération est un processus à espace de temps discret et à espace d'états continu

$$\mathcal{F} = \{1, 2, ...\} \text{ et } S = \{x \mid 0 \leqslant x < \infty\}.$$

– Le temps total cumulé de toutes les opérations en cours au temps t forme un processus à temps continu et à espace d'états continu

$$\mathcal{F} = \{t \mid 0 \leqslant t < \infty\} \text{ et } S = \{x \mid 0 \leqslant x < \infty\}.$$

2.2 Processus Markovien

Un processus de Markov $\{X(t); t \in \mathcal{F}\}$ à temps continu et à espace d'états fini est un processus dont l'évolution future $\{X(t); t \geqslant s\}$ ne dépend de son passé qu'à travers son état à l'instant s, pour tout $t \geqslant s$

$$\Pr\{X(t) = j \mid X(r) = x_r, r \leqslant s\} = \Pr\{X(t) = j \mid X(s) = x_s\}.$$

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

Définition 2 (Chaîne de Markov) Un **processus de Markov** à temps continu et à espace d'états fini $S = \{1, ..., k\}$ est complètement défini par

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

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

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

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

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

avec

$$\mathbf{P}(s,s) = \mathbf{Id}, \text{ et } \sum_{j=1}^{k} p_{ij}(s,t) = 1 \text{ pour tout } h \text{ et } 0 \leqslant s \leqslant t.$$

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

$$p_{ij}(s,t) = \sum_{k \in S} p_{ik}(s,u) p_{kj}(u,t), \tag{II.1}$$

Cette propriété est appelée équation de Chapman-Kolmogorov. Sous forme matricielle, l'équation (II.1) s'écrit

$$\mathbf{P}(s,t) = \mathbf{P}(s,u)\mathbf{P}(u,t) \qquad \forall\ s \leqslant u \leqslant t.$$

Les intensités de transition sont d'autres paramètres qui permettent de définir un processus de Markov. Soit $\mathbf{Q}(.) = (q_{ij}(.))_{i,j}$ la matrice $k \times k$ des intensités de transition,

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

Remarque 3 *La propriété de Markov définie ici est dite d'ordre 1 car seul l'état précédent résume le passé. De manière similaire, on définit les chaînes de Markov d'ordre r pour lesquelles le passé sera résumé par les r états précédents. Ces processus peuvent être intéressants dans les cas où l'évolution du processus est fortement liée au passé (étude d'une maladie par exemple). Cependant, ils sont peu étudiés et peu utilisés car les définitions des paramètres font intervenir des temps de transition supplémentaires, ce qui complique considérablement les définitions et la méthodologie. Dans le cas des chaînes de Markov à temps discret, il est plus simple d'utiliser des ordres supérieurs mais le nombre de paramètres augmente exponentiellement avec la mémoire de la chaîne ce qui rend la méthode peu exploitable en pratique.*

2.3 Processus Markovien homogène

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

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

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

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

A l'aide de l'équation de Chapman-Kolmogorov (II.1), on obtient une équation différentielle qui fait le lien entre la matrice des probabilités de transition et les intensités de transition,

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

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

$$\mathbf{P}(0,t) = \exp(\mathbf{Q} \times t). \tag{II.2}$$

Ainsi, pour le terme $P_{ii}(.)$, $i \in S$, cette équation donne,

$$\begin{aligned} P_{ii}(t) &= \exp(q_{ii} \times t) \\ &= \exp(-\sum_{j \neq i} q_{ij} \times t). \end{aligned}$$

$P_{ii}(t) = p_{ii}(0,t)$ correspond à la probabilité que le processus soit dans l'état i au temps t sachant qu'il était dans l'état i au temps 0 (L'équation de Chapman-Kolmogorov (II.1) exprime le fait que le processus a pu transiter par d'autres états pendant cette période). En pratique, les observations des sujets sont discrètes : on ne dispose d'aucune information sur l'état du patient entre deux consultations, ainsi, un nombre quelconque de transitions (non observées) peuvent avoir eu lieu. Par conséquent, dans le cas des modèles avec des états réversibles, les durées de séjour dans un état donné ne sont pas disponibles. Par contre, si les états du modèle sont hiérarchiques (pas de retour possible), ce problème ne se pose plus (car si l'individu est observé dans le même état, il n'y a pas eu de transitions non observées). Dans certaines applications, on peut régler le problème de la réversibilité en supposant que les temps d'observation sont suffisamment rapprochés pour pouvoir considérer qu'il s'agit d'une observation continue du sujet. Dans ce cas, tous les changements d'états sont observés et les durées de séjour dans les états sont disponibles. La variable aléatoire $T =$ « temps passé dans l'état i avant de le quitter » suit une loi exponentielle de paramètre $\sum_{j \neq i} q_{ij}$. Dans les modèles Markoviens homogènes, les distributions du temps d'attente 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 $\mathbb{E}(T) = 1/\sum_{j \neq i} q_{ij}$.

3 Modèle de Markov homogène

En épidémiologie, par exemple, les processus de Markov homogènes peuvent être utilisés pour modéliser l'évolution d'une maladie. L'hypothèse d'homogénéité permet d'avoir une définition simplifiée des probabilités de transition à partir des intensités de transition (constantes).

3.1 Vraisemblance

Soit $\{X(t), t \in \mathcal{F}\}$ un processus de Markov homogène à espace d'états fini $S = \{1, ..., k\}$. Supposons que chaque individu se déplace indépendamment entre les états suivant le processus X. Les données observées pour le sujet h , $h = 1, ..., n$ sont

– les temps successifs de suivi $T_{h,0} < T_{h,1} < ..., < T_{h,n_h}$

– les états occupés $x_{h,j} = X(T_{h,j})$, $j = 0, 1, ..., n_h$.

La contribution à la vraisemblance pour un individu, est le produit des probabilités associées à chaque transition observée. Ainsi, la contribution de l'individu h à la vraisemblance est

$$l_h = \mathbf{P}_0[x_{h,0}] \times \prod_{j=1}^{n_h} P_{x_{h,j-1},x_{h,j}}(T_{h,j} - T_{h,j-1}).$$

La vraisemblance totale est le produit des contributions individuelles,

$$L = \prod_{h=1}^{n} l_h. \tag{II.3}$$

Les probabilités dans la vraisemblance s'obtenant par l'équation (II.2), la vraisemblance dépend alors uniquement des intensités de transition. La méthode du maximum de vraisemblance permet d'obtenir les estimations des paramètres. La méthode de la diagonalisation ou un développement en série entière peuvent être utilisés pour calculer l'exponentielle matricielle. L'algorithme de quasi-Newton est une méthode simple et efficace pour calculer l'estimateur du maximum de vraisemblance. Les estimations des écarts-types asymptotiques peuvent être obtenues à partir de la matrice d'information empirique. Pour une discussion plus approfondie sur l'estimation dans ces modèles, on pourra consulter Kalbfleisch et Lawless [1985].

3.2 Incorporation de covariables

Dans bien des applications, on dispose de covariables sur chaque individu et il est particulièrement intéressant d'étudier l'impact de ces covariables sur les paramètres du modèle. Le modèle peut être étendu de manière simple en considérant un modèle de régression à intensités proportionnelles (Cox [1972]). Les intensités de transition peuvent s'écrire

$$q_{ij}(\mathbf{z}) = q_{ij0} \exp(\boldsymbol{\beta}_{ij}^{'} \mathbf{z}) \qquad i \neq j, \tag{II.4}$$

avec $\mathbf{z}$ un vecteur de covariables indépendantes du temps de dimension s, $\boldsymbol{\beta}_{ij}$ un vecteur de s coefficients de régression et q_{ij0} l'intensité de transition de base associée à la transition de l'état i vers l'état j. Ce modèle log-linéaire pour les intensités de transition est souvent utilisé dans la littérature (Andersen et al. [1991], Marshall et Jones [1995]). En effet, les estimations des intensités de transition sont toujours positives quelles que soient les valeurs de $\mathbf{z}$ et de $\boldsymbol{\beta}_{ij}$. De plus, ce modèle fournit des résultats en terme de risques relatifs qui sont facilement interprétables (comme dans le modèle de Cox à risques proportionnels). D'autres paramétrisations peuvent être plus appropriées dans des applications particulières.

Le modèle peut être adapté afin de prendre en compte des covariables dépendantes du temps. En effet, la vraisemblance est le produit des contributions associées à chaque

transition observée dans la base. Le terme $P_{ij}(t-s \mid \mathbf{z})$, peut être remplacé par $P_{ij}(t-s \mid \mathbf{z}(s))$ en supposant que les valeurs des covariables dépendantes du temps restent constantes entre les deux temps consécutifs s et t. Cette hypothèse de constance des covariables entre deux consultations peut cependant être forte dans certaines situations, en particulier, lorsque le temps entre deux consultations est important.

3.3 Modèle de Markov homogène par périodes

Le modèle de Markov homogène avec covariables dépendantes du temps permet de considérer un modèle non-homogène ou plus précisément un modèle homogène par périodes. L'hypothèse d'homogénéité étant une hypothèse forte, il est utile de la relâcher en considérant un modèle où les intensités de transition sont constantes au sein d'une même période mais sont différentes d'une période à une autre.

Considérons une subdivision de l'axe du temps $[\tau_{l-1}, \tau_l)$, où $l = 1, ..., r+1$, $\tau_0 = 0$ et $\tau_{r+1} = +\infty$, et supposons que pour chaque transition, les intensités sont constantes dans chaque intervalle. Le temps est mesuré depuis l'origine du processus. Soit $\mathbf{z}^*(t) = (z_0^*(t), z_1^*(t), ..., z_r^*(t))'$ un vecteur de covariables artificielles défini par

$$\begin{aligned} z_0^*(t) &= 0 \qquad\qquad \forall t \\ z_l^*(t) &= \begin{cases} 0 & \text{si } \tau_0 \leqslant t < \tau_l \\ 1 & \text{si } t \geqslant \tau_l \end{cases} \qquad \text{pour } l = 1, 2, ..., r. \end{aligned}$$

Les intensités de transition sont données par $(i \neq j)$,

$$\begin{aligned} q_{ij}(t \mid \mathbf{z}^*(t)) &= q_{ij0} \exp\{\boldsymbol{\alpha}_{ij}' \mathbf{z}^*(t)\} \\ &= \begin{cases} q_{ij0} & \text{if } \tau_0 \leqslant t < \tau_1 \\ q_{ij1} = q_{ij0} \exp\{\alpha_{ij,1}\} & \text{if } \tau_1 \leqslant t < \tau_2 \\ \vdots & \\ q_{ijr} = q_{ij0} \exp\{\alpha_{ij,1} + \alpha_{ij,2} + ... + \alpha_{ij,r}\} & \text{if } t \geqslant \tau_r. \end{cases} \end{aligned}$$

Les paramètres du modèle sont l'intensité de transition de base q_{ij0} et le vecteur des coefficients de régression $\boldsymbol{\alpha}_{ij}$ Les intensités de transition sont des fonctions en escalier définies sur les intervalles pré-spécifiés. On note que ce modèle généralise le modèle homogène en considérant $r = 0$.

Pour ajuster ce modèle, la vraisemblance doit être modifiée. En effet, les intensités de transition sont constantes sur un même intervalle mais sont différentes d'un intervalle à l'autre. Pour tout $i, j \in S$, $p_{ij}^{(l)}(t)$ représente la probabilité de transition associée à l'intervalle $[\tau_{l-1}, \tau_l)$, ainsi pour $l = 1, ..., r+1$,

$$p_{ij}(s, s+t) = p_{ij}^{(l)}(t) \qquad \text{si } \tau_{l-1} \leqslant s < s+t < \tau_l.$$

Pour tout t, considérons $I_t \in \{1, 2, ..., r+1\}$ qui renseigne sur l'intervalle de temps de la forme $[\tau_{l-1}, \tau_l)$ contenant t. Pour simplifier les notations, considérons X_1 et X_2 les états occupés par un individu pour deux temps consécutifs de consultation T_1 et T_2. Les contributions à la vraisemblance pour lesquelles les temps consécutifs T_1 et T_2 n'appartiennent pas aux mêmes intervalles doivent être réécrites à l'aide de l'équation de Chapman-Kolmogorov (II.1). De manière générale, la contribution à la vraisemblance pour cette observation s'écrit

$$p_{X_1X_2}\{T_1,T_2 \mid \mathbf{z}^*(t), T_1 \leqslant t < T_2\} = \begin{array}{l} \sum_{k_1\in S}\sum_{k_2\in S}\cdots\sum_{k_v\in S}[p^{(I_{T_1})}_{X_1k_1}\{\tau_{I_{T_1}} - T_1 \mid \mathbf{z}^*(T_1)\} \\ \times p^{(I_{T_1}+1)}_{k_1k_2}\{\tau_{I_{T_1}+1} - \tau_{I_{T_1}} \mid \mathbf{z}^*(\tau_{I_{T_1}})\} \times \cdots \\ \times p^{(I_{T_2})}_{k_vX_2}\{T_2 - \tau_{I_{T_2}-1} \mid \mathbf{z}^*(\tau_{I_{T_2}-1})\}], \end{array}$$

où $v = I_{T_2} - I_{T_1}$. La vraisemblance s'obtient comme précédemment et l'estimation des paramètres par maximisation de cette vraisemblance (Kay [1986], Alioum et Commenges [2001]).

Il est important de noter que le modèle homogène par périodes peut aisément être combiné avec le modèle avec covariables de manière à prendre en compte dans la modélisation à la fois des covariables et des intensités constantes par périodes. Cette extension est naturelle car les deux modèles sont pertinents et mieux adaptés que le modèle homogène simple. Cependant, en pratique, on peut vite être confronté à un nombre trop important de paramètres à estimer.

L'utilisation des modèles de rupture peut permettre de déterminer les seuils de la partition. Par exemple, dans le cas où $r = 1$ (modèles avec deux périodes), ces modèles peuvent permettre de déterminer de manière statistique la valeur optimale pour la partition. Ce seuil permet d'obtenir la meilleure approximation des intensités de transition par une fonction constante par morceaux.

3.4 Tests d'hypothèses et d'adéquation

Au cours de l'analyse et dans l'interprétation des résultats, il est souvent intéressant et nécessaire de tester les paramètres et les hypothèses du modèle. Nous rappelons ici succinctement certains tests possibles dans ce type de modèle.

Afin de simplifier le modèle et de réduire le nombre de paramètres, on peut tester des hypothèses de la forme $H_0 : q_{ij} = 0$, $H_0 : q_{ij} = q_{hk}$ ou encore $H_0 : \beta_{ij,k} = 0$ à l'aide du test du rapport de vraisemblance ou du test de Wald (Cook et al. [2002], Marshall et Jones [1995], Self et Liang [1987]). Il est particulièrement intéressant d'étudier les coefficients de régression. En effet, si le coefficient $\beta_{ij,k}$ est statistiquement différent de zéro, alors il y aura une relation entre la transition de l'état i vers l'état j et la $k^{ième}$ covariable.

Une hypothèse importante du modèle est l'hypothèse d'homogénéité (intensités de transitions constantes au cours du temps). Cette hypothèse peut par exemple être testée en comparant le modèle homogène par périodes avec un modèle homogène : si le modèle homogène par périodes s'ajuste mieux, alors l'hypothèse d'homogénéité sera trop restrictive (Kalbfleisch et Lawless [1985], de Stavola [1988]). L'hypothèse de Markov peut être testée en considérant par exemple une covariable « durée de séjour dans l'état » (Dans un modèle Markovien, les intensités de transition ne dépendent pas du temps écoulé dans l'état avant de transiter). L'hypothèse de proportionnalité des risques peut être vérifiée en utilisant des covariables artificielles par exemple (*cf.* chapitre IV page 96). Plusieurs autres hypothèses peuvent également être testées en incluant des covariables spécifiques dans le modèle (Kay [1986]). Enfin, il est possible de se faire une idée de l'adéquation du modèle en comparant les effectifs observés et les effectifs théoriques obtenus par le modèle (Kalbfleisch et Lawless [1985], Aguirre-Hernandez et Farewell [2002]).

4 Application à l'asthme

L'ensemble de l'analyse statistique présente dans ce document a été réalisé avec le logiciel *S-Plus* (et/ou avec le logiciel *R*). Les fonctions *optim*() de *R* et *nlminb*() de *S-Plus* sont utilisées pour obtenir les estimateurs du maximum de vraisemblance. La fonction *optim*() fournit une estimation de la matrice hessienne. Ces fonctions font appel à l'algorithme de quasi-Newton (Nocedal et Wright [1999]). En ce qui concerne l'initialisation des paramètres, ceux associés aux intensités de transition sont estimés à partir de simples proportions et ceux associés aux coefficients de régression sont initialisés à 0 (aucun effet).

4.1 Modèle à trois états

Dans cette partie, un modèle de Markov à trois états (Figure II.1) est considéré pour modéliser l'évolution du contrôle de l'asthme. La notion de contrôle permettant de définir les états de santé et le modèle sont présentés au chapitre I page 9.

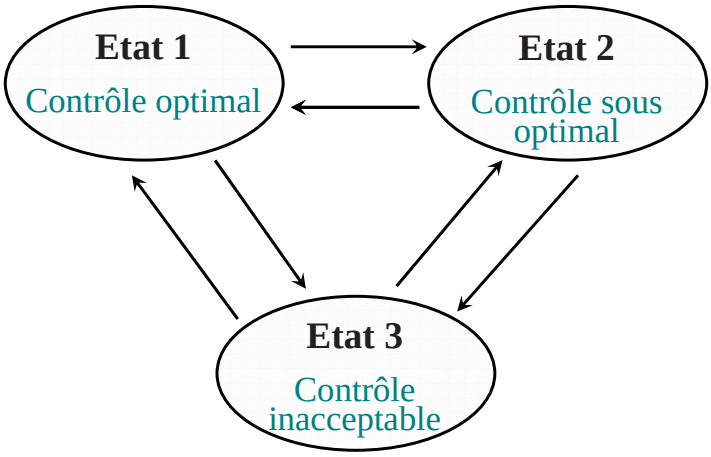


FIG. II.1 – Modèle à trois états de contrôle pour l'asthme.

Les résultats présentés concernent les covariables dépendantes du temps suivantes :
- l'indice de masse corporelle à chaque consultation : codée 0 si IMC < 25, 1 sinon,
- La sévérité de l'asthme à chaque consultation : codée 0 si le patient est non sévère, 1 sinon,
- La dose de corticoïdes oraux à chaque consultation : codée 0 si la dose est égale à 0 mg, 1 sinon.

Dans un premier temps, un modèle avec une covariable est ajusté pour chacune des covariables décrites précédemment. Le tableau II.2 fournit les résultats des estimations des coefficients de régression, des écarts-types et le maximum des p obtenus en testant $\beta_{ij} = 0$ par le test de Wald et par le test du maximum de vraisemblance (LRT).

Transition	IMC			Sévérité			Corticoïdes Oraux		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
$1 \to 2$	-0.835	(0.537)	(0.12)	-0.974	(0.439)	(0.03)	-1.194	(0.482)	(0.01)
$1 \to 3$	-0.227	(0.463)	(0.62)	-0.581	(0.473)	(0.22)	-1.229	(0.506)	(0.01)
$2 \to 1$	-0.693	(0.510)	(0.17)	-1.615	(0.453)	(<0.01)	-2.080	(0.480)	(<0.01)
$2 \to 3$	-0.029	(0.467)	(0.95)	4.715	(0.524)	(<0.01)	3.995	(0.545)	(<0.01)
$3 \to 1$	-1.424	(0.370)	(<0.01)	-1.805	(0.306)	(<0.01)	-1.350	(0.402)	(<0.01)
$3 \to 2$	-0.158	(0.354)	(0.65)	0.746	(0.438)	(0.09)	-0.761	(0.487)	(0.11)

[1] estimations des écarts-types.
[2] maximum de p avec le test de Wald et du LRT pour $H_0 : \beta_{ij} = 0$.

TAB. II.2 – Estimations des coefficients de régression dans un modèle homogène avec une covariable.

Plusieurs covariables influencent de manière significative les différentes transitions du modèle, en particulier l'indice de masse corporelle : le coefficient associé à la transition $3 \to 1$ est le plus important et il est négatif ($\beta_{31} = -1.424$). Ce résultat peut être interprété en terme de risques relatifs : en effet, le risque associé à la transition $3 \to 1$ est divisé par $exp(1.424)$ pour les patients en surpoids. Autrement dit, les patients en surpoids ont moins de chances de passer d'un état inacceptable vers un état optimal. Les patients sévères ont plus de chances de passer de $2 \to 3$ ($\beta_{23} = 4.715$) et moins de chances de passer de $3 \to 1$ ($\beta_{31} = -1.805$). Ces résultats sur la sévérité confirment ceux obtenus précédemment (Combescure et al. [2003]) en stratifiant la base de données (résultats non présentés). Les effets négatifs du traitement par corticoïdes oraux s'expliquent par le fait que ce sont souvent les patients avec un asthme difficile à contrôler qui ont ce type de traitement. Il serait intéressant d'utiliser un modèle avec plusieurs covariables pour avoir des résultats ajustés et pour éviter les biais de confusion. Mais, avec cette base de données, on rencontre des difficultés dans les procédures de maximisation de la vraisemblance. Dans un modèle avec deux covariables, les estimations sont peu fiables et dépendent fortement des conditions initiales.

Transition	Modèle deux périodes ($\tau_1 = 100$)		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$
$1 \to 2$	-0.945	(0.417)	(0.02)
$1 \to 3$	-1.232	(0.384)	(<0.01)
$2 \to 1$	-1.703	(0.437)	(<0.01)
$2 \to 3$	2.923	(0.535)	(<0.01)
$3 \to 1$	-2.825	(0.474)	(<0.01)
$3 \to 2$	0.025	(0.378)	(0.93)

[1] estimations des écarts-types.
[2] maximum des p avec le test de Wald et du LRT pour $H_0 : \beta_{ij} = 0$.

TAB. II.3 – Estimations des coefficients de régression dans un modèle homogène par période (2 périodes, $\tau_1 = 100$).

Dans un second temps, un modèle homogène par périodes est ajusté afin de vérifier la validité de l'hypothèse d'homogénéité. Un modèle avec deux périodes est utilisé : les intensités de transition sont constantes sur $[0, \tau_1[$ et sur $[\tau_1, \infty[$. Le tableau II.3 donne les estimations des coefficients de régression dans un modèle homogène par périodes (deux périodes, $\tau_1 = 100$ jours). Seul le coefficient associé à la transition $3 \rightarrow 2$ n'est pas significatif. Le coefficient β_{23} est statistiquement positif donc le risque pour cette transition est accéléré dans l'intervalle $[100, \infty[$, les autres coefficients sont négatifs $i.e$, le risque diminue avec le temps. Le test du rapport de vraisemblance et le test de Wald peuvent être utilisés pour comparer le modèle homogène par périodes et le modèle homogène (modèles emboités). Pour les seuils suivants : $\tau_1 = 50, 100, 150, 200, 250$ et 300 jours, le modèle homogène par périodes s'ajuste toujours mieux que le modèle homogène. Ainsi, il semble que l'hypothèse de constance des intensités de transition au cours du temps (homogénéité) soit trop restrictive dans le cas de l'asthme.

En supposant que l'information est disponible en continu, tous les changements d'états sont observés et les durées de séjours sont connues (cette hypothèse entraîne une surestimation des durées de séjours). Ainsi, on peut tester l'hypothèse de Markov en considérant une covariable « temps de séjour dans l'état avant de transiter ». Si cette variable influence significativement l'évolution du processus alors l'hypothèse de Markov sera abusive. En effet, l'hypothèse de Markov implique que les intensités de transition ne dépendent pas du temps de séjour. On peut par exemple considérer une covariable binaire pour coder la durée de séjour avec plusieurs seuils différents ($t = 50, 100, 150, 200, 250, 300$). Dans le cas de l'asthme, si on suppose que les patients viennent consulter quand ils ressentent un changement d'état et que les temps de consultation sont suffisament proches, on peut considérer que toutes les transitions sont observées (durées de séjour connues). Dans notre cas, la covariable artificielle binaire renseignant sur la durée dans l'état avant de transiter modifie de manière significative les intensités de transition (pour plusieurs seuils) : le plus souvent, plus la durée écoulée avant de transiter est grande moins l'individu a de chances de changer d'état (résultats non présentés). Ces résultats montrent les limites de l'hypothèse de Markov et suggèrent l'utilisation de modèles de semi-Markov.

4.2 Modèle à deux états

L'objectif de cette section est de :

(i) compléter la présentation des résultats obtenus en appliquant les modèles décrits dans ce chapitre,

(ii) fournir des résultats aboutissant à des orientations cliniques.

Dans ce qui suit, les résultats portent sur un modèle avec uniquement deux états de contrôle présenté au chapitre I page 14 (Figure II.2). En effet, un récent rapport de l'ANAES (Agence National d'Accréditation et d'Evaluation en Santé) (ANAES [2004]) préconisent de définir le contrôle de l'asthme en deux états afin de bien différencier le stade le plus grave de la maladie.

Cette considération clinique fournit un modèle plus simple ce qui permet en contre partie d'introduire plus de covariables dans la modélisation. Le modèle à deux états permet

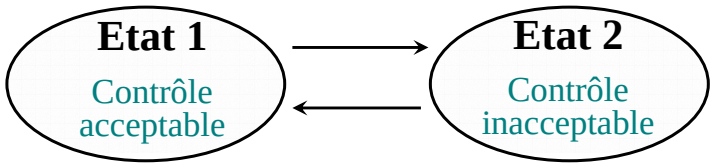


FIG. II.2 – Modèle à deux états de contrôle pour l'asthme.

d'inclure jusqu'à quatre covariables tout en conservant des estimations fiables. Les résultats ainsi obtenus sont ajustés sur plusieurs covariables ce qui permet de réduire les biais de confusion et d'avoir des résultats plus intéressants d'un point de vue clinique.

L'impact du surpoids sur l'évolution de l'asthme nous a particulièrement intéressé. En effet, plusieurs publications récentes montrent que le surpoids semble avoir un lien avec l'évolution de l'asthme. Cependant, un des biais les plus communs quand on étudie le surpoids est lié au traitement par corticoïdes oraux. En effet, la prise de poids est un effet secondaire de ces traitements. De ce fait, il peut y avoir une confusion entre l'influence des médicaments et l'influence du surpoids. L'utilisation d'un modèle avec plusieurs covariables permet d'observer l'effet du surpoids en ajustant les résultats sur ces traitements. Les résultats obtenus ont fait l'objet d'un article soumis dans une revue de pneumologie (Saint-Pierre et al. [2005a]). Cet article intitulé « Overweighted asthmatics are more difficult to control », approche le problème d'un point de vue clinique : il donne entre autres, plusieurs références bibliographiques sur la relation entre l'asthme et le surpoids et présente une discussion clinique détaillée des résultats.

L'objectif étant d'étudier l'impact du surpoids en tenant compte des possibles facteurs de confusion, les covariables suivantes sont sélectionnées :

- l'indice de masse corporelle à chaque consultation : codée 0 si IMC < 25, 1 sinon,
- la sévérité à chaque consultation : codée 0 si le patient est non sévère, 1 sinon,
- la dose de corticoïdes oraux à chaque consultation : codée 0 si la dose est égale à 0 mg, 1 sinon,
- la dose cumulée de corticoïdes oraux pendant l'année avant l'inclusion : codée 0 si le patient a une dose cumulée inférieure ou égale à 2 grammes, 1 sinon.

Le tableau II.4 présente les résultats des estimations des coefficients de régression dans un modèle avec uniquement l'IMC en covariable (univarié) et dans un modèle avec les quatre covariables (multivarié).

Le modèle univarié montre un effet significatif du surpoids sur l'évolution de la maladie. Les patients en surpoids ont un risque plus faible (divisé par $exp(0.801) \approx 2.2$) de passer d'un état de contrôle inacceptable vers un état acceptable ($2 \rightarrow 1$). Dans le modèle multivarié, l'effet du surpoids est toujours significatif même s'il est légèrement atténué ($\beta_{21} = -0.637$). Ainsi après ajustement sur certains facteurs de confusion connus, le surpoids diminue de manière significative le risque de retour vers un état stable (divisé par 1.9). Ces résultats confirment ceux obtenus avec un modèle à trois états à savoir que les patients en surpoids avaient un risque plus faible de quitter l'état inacceptable.

Transition	Covariables	Modèle univarié (IMC)			Modèle multivarié		
		$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
$1 \rightarrow 2$	IMC	-0.129	(0.247)	(0.60)	-0.174	(0.278)	(0.53)
	Sévérité	0.665	(0.272)	(0.04)	0.820	(0.305)	(<0.01)
	Corticoïdes Oraux	0.110	(0.269)	(0.68)	-0.422	(0.305)	(0.17)
	Antécédent Corticoïde	0.651	(0.262)	(0.01)	0.498	(0.299)	(0.10)
$2 \rightarrow 1$	IMC	-0.801	(0.184)	(<0.01)	-0.637	(0.219)	(<0.01)
	Sévérité	-0.726	(0.203)	(<0.01)	-0.062	(0.255)	(0.81)
	Corticoïdes Oraux	-1.002	(0.209)	(<0.01)	-0.693	(0.248)	(<0.01)
	Antécédent Corticoïde	-0.852	(0.212)	(<0.01)	-0.312	(0.266)	(0.24)

[1] estimations des écarts-types.
[2] maximum des p avec le test de Wald et du LRT pour $H_0 : \beta_{ij} = 0$.

TAB. II.4 – Estimations des coefficients de régression dans un modèle homogène avec une covariable (IMC) et dans un modèle avec quatre covariables.

Dans les modèles de Markov, il est particulièrement intéressant d'étudier les probabilités de transition au cours du temps. Ces courbes sont facilement interprétables et permettent de comparer les probabilités de transition en fonction des caractéristiques des patients. Les probabilités sont obtenues par l'équation suivante :

$$\mathbf{P}(0,t) = \exp(\mathbf{Q}t)$$

avec $\mathbf{P}(0,t) = (\Pr\{X(t) = j \mid X(0) = i\})_{i,j}$, $\mathbf{Q} = (q_{ij})_{i,j}$ et

$$q_{ij} = q_{ij0} \exp(\beta_{ij}^T Cov)$$

La figure II.3 permet d'observer l'impact du surpoids sur la probabilité $\Pr(X(t) = 1 \mid X(0) = 2)$. Soit Cov, le vecteur composé des quatres covariables suivantes : l'IMC, la sévérité, le traitement par corticoïdes oraux et les antécédent de corticoïdes oraux. Les courbes correspondent à différentes valeurs du vecteur Cov :

- les deux courbes du haut représentent les probabilités de transition pour
  - $Cov = (0,0,0,0)$: courbe en trait plein
  - $Cov = (1,0,0,0)$: courbe en pointillé
- les deux courbes du bas représentent les probabilités de transition pour
  - $Cov = (0,1,1,1)$: courbe en trait plein
  - $Cov = (1,1,1,1)$: courbe en pointillé

Sur les deux paires de courbes, la probabilité associée à la transition $2 \rightarrow 1$ est toujours plus faible pour les patients en surpoids.

5 Discussion

Ce chapitre présente la méthodologie des modèles de Markov homogènes. En particulier, une extension de ce modèle à un modèle de Markov non-homogène par périodes est proposée.

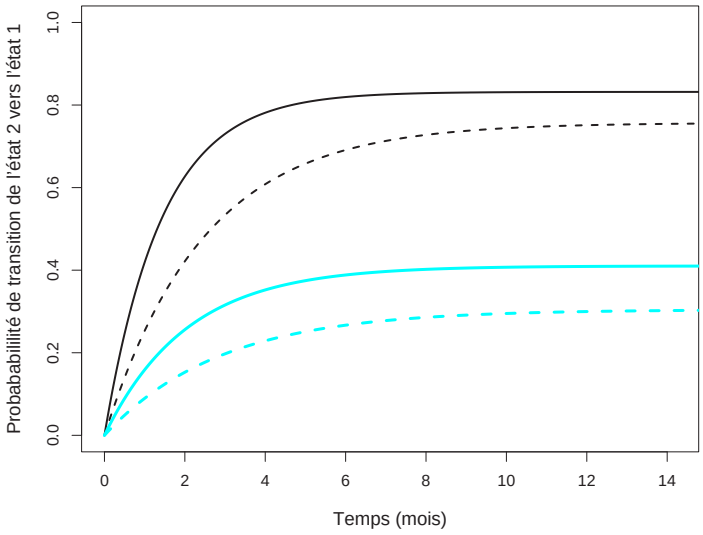


FIG. II.3 – Probabilités de transition de l'état inacceptable vers l'état acceptable. Les deux courbes du haut (foncées) sont associées aux patients non sévères, sans corticothérapie orale et sans antécédents majeurs de corticoïdes oraux. Les deux courbes du bas (claires) sont associées aux patients sévères, avec une corticothérapie orale et avec des antécédents de corticoïdes oraux. IMC < 25 : courbes en trait plein, IMC $\geqslant 25$: courbes en pointillé.

Cette extension permet d'affaiblir l'hypothèse d'homogénéité dans la modélisation et de pouvoir tester cette hypothèse en comparant les deux modèles. Nous avons ensuite discuté l'application de ces méthodes au cas de l'asthme et particulièrement étudié l'impact du surpoids sur l'évolution de la maladie dans un modèle à deux états de contrôle. Les résultats sont très intéressants, en effet, un patient en surpoids dans un état de contrôle inacceptable a moins de chances de revenir vers un contrôle acceptable indépendamment de son traitement et de sa sévérité. Ce résultat qui fait l'objet d'une publication dans une revue médicale corrobore et renforce les conclusions de plusieurs travaux visant à montrer un lien entre l'asthme et le surpoids.

Ces modèles s'avèrent être des outils attractifs pour la modélisation des données longitudinales. Les modèles Markoviens continus permettent de modéliser des données où les individus sont observés en des temps différents, c'est-à-dire que les dates de consultation peuvent être quelconques pour chaque patient et que le temps entre deux consultations n'est pas constant. L'hypothèse d'homogénéité facilite l'approche méthodologique et la programmation des méthodes. De plus, dans le modèle homogène, les résultats obtenus à partir des coefficients de régression et des probabilités de transition sont facilement interprétables d'un point de vue clinique. Ces modèles représentent ainsi une alternative de plus en plus utilisée dans la modélisation des suivis médicaux.

Cependant, on rencontre quelques difficultés dans l'estimation quand le nombre de paramètres du modèle est trop important ce qui nous conduit à restreindre la complexité du modèle. Il y a aussi des limites dues aux hypothèses du modèle. En particulier, l'hypothèse d'homogénéité est souvent trop restrictive : en effet la durée du suivi peut comme dans

l'asthme influencer les probabilités de transition. L'hypothèse de Markov peut aussi être une hypothèse forte, puisqu'elle suppose que le temps écoulé dans un état avant de transiter n'influence pas les intensités de transition.

Au vue des résultats obtenus sur la base, l'hypothèse d'homogénéité est trop contraignante. Ainsi, les modèles Markoviens non-homogènes semblent une alternative intéressante puisque qu'ils considèrent que les intensités de transition d'un état à un autre de la maladie dépendent de la durée du suivi. De même, il semble que la durée écoulée dans un état avant de transiter influence l'évolution de l'asthme. Cela suggère l'ajustement d'un modèle semi-Markovien afin d'accorder de l'importance à cette échelle de temps. Dans ce modèle, les distributions des temps de séjour ne sont plus des lois exponentielles comme dans le cas Markovien homogène : on peut choisir d'autres distributions paramétriques ce qui conduit à augmenter le nombre de paramètres à estimer.

Chapitre III

Modèle semi-Markovien homogène

1 Introduction

Les modèles Markoviens homogènes ont été appliqués avec succès dans de nombreux domaines et sont utilisés de plus en plus fréquemment. Cependant, dans ces modèles, l'évolution du processus est indépendante du temps déjà passé dans l'état actuel. Dans le domaine clinique par exemple, cette hypothèse correspond rarement à la réalité. Les processus semi-Markoviens constituent alors une alternative intéressante puisqu'ils intègrent dans la définition du modèle les lois de temps de séjour dans l'état. Un processus semi-Markovien dont les temps de séjour suivent des lois exponentielles devient un processus Markovien homogène. Les modèles semi-Markoviens généralisent ainsi les modèles Markoviens dans le sens où ils permettent de définir explicitement les lois des temps de séjour dans les états.

Les modèles semi-Markoviens commencent à être utilisés dans plusieurs domaines. En épidémiologie, Huber-Carol et Pons [2004] ont appliqué ces modèles à la transplantation cardiaque, Heutte et al. [2001] ont modélisé l'évolution d'un patient atteint du VIH, alors que Dabrowska et al. [1994] ont étudié la greffe de moelle osseuse. Ils sont aussi appliqués en fiabilité, Perez-Ocon et Torres-Castro [2002], dans les sciences sociales pour la recherche d'emploi, par exemple Vassiliou et Papadopoulou [1992], et en finance, Janssen et al. [1997]. Dans la littérature, on rencontre plusieurs méthodes d'estimation correspondant à différentes utilisation de ces modèles dans un cadre discret ou continu, à espace d'états fini ou non. On peut citer par exemple les ouvrages de Janssen [1986] et de Janssen et Limnios [1999] qui présentent de nombreuses méthodes d'estimation dans un cadre paramétrique et non-paramétrique. Les modèles semi-Markoviens ont été étudiés dans un cadre non-homogène par Vassiliou et Papadopoulou [1992] et Papadopoulou et Vassiliou [1994], alors que Sternberg et Satten [1999] se sont intéressés aux problèmes de données censurées par intervalles ou tronquées. On peut ajouter qu'il est aussi possible d'obtenir des estimations non-paramétriques dans les modèles semi-Markoviens en utilisant la théorie des processus de comptage comme nous le verrons au chapitre suivant (Gill [1980], Andersen et al. [1993]).

Après avoir présenté les processus semi-Markoviens, nous étudierons deux méthodes pour estimer les paramètres de tels modèles. Dans un premier temps, nous étudierons une méthode d'estimation non-paramétrique des intensités de transition du processus. L'estimation consiste à approximer ces intensités par des fonctions constantes par morceaux et

à maximiser la vraisemblance modifiée. Après avoir obtenu les estimations des intensités du processus, il sera possible de déduire un estimateur de la matrice des probabilités du processus semi-Markovien. Les estimateurs présentés dans ce travail sont une généralisation des estimateurs obtenus par Ouhbi et Limnios [1999] dans le cas d'un seul processus censuré. Les estimateurs sont adaptés au cas de plusieurs processus afin d'obtenir des estimations à partir d'un échantillon de processus. De plus, ces estimateurs sont présentés sous une forme plus générale permettant de prendre en compte des modèles avec un ou plusieurs états absorbants. Ces généralisations sont très utiles dans l'étude des maladies où l'on utilise souvent des modèles avec des états absorbants et où l'on dispose d'un processus pour chaque individu inclus dans l'étude.

Nous présenterons ensuite une méthode d'estimation dite paramétrique qui consiste à modéliser la distribution du temps de séjour dans l'état par une fonction paramétrique (Perez-Ocon et Ruiz-Castro [1999], Foucher et al. [2004]). L'approche présentée permet de choisir une distribution de temps de séjour et un nombre spécifique de covariables pour chaque transition. Plusieurs distributions sont utilisées pour la modélisation des temps de séjours : exponentielle, Weibull et Weibull généralisée qui sont particulièrement bien adaptées pour la modélisation des données médicales. La méthode du maximum de vraisemblance permet d'obtenir les estimations des distributions des temps de séjour et des intensités de transition du processus semi-Markovien.

Ce chapitre se terminera par l'application des modèles semi-Markoviens à la base de données de patients asthmatiques. En effet, comme nous l'avons vu dans le chapitre précédent, l'hypothèse d'intensités de transition constantes au cours du temps des modèles Markoviens homogènes semble abusive. De plus, dans le cas de l'asthme, il semble que le temps passé dans un état ait un impact sur l'évolution des patients. Les cliniciens pensent qu'un individu qui a déjà passé un temps important dans un état stable sera d'autant plus stable dans son évolution future. Nous discuterons l'application des méthodes étudiées, nous comparerons les méthodes d'estimations et nous interprèterons les résultats obtenus afin de mieux comprendre l'apport de ces modèles dans le cas de l'asthme.

2 Préliminaires

Cette partie, inspirée des travaux de Foucher [2004], Ouhbi et Limnios [1999] et Dabrowska et al. [1994], présente les processus semi-Markoviens homogènes à temps continu. Sont rappelées notamment les définitions et les propriétés de ces processus, les équations permettant d'obtenir les probabilités de transition du processus et l'écriture de la vraisemblance dans les modèles semi-Markoviens.

2.1 Définitions

On considère $(J_n, S_n)_{n\geqslant 0}$ un processus semi-Markovien, où $0 = S_0 < S_1 < \cdots < S_n < \cdots$ sont les temps consécutifs d'entrée dans les états $J_0, J_1, ..., J_n, ...$ avec $J_n \neq J_{n+1}, \forall n \geqslant 0$. $(J_n)_{n\geqslant 0}$ est une chaîne de Markov homogène à valeurs dans l'espace d'état fini $E = \{1, ..., s\}$.

$X_0, X_1, X_2, ...$ définis par $X_0 = 0$ et $X_n = S_n - S_{n-1}$, représentent les temps de séjour dans ces états.

Un processus semi-Markovien homogène peut être entièrement déterminé par

(i) sa loi initiale : $\Pr(J_0 = k) = p(k)$

(ii) le noyau semi-Markovien

$$\begin{aligned} Q_{ij}(d) &= \Pr(J_{n+1} = j, X_{n+1} \leqslant d \mid J_0, ..., J_n = i, X_1, ..., X_n) \\ &= \Pr(J_{n+1} = j, X_{n+1} \leqslant d \mid J_n = i), \end{aligned} \tag{III.1}$$

Les probabilités de transition de la chaîne de Markov $(J_n)_{n\geqslant 0}$ sont définis par

$$\begin{aligned} p_{ij} &= \lim_{d\to\infty} Q_{ij}(d) \\ &= \Pr(J_{n+1} = j \mid J_0, J_1, ..., J_n = i) \\ &= \Pr(J_{n+1} = j \mid J_n = i) \end{aligned} \tag{III.2}$$

Le processus $(J_n)_{n\geqslant 0}$ est une chaîne de Markov sous-jacente qui ne gère pas le temps, mais seulement la séquence des états. Les temps de séjour dans les états sont renseignés par le processus $(X_n)_{n\geqslant 0}$. D'après (III.1), les distributions des temps de séjour dépendent uniquement des états contigus : en effet, le passé de l'individu est résumé uniquement par le dernier état visité (Markov d'ordre 1). De plus, les distributions des temps de séjour sont des variables aléatoires positives indépendantes sachant la séquence des états. Notons que dans les définitions (III.1) et (III.2), l'indice n n'a pas d'importance, ce qui signifie que le processus est homogène sur le temps chronologique. Grâce à l'hypothèse d'homogénéité, ces quantités dépendent de l'état précédent et non du couple (état, temps d'entrée). Notons que $J_n \neq J_{n+1}$, c'est-à-dire qu'une transition vers le même état est impossible, ainsi, $Q_{ii}(d) \equiv 0$ $\forall i \in E$, $(p_{ii} = 0, \forall i \in E)$.

La fonction de distribution du temps de séjour dans l'état i avant d'aller dans l'état j, est définie par

$$\begin{aligned} F_{ij}(d) &= \Pr(X_{n+1} \leqslant d \mid J_n = i, J_{n+1} = j) \\ &= \begin{cases} \frac{\Pr(X_{n+1}\leqslant d, J_{n+1}=j \mid J_n=i)}{\Pr(J_{n+1}=j \mid J_n=i)} & \text{si } \Pr(J_{n+1} = j \mid J_n = i) > 0 \\ 0 & \text{sinon} \end{cases} \\ &= \begin{cases} \frac{Q_{ij}(d)}{p_{ij}} & \text{si } p_{ij} > 0 \\ 0 & \text{sinon} \end{cases} \end{aligned}$$

$S_{ij}(.) = 1 - F_{ij}(.)$ est la fonction de survie correspondant à $F_{ij}(.)$. Dans la pratique, l'état dans lequel va passer le processus est incertain. Ainsi, on définit la fonction de distribution du temps de séjour dans l'état i,

$$\begin{aligned} H_i(d) &= \Pr(X_{n+1} \leqslant d \mid J_n = i) \\ &= \sum_{j=1}^{s} \Pr(X_{n+1} \leqslant d, J_{n+1} = j \mid J_n = i) \\ &= \sum_{j=1}^{s} Q_{ij}(d) \\ &= \sum_{j=1}^{s} F_{ij}(d) p_{ij} \end{aligned} \tag{III.3}$$

On peut définir les fonctions de survie correspondantes, $S_i(.) = 1 - H_i(.)$ et $S_{ij}(.) = 1 - F_{ij}(.)$,

$$\begin{aligned} S_i(d) &= \Pr(X_{n+1} > d \mid J_n = i) \\ &= \sum_{j=1}^{s} p_{ij} S_{ij}(d) \end{aligned} \tag{III.4}$$

$S_i(d)$ représente la probabilité que l'individu survive dans l'état $J_n = i$ jusqu'au temps $S_n + d$.

En supposant que $Q_{ij}(.)$ est absolument continue par rapport à la mesure de Lebesgue, on peut définir $q_{ij}(.)$ la densité de $Q_{ij}(.)$,

$$q_{ij}(d) = \lim_{\Delta d \to 0} \frac{1}{\Delta t} \Pr(J_{n+1} = j, d < X_{n+1} \leqslant d + \Delta d \mid J_n = i),$$

et $f_{ij}(.)$ la densité de $F_{ij}(.)$,

$$\begin{aligned} f_{ij}(d) &= \lim_{\Delta d \to 0} \frac{1}{\Delta d} \Pr(d < X_{n+1} \leqslant d + \Delta d \mid J_n = i, J_{n+1} = j) \\ &= \begin{cases} \frac{q_{ij}(d)}{p_{ij}} & \text{si } p_{ij} > 0 \\ 0 & \text{sinon.} \end{cases} \end{aligned} \tag{III.5}$$

Par définition, les intensités de transition instantanées du noyau semi-Markovien, $\forall i, j \in E$, sont données par l'expression suivante

$$\begin{aligned} \lambda_{ij}(d) &= \lim_{\Delta d \to 0} \frac{1}{\Delta d} \Pr(J_{n+1} = j, d < X_{n+1} \leqslant d + \Delta d \mid J_n = i, X_{n+1} > d) \\ &= \lim_{\Delta d \to 0} \frac{1}{\Delta d} \frac{\Pr(J_{n+1} = j, d < X_{n+1} \leqslant d + \Delta d, X_{n+1} > d \mid J_n = i)}{\Pr(X_{n+1} > d \mid J_n = i)} \\ &= \begin{cases} \frac{q_{ij}(d)}{1 - H_i(d)} & \text{si } p_{ij} > 0 \text{ et } H_i(d) < 1 \\ 0 & \text{sinon.} \end{cases} \\ &= \begin{cases} \frac{p_{ij} f_{ij}(d)}{S_i(d)} & \text{si } p_{ij} > 0 \text{ et } S_i(d) > 0 \\ 0 & \text{sinon.} \end{cases} \end{aligned} \tag{III.6}$$

En fait, $\lambda_{ij}(d)\Delta d+o(\Delta d)$, $i \neq j$, représente la probabilité que le processus transite dans l'état j dans $]d, d+\Delta d]$ sachant qu'il est resté un temps d dans l'état i. La force de changement d'état $\lambda_{ij}(.)$ sera d'autant plus grande que la densité $f_{ij}(.)$ et la probabilité p_{ij} seront grandes et que la survie $S_i(.)$ sera petite.

Cette fonction de risque du processus semi-Markovien ne doit pas être confondue avec la fonction de risque de la loi des temps de séjour qui suppose, par définition, que l'état d'arrivée est connu. Les fonctions de risque des temps de séjour sont données par

$$\tilde{\alpha}_{ij}(d) = \lim_{\Delta d \to 0} \frac{1}{\Delta d} \Pr(d < X_{n+1} \leqslant d + \Delta d \mid X_{n+1} > d, J_n = i, J_{n+1} = j). \tag{III.7}$$

Remarque 4 *Dans les définitions, l'indice n renseigne sur le numéro de la transition, S_n sur les temps de sauts, J_n sur les états et X_{n+1} sur la durée écoulé dans l'état J_n. D'autres notations issues des processus de comptage peuvent être utilisées. On considère $N_{ij}(t)$ un processus de comptage du nombre de transitions observées de l'état i vers l'état j dans $[0, t]$, où t est le temps chronologique,*

$$N_{ij}(t) = \sum_{n \geqslant 0} \mathbb{1}_{\{S_n \leqslant t, J_n = i, J_{n+1} = j\}}(t).$$

$N_{ij}(.)$ est un processus cadlag avec des sauts de 1, $N_{ij}(0) = 0$ (cf. chapitre IV). On définit ensuite

$$N(t) = \sum_{i,j \in E} N_{ij}(t),$$

comptant le nombre total de transitions observées dans $[0, t]$. Ainsi, l'état du processus au temps t (souvent noté $J(t)$ dans la littérature) peut s'écrire $J_{N(t)}$. $(J_{N(t)}, t \in \bar{\mathbb{R}}^+)$ contient la même information que le processus $Z = (J_n, S_n)_{n \geqslant 0}$.

2.2 Probabilités de transition du processus semi-Markovien

Dans cette section, on s'intéresse à la probabilité de transition du processus semi-Markovien. Comme nous l'avons vu dans la section précédente ($cf.$ Remarque 4), le processus semi-markovien $Z = (J_n, S_n)_{n \geqslant 0}$ peut aussi s'écrire $Z(t) = J_{N(t)}$. $J_{N(t)}$ représente l'état du processus au temps t, ou l'état après $N(t)$ sauts. Les probabilités de transition du processus $P_{ij}(s, t)$ sont définies par

$$P_{ij}(s, t) = \Pr(J_{N(t)} = j \mid J_{N(s)} = i)$$

Par la propriété d'homogénéité de la chaîne de Markov sous-jacente $(J_n)_{n \geqslant 0}$, les probabilités de transition vérifient la propriété suivante,

$$\begin{aligned} P_{ij}(t, t+d) &= \Pr(J_{N(t+d)} = j \mid J_{N(t)} = i) \\ &= \Pr(J_{N(d)} = j \mid J_{N(0)} = i) \\ &= P_{ij}(0, d) \\ &= \Psi_{ij}(d). \end{aligned}$$

La propriété d'homogénéité de la chaîne de Markov est transmise au processus semi-Markovien.

Les probabilités de transition du processus semi-Markovien homogène sont définies par l'équation suivante,

$$\begin{aligned}\Psi_{ij}(t) &= \Pr(Z(t) = j \mid Z(0) = i) \\ &= \sum_{r=1}^{s} \int_0^t p_{ir} f_{ir}(u) \Psi_{rj}(t-u) du + \delta_{ij} \sum_{r=1}^{s} p_{ir} S_{ir}(t) \end{aligned} \tag{III.8}$$

où $\delta_{ij} = 1$ si $i = j$ et 0 sinon. Afin d'obtenir l'écriture des probabilités $\Psi_{ij}(t)$, on peut considérer les deux cas suivants :

– Cas $i \neq j$: dans ce cas, au moins un événement s'est produit entre $[0, t]$ (car une transition vers le même état est impossible). Pour écrire la probabilité de transition, on considère un conditionnement sur le premier événement qui est couramment utilisé dans la théorie du renouvellement. Ainsi la probabilité $\Psi_{ij}(t)$ peut s'écrire,

$$\begin{aligned}\Psi_{ij}(t) &= \Pr(J_{N(t)} = j \mid J_{N(0)} = i) \\ &= \Pr(J_{N(t)} = j, X_1 \leqslant t \mid J_0 = i) \\ &= \sum_{r=1}^{s} \Pr(J_{N(t)} = j, X_1 \leqslant t, J_1 = r \mid J_0 = i) \end{aligned} \tag{III.9}$$

Il faut déterminer la probabilité que le processus soit dans l'état j sachant que l'état initial est i et que le premier nouvel état est r. Le calcul de cette probabilité peut s'obtenir dans un cadre formel à l'aide du produit de convolution. Cependant, l'écriture de cette probabilité peut aussi s'obtenir de manière plus intuitive. En effet, en considérant x le temps d'entrée dans le premier état, la probabilité cherchée s'écrit comme le produit des probabilités suivantes :
- la probabilité que le sujet reste dans l'état i un temps x, *i.e*, $S_i(x)$,
- la probabilité qu'il transite de l'état i vers l'état r au temps x, *i.e*, $\lambda_{ir}(x)$,
- la probabilité que l'individu soit dans l'état j au temps t sachant qu'il était dans l'état r au temps x, *i.e*, $P_{rj}(x, t) = \Psi_{rj}(t - x)$.

La probabilité $\Psi_{ij}(t)$ s'exprime ensuite à l'aide d'une intégrale afin de prendre en compte tous les temps $x \in [0, t]$,

$$\begin{aligned}\Pr(J_{N(t)} = j, X_1 \leqslant t, J_1 = r \mid J_0 = i) &= \int_0^t S_i(u) \lambda_{ir}(u) \Psi_{rj}(t-u) du \\ &= \int_0^t p_{ir} f_{ir}(u) \Psi_{rj}(t-u) du.\end{aligned}$$

L'équation (III.9) permet ensuite d'obtenir la probabilité cherchée.

– Cas $i = j$: dans ce cas, soit au moins deux événements se produisent dans $[0, t]$ soit aucun événement ne se produit. $\Psi_{ij}(t)$ peut alors s'écrire

$$\begin{aligned}\Psi_{ij}(t) &= \Pr(J_{N(t)} = i \mid J_{N(0)} = i) \\ &= \underbrace{\Pr(J_{N(t)} = i,\ N(t) > 0 \mid J_0 = i)}_{A} + \underbrace{\Pr(J_{N(t)} = i,\ N(t) = 0 \mid J_0 = i)}_{B}\end{aligned}$$

A se calcule de manière identique au cas $i \neq j$. La probabilité B s'exprime en terme de survie (III.4),

$$B = \Pr(X_1 > t \mid J_{N(0)} = i) = S_i(t) = \sum_{r=1}^{s} p_{ir} S_{ir}(t).$$

Ainsi la probabilité cherchée est,

$$\Psi_{ij}(t) = \sum_{r=1}^{s} \int_0^t p_{ir} f_{ir}(u) \Psi_{rj}(t-u) du + \sum_{r=1}^{s} p_{ir} S_{ir}(t).$$

L'utilisation du symbole de Kroneker δ_{ij} permet de généraliser l'écriture des $\Psi_{ij}(t)$ aux deux cas possibles.

2.3 Fonction de vraisemblance

Considérons un échantillon de n individus ($h = 1, ..., n$). Pour l'individu h, considérons $0 = S_{h,0} < S_{h,1} < \cdots < S_{h,N_h}$ les temps de sauts. A ces temps, l'individu a successivement occupé les états $J_{h,0}, J_{h,1}, ..., J_{h,N_h}$ avec $J_{h,p} \neq J_{h,p+1}$, où N_h est le nombre de sauts du processus associé à l'individu h à la date de point ($N_h \geqslant 0$). En utilisant ces notations, S_{h,N_h} représente le dernier temps d'entrée dans un état et J_{h,N_h} le dernier état occupé par l'individu h. Considérons $X_{h,p} = S_{h,p} - S_{h,p-1}$, le temps de séjour dans l'état $J_{h,p-1}$, $p = 1, ..., N_h$.

Dans tout ce chapitre, on considère le cas de données censurées à droite. De plus, on suppose que la censure n'apporte aucune information sur l'événement étudié (censure indépendante). Le phénomène de censure à droite est étudié plus en détails aux chapitres IV page 76. Dans le cas semi-Markovien, la censure à droite empêche l'observation du temps de séjour dans le dernier état visité. Notons qu il est supposé que les temps de transition entre les états correspondent aux temps d'observation (de la base de données). Dans le cas des modèles avec états absorbants, on rencontre de deux types d'« histoires ».

– Soit l'individu h rentre dans un état absorbant, son « histoire » n'est pas censurée,

$$\mathcal{H}_h(t) = (J_{h,0}, J_{h,1}, ..., J_{h,N_h}, X_{h,1}, X_{h,2}, \cdots, X_{h,N_h}).$$

– Soit l'individu h n'entre pas dans un état absorbant. Dans ce cas, le temps de séjour dans le dernier état visité (J_{h,N_h}) est censuré à la date de fin d'étude. L'« histoire » de l'individu est

$$\mathcal{H}_h(t) = (J_{h,0}, J_{h,1}, ..., J_{h,N_h}, X_{h,1}, X_{h,2}, \cdots, X_{h,N_h}, U_h).$$

où U_h représente la durée écoulée entre S_{h,N_h} et la date de fin d'étude.

Ainsi, deux types de contribution à la vraisemblance sont possibles.

(i) Soit l'individu reste un temps d dans l'état i et ensuite, il transite dans l'état j ($j \neq i$), dans ce cas où la transition est observée, la contribution à la vraisemblance est

$$S_i(d)\lambda_{ij}(d) = p_{ij} f_{ij}(d) = q_{ij}(d).$$

(ii) Soit l'individu reste un temps d dans l'état i et il est censuré. Dans ce cas de censure à droite où le temps de séjour est censuré, la contribution à la vraisemblance s'exprime en terme de survie par

$$S_i(d) = \sum_{j=1}^{s} p_{ij} S_{ij}(d).$$

Ainsi, en considérant δ_h qui vaut 1 si l'individu est censuré et 0 sinon, la contribution à la vraisemblance de l'individu h s'écrit

$$L_h = \prod_{k=1}^{N_h} p_{J_{h,k-1}J_{h,k}} f_{J_{h,k-1}J_{h,k}}(X_{h,k}) \times \left[S_{J_{h,N_h}}(U_h)\right]^{\delta_h}$$

La vraisemblance totale est obtenue en faisant le produit des contributions individuelles,

$$L = \prod_{h=1}^{n} L_h. \tag{III.10}$$

3 Estimation paramétrique des temps de séjour

3.1 Introduction

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

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

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

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

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

$$S_{ij}(d) = \exp(-\int_0^d \alpha_{ij}(u)du). \tag{III.11}$$

De plus, comme $S_{ij}(.) = 1 - F_{ij}(.)$ et comme $f_{ij}(.)$ est la densité de $F_{ij}(.)$, on peut déduire l'écriture suivante pour $f_{ij}(.)$

$$f_{ij}(d) = -\frac{\partial S_{ij}(d)}{\partial d} = S_{ij}(d)\alpha_{ij}(d). \tag{III.12}$$

3.2 Modèle à risques proportionnels

Dans l'étude des maladies chroniques, il est important d'étudier l'impact de divers facteurs sur l'évolution de la maladie. En effet, l'utilisation de covariables permet de prendre en compte l'hétérogénéité de la population et d'obtenir des résultats adaptés aux caractéristiques des patients. L'utilisation de cette méthode paramétrique permet d'incorporer des covariables dans la modélisation des fonctions de risque des temps d'attente. Nous utiliserons à cet effet, un modèle à risques proportionnels (Cox [1972]).

Considérons $(J_n, S_n)_{n\geqslant 0}$ un processus semi-Markovien. Les covariables sont introduites dans les fonctions de risque des temps d'attente dans les états. La chaîne de Markov sous-jacente $(J_n)_{n\geqslant 0}$ ne dépend pas du vecteur de covariables et ainsi la chaîne conserve la probabilité de transition $p_{ij} = \Pr(J_{n+1} = j \mid J_n = i)$. Les fonctions de risque du processus semi-Markovien ne dépendent pas directement des covariables. Cependant comme ils sont définis à l'aide des risques de temps d'attente, l'effet des covariables sera quand même répercuté même s'il ne pourra pas s'interpréter en termes de risque relatif. En reprenant les notations utilisant le processus de comptage $N(t)$, les fonctions d'intensité $\tilde{\alpha}_{ij}(t)$ peuvent s'écrire

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

où t représente le temps chronologique.

Considérons $\mathbf{z}_{ij}(.) = (z_{ij}^1(.), z_{ij}^2(.), ..., z_{ij}^{n_{ij}}(.))$ un vecteur de covariables associé à la transition de l'état i vers l'état j, tel que n_{ij} soit le nombre de covariables pour cette transition. Cette notation permet d'ajuster un modèle où le nombre de covariables est spécifique à chaque transition. Les covariables peuvent être dépendantes du temps, cependant il est nécessaire de supposer que la valeur des covariables ne change pas entre deux consultations. Afin de simplifier les calculs et les écritures, on supposera dans ce qui suit que les covariables sont fixées au cours du temps d'attente : $\mathbf{z}_{ij}(t) = \mathbf{z}_{ij}$. On suppose que les covariables vont modifier les fonctions d'intensité en suivant un modèle à risques proportionnels de Cox.

$$\tilde{\alpha}_{ij}(t - S_{N(t^-)}) = \mathbb{1}_{\{J_{N(t^-)}=i\}}\alpha_{ij,0}(t - S_{N(t^-)})\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}), \tag{III.13}$$

où $\boldsymbol{\beta}_{ij}$ est le vecteur des coefficients de régression associés à $\mathbf{z}_{ij}$ et $\alpha_{ij,0}(.)$ est le risque de base. Dans ce modèle, la proportionnalité des risques est supposée au sein d'une même transition.

Intéressons nous maintenant aux fonctions de survie et de densité correspondant à des fonctions de risque de temps d'attente dépendantes de covariables. Considérons $\forall i, j \in E$, $\alpha_{ij}(d, \mathbf{z}) = \alpha_{ij,0}(d)e^{\boldsymbol{\beta}_{ij}^T \mathbf{z}_{ij}}$, alors d'après les équations (III.11) et (III.12) les fonctions de survie correspondantes sont données par

$$\begin{aligned} S_{ij}(d, \mathbf{z}) &= \exp(-\int_0^d \alpha_{ij}(u)du) \\ &= \exp(-\int_0^d \alpha_{ij,0}(u)e^{\boldsymbol{\beta}_{ij}^T \mathbf{z}_{ij}}du) \\ &= S_{ij,0}(d)^{e^{\boldsymbol{\beta}_{ij}^T \mathbf{z}_{ij}}} \end{aligned} \tag{III.14}$$

où $S_{ij,0}(d) = \exp(-\int_0^d \alpha_{ij,0}(u)du)$ et les fonctions de densité sont

$$\begin{aligned} f_{ij}(d, \mathbf{z}) &= S_{ij}(d, \mathbf{z})\alpha_{ij}(d, \mathbf{z}) \\ &= \alpha_{ij,0}(d)e^{\boldsymbol{\beta}_{ij}^T \mathbf{z}_{ij}} S_{ij,0}(d)^{e^{\boldsymbol{\beta}_{ij}^T \mathbf{z}_{ij}}}. \end{aligned} \tag{III.15}$$

Remarque 5 *Afin de prendre en compte les covariables, on utilise un modèle log-linéaire utilisant la fonction exponentielle. Ce modèle est attractif car il permet d'avoir des fonctions définies positives et permet d'éviter les estimations sous contraintes. De plus, les coefficients de régression peuvent être interprétés en terme de risques relatifs. Ce modèle log-linéaire est le modèle le plus couramment utilisé dans la littératurre (Cox [1972], Andersen et al. [1993]) Cependant, d'autres choix de fonctions sont possibles.*

3.3 Modélisation paramétrique de la loi de séjour dans l'état

La modélisation paramétrique consiste à estimer les fonctions de risque des temps d'attente dans les états par des fonctions paramétriques. Ainsi, on suppose que $\alpha_{ij}(t) = g_{ij}(t, \boldsymbol{\theta}_{ij})$, où g_{ij} est une fonction paramétrique intégrable. L'estimation de $\alpha_{ij}(.)$ consiste à estimer le vecteur de paramètres $\boldsymbol{\theta}_{ij}$ (qui inclut les coefficients de régression).

D'après l'équation (III.4), la vraisemblance (III.10) peut s'écrire

$$L = \prod_{h=1}^{n} \left\{ \prod_{k=1}^{N_h} p_{J_{h,k-1}J_{h,k}} f_{J_{h,k-1}J_{h,k}}(X_{h,k}) \times \left[\sum_{j=1}^{s} p_{J_{h,N_h}j} S_{J_{h,N_h}j}(U_h) \right]^{\delta_h} \right\}. \tag{III.16}$$

A partir des équations (III.14) et (III.15), la vraisemblance (III.16) peut s'écrire en fonction des paramètres p_{ij} et $\boldsymbol{\theta}_{ij}$. L'estimation des paramètres se fait ensuite par maximisation de la vraisemblance. On obtient ainsi les estimations $\hat{p}_{ij}$ des probabilités de la chaîne de Markov et les estimations $\hat{\alpha}_{ij}(.)$ des fonctions de risque des temps de séjour. On en déduit les estimateurs $\hat{f}_{ij}(.)$ des fonctions de densité et $\hat{S}_{ij}(.)$ des fonctions de survie. De plus, d'après

(III.6) et (III.4), il est possible de déduire les estimeurs $\hat{\lambda}_{ij}(.)$ des intensités du processus semi-Markovien par la formule suivante

$$\hat{\lambda}_{ij}(d) = \frac{\hat{p}_{ij}\hat{f}_{ij}(d)}{\sum_{j=1}^{s}\hat{p}_{ij}\hat{S}_{ij}(d)}. \tag{III.17}$$

On peut noter que la notation $\alpha_{ij}(d) = g_{ij}(d, \boldsymbol{\theta}_{ij})$ permet de considérer des natures de fonctions différentes suivant la transition étudiée. De plus, le nombre de paramètres qui définissent la fonction est spécifique à chaque transition. En pratique, cette écriture est très utile car elle permet d'adapter la modélisation à chaque transition et elle permet ainsi d'optimiser le nombre de paramètres. En effet, un des problèmes majeurs de l'estimation est dû au nombre de paramètres : s'il est trop grand (pour la base de données), les estimations seront peu fiables. Il est donc important de conserver uniquement les paramètres nécessaires en considérant une loi et un vecteur de covariables spécifiques à chaque transition.

Dans les études de survie de données épidémiologiques, les familles de fonctions les plus couramment utilisées pour modéliser les risques sont les lois exponentielles, les lois de Weibull et les lois de Weibull généralisées. Pour la modélisation de l'asthme, nous utiliserons ces mêmes familles de lois car elles sont bien adaptées aux problèmes épidémiologiques et elles ont l'avantage d'être « emboîtées », dans le sens où la loi de Weibull généralisée généralise la loi de weibull qui généralise la loi exponentielle. Il sera ainsi possible de juger la pertinence d'une loi en analysant les coefficients. D'autres choix de distributions sont possibles, mais il est important de faire un compromis entre la taille de la base et le nombre de paramètres qui définissent la loi. Dans la suite, nous présentons les fonctions de risque, de densité et de survie associées aux lois utilisées.

3.3.1 Loi de Weibull

La loi de Weibull possède de bonnes propriétés pour la modélisation des données de survie. Elle permet de prendre en compte une évolution monotone du risque instantané au cours du temps. Si la loi de Weibull sans covariable est utilisée pour modéliser le risque alors, $\forall i, j \in E$, $i \neq j$,

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

La distribution exponentielle, qui est sous jacente à une modélisation Markovienne homogène est obtenue pour $\nu_{ij} = 1$. Le modèle semi-Markovien avec loi de Weibull constitue ainsi une généralisation du modèle Markovien homogène à temps continu. En supposant le modèle à risques proportionnels, la fonction de risque avec covariables s'écrit,

$$\alpha_{ij}(d, \mathbf{z}) = \nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1}\exp(\boldsymbol{\beta}_{ij}^{T}\mathbf{z}_{ij}).$$

En suivant la définition (III.14), la fonction de survie est

$$\begin{aligned} S_{ij}(d,\mathbf{z}) &= S_{ij}(d)^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}} \\ &= \left[\exp(-\int_0^d \nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} u^{\nu_{ij}-1}du)\right]^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}} \\ &= \left[\exp(-(\frac{d}{\sigma_{ij}})^{\nu_{ij}})\right]^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}}. \end{aligned}$$

avec $S_{ij}(d)$ la fonction de survie associée à une Weibull sans covariable. D'après (III.15), la densité correspondante est

$$\begin{aligned} f_{ij}(d,\mathbf{z}) &= S_{ij}(d,\mathbf{z})\alpha_{ij}(d,\mathbf{z}) \\ &= \left[\exp(-(\frac{d}{\sigma_{ij}})^{\nu_{ij}})\right]^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}} \nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1}\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}). \end{aligned}$$

Remarque 6 *Si $\nu_{ij} = 1$, on retrouve les fonctions associées à la loi exponentielle avec covariables :*

$$\begin{aligned} \alpha_{ij}(d,\mathbf{z}) &= \frac{1}{\sigma_{ij}}\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}) \\ S_{ij}(d,\mathbf{z}) &= \exp(-\frac{d}{\sigma_{ij}})^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}} \\ f_{ij}(d,\mathbf{z}) &= \frac{1}{\sigma_{ij}}\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij})\exp(-\frac{d}{\sigma_{ij}})^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}} \end{aligned}$$

3.3.2 Loi de Weibull généralisée

La loi de Weibull est intéressante pour modéliser des risques monotones. Cependant, elle devient mal adaptée quand les risques ne sont pas monotones : par exemple les formes en cloches qui sont souvent présentes dans les études du vivant. Dans ces cas là, une alternative est l'utilisation de la loi de Weibull généralisée qui permet de modéliser des fonctions de risque instantané en forme de $\cup$ ou $\cap$. La fonction de risque (sans covariable) est donnée par

$$\alpha_{ij}(d) = \frac{1}{\theta_{ij}}\left(1+\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right)^{\frac{1}{\theta_{ij}}-1}\nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1},$$

$\forall d \geqslant 0, \forall \nu_{ij} > 0, \forall \sigma_{ij} > 0, \forall \theta_{ij} > 0$. En supposant les risques proportionnels, on a

$$\alpha_{ij}(d,\mathbf{z}) = \frac{1}{\theta_{ij}}\left(1+\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right)^{\frac{1}{\theta_{ij}}-1}\nu_{ij}\left(\frac{1}{\sigma_{ij}}\right)^{\nu_{ij}} d^{\nu_{ij}-1}\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}).$$

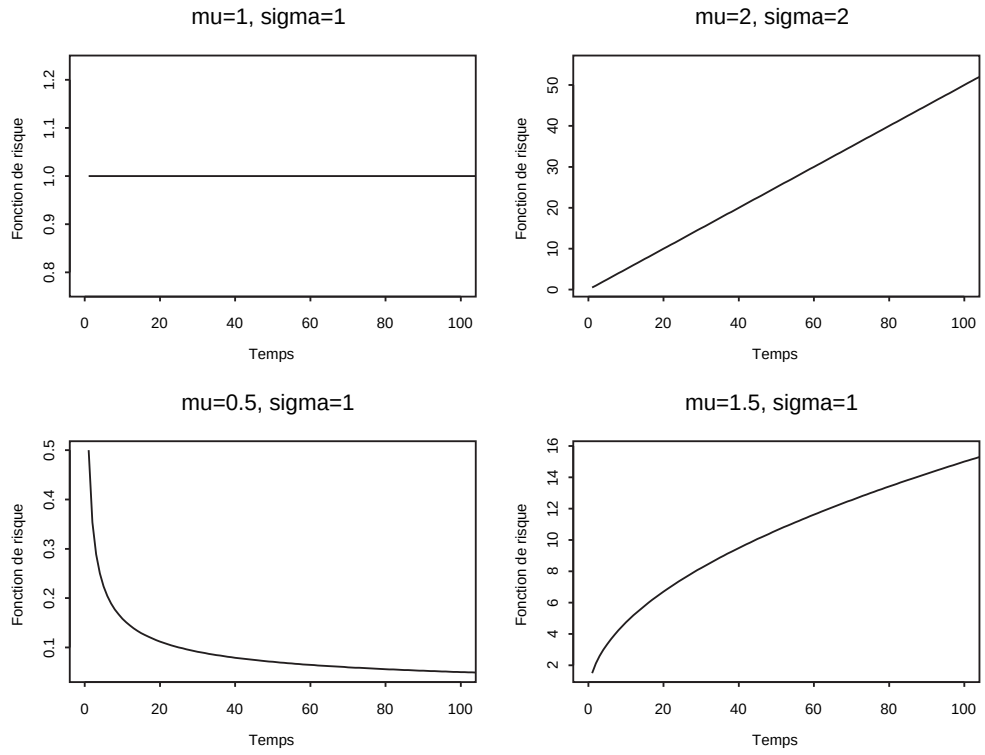


FIG. III.1 – Exemple de fonctions de risque d'une loi de Weibull.

En utilisant les définitions (III.14) et (III.15), la fonction de survie est

$$\begin{aligned} S_{ij}(d,\mathbf{z}) &= S_{ij}(d)^{e^{\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij}}} \\ &= \exp\left(1-\left(1+\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right)^{\frac{1}{\theta_{ij}}}\right)^{\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij})} \end{aligned}$$

avec $S_{ij}(d)$, la fonction de survie sans covariable et la densité correspondante est

$$\begin{aligned} f_{ij}(d,\mathbf{z}) &= S_{ij}(d,\mathbf{z})\alpha_{ij}(d,\mathbf{z}) \\ &= \frac{1}{\theta_{ij}}\left(1+\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right)^{\frac{1}{\theta_{ij}}-1}\frac{\nu_{ij}}{\sigma_{ij}}\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}-1}\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij})\exp\left(1-\left(1+\left(\frac{d}{\sigma_{ij}}\right)^{\nu_{ij}}\right)^{\frac{1}{\theta_{ij}}}\right)^{\exp(\boldsymbol{\beta}_{ij}^T\mathbf{z}_{ij})} \end{aligned}$$

Remarque 7 *Si $\theta_{ij}=1$, on retrouve les fonctions associées à la loi de Weibull. Si $\theta_{ij}=1$ et $\nu_{ij}=1$ on retrouve les fonctions associées à la loi exponentielle.*

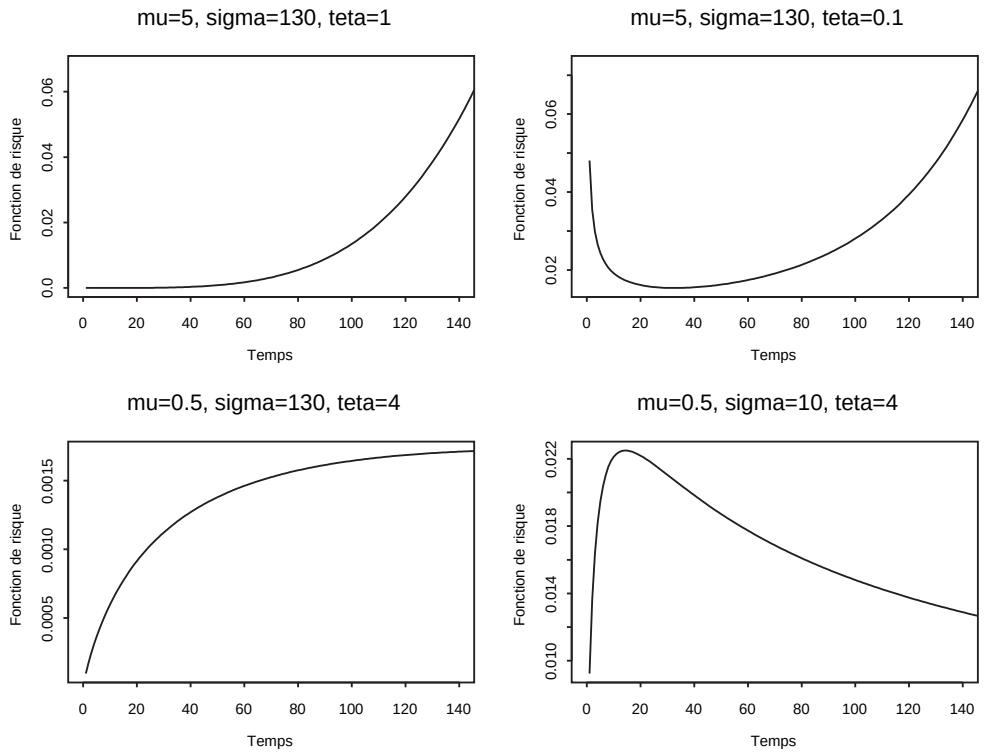


FIG. III.2 – Exemple de fonctions de risque d'une loi de Weibull généralisée.

3.4 Extension à un modèle semi-Markovien non-homogène

Dans le modèle semi-Markovien homogène, l'évolution du processus dépend uniquement de la durée écoulée dans l'état. En effet, comme la chaîne de Markov sous-jacente est homogène (probabilités de transition indépendantes du temps chronologique), les probabilités de transition du processus sont aussi homogènes dans le sens où elles ne dépendent pas du temps chronologique où se produit la transition. Cette hypothèse d'homogénéité peut, dans certaines applications, être trop restrictive. Dans le cas de l'asthme par exemple, les patients sont traités et éduqués, ainsi on peut espérer que les probabilités de rester dans un état stable augmentent avec la durée du suivi.

Afin de rendre le modèle non-homogène, on peut considérer que les probabilités de la chaîne de Markov sous-jacente dépendent de la durée du suivi (temps depuis l'inclusion dans l'étude). Ainsi, l'évolution du processus semi-Markovien dépend de la durée écoulée dans l'état avant de transiter d, mais aussi de la durée du suivi t. On définit alors les probabilités de transition de la chaîne de Markov sous-jacente, qui dépendent du temps t,

$$p_{ij}(t) = \Pr(J(t) = j \mid J(t^-) = i),$$

où $J(t)$ représente l'état occupé par le processus au temps t. Il est ensuite possible de redéfinir les fonctions de survie, la fonction de vraisemblance et les probabilités de transition du processus de manière à prendre en compte la durée de suivi dans ces quantités.

La méthode d'estimation paramétrique présentée précédemment peut ensuite être étendue afin d'estimer de manière paramétrique les probabilités $p_{ij}(.)$. On peut supposer, par exemple, que les probabilités sont de la forme

$$\hat{p}_{ij}(t) = \frac{V_{ij}(t)}{\sum_{\{(i,j)\in E\times E\}} V_{ij}(t)},$$

avec

$$V_{ij}(t) = \begin{cases} \frac{\exp(a_{ij}\times t+b_{ij})}{1+\exp(a_{ij}\times t+b_{ij})} & i \neq j \\ 0 & i = j. \end{cases}$$

Le choix de ces fonctions est motivé par le fait que $\hat{p}_{ij}(.)$ est compris entre 0 et 1 et que $\sum_{j=1}^{n} \hat{p}_{ij}(.) = 1$. L'estimation des $p_{ij}(.)$ consiste alors à estimer les paramètres a_{ij} et b_{ij} par maximisation de la vraisemblance.

En pratique, ce modèle et cette méthode d'estimation sont difficilement utilisables. En effet, le nombre de paramètres devient très important et les estimations deviennent difficiles. Dans le cas de l'asthme par exemple, le nombre d'observations dans la base de données ne permet pas d'appliquer ce modèle qui comprend trop de paramètres. L'augmentation du nombre de paramètres pour définir les lois et les probabilités a aussi pour conséquence de restreindre le nombre de covariables pouvant être incluses dans la modélisation. De plus, les résultats obtenus avec ce type de modèle dépendent de deux échelles de temps ce qui rend l'interprétation délicate. Ces modèles peuvent cependant être intéressants pour observer l'impact de la durée du suivi dans un modèle semi-Markovien et pour étudier des maladies où la durée du suivi et le temps de séjour dans l'état influencent l'évolution de la maladie. Pour plus d'informations sur les modèles semi-Markoviens non-homogènes, on pourra consulter Vassiliou et Papadopoulou [1992], Papadopoulou et Vassiliou [1994] et Janssen et Limnios [1999].

4 Estimation non-paramétrique des intensités du processus semi-Markovien

4.1 Introduction

L'objectif de cette section est de proposer un estimateur du maximum de vraisemblance non-paramétrique des intensités du processus semi-Markovien. L'estimateur est dit non-paramétrique car il repose sur une approximation des intensités du processus par des fonctions constantes par morceaux. Cette méthode ne fait aucune hypothèse sur la forme des distributions, cependant, elle nécessite de définir une subdivision sur laquelle les intensités sont constantes. L'estimation consiste à obtenir les valeurs de la fonction sur chaque intervalle. Ses valeurs sont estimées par la méthode du maximum de vraisemblance.

L'estimateur présenté est une généralisation de l'estimateur de Ouhbi et Limnios [1999]. En effet, Ouhbi et Limnios [1999] ont obtenu un estimateur non-paramétrique des intensités du processus semi-Markovien dans le cas d'un seul processus censuré. La généralisation consiste à adapter cet estimateur :

- au cas de processus pouvant être censurés ou non (pour prendre en compte des modèles avec états absorbants),
- au cas de plusieurs processus.

Cette généralisation a été motivée par des raisons pratiques. En effet, dans de nombreuses applications et en particulier dans l'étude des maladies, on utilise souvent des modèles avec états absorbants pour prendre en compte par exemple le décès ou le rejet de greffe. De plus, on dispose souvent d'un échantillon où un processus est associé à chaque individu. Il était alors important d'adapter l'estimateur proposé par Ouhbi et Limnios [1999] afin qu'il soit applicable à des problèmes pratiques où l'on dispose de plusieurs processus. L'estimateur proposé ici permet d'obtenir un estimateur non-paramétrique des intensités du processus semi-Markovien dans un modèle avec (ou sans) état absorbant quand on dispose d'un échantillon de processus.

Remarque 8 *La différence avec la méthode de Ouhbi et Limnios [1999] (dans le cas d'un seul processus censuré) tient à ce que les résultats de convergence asymptotique sont dus à $t \to \infty$ alors que pour plusieurs processus censurés ou non, les résultats de convergence asymptotique sont dus à $n \to \infty$ (n nombre d'individus).*

Remarque 9 *La méthode d'estimation non-paramétrique considérée dans cette section repose sur le choix déterministe d'une subdivision. Il serait préférable d'utiliser une méthode NPML (Non Parametric Maximum Likelihood) qui considère une subdivision aléatoire fonction des données (Huber-Carol et Pons [2004], Chang et al. [2000], Andersen et al. [1993], Gill [1980]).*

4.2 Ecriture de la vraisemblance

D'après la définition (III.5), la vraisemblance (III.10) est donnée par :

$$L = \prod_{h=1}^{n} \prod_{k=1}^{N_h} q_{J_{h,k-1}J_{h,k}}(X_{h,k}) \times \left[S_{J_{h,N_h}}(U_h)\right]^{\delta_h} \tag{III.18}$$

Afin d'obtenir la vraisemblance en fonction des $\lambda_{ij}(.)$, les paramètres $S_{ij}(.)$ et $q_{ij}(.)$ doivent être écrits sous une autre forme. Par la définition de $S_{ij}(.)$,

$$\begin{aligned}
\frac{\partial S_i(d)}{\partial d} &= -\lim_{\Delta d\to 0}\frac{1}{\Delta d}\Pr(d < X_{n+1} \leqslant d+\Delta d \mid J_n = i) \\
&= -\lim_{\Delta d\to 0}\frac{1}{\Delta d}\Pr(d < X_{n+1} \leqslant d+\Delta d \mid J_n = i, X_{n+1} > d) \times \Pr(X_{n+1} > d \mid J_n = i) \\
&= -S_i(d) \times \lim_{\Delta d\to 0}\frac{1}{\Delta d}\sum_{j=1}^{s}\Pr(J_{n+1} = j, d < X_{n+1} \leqslant d+\Delta d \mid J_n = i, X_{n+1} > d) \\
&= -S_i(d) \times \sum_{j=1}^{s}\lambda_{ij}(d).
\end{aligned}$$

La solution de cette équation différentielle sachant que $S_i(0) = 1$ est

$$\begin{aligned}
S_i(d) &= \exp(-\sum_{j=1}^{s}\int_0^d \lambda_{ij}(u)du) \\
&= \exp(-\Lambda_i(d)).
\end{aligned} \tag{III.19}$$

où $\Lambda_i(.) = \sum_{j=1}^{s}\Lambda_{ij}(.)$ avec $\Lambda_{ij}(d) = \int_0^d \lambda_{ij}(u)du$ est l'intensité cumulée de l'état i vers l'état j au temps d.

En utilisant la définition de l'intégrale et l'équation (III.19), le noyau de transition du processus semi-Markovien peut s'écrire $\forall i, j \in E$ et $d \in \bar{\mathbb{R}}^+$,

$$\begin{aligned}
Q_{ij}(d) &= \Pr(J_{n+1} = j, X_{n+1} \leqslant d \mid J_n = i) \\
&= \int_0^d \Pr(J_{n+1} = j, X_{n+1} \in]u, u+du] \mid J_n = i) \\
&= \int_0^d \Pr(X_{n+1} > u \mid J_n = i) \times \Pr(J_{n+1} = j, X_{n+1} \in]u, u+du] \mid X_{n+1} > u, J_n = i) \\
&= \int_0^d S_i(u) \times \lambda_{ij}(u)du \\
&= \int_0^d \exp(-\Lambda_i(u)) \times \lambda_{ij}(u)du.
\end{aligned} \tag{III.20}$$

Ainsi,

$$\frac{\partial Q_{ij}(d)}{\partial d} = q_{ij}(d) = \exp(-\Lambda_i(d)) \times \lambda_{ij}(d). \tag{III.21}$$

En utilisant les équations (III.19) et (III.21) on peut, sans perte d'information, réécrire la vraisemblance (III.18) sous la forme suivante,

$$L = \prod_{h=1}^{n}\prod_{k=1}^{N_h}\exp(-\Lambda_{J_{h,k-1}}(X_{h,k})) \times \lambda_{J_{h,k-1}J_{h,k}}(X_{h,k}) \times \left[\exp(-\Lambda_{J_{h,N_h}}(U_h))\right]^{\delta_h}$$

et par conséquent la log-vraisemblance est

$$l = \sum_{h=1}^{n}\left[\sum_{k=1}^{N_h}\left(\log\lambda_{J_{h,k-1}J_{h,k}}(X_{h,k}) - \Lambda_{J_{h,k-1}}(X_{h,k})\right) - \delta_h\Lambda_{J_{h,N_h}}(U_h)\right] \tag{III.22}$$

avec δ_h vaut 1 si l'individu est censuré en S_{h,N_h} et 0 sinon.

4.3 Estimation non-paramétrique des intensités

L'intensité de transition du processus semi-Markovien $\lambda_{ij}(d)$, peut être approximée par une fonction constante par morceaux $\lambda^*_{ij}(d)$ définie par $\lambda^*_{ij}(d) = \lambda_{ij}(\upsilon_l) = \lambda_{ijl}$ pour $d \in I_l =]\upsilon_l, \upsilon_{l+1}]$, où $(\upsilon_l)_{0 \leqslant l \leqslant M-1}$ est une subdivision régulière de $[0, D]$, avec

$$D = \max_{h=1,...,n} \left[\max(X_{h,1}, ..., X_{h,N_h}, U_h)\right].$$

D représente la plus grande durée écoulée dans un état parmi tous les individus. Notons que, dans le cas d'observations discrètes, les durées de séjour sont disponibles uniquement si les états du modèle sont hiérarchiques. Pour considérer que les durées de séjour sont observées dans le cas des modèles avec états réversibles, il faut supposer que le sujet ne change pas d'état entre deux consultations consécutives (observation continue du sujet).

Dans le cas où D est connue, on peut considérer la subdivision suivante,

$$0 = \upsilon_0 < \upsilon_1 < \upsilon_2 < \cdots < \upsilon_{M-1} < \upsilon_M = D,$$

avec des sauts $\Delta_D = D/M$ où $M = [D^{1+\alpha}]$ avec $0 < \alpha < 1$ et $[x]$ représente la partie entière de x. Ouhbi et Limnios [1999] montrent que Δ_D est asymptotiquement équivalent à $D^{-\alpha}$ quand D tend vers l'infini. Cette considération est aussi utilisée dans Colvert et Boardman [1999]. Ainsi, l'approximation de $\lambda_{ij}(d)$ est donnée par

$$\lambda^*_{ij}(d) = \sum_{l=0}^{M-1} \lambda_{ijl} \mathbb{1}_{]\upsilon_l, \upsilon_{l+1}]}(d).$$

En remplaçant $\lambda_{ij}(.)$ par $\lambda^*_{ij}(.)$ dans (III.22), la log-vraisemblance peut s'écrire sous la forme suivante :

$$l = \sum_{h=1}^{n} \sum_{i,j \in E} \sum_{l=0}^{M-1} \left(\log\left(\lambda_{ijl}\right) d^h_{ijl} - \lambda_{ijl} \nu^h_{il} \right), \tag{III.23}$$

où ν^h_{il} est la fonction du temps de séjour dans l'état i, pour l'individu h, sur l'intervalle de temps I_l, *i.e*

$$\nu^h_{il} = \sum_{k=1}^{N_h} (X_{h,k} \wedge \upsilon_{l+1} - \upsilon_l) \mathbb{1}_{\{J_{h,k-1}=i, X_{h,k} \geqslant \upsilon_l\}} + \delta_h \times (U_h \wedge \upsilon_{l+1} - \upsilon_l) \mathbb{1}_{\{J_{h,N_h}=i, U_h \geqslant \upsilon_l\}} \tag{III.24}$$

et d^h_{ijl} est le nombre de transitions, pour l'individu h, de l'état i vers l'état j pour lesquelles le temps de séjour observé dans l'état i appartient à I_l, *i.e*

$$d^h_{ijl} = \sum_{k=1}^{N_h} \mathbb{1}_{\{J_{h,k-1}=i, J_{h,k}=j, X_{h,k} \in I_l\}}. \tag{III.25}$$

Preuve *D'après la définition de* $\lambda_{ij}^*(.)$, *on peut définir*

$$\begin{aligned}\Lambda_i^*(d) &= \sum_{j=1}^{s}\int_0^d \lambda_{ij}^*(u)du \\ &= \sum_{j=1}^{s}\sum_{l=0}^{M-1}\lambda_{ijl}\int_0^d \mathbb{1}_{]\upsilon_l,\upsilon_{l+1}]}(u)du \\ &= \sum_{j=1}^{s}\sum_{l=0}^{M-1}\lambda_{ijl}\times(d\wedge\upsilon_{l+1}-\upsilon_l)\mathbb{1}_{\{d\geqslant\upsilon_l\}}. \end{aligned} \tag{III.26}$$

Ainsi,

$$\begin{aligned}\Lambda_{J_{h,k}}^*(d) &= \sum_{j=1}^{s}\sum_{l=0}^{M-1}\lambda_{J_{h,k}jl}\times(d\wedge\upsilon_{l+1}-\upsilon_l)\mathbb{1}_{\{d\geqslant\upsilon_l\}} \\ &= \sum_{i,j\in E}\sum_{l=0}^{M-1}\lambda_{ijl}\times(d\wedge\upsilon_{l+1}-\upsilon_l)\mathbb{1}_{\{J_{h,k}=i,d\geqslant\upsilon_l\}}. \end{aligned} \tag{III.27}$$

De plus,

$$\begin{aligned}\log\lambda_{J_{h,k-1}J_{h,k}}^*(d) &= \log\left(\sum_{l=0}^{M-1}\lambda_{J_{h,k-1}J_{h,k}l}\mathbb{1}_{]\upsilon_l,\upsilon_{l+1}]}(d)\right) \\ &= \sum_{i,j\in E}\log\left(\sum_{l=0}^{M-1}\lambda_{ijl}\mathbb{1}_{]\upsilon_l,\upsilon_{l+1}]}(d)\right)\mathbb{1}_{\{J_{h,k-1}=i,J_{h,k}=i\}} \\ &= \sum_{i,j\in E}\sum_{l=0}^{M-1}\log(\lambda_{ijl})\,\mathbb{1}_{\{J_{h,k-1}=i,J_{h,k}=i,d\in I_l\}}, \end{aligned} \tag{III.28}$$

comme λ_{ijl} *est constant sur* I_l. *En utilisant les équations (III.27) et (III.28) et en remplaçant* $\lambda_{ij}(.)$ *par* $\lambda_{ij}^*(.)$ *et* $\Lambda_i(.)$ *par* $\Lambda_i^*(.)$ *dans la log-vraisemblance (III.22) on a*

$$\begin{aligned} l &= \sum_{h=1}^{n}\left[\sum_{k=1}^{N_h}\left(\log\lambda_{J_{h,k-1}J_{h,k}}^*(X_{h,k})-\Lambda_{J_{h,k-1}}^*(X_{h,k})\right)-\delta_h\Lambda_{J_{h,N_h}}^*(U_h)\right] \\ &= \sum_{h=1}^{n}\left\{\sum_{i,j\in E}\sum_{l=0}^{M-1}\left[\sum_{k=1}^{N_h}\left(\log(\lambda_{ijl})\,\mathbb{1}_{\{J_{h,k-1}=i,J_{h,k}=i,X_{h,k}\in I_l\}}-\lambda_{ijl}(X_{h,k}\wedge\upsilon_{l+1}-\upsilon_l)\mathbb{1}_{\{J_{h,k-1}=i,X_{h,k}\geqslant\upsilon_l\}}\right)\right.\right. \\ &\quad \left.\left.-\ \delta_h\times\lambda_{ijl}(U_h\wedge\upsilon_{l+1}-\upsilon_l)\mathbb{1}_{\{J_{h,N_h}=i,U_h\geqslant\upsilon_l\}}\right]\right\}. \end{aligned}$$

Pour simplifier les écritures, on utilise les quantités suivantes

$$d_{ijl}^h=\sum_{k=1}^{N_h}\mathbb{1}_{\{J_{h,k-1}=i,J_{h,k}=j,X_{h,k}\in I_l\}}.$$

et

$$\nu_{il}^h = \sum_{k=1}^{N_h} (X_{h,k} \wedge \upsilon_{l+1} - \upsilon_l) \mathbb{1}_{\{J_{h,k-1}=i, X_{h,k} \geqslant \upsilon_l\}} + \delta_h (U_h \wedge \upsilon_{l+1} - \upsilon_l) \mathbb{1}_{\{J_{h,N_h}=i, U_h \geqslant \upsilon_l\}}$$

En utilisant les quantités précédentes, la log-vraisemblance devient

$$l = \sum_{h=1}^{n} \sum_{i,j \in E} \sum_{l=0}^{M-1} \left(\log(\lambda_{ijl})\, d_{ijl}^h - \lambda_{ijl} \nu_{il}^h \right).$$

■

A partir de cette écriture, on peut obtenir les dérivées de la log-vraisemblance (III.23), $\forall i, j \in E$ et $\forall l = 0, ..., M-1$,

$$\frac{\partial l}{\partial \lambda_{ijl}} = \frac{\sum_{h=1}^n d_{ijl}^h}{\lambda_{ijl}} - \sum_{h=1}^n \nu_{il}^h$$

Un estimateur du maximum de vraisemblance de λ_{ijl} est donné pour $\forall i, j \in E$ et $\forall l = 0, ..., M-1$, par

$$\hat{\lambda}_{ijl} = \begin{cases} \frac{\sum_{h=1}^n d_{ijl}^h}{\sum_{h=1}^n \nu_{il}^h} & \text{si } \sum_{h=1}^n \nu_{il}^h > 0 \\ 0 & \text{sinon,} \end{cases} \tag{III.29}$$

avec la convention $0/0 = 0$ (le numérateur sera nul quand $\sum_{h=1}^n \nu_{il}^h = 0$) et δ_h égale à 1 si l'individu h est censuré, 0 sinon. L'estimateur de $\lambda_{ij}(.)$ est finalement défini par

$$\hat{\lambda}_{ij}(d) = \sum_{l=0}^{M-1} \hat{\lambda}_{ijl} \mathbb{1}_{]\upsilon_l, \upsilon_{l+1}]}(d).$$

Remarque 10

- *L'estimateur non-paramétrique présenté ici est bien une généralisation de l'estimateur proposé par Ouhbi et Limnios [1999]. En effet, si on considère un seul processus censuré, alors $n = 1$ et $\delta_h = 1$. On retrouve ainsi l'estimateur $\hat{\lambda}_{ijl} = d_{ijl}/\nu_{il}$ obtenu par Ouhbi et Limnios [1999].*
- *L'estimateur du maximum de vraisemblance utilisé est une généralisation aux modèles semi-Markoviens de l'estimateur du maximum de vraisemblance de la fonction de risque associée à une variable aléatoire à densité continue à partir d'un échantillon de variable i.i.d (Singpurwalla et Wong [1983]).*

4.4 Estimateurs dérivés

A partir de l'estimateur non-paramétrique des intensités de transition du processus semi-Markovien, $\hat{\lambda}_{ij}(.)$, on peut déduire un estimateur du noyau semi-Markovien $Q_{ij}(.)$ et un estimateur de la matrice des probabilités de transition du processus semi-Markovien.

4.4.1 Estimateur du noyau semi-Markovien

En utilisant l'estimateur constant par période $\hat{\lambda}_{ij}(.)$ et la relation (III.20), un estimateur du noyau semi-Markovien (III.1) est donné par

$$\hat{Q}_{ij}(d) = \int_0^d \exp(-\hat{\Lambda}_i(u))\hat{\lambda}_{ij}(u)du.$$

où $\hat{\Lambda}_i(d) = \sum_{j=1}^s \int_0^d \hat{\lambda}_{ij}(u)du$. Cet estimateur peut s'écrire pour tous les temps d de la subdivision $\upsilon_0 < \upsilon_1 \cdots < \upsilon_M$. En utilisant le fait que $\exp(-\hat{\Lambda}_i(u))\hat{\lambda}_{ij}(u)$ est constant $\forall u \in I_k$, on obtient

$$\begin{aligned}\hat{Q}_{ij}(d) &= \int_0^d \exp(-\hat{\Lambda}_i(u))\hat{\lambda}_{ij}(u)du. \\ &= \sum_{k:0\leqslant\upsilon_k\leqslant d} \int_{I_k} \exp(-\hat{\Lambda}_i(u))\hat{\lambda}_{ij}(u)du \\ &= \Delta_D \sum_{k:0\leqslant\upsilon_k\leqslant t} \exp(-\hat{\Lambda}_{ik})\hat{\lambda}_{ijk}\end{aligned}$$

avec $\forall k = 1, ..., M$, $\hat{\Lambda}_{ik} = \sum_{j=1}^s \sum_{l=0}^{k-1} \hat{\lambda}_{ijl}\Delta_D$ et $\Delta_D = \upsilon_k - \upsilon_{k-1}$ (d'après l'équation (III.26)).

4.4.2 Estimateur des probabilités du processus semi-Markovien

L'estimateur $\hat{\lambda}_{ij}(.)$ permet aussi de déduire un estimateur de la matrice des probabilités de transition du processus semi-Markovien. A partir de la définition des probabilités du processus (III.8) et des équations (III.3), (III.4) et (III.5), on peut écrire,

$$\begin{aligned}\Psi_{ij}(t) &= \sum_{r=1}^s \int_0^t p_{ir} f_{ir}(u)\Psi_{rj}(t-u)du + \delta_{ij}S_i(t) \qquad \text{(III.30)} \\ &= \sum_{r=1}^s \int_0^t Q_{ir}(du)\Psi_{rj}(t-u) + \delta_{ij}(1 - \sum_{j=1}^s Q_{ij}(t))\end{aligned}$$

Considérons le produit de convolution entre deux matrices $\mathbf{A} = \{a_{ij}(t)\}$ et $\mathbf{B} = \{b_{ij}(t)\}$ noté $\mathbf{C} = \{c_{ij}(t)\}$,

$$(\mathbf{A} * \mathbf{B})(t) = \mathbf{C}(t) \text{ avec } c_{ij}(t) = \sum_{r=1}^s \int_0^t a_{ir}(t-u)b_{rj}(du)$$

Ainsi, en considérant, les matrices $\mathbf{Q}(.) = \{Q_{ij}(.)\}_{i,j\in E}$, $\mathbf{P}(.) = \{\Psi_{ij}(.)\}_{i,j\in E}$ et la matrice diagonale $\mathbf{S}(.) = diag(S_i(.))_{i\in E}$, les probabilités du processus peuvent s'écrire sous forme matricielle

$$\mathbf{P}(t) = (\mathbf{P} * \mathbf{Q})(t) + \mathbf{S}(t).$$

En définissant $\mathbf{A}^{(-1)}$ l'inverse de la matrice $\mathbf{A}$ au sens de la convolution et $\mathbf{I} = diag(1\ 1_{\{t\geqslant 0\}}(t))$, alors si les fonctions $S_i(.)$ sont intégrables, on a

$$\mathbf{P}(t) = (\mathbf{I}(t) - \mathbf{Q}(t))^{(-1)} * \mathbf{S}(t).$$

A partir de cette écriture, on peut en déduire un estimateur non-paramétrique de la matrice des probabilités $\mathbf{P}(.)$

$$\hat{\mathbf{P}}(t) = (\mathbf{I} - \hat{\mathbf{Q}}(t))^{(-1)} * \hat{\mathbf{S}}(t),$$

où $\hat{\mathbf{Q}}(.)$ est l'estimateur de $\mathbf{Q}(.)$ et $\hat{\mathbf{S}}(.)$ l'estimateur de $\mathbf{S}(.)$. La matrice $(\mathbf{I} - \hat{\mathbf{Q}}(t))^{(-1)}$, qui est l'inverse au sens de la convolution, peut être soit approximée par un développement en serie entière (Limnios [1997]), soit calculée par la formule d'algèbre suivante

$$(\mathbf{I} - \hat{\mathbf{Q}}(t))Com(\mathbf{I} - \hat{\mathbf{Q}}(t))^t = \det(\mathbf{I} - \hat{\mathbf{Q}}(t))\mathbf{Id}$$

où $\mathbf{Id}$ est la matrice identité, $Com(\mathbf{A})$ est la comatrice de $\mathbf{A}$. La seule différence avec le cas classique d'inversion de matrice réside dans le calcul du déterminant où le produit usuel est remplacé par le produit de convolution (Limnios [1997]).

4.5 Propriétés asymptotiques des estimateurs

Dans le cas d'un seul processus censuré, Ouhbi et Limnios [1999] ont démontré que les estimateurs $\hat{\lambda}_{ij}(.)$, $\hat{Q}_{ij}(.)$, $\hat{P}_{ij}(.)$ étaient uniformément consistants et convergeaient faiblement vers des variables aléatoires normales quand le temps de censure tend vers l'infini (Les auteurs obtiennent les variances asymptotiques).

Dans le cas de plusieurs processus censurés ou non, les mêmes types de résultats asymptotiques peuvent être obtenus en considérant que n (nombre d'individus) tend vers l'infini. L'adaptation des démonstrations au cas de plusieurs processus semble possible en considérant que les n processus forment un seul « super » processus.

5 Application à l'asthme

Dans la plupart des maladies chroniques comme l'asthme, le temps passé dans un état de santé représente un facteur important de l'évolution de la maladie. Les modèles semi-Markoviens prennent en compte cette échelle de temps et représentent ainsi un outil intéressant pour l'étude de l'asthme.

Le modèle à trois états de contrôle (Figure III.3) est considéré pour étudier la base de données présentée au chapitre I page 9.

L'écriture de la vraisemblance dans les modèles semi-Markoviens fait intervenir, pour chaque individu censuré, la durée entre la dernière consultation et la date de fin de l'étude. On définit la date de fin d'étude comme la dernière consultation renseignée dans la base.

L'objectif de ce travail est d'étudier l'évolution de l'asthme en prenant en compte certains facteurs de risque. Nous considérons des covariables indépendantes du temps afin de pouvoir stratifier la base de données. Les covariables étudiées sont les suivantes :

- l'indice de masse corporelle à la première consultation : codée 0 si IMC < 25, 1 sinon ;
- la sévérité de l'asthme à la première consultation : codée 0 si le patient est non sévère, 1 sinon.

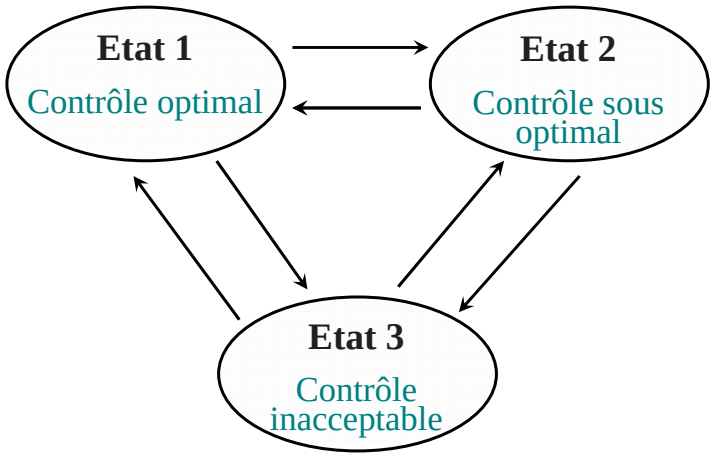


FIG. III.3 – Modèle à trois états de contrôle pour l'asthme.

5.1 Application de l'estimation paramétrique

Les lois de Weibull et de Weibull généralisée sont utilisées pour la modélisation paramétrique des distributions des temps de séjour. Afin d'interpréter les résultats obtenus, nous étudierons particulièrement les **intensités des temps de séjour** (Equation (III.7)) et les **intensités du processus semi-Markovien** (Equation (III.6)).

- **Les intensités des temps de séjour** sont modélisées par des lois paramétriques. Des covariables peuvent être incorporées dans la définition de ces intensités par l'intermédiaire d'un modèle à risques proportionnels (Equation III.13). Ainsi, l'effet des covariables s'interprète en termes de risques relatifs par l'intermédiaire des coefficients de régression. Cependant, la définition de ces intensités fait apparaître un conditionnement sur le fait que l'individu est dans l'état i et qu'il sera ensuite dans l'état j.

- **Les intensités du processus semi-Markovien** sont les paramètres que l'on cherche à estimer. Ces intensités sont plus intéressantes pour l'interprétation car elles représentent le risque d'un individu de passer dans l'état j sachant qu'il est dans l'état i. Dans l'estimation non-paramétrique, ces intensités sont estimées (directement) par des fonctions constantes par morceaux. Dans le cadre de l'estimation paramétrique, d'après les équations (III.6), (III.15) et (III.4), les intensités de transition du processus semi-Markovien sont estimées par

$$\hat{\lambda}_{ij}(d) = \frac{\hat{p}_{ij}\hat{S}_{ij}(d)\hat{\alpha}_{ij}(d)}{\hat{S}_i(d)} = \frac{\hat{p}_{ij}\hat{S}_{ij}(d)\hat{\alpha}_{ij}(d)}{\sum_{j=1}^{s}\hat{p}_{ij}\hat{S}_{ij}(d)}. \tag{III.31}$$

où $\hat{\alpha}_{ij}(.)$ est l'estimation de l'intensité de temps de séjour. Pour ces intensités, l'interprétation des coefficients de régression en terme de risque relatif devient complexe (III.31). Cependant, si les estimations obtenues pour chaque valeur des covariables sont à peu près parallèles, il est possible en faisant le rapport entre les deux courbes d'obtenir un facteur multiplicatif constant qui est utile pour l'interprétation de ces risques. Si les courbes ne sont pas parallèles, on obtient un facteur multiplicatif dépendant du temps qui peut également être interprété.

Les fonctions *optim*() de *R* et *nlminb*() de *S-Plus* sont utilisées pour obtenir les estimations par maximum de vraisemblance. En ce qui concerne l'initialisation des paramètres, ceux associés à la définition des lois sont initialisés à 1, ceux associés à la chaîne de Markov sous-jacente sont estimés à partir de simples proportions et ceux associés aux coefficients de régression sont initialisés à 0 (aucun effet).

Dans un premier temps, un modèle sera ajusté pour chaque modalité des covariables. Cette étape permettra d'identifier les facteurs influençant les intensités de temps de séjour et d'évaluer graphiquement la validité de l'hypothèse de proportionnalité des risques pour chaque transition et chaque covariable. Dans un second temps, on étudiera un modèle avec une seule covariable influençant toutes les transitions. L'impact des covariables sur les distributions des temps de séjour pourra ainsi être mesuré par l'intermédiaire des coefficients de régression. Enfin, il sera possible d'étudier un modèle avec des distributions et des effets de covariables spécifiques à chaque transition. L'application de cette méthode d'estimation à une base de données de patients atteints du VIH fait l'objet d'un travail soumis (Foucher et al. [2004]).

5.1.1 Modèle stratifié

Dans un premier temps, un modèle est estimé pour chaque modalité des covariables (analyse en sous-groupe). Cette étape possède plusieurs intérêts : elle permet
(*i*) d'identifier les variables qui semblent avoir un effet sur les estimations des intensités des temps de séjour,
(*ii*) de vérifier que la loi utilisée est justifiée par rapport à une loi exponentielle (loi la plus simple),
(*iii*) d'évaluer graphiquement la validité de l'hypothèse de proportionnalité des risques, propre à chaque covariable et à chaque transition.

Des modèles de type Weibull et Weibull généralisé sont ajustés dans chaque strate. Les graphiques présentant les estimations pour chaque transition ne sont pas présentés exhaustivement pour des raisons de clarté. Afin de discuter les résultats, nous nous intéresserons particulièrement à la transition $3 \rightarrow 1$ stratifiée sur l'IMC et à la transition $2 \rightarrow 3$ stratifiée sur la sévérité.

De manière générale, les résultats de l'estimation des risques des temps de séjour (Equation (III.7)) montrent des écarts entre les courbes pour certaines transitions et pour certaines covariables, ce qui souligne l'intérêt de prendre en compte ces facteurs dans l'étude des forces de transition. Par exemple, les figures III.4 **(a)** et III.4 **(b)** montrent les écarts dans les estimations pour une stratification sur l'IMC et sur la sévérité. Pour une modélisation par loi de Weibull ou Weibull généralisée, il semble que les patients en surpoids aient un risque plus faible que les patients normaux pour la transition de l'état inacceptable vers l'état optimal. Les patients sévères ont un risque plus important pour la transition de l'état sous-optimal vers l'état inacceptable.

On peut aussi noter une différence entre les estimations par une loi de Weibull et par une loi de Weibull généralisée, en particulier sur la forme des estimations. En effet, la majeure partie des transitions répondent à des fonctions de risque en forme de $\cap$ qui ne sont pas prises

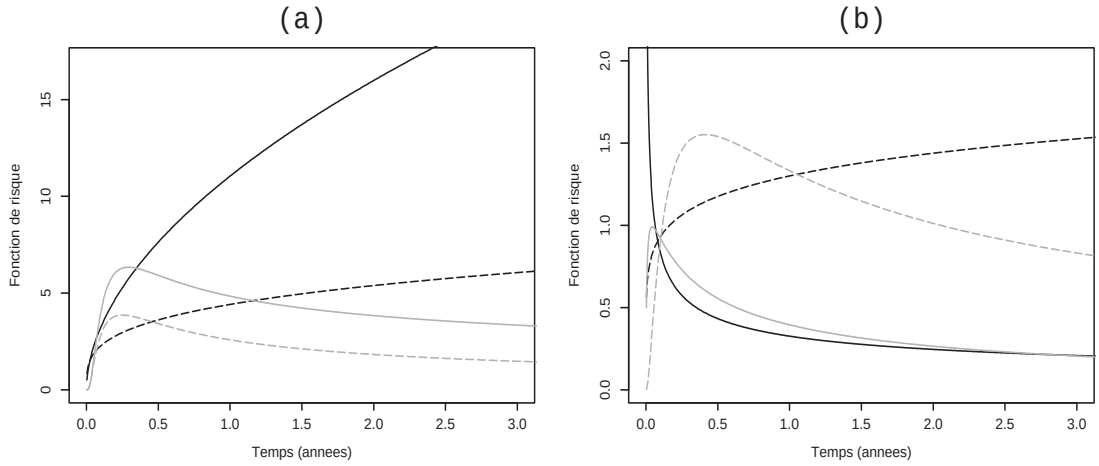


FIG. III.4 – Estimations des intensités du temps de séjour par des lois de Weibull (courbes noires) et Weibull généralisée (courbes grises). **(a)** Temps de séjour dans un état inacceptable vers un état optimal ($3 \to 1$) dans les strates IMC < 25 (——) et IMC $\leqslant 25$ (- - -). **(b)** Temps de séjour dans un état sous-optimal vers un état inacceptable ($2 \to 3$) dans les strates non sévère (——) et sévère (- - -).

en compte par les lois de Weibull. Peu de temps après l'entrée dans un état, le risque croît rapidement ce qui reflète le caractère d'instabilité du patient qui vient de changer d'état. Ensuite, après un délai variable, le risque diminue, signe d'une stabilisation de la maladie : plus la personne reste dans un état, moins elle a de chance d'en sortir. Par exemple, sur les figures III.4 **(a)** et III.4 **(b)**, la loi de Weibull généralisée est mieux adaptée pour modéliser ces risques qui semblent être en forme de $\cap$. En effet, il apparaît (graphiquement) que la monotonie de la loi de Weibull ne correspond pas aux données et fournit ainsi des estimations peu fiables des risques. Ce résultat est confirmé par les estimations des paramètres associés à la définition des lois : le paramètre θ_{ij} de la loi de Weibull généralisée est toujours différent de la valeur 1. Ceci signifie (*cf.* remarque page 50) que dans le cas de l'asthme, la loi de Weibull généralisée semble mieux adaptée que la loi de Weibull (et exponentielle).

Afin d'étudier un modèle avec covariables pour les intensités des temps de séjour, il est important de vérifier que l'hypothèse de proportionnalité des risques ne soit pas trop contraignante. Une comparaison graphique des intensités des temps de séjour dans chaque strate permet de sélectionner les transitions qui semblent vérifier l'hypothèse. Si les risques dans chaque strate ne se croisent pas et sont à peu près parallèles, alors l'hypothèse de proportionnalité sera considérée comme étant vérifiée. Pour chaque covariable et pour chacune des lois utilisées, le tableau III.1 résume les transitions qui semblent vérifier l'hypothèse de risques proportionnels. Le symbole $\times$ signifie que la transition ne vérifie pas l'hypothèse et le symbole O signifient que la transition vérifie l'hypothèse. Par exemple, pour la transition $2 \to 3$ avec la sévérité comme variable de stratification (Figure III.4 **(b)**), les fonctions de risque des deux strates se croisent ce qui reflète la non proportionnalité des risques. Inversement, la transition $3 \to 1$ (Figure III.4 **(a)**) semble être à risques proportionnels dans le cas d'une loi de Weibull généralisée. On peut aussi noter, que pour certaines transitions, la forme

Transition	Loi	Covariable	
		IMC	Sévérité
$1 \rightarrow 2$			
	Weibull	O	$\times$
	Weibull G	$\times$	$\times$
$1 \rightarrow 3$			
	Weibull	$\times$	$\times$
	Weibull G	$\times$	$\times$
$2 \rightarrow 1$			
	Weibull	$\times$	$\times$
	Weibull G	$\times$	$\times$
$2 \rightarrow 3$			
	Weibull	$\times$	$\times$
	Weibull G	$\times$	$\times$
$3 \rightarrow 1$			
	Weibull	O	$\times$
	Weibull G	O	O
$3 \rightarrow 2$			
	Weibull	O	O
	Weibull G	$\times$	$\times$

TAB. III.1 – Transitions qui semblent vérifier l'hypothèse de proportionnalité des risques.

de la loi de distribution diffère selon les modalités de la covariable. Dans de telles situations, seule la stratification permet d'étudier les impacts des covariables sur les intensités.

Les figures III.5 **(a)** et III.5 **(b)** montrent les risques du processus semi-Markovien obtenus avec une loi de Weibull et Weibull généralisée pour la transition $3 \rightarrow 1$ avec stratification sur l'IMC et les risques pour la transition $2 \rightarrow 3$ avec stratification sur la sévérité. Dans l'ensemble, les estimations avec les deux lois sont assez proches même si les estimations par loi de Weibull semblent sous-estimer et étaler la forme en $\cap$ du risque semi-Markovien. Notons qu'une modélisation des risques de temps de séjour par une loi de Weibull n'empêche pas d'obtenir des formes en $\cap$ pour les risques du processus (III.31).

D'un point de vue clinique, les résultats (Figure III.5 **(a)**) confirment ceux obtenus au chapitre II, à savoir un effet négatif du surpoids sur un retour vers un état de contrôle optimal. En effet, avec une loi de Weibull et avec une loi de Weibull généralisée, le risque de transiter vers un état optimal après être resté trois mois dans un état inacceptable est deux fois plus grand pour les patients qui ne sont pas en surpoids. Ces risques sont différents pendant les six premiers mois où la forme des risques est en cloche et sont voisins ensuite quand les risques se stabilisent. Pour la sévérité (Figure III.5 **(b)**), la forme des risques est différente dans chacune des strates, en cloche pour les patients sévères alors que le risque est presque constant pour les patients non sévères. Les patients sévères dans un état de contrôle sous-optimal, ont un risque plus important (surtout au début) de devenir mal contrôlés.

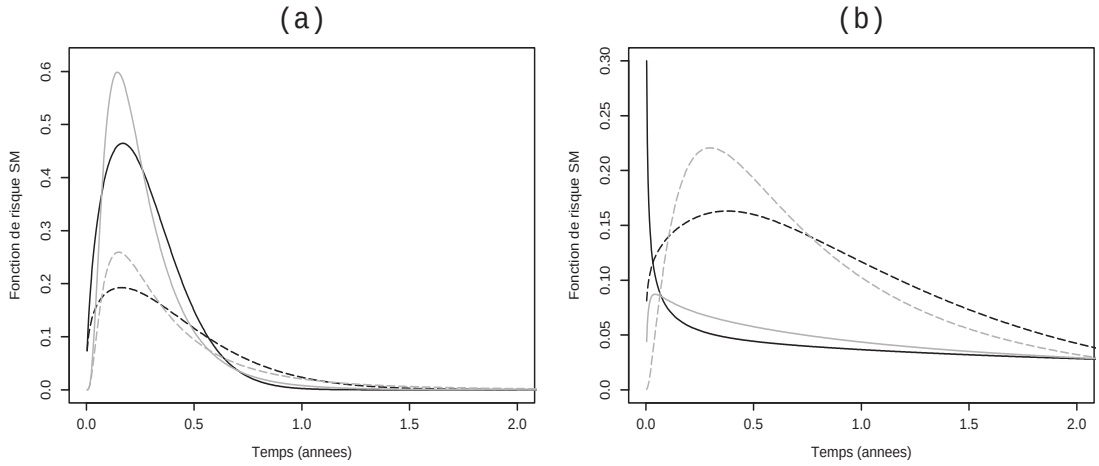


FIG. III.5 – Estimations des intensités du processus semi-Markovien en utilisant des lois de Weibull (courbes noires) et Weibull généralisée (courbes grises). **(a)** Intensité de transition d'un état inacceptable vers un état optimal ($3 \to 1$) dans les strates IMC < 25 (——) et IMC $\leqslant 25$ (- - -). **(b)** Intensité de transition d'un état sous-optimal vers un état inacceptable ($2 \to 3$) dans les strates non sévère (——) et sévère (- - -).

5.1.2 Modèle univarié

Pour chaque covariable, nous ajustons un modèle « univarié » dans le sens où une seule covariable est prise en compte dans la modélisation. Nous considérons un modèle à risques proportionnels pour inclure les covariables dans la définition des intensités de temps de séjour. Cette étape permet de tester l'impact des covariables sur chaque transition à l'aide des tests de Wald. On pourra ensuite interpréter les estimations des coefficients de régression en termes de risques relatifs. En pratique, afin d'éviter les interprétations abusives, on utilisera les résultats obtenus par stratification pour sélectionner de manière graphique les transitions qui semblent vérifier l'hypothèse de risques proportionnels.

Les résultats des estimations des coefficients de régression, des écarts-types et les p-values (test de $\beta_{ij} = 0$ avec le test de Wald) pour chaque covariable et pour chaque loi utilisée sont donnés dans le tableau III.2.

D'une manière générale, les estimations des coefficients avec des lois de Weibull et de Weibull généralisée sont proches et vont dans le même sens. Les coefficients de régression qui sont statistiquement différents de zéro et pour lesquels la transition semble vérifier l'hypothèse de proportionnalité (Tableau III.1) peuvent être interprétés. Parmi ces coefficients, on note en particulier que le coefficient associé à l'IMC pour la transition $3 \to 1$ est négatif (avec les deux lois). Ainsi, sachant que le patient est dans l'état inacceptable et qu'il va transiter dans un état optimal, le risque de transiter est plus faible pour les patients en surpoids comparativement aux autres patients (le risque est divisé par $\exp(0.53) = 1,7$). Ces résultats confirment ceux obtenus par stratification.

La comparaison des figures III.4 et III.6 reflète les difficultés d'interprétation. Les estimations des intensités des temps de séjour pour la transtion $3 \to 1$ avec l'IMC en covariable

Transition	Loi	IMC			Sévérité		
		$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
$1 \to 2$							
	Weibull	-0.36	(0.24)	(0.12)	-0.42	(0.24)	(0.09)
	Weibull G	-0.38	(0.23)	(0.10)	-0.55	(0.23)	(0.02)
$1 \to 3$							
	Weibull	-0.84	(0.35)	(0.02)	0.03	(0.33)	(0.94)
	Weibull G	-0.65	(0.35)	(0.06)	-0.07	(0.36)	(0.84)
$2 \to 1$							
	Weibull	-0.02	(0.20)	(0.90)	-0.38	(0.26)	(0.15)
	Weibull G	-0.06	(0.20)	(0.78)	-0.54	(0.24)	(0.02)
$2 \to 3$							
	Weibull	-0.22	(0.28)	(0.42)	-0.97	(0.24)	(<0.01)
	Weibull G	-0.14	(0.32)	(0.65	-1.03	(0.24)	(<0.01)
$3 \to 1$							
	Weibull	-0.58	(0.21)	(<0.01)	0.52	(0.21)	(0.01)
	Weibull G	-0.53	(0.22)	(0.01)	0.38	(0.20)	(0.07)
$3 \to 2$							
	Weibull	-0.16	(0.18)	(0.39)	0.10	(0.19)	(0.59)
	Weibull G	-0.24	(0.18)	(0.18)	0.02	(0.18)	(0.91)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta_{ij} = 0$.

TAB. III.2 – Estimations des coefficients de régression dans un modèle semi-Markovien avec une covariable.

(Figure III.6 **(a)**) et par stratification sur l'IMC (Figure III.4 **(a)**) sont assez proches et vont dans le même sens. Par contre, pour la transition $2 \to 3$ avec la sévérité (Figures III.4 **(b)** et III.6 **(b)**), les estimations par stratification et avec covariables conduisent à des résultats contradictoires sur l'effet de la sévérité. Cette différence peut s'expliquer par le fait que, pour la sévérité, l'hypothèse de proportionnalité ne semble vérifiée pour cette transition. Cette exemple montre la prudence nécessaire pour l'interprétation clinique des résultats obtenus avec un modèle à risques proportionnels.

Il est intéressant d'observer l'impact des covariables sur les intensités du processus semi-Markovien. La figure III.7 **(a)** confirme l'impact négatif du surpoids pour un retour vers un état de contrôle optimal même s'il est légèrement atténué par rapport au modèle stratifié (Figure III.5 **(a)**). Les patients en surpoids ont un risque plus faible (divisé par 1.5 à $t = 2$ *mois*) les cinq premiers mois et ont ensuite un risque légèrement supérieur les six mois qui suivent. Les résultats sur la sévérité (Figure III.7 **(b)**) sont différents de ceux obtenus par stratification (III.5 **(b)**). Cependant, comme on l'a vu précédemment, ces résultats ne doivent pas être retenus.

Le choix des distributions des temps de séjour peut s'effectuer en observant les coefficients associés à la définition des lois. Dans notre cas, avec le modèle stratifié et avec le modèle

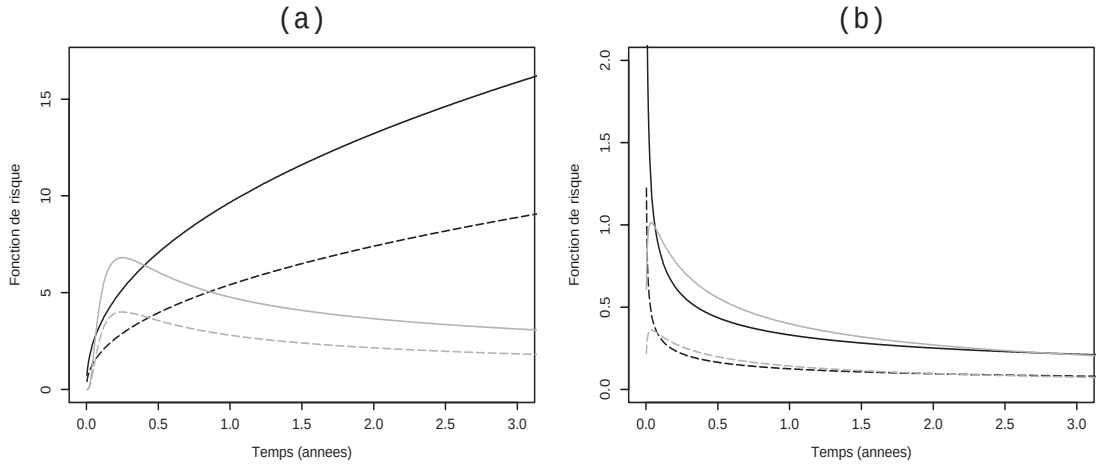


FIG. III.6 – Estimations des intensités du temps de séjour par des lois de Weibull (courbes noires) et Weibull généralisée (courbes grises). **(a)** Temps de séjour dans un état inacceptable vers un état optimal ($3 \to 1$) avec l'IMC en covariable (IMC < 25 (——) ; IMC $\geqslant 25$ (- - -)). **(b)** Temps de séjour dans un état sous-optimal vers un état inacceptable ($2 \to 3$) avec la sévérité en covariable (non sévère (——) ; sévère (- - -)).

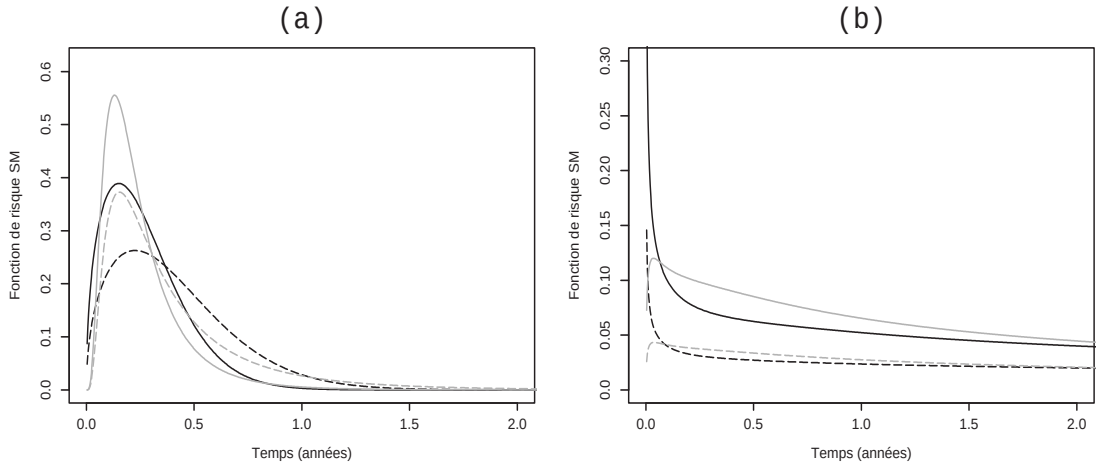


FIG. III.7 – Estimations des intensités du processus semi-Markovien en utilisant des lois de Weibull (courbes noires) et Weibull généralisée (courbes grises). **(a)** Intensité de transition d'un état inacceptable vers un état optimal ($3 \to 1$) avec l'IMC en covariable (IMC < 25 (——) ; IMC $\geqslant 25$ (- - -)). **(b)** Intensité de transition d'un état sous-optimal vers un état inacceptable ($2 \to 3$) avec la sévérité en covariable (non sévère (——) ; sévère (- - -)).

univarié, les coefficients θ_{ij} sont tous différents de la valeur 1. Ce résultat montre bien l'utilité de la loi de Weibull généralisée dans le cas de l'asthme où les lois exponentielles et les lois de Weibull ne sont pas adaptées.

Remarque 11 *L'hypothèse de proportionnalité des risques peut être testée et contournée en considérant un modèle avec des coefficients de régresssion dépendants du temps. Cette méthode est discutée au chapitre IV page 96. Dans notre cas, le nombre d'observations dans la base de données n'est pas suffisant pour une estimation satisfaisante des paramètres.*

5.1.3 Modèle multivarié avec transition spécifique

Suite à l'analyse des covariables par stratification et par un modèle univarié, il est possible de définir un modèle avec plusieurs covariables pour étudier d'éventuels facteurs de confusion. Ce modèle pourra prendre en compte des lois de temps de séjour et des effets de covariables spécifiques à chaque transition. Dans un premier temps, il est nécessaire de sélectionner, pour chaque covariable, les transitions qui semblent à risques proportionnels. Par cette sélection, on exclut inévitablement de la modélisation un certain nombre d'effets des covariables. Ensuite, seuls les effets statistiquement différents de zéro dans le modèle univarié sont pris en compte dans la modélisation. Il est ensuite possible de choisir une loi spécifique à chaque transition en faisant un compromis entre le nombre de paramètres et l'adéquation de la loi.

Dans le cas de l'asthme, les lois de Weibull généralisées sont les mieux adaptées à la modélisation des risques. Cependant, pour un modèle de type de Weibull généralisé avec deux covariables ($5 \times 6 + 3$ paramètres), les estimations sont mauvaises et peu fiables. On est confronté à un nombre trop important de paramètres pour le nombre d'observations de la base de données. De plus, pour chaque covariable, peu de transitions vérifient l'hypothèse de proportionnalité. Ainsi, les résultats obtenus avec ce type d'analyse ne sont pas présentés car moins intéressants et moins fiables que ceux des analyses univariée et stratifiée.

5.2 Application de l'estimation non-paramétrique

Cette partie présente l'application de la méthode d'estimation non-paramétrique des **intensités du processus semi-Markovien** (Equation III.6).

Afin d'estimer les intensités par des fonctions constantes par morceaux, la méthode nécessite de définir une subdivision régulière de $[0, D]$, avec D représente la plus grande durée écoulée dans un état parmi tous les individus. Le pas de la subdivision est donné par $\Delta_D = D/M$ où $M = [D^{1+\alpha}]$ avec $0 < \alpha < 1$ et $[x]$ représente la partie entière de x. Le choix de la subdivision revient alors à fixer une valeur de $\alpha \in]0, 1[$. D'après (III.29) et (III.25), la valeur de la fonction sur un intervalle est nulle si aucune transition ne se produit dans cet intervalle. Afin d'améliorer la lisibilité des courbes, seules les valeurs différentes de zéro sont conservées dans la représentation graphique.

La figure III.8 présente les estimations des intensités du processus semi-Markovien associées à la transition $3 \to 1$ dans les strates IMC < 25 et IMC $\geqslant 25$. Chaque graphique

correspond à différentes valeurs de α ($\alpha = 0.05$, $\alpha = 0.1$, $\alpha = 0.3$, $\alpha = 0.5$). Suivant les valeurs de α, les risques ont sensiblement la même allure, cependant, ils comportent des différences importantes. Les valeurs des risques augmentent de manière significative avec la valeur de α. L'impact du surpoids semble également différent : pour $\alpha = 0.05$ et $\alpha = 0.1$, les patients en surpoids semblent avoir un risque plus faible alors que ce n'est plus le cas pour $\alpha = 0.3$ et $\alpha = 0.5$. Notons que les valeurs des fonctions peuvent varier fortement d'un intervalle à un autre en fonction du nombre de transitions. D'un point de vue numérique, le temps de calcul augmente considérablement avec la valeur de α.

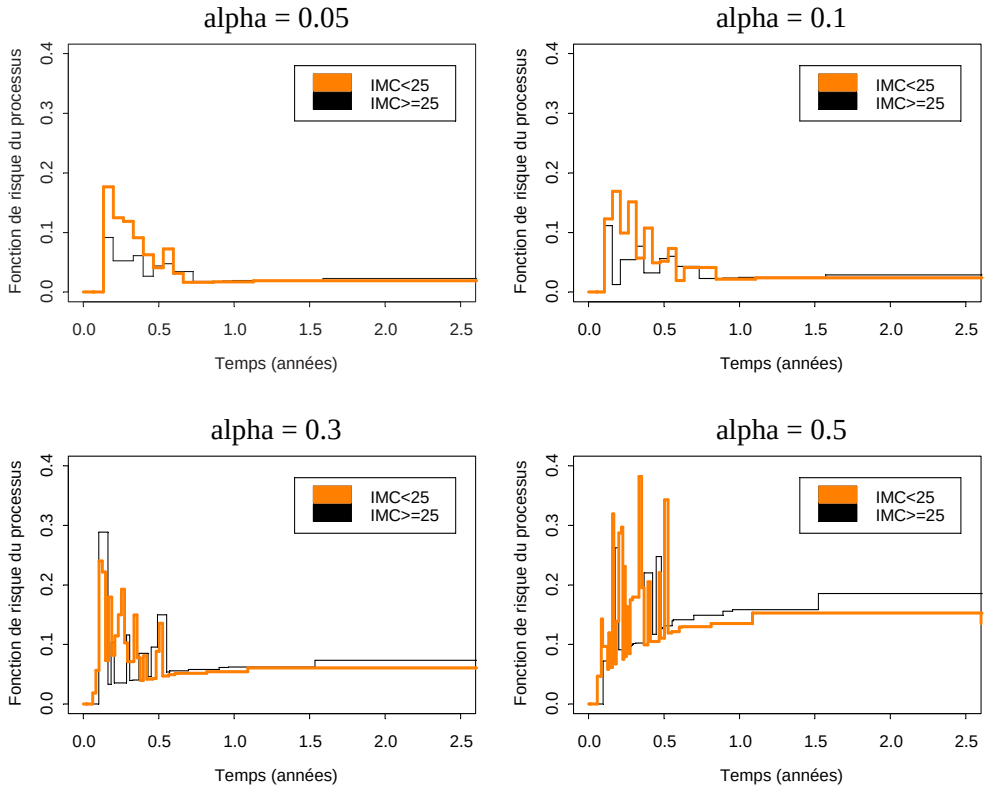


FIG. III.8 – Estimations non-paramétrique des intensités du processus semi-Markovien associée à la transition $3 \to 1$ dans les strates IMC < 25 et IMC $\geqslant 25$. Chaque figure correspond à une valeur de α.

Afin de comparer les estimations paramétrique et non-paramétrique, la valeur de α est fixée à 0.3. Ce choix correspond à un compromis mais reste arbitraire. Les figures III.9 **(a)** et III.9 **(b)** présentent les estimations paramétrique et non-paramétrique des intensités du processus semi-Markovien. Les deux estimations des risques pour la strate IMC < 25 (Figure III.9 **(a)**), diffèrent dans leurs valeurs maximales. La valeur de stabilisation est différente mais ne peut être interprétée car les valeurs nulles (nombreuses au delà de 7 ou 8 mois) sont supprimées. Cependant, les deux estimations ont en commun leurs formes en cloche sur les huit premiers mois. Dans la strate des patients en surpoids (Figure III.9 **(b)**), les valeurs maximales des estimations sont proches et se produisent pour le même temps. Les deux

estimations restent néanmoins différentes, en particulier à cause des natures paramétrique et non-paramétrique.

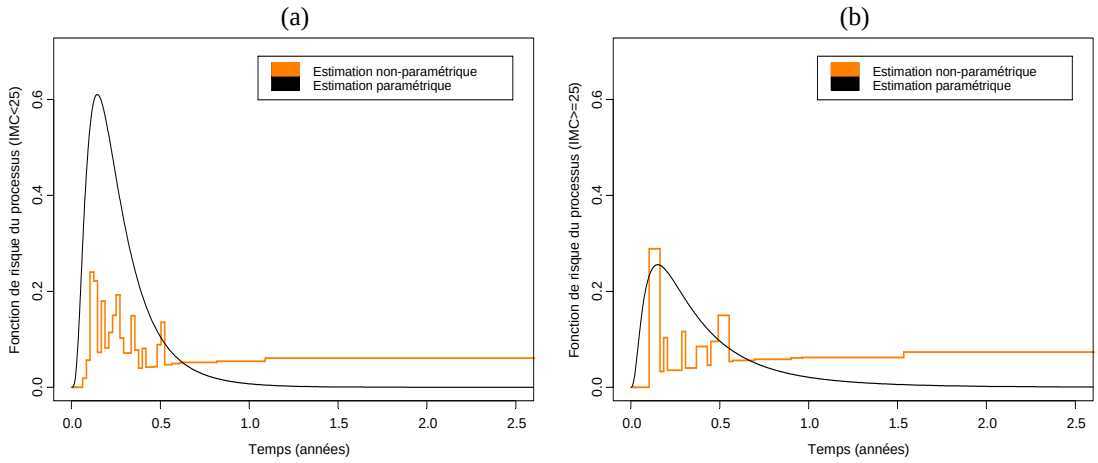


FIG. III.9 – Comparaison des estimations paramétrique et non-paramétrique ($\alpha = 0.3$) des intensités du processus semi-Markovien associées à la transition $3 \to 1$. **(a)** strate IMC < 25; **(b)** strate IMC $\geqslant 25$.

6 Discussion

Dans ce chapitre, deux méthodes d'estimation pour modèles semi-Markoviens homogènes ont été présentées. Nous avons proposé une extension de la méthode non-paramétrique étudiée par Ouhbi et Limnios [1999]. Cette méthode permet d'estimer de manière non-paramétrique les intensités d'un processus semi-Markovien (dans un modèle sans ou avec états absorbants) à partir d'un échantillon $i.i.d$ de processus. Nous avons également présenté une méthode d'estimation paramétrique qui consiste à estimer les distributions des temps de séjour et les probabilités de la chaîne de Markov sous-jacente par des fonctions paramétriques. Il est ensuite possible de déduire des estimations des intensités du processus semi-Markovien. La méthode est adaptée afin de pouvoir choisir des lois de temps de séjour et un nombre de covariables spécifiques à chaque transition. Outre les lois exponentielles et Weibull qui sont généralement utilisées, nous avons considéré pour la modélisation des lois de Weibull généralisées pour prendre en compte les formes en $\cup$ et $\cap$ des distributions des temps de séjour.

6.1 Application

Dans un premier temps, la méthode d'estimation paramétrique est appliquée au cas de l'asthme. D'un point de vue clinique, les résultats sur l'impact de l'IMC sur l'évolution de la maladie sont les plus intéressants. En effet, tous les résultats obtenus avec la stratification

et avec les modèles univariés montrent qu'un patient en surpoids a moins de chances de passer d'un état inacceptable vers un état optimal. Ces résultats sont à mettre en relation avec ceux obtenus avec un modèle de Markov homogène ($cf.$ chapitre II). En effet, avec les deux modèles, le surpoids diminue la possibilité d'une transition de l'état inacceptable vers un état optimal.

L'utilisation des modèles semi-Markoviens semble justifiée dans le cas de l'asthme. En effet, les lois exponentielles (implicites à une modélisation Markovienne homogène) ne sont pas retenues pour la modélisation des intensités de temps de séjour. Les lois de Weibull généralisées sont toujours les mieux adaptées pour modéliser les formes en cloche des intensités de temps de séjour dans les états de contrôle de l'asthme. Par conséquent, l'estimation des risques de temps de séjour et des risques du processus est plus précise avec des lois de Weibull généralisées. Cependant, dans notre cas, l'impact de la distribution est à relativiser. En effet, les estimations des coefficients de régression et des intensités du processus obtenues avec des lois de Weibull sont relativement proches de celles obtenues avec des lois de Weibull généralisées.

Dans le cas de l'asthme, la méthode d'estimation non-paramétrique fournit moins de résultats intéressants. En effet, les estimations des intensités varient fortement en fonction du choix de la subdivision. De plus, la nature non-paramétrique des estimations rend les résultats difficilement interprétables d'un point de vue clinique. Cependant, la comparaison des méthodes paramétrique et non-paramétrique montre que même si les estimations diffèrent en plusieurs points, la forme en cloche de ces estimations est sensiblement la même.

6.2 Méthodes

La méthode d'estimation paramétrique fournit plusieurs résultats pouvant être interprétés d'un point de vue clinique. Tout d'abord, l'effet des covariables sur les risques des temps de séjour peut être quantifié par l'intermédiaire des coefficients de régression (même si les risques des temps de séjour sont difficiles à interpréter de par leurs définitions). Ensuite, les risques du processus semi-Markovien fournissent des résultats visuels faciles à interpréter en terme d'évolution de la maladie.

Le modèle à risques proportionnels est très utile pour incorporer des covariables dans la modélisation. Cependant, l'hypothèse de proportionnalité est souvent contraignante et impose une prudence dans l'interprétation des résultats. L'utilisation de la stratification permet d'observer l'impact des covariables sans avoir des résultats biaisés par l'hypothèse de proportionnalité. Cependant, avec cette méthode, l'impact des covariables n'est pas quantifié et on est confronté aux problèmes liés à l'analyse en sous-groupes.

Dans les différents choix de modélisation (distributions des temps de séjour, nombre de covariables), on doit faire un compromis entre le nombre de paramètres du modèle et le nombre d'observations de la base de données : si le nombre de paramètres est trop important, les estimations sont peu fiables et difficilement interprétables.

La méthode d'estimation non-paramétrique est attractive car elle permet d'obtenir directement les estimations des intensités du processus semi-Markovien sans faire aucune hypothèse sur la forme des distributions des temps de séjour. Cependant, la nature non-paramétrique de l'estimation rend l'interprétation délicate. De plus, le choix de la subdivision nécessaire à l'application de la méthode peut influencer fortement les estimations.

Notons également que la présentation actuelle de la méthode ne permet pas d'introduire des covariables. L'estimation non-paramétrique reste néanmoins intéressante pour une première analyse des données. En effet, cette méthode permet de se faire une idée générale de la forme des intensités du processus semi-Markovien.

Afin de développer la méthode d'estimation non-paramétrique, il serait intéressant d'étudier le choix optimal de la subdivision (par exemple, en choisissant la valeur de α qui maximise la vraisemblance). De même, on pourrait essayer d'introduire des covariables dans la méthode. Il sera également important d'étudier les propriétés asymptotiques des estimateurs non-paramétriques en adaptant les démonstrations de Ouhbi et Limnios [1999]. Dans un contexte paramétrique, on pourrait développer une méthode d'estimation des distributions des temps de séjour qui utiliserait les bases de fonctions splines (à noeuds fixes ou variables) ou encore les bases de Fourier.

Chapitre IV

Modèle de Markov non-homogène

1 Introduction

Les modèles de Markov homogènes (*cf.* chapitre II page 21) sont de plus en plus appliqués en épidémiologie. La méthodologie est bien connue et plusieurs programmes sont disponibles pour ajuster ce type de modèles (Alioum et Commenges [2001], Jackson [2005]). Cependant, dans de nombreuses applications, l'hypothèse d'homogénéité (intensités de transition constantes au cours du temps) n'est pas adaptée et s'avère trop restrictive.

Les modèles de Markov non-homogènes comme les modèles semi-Markoviens, peuvent être ajustés pour complexifier la modélisation de l'évolution du processus. Ces deux modèles sont comparables en terme de flexibilité et le choix du modèle dépend de l'échelle de temps la plus importante dans une application donnée. Un modèle semi-Markovien sera adapté pour étudier l'impact de la durée écoulée dans un état avant de transiter alors qu'un modèle non-homogène accordera de l'importance à la durée du suivi.

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 notamment, par l'intermédiaire de la théorie des martingales, un cadre rigoureux à de nombreuses problématiques (en particulier concernant les propriétés asymptotiques des estimateurs). Le nombre de publications lié au sujet est très important : on peut sans être exhaustif noter les nombreuses contributions de Aalen O., Andersen P.K., Gill R.D., Keiding N., Borgan Ø., Dabrowska D.M. ou encore de Hougaard P. La théorie des processus de comptage permet d'obtenir des estimateurs non-paramétriques dans les modèles de Markov non-homogènes. En particulier, l'estimateur 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 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, des estimateurs de Kaplan-Meier et de la vraisemblance partielle de Cox. Cette théorie confère une justification mathématique à l'approche « heuristique » permettant la construction de ces estimateurs.

Ces méthodes constituent une alternative 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 complexes et la programmation des estimateurs apparaît compliquée. Même si, récemment, une macro SAS permettant d'ajuster un modèle à deux états a été publiée (Paes et de Lima [2004]), il n'existe pas de programme permettant d'obtenir ces estimateurs pour des modèles multi-états spécifiques.

Ce chapitre aborde des méthodes d'estimation basées sur la théorie des processus de comptage. Afin de rendre ce travail accessible tant au mathématicien qu'à l'épidémiologiste, les notions mathématiques complexes relatives à la théorie des processus de comptage sont présentées en annexe page 161.

La première partie de ce chapitre, présente les résultats essentiels liés aux processus de comptage et leur relation avec les processus de Markov. La prise en compte de la censure et le cas particulier du modèle homogène sont également considérés. La deuxième partie présente les estimateurs non-paramétriques de Nelson-Aalen et de Aalen-Johansen. La troisième partie décrit l'adaptation de ces estimateurs à un modèle semi-paramétrique de type Cox. Ensuite, l'adaptation de cette théorie au cas des modèles semi-Markoviens est abordée. La dernière partie du chapitre considère l'application de ces méthodes à la base de données sur l'asthme afin de modéliser l'évolution en prenant en compte la durée du suivi. Les résultats de ce chapitre font l'objet d'une publication (Saint-Pierre et al. [2005c]) acceptée dans la revue *Far East Journal of Theoritical Statistics*.

Nous présentons également en annexe (page 177), un « guide » détaillant la programmation des estimateurs à partir d'une matrice décrivant les données. L'objet de ce guide est de permettre à chacun d'implémenter (dans tous les langages de programmation) un modèle spécifique à l'application donnée (choix du nombre d'états, du nombre de transition, du nombre de covariables). Notons que les programmes ont été validés sur le jeu de données utilisé par Andersen et al. [1991] et Andersen et al. [1993]. Cette méthode de programmation fait l'objet d'un travail (Saint-Pierre et al. [2004]) soumis dans la revue *Computer Methods and Programs in Biomedicine.*

2 Processus de Markov et processus de comptage

2.1 Processus de Markov

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

Définition 3 Un processus de Markov à temps continu est complètement défini par

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

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

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

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

$$p_{hj}(s,t) = \mathbb{P}\{X(t) = j \mid X(s) = h\}, \qquad \forall 0 \leqslant s \leqslant t, \forall h, j \in S,$$

avec

$$\mathbf{P}(s,s) = \mathbf{Id}, \ \sum_{j=1}^{k} p_{hj}(s,t) = 1 \qquad \text{pour tout } h \text{ et } 0 \leqslant s \leqslant t.$$

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

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

et telle que pour $h \neq j$, $A_{hj}(.)$ soit une fonction cadlag (continue à droite avec une limite à gauche) non décroissante, nulle en zéro. A_{hj} est la fonction d'intensité cumulée pour les transitions de l'état h vers l'état j, alors que A_{hh} est l'opposée de la fonction d'intensité cumulée pour les transitions qui quittent l'état h. Les équations différentielles de Kolmogorov définissent le lien entre la matrice de probabilité de transition et la matrice d'intensité cumulée

– équation « forward » de Kolmogorov

$$\frac{\partial \mathbf{P}(s,t)}{\partial t} = \mathbf{P}(s,t)\,\mathbf{A}(dt), \tag{IV.1}$$

– équation « backward » de Kolmogorov

$$\frac{\partial \mathbf{P}(s,t)}{\partial s} = \mathbf{A}(ds)\,\mathbf{P}(s,t). \tag{IV.2}$$

Remarque 12 *La notation $A(dt)$ renvoie à l'écriture formelle de l'intégrale stochastique (cf. page 161). Si $A(t)$ est un processus croissant alors,*

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

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, la notation $\mathcal{P}_{]s,t]}$ suggère une version continue du produit ordinaire comme l'intégrale $\int$ généralise la somme $\sum$.

Théorème 1 Soit $\mathbf{A}$ une matrice de fonctions de dimension $k \times k$ correspondant à une mesure d'intensité (matrice des intensités cumulées). Soit $\mathbf{Id}$ la matrice identité.

Alors la matrice,

$$\mathbf{P}(s,t) = \underset{u\in]s,t]}{\mathcal{P}} (\mathbf{Id} + \mathbf{A}(du)), \qquad s \leqslant t,\ t,s \in \mathcal{T}, \tag{IV.3}$$

est la matrice de probabilité de transition d'un processus de Markov à espace d'états fini $\{1,\ldots,k\}$.

Quand la fonction $\mathbf{A}$ est une fonction en escalier (cadlag), le produit intégral (IV.3) devient un produit fini sur les temps de sauts de $\mathbf{A}$; ainsi

$$\mathbf{P}(s,t) = \prod_{i=1}^{n} (\mathbf{Id} + \Delta\mathbf{A}(T_k)), \qquad s \leqslant t,$$

où $s < T_1 < \cdots < T_n = t$ sont les temps de sauts et $\Delta\mathbf{A}(T_k) = \mathbf{A}(T_k) - \mathbf{A}(T_{k-1})$ et $T_0 = 0$.

Des résultats complémentaires sur le produit intégral sont données en annexe page 162.

2.2 Processus de comptage

La notion de processus de comptage constitue la base de la théorie des processus stochastiques. A partir de ces processus ponctuels, différentes classes de processus avec des trajectoires de saut peuvent être définies à savoir les processus de Markov ou semi-Markov.

Définition 4 Un **processus aléatoire de comptage** $N(.)$ sur $\mathcal{T}$ est une fonction aléatoire

$$\begin{array}{rcl} N : \mathcal{T} \times \Omega & \longrightarrow & \mathbb{N} \\ (t,\omega) & \longmapsto & N(t,\omega), \end{array}$$

cadlag, adaptée (*cf.* page 161), 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** $\{\mathcal{F}_t : t \in \mathcal{T}\}$ associée au processus N est définie par la σ-algèbre

$$\mathcal{F}_t = \sigma(N(u), 0 \leqslant u \leqslant t).$$

Définition 5 Un processus de comptage k-dimensionnel $\mathbf{N} = (N_1, N_2, ..., N_k)$ est appelé **processus de comptage multivarié** si chacune de ses composantes est un processus de comptage univarié et s'il ne peut y avoir simultanéité des sauts de deux (ou plus) de ses composantes.

La proposition qui suit, permet de faire le lien entre la théorie des processus de comptage et l'écriture usuelle des processus Markoviens. Un processus de Markov à temps continu et à espace d'états discret peut s'écrire comme un processus de comptage.

Proposition 1 Soit A la mesure d'intensité d'un processus de Markov $\{X(t), t \in \mathcal{T}\}$. Définissons les fonctions suivantes

$$Y_h(t) = \mathbb{1}_{\{X(t-)=h\}},$$

$$N_{hj}(t) = card\left\{s \leqslant t : X\left(s^-\right) = h, X(s) = j\right\} \qquad h \neq j.$$

$\mathbf{N} = (N_{hj}; h \neq j)$ est un processus de comptage multivarié. Le **processus d'intensité cumulée** (ou le **compensateur** de $\mathbf{N}$) par rapport à $\sigma\{X(s), s \leqslant t\}$ est $\mathbf{\Lambda} = (\Lambda_{hj}; h \neq j)$ avec

$$\Lambda_{hj}(t) = \int_0^t Y_h(s) A_{hj}(ds). \tag{IV.4}$$

De plus, les processus M_{hj} définis par

$$M_{hj} = N_{hj} - \Lambda_{hj} \tag{IV.5}$$

sont des **martingales**.

Le processus de comptage $N_{hj}(t)$ compte le nombre de transitions observées (du processus X) de l'état h vers l'état j dans l'intervalle de temps $[0,t]$. La quantité $Y_h(t)$ renseigne sur l'état du processus : si le processus X est dans l'état h juste avant l'instant t alors $Y_h(t) = 1$; $Y_h(t) = 0$ sinon.

La décomposition (IV.5) du processus de comptage en son compensateur et une martingale est un résultat essentiel de la théorie des processus de comptage. Cette décomposition permet notamment d'obtenir les résultats asymptotiques par l'intermédiaire du théorème de la limite centrale pour martingales.

Remarque 13

- *La connaissance du processus de comptage* $\mathbf{N} = (N_{hj}; h \neq j)$ *sur* $[0,t]$, *apporte la même information que l'observation du processus* $X(u), 0 \leqslant u \leqslant t$.
- *Des définitions complémentaires concernant les processus de comptages, les martingales et les processus ponctuels marqués sont données en annexe page 163.*

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

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

Dans ce cas, la proposition précédente, implique que le processus $\mathbf{N}$ a un **processus d'intensité multiplicative** $\boldsymbol{\lambda} = (\lambda_{hj}; h \neq j)$ par rapport à $\mathcal{F}_t = \sigma(\mathbf{N}(u), 0 \leqslant u \leqslant t)$ telle que

$$\lambda_{hj}(t) = \alpha_{hj}(t)Y_h(t).$$

Notons que $\lambda_{hj}(.)$ est une variable aléatoire par l'intermédiaire de $Y_h(t)$; $\alpha_{hj}(.)$ est déterministe. Les fonctions $\alpha_{hj}(.)$ sont appelées les **intensités de transition** et sont définies par

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{p_{hj}(t, t+\Delta t)}{\Delta t}, \qquad h \neq j,$$
$$\alpha_{hh}(t) = -\sum_{h \neq j} \alpha_{hj}(t), \qquad h = 1, ..., k.$$

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

2.3 Vraisemblance

Considérons, le processus de comptage $\mathbf{N} = \{N_{hj}(t), h, j \in S, h \neq j\}$, sa filtration naturelle $\mathcal{F}_t$ et $\mathbf{\Lambda} = (\Lambda_{hj}; h, j \in S, h \neq j)$ son compensateur par rapport à $\mathcal{F}_t$.

La $\mathcal{F}_t$-vraisemblance basée sur l'observation de $\mathbf{N}(t)$ s'obtient par le théorème 6 page 166. Conditionnellement à l'état initial, la vraisemblance complète est donnée par

$$\mathcal{L} = \underset{t \in [0,\tau]}{\mathcal{P}} \left\{ \prod_{h=1}^{k} \prod_{j \neq h} \left[d\Lambda_{hj}(t) \right]^{\Delta N_{hj}(t)} \left[1 - \sum_{h=1}^{k} \sum_{j \neq h} \Lambda_{hj}(t) \right]^{1 - \sum_{h=1}^{k} \sum_{j \neq h} \Delta N_{hj}(t)} \right\}. \tag{IV.6}$$

où $\Delta Z(t) = Z(t) - Z(t^-)$. D'après l'équation (IV.4) et comme $A_{hj}(t) = \int_0^t \alpha_{hj}(u)du$, la vraisemblance s'écrit,

$$\mathcal{L} = \underset{t \in \mathcal{T}}{\mathcal{P}} \left\{ \prod_{h=1}^{k} \prod_{j \neq h} \left(\alpha_{hj}(t) Y_h(t) \right)^{\Delta N_{hj}(t)} \left(1 - \sum_{h=1}^{k} \sum_{j \neq h} \alpha_{hj}(t) Y_h(t) dt \right)^{1 - \sum_{h=1}^{k} \sum_{j \neq h} \Delta N_{hj}(t)} \right\} \tag{IV.7}$$

2.4 Processus de comptage et censure à droite

2.4.1 Définitions

Définition 6 Une variable de censure U est définie par la possible non-observation de l'événement. Si l'on observe U, et non T, et que l'on sait que $T > U$ (respectivement $T < U$, $U_1 < T < U_2$), on dit qu'il y a censure à droite (respectivement censure à gauche, censure par intervalle).

X peut être considérée comme la durée séparant un événement initial A d'un événement final B, ou comme la durée pendant laquelle un sujet reste dans un état donné (auquel cas A désigne l'entrée dans cet état et B la sortie de cet état – par exemple le chômage). La censure à droite, dont il sera essentiellement question par la suite, est due à la non-observation de B, dont on sait seulement qu'il sera postérieur à la dernière date d'observation du sujet.

Dans les études de cohorte, on rencontre la censure à droite pour deux raisons :

(i) le sujet n'est pas rentré dans un état absorbant avant la fin de l'étude (date de point), on parle « d'exclu vivant »,

(ii) le sujet a quitté l'étude en cours pour différentes raisons, par exemple un déménagement ou un refus de continuer à participer à la cohorte, c'est ce que l'on nomme des « perdus de vue ».

Par ailleurs, la censure se distingue de la troncature : une observation est dite tronquée si elle est conditionnelle à un autre événement. Par exemple, pour des données de survie, on dit qu'il y a troncature à gauche (respectivement à droite) lorsque la variable d'intérêt T_i (durée de vie du $i^{ème}$ individu) n'est observable qu'à la condition $T_i > c$ (respectivement $T_i < c$), où $c > 0$ est un seuil fixé. La troncature élimine de l'étude une partie des T_i, ce qui a pour conséquence de faire porter l'analyse uniquement sur la loi de T conditionnellement à l'événement $\{T < c\}$ (respectivement $\{T > c\}$)

Dans ce qui suit, nous verrons comment la censure à droite, une forme de données incomplètes parmi les plus courantes, peut être prise en compte dans les modèles basés sur les processus de comptage.

2.4.2 Notations

Afin d'aborder la censure par les processus de comptage, considérons :

– $\mathbf{N}(t)$ le processus de comptage multivarié non censuré par rapport à la filtration naturelle $\mathcal{F}_t = \sigma(\mathbf{N}(u),\ 0 \leqslant u \leqslant t)$,

$$\mathbf{N}(t) = (N_{hj}(t), h, j = 1, ..., k, h \neq j),$$

où

$$N_{hj}(t) = \mathbb{1}_{\{T_{hj} \leqslant t\}}$$

avec T_{hj} le temps d'apparition de la transition de l'état h vers j.

– $Y_h(t)$ les indicateurs permettant de savoir si le processus X est dans l'état h juste avant le temps t,

$$Y_h(t) = \mathbb{1}_{\{X(t-)=h\}}, \qquad h = 1, \ldots, k.$$

– $\mathbf{C}(t)$ le processus de censure à droite,

$$\mathbf{C}(t) = (C_h(t), h = 1, ..., k),$$

avec

$$C_h(t) = I_{\{t \leqslant U_h\}},$$

où U_h est le temps de censure par rapport à l'état h.

Remarque 14 *Dans la plupart des études épidémiologiques, le processus de censure ne dépend que de l'individu et non du type d'événement. C'est pourquoi, dans la suite, l'indice h est suprimé . Les résultats obtenus peuvent facilement être adaptés au cas de différents mécanismes de censure pour différents types d'événements.*

L'objectif est d'étudier le processus de comptage $\mathbf{N}(.)$ en tenant compte de la censure à droite. Celle-ci introduit une variation aléatoire supplémentaire que la tribu $\mathcal{F}_t$ ne prend pas en compte. Cette dernière a donc besoin d'être élargie. Notons $\mathcal{G}_t$ cette nouvelle filtration :

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(C(u), 0 \leqslant u \leqslant t),$$

Considérons, les processus réellement observés après censure :

– $\mathbf{N}^c(.)$ le processus de comptage censuré à droite qui représente la partie observable de $\mathbf{N}(.)$:

$$\mathbf{N}^c(t) = \left(N^c_{hj}(t), h, j = 1, ..., k, h \neq j\right),$$

avec

$$N^c_{hj}(t) = \int_0^t C(s)\, dN_{hj}(s),$$

– $Y^c_h(t)$ tel que,

$$Y^c_h(t) = C(t)\, Y_h(t), \qquad h = 1, \ldots, k.$$

2.4.3 Censure à droite indépendante

Un problème fréquent avec la censure à droite est qu'elle peut modifier les intensités des événements d'interêt. Si, par exemple, dans un essai clinique, les patients particulièrement malades sont enlevés de l'étude, les patients qui restent à risque ne sont plus représentatifs de l'échantillon total et ainsi les intensités sont différentes d'une situation sans censure. La notion de censure à droite indépendante permet d'éviter ce problème.

Définition 7 Soit $\mathbf{N}(.)$ un processus de comptage multivarié de compensateur $\mathbf{\Lambda}(.)$ par rapport à la filtration $\mathcal{F}_t$. Soit $C(.)$ un processus de censure à droite prévisible par rapport à $\mathcal{G}_t \supseteq \mathcal{F}_t$.

Alors, la censure à droite générée par $C(.)$ est **indépendante** si le compensateur de $\mathbf{N}(.)$ par rapport à $\mathcal{G}_t$ est aussi $\mathbf{\Lambda}(.)$.

Autrement dit, la connaissance des temps de censure juste avant t ne modifie pas l'intensité du processus $\mathbf{N}$ au temps t. Lorsque la censure est indépendante, la répartition des temps de décès est la même pour les patients censurés et pour les patients non censurés. En pratique, cela signifie, que les individus ne doivent pas être censurés parce qu'ils ont un risque de décès particulièrement élévé ou faible. La censure devient dépendante si, par exemple, certains patients ne sont plus suivis car leur état s'est sérieusement dégradé (les personnes les plus à risque sont enlevées de l'étude).

Proposition 2 Sous l'hypothèse de censure à droite indépendante,

(1) la mesure d'intensité (le compensateur) de $\mathbf{N}^c(.)$ par rapport à $\mathcal{G}_t$ est

$$\Lambda_{hj}^c(t) = \int_0^t C(s)\, d\Lambda_{hj}(s).$$

(2) si le processus $\mathbf{N}(.)$ a une intensité multiplicative par rapport à $\mathcal{G}_t$ alors $\mathbf{N}^c(.)$ a aussi une intensité multiplicative par rapport à $\mathcal{G}_t$.

Proposition 3 Soit $\mathbf{N}^c(t) = \left(N_{hj}^c(t); h, j = 1, ..., k, h \neq j\right)$ un processus de comptage censuré. Si la censure à droite est indépendante alors la vraisemblance du modèle s'écrit

$$\mathcal{L}^c = \underset{t\in[0,\tau]}{\mathcal{P}} \prod_{h=1}^{k} \prod_{j\neq h} \left[d\Lambda_{hj}^c(t)\right]^{\Delta N_{hj}^c(t)} \left[1 - d\Lambda_{\cdot\cdot}^c(t)\right]^{1-\Delta N_{\cdot\cdot}^c(t)}. \tag{IV.8}$$

où

$$\Lambda_{\cdot\cdot}^c = \sum_{h=1}^{k} \sum_{j\neq h} \Lambda_{hj}^c,$$

et

$$N_{\cdot\cdot}^c(t) = \sum_{h=1}^{k} \sum_{j\neq h} N_{hj}^c(t).$$

A la vue des propositions précédentes, la vraisemblance partielle avec censure a la même forme que la vraisemblance sans censure (IV.6). Le fait que la forme de la vraisemblance soit préservée par la censure indépendante implique que les propriétés des martingales restent les mêmes. Ainsi, lorsque la censure est indépendante, l'inférence et la théorie asymptotique sont toujours applicables de la même manière que pour des données non censurées (par la structure de martingale).

Des résultats complémentaires sur la censure indépendante sont donnés page 172. En particulier, le principe général de la vraisemblance partielle est décrit page 167. Ce dernier est utilisé pour construire la vraisemblance associée à un processus censuré (IV.8). La notion de censure non-informative est également définie page 174.

Désormais, nous nous placerons dans le cadre d'un **mécanisme de censure indépendante**. Dans tout ce qui suit, les processus considérés seront des versions censurées mêmes si les notations ne le feront plus apparaître.

Remarque 15 *Dans certaines situations, la censure apporte une information sur l'état de santé du patient est devient dépendante. L'expression de la vraisemblance ne correspond plus à une vraisemblance complète car toute l'information n'est pas utilisée. Cette vraisemblance peut toujours être utilisée mais cela entraîne un biais dans l'estimation.*

2.4.4 Caractéristique de la censure à droite

Afin de simplifier les différentes définitions et les exemples, les mécanismes de censure à droite sont décrits dans le cadre des données de survie. En effet, les mêmes principes s'appliquent aux processus Markov.

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

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

avec

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

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

Censure non aléatoire de type I :

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

$$\begin{cases} \tilde{T}_i = \min(T_i, u) \\ D_i = \mathbb{1}_{\{T_i \leqslant u\}} \end{cases}$$

Ce mécanisme de censure est couramment utilisé dans l'industrie pour tester la durée de vie de n objets identiques sur un intervalle d'observation fixé $[0, u]$. En biologie, afin de tester l'efficacité d'une molécule sur un lot de souris, les souris survivantes sont sacrifiées au bout d'un temps déterminé u.

Remarque 16 *Bien que similaires dans l'écriture de leur définition, la censure non-aléatoire de type I à droite et la troncature à droite doivent être distinguées : en effet, l'inférence statistique diffère selon qu'elle s'applique à l'un ou l'autre de ces deux types de données de survie. Dans le cas de données censurées, la vraisemblance sera le produit d'un nombre aléatoire de facteurs ; dans le cas de données tronquées, la vraisemblance sera le produit d'un nombre fixe de facteurs.*

Censure aléatoire de type I :

Les observations pour chaque individu sont :

$$\begin{cases} \tilde{T}_i = \min(T_i, U_i) \\ D_i = I_{\{T_i \leqslant U_i\}} \end{cases}$$

où U_i est un temps de censure aléatoire indépendant de T_i. Si, de plus, les $(U_i)_{i=1,...,n}$ sont *i.i.d.* la censure aléatoire de type I est dite « classique ».

Cette censure est l'une des plus utilisées pour l'analyse de données de survie. Par exemple, dans un essai thérapeutique, on admet que les sujets entrent de façon aléatoire. Si la date d'analyse est fixée a priori, le délai entre la date d'entrée et la date de point, c'est-à-dire le temps de participation des exclus-vivants, est aléatoire. La censure aléatoire de type I est indépendante quand le processus d'événement et le processus de censure sont indépendants.

Censure de type II :

Les observations pour chaque individu sont arrêtées à un temps aléatoire commun à tous, *i.e.*

$$\begin{cases} \tilde{T}_i = \min\left(T_i, T_{(r)}\right) \\ D_i = I_{\left\{T_i \leqslant T_{(r)}\right\}} \end{cases}$$

où r est un entier fixé, $1 \leqslant r \leqslant n$. $T_{(1)}, ..., T_{(r)}, ..., T_{(n)}$ est la statistique d'ordre, *i.e.* $T_{(r)}$ correspond au $r^{ième}$ temps de décès.

Ce mécanisme de censure est utilisé dans le monde industriel, par exemple lorsque l'on veut observer la durée de fonctionnement de n machines tant que r d'entre elles ne tombent pas en panne. En biologie, pour tester l'efficacité d'un poison sur un lot de souris, la durée de l'étude correspond au temps que mettent r souris à mourir.

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

Par la suite, nous considérerons un **mécanisme de censure aléatoire de type I** et **indépendante**.

3 Estimation non-paramétrique

3.1 Observations et notations

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

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

$$N_{hji}(t) = card\left\{s \leqslant t : X_i\left(s^-\right) = h, X_i(s) = j\right\}, \qquad \forall h, j = 1, \ldots, k, h \neq j.$$

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

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

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

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

– $Y_{h+}(t)$ renseigne sur le nombre total de personne « à risque » dans l'état h juste avant l'instant t, c'est-à-dire le nombre de personnes succeptibles de subir une transition à partir de l'état h,

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

Proposition 4 Le processus $N_{hji}(t)$ satisfait aux conditions d'un modèle à intensité multiplicative, *i.e.* $\forall i = 1, \ldots, n$; $\forall h, j = 1, \ldots, k, h \neq j$,

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

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

$$\lambda_{hji}(t) = \alpha_{hj}(t) Y_{hi}(t). \tag{IV.9}$$

D'après l'équation (IV.7), la vraisemblance complète associée au processus de comptage $\mathbf{N} = \{N_{hji}(t), i = 1, \ldots, n; h, j \in S, h \neq j\}$ conditionnellement aux données initiales est donnée par,

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

Par conséquent, le processus agrégé $N_{hj+}(t)$ est un processus de comptage ayant pour intensité

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

où le premier terme est une intensité au niveau individuel et le second est un processus renseignant sur les personnes à risque.

3.2 Estimation des intensités cumulées

Pour les démonstrations des propriétés de cette section, nous renvoyons le lecteur à l'ouvrage de référence (Andersen et al. [1993]).

Dans la suite, nous considérons le processus $\mathbf{N} = \{N_{hj+}(t), h, j \in S, h \neq j\}$ obtenu après agrégation des observations (censurées ou non). Ce processus a une intensité $\boldsymbol{\lambda} = \{\lambda_{hj}(t), h, j \in S, h \neq j\}$ satisfaisant aux conditions d'un modèle à intensité multiplicative décrit précédemment.

Les intensités de transition vérifient $\int_0^t \alpha_{hj}(u)du < \infty$ pour tout $h \neq j$ et pour tout $t \in \mathcal{T}$. On considère une estimation non-paramétrique des intensités cumulées. Ainsi, aucune hypothèse supplémentaire n'est faite sur les fonctions α_{hj}.

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

Définition 8 L'estimateur de **Nelson-Aalen** des fonctions d'intensité cumulée est défini

$$\hat{A}_{hj}(t) = \int_0^t \frac{J_h(u)}{Y_{h+}(u)} dN_{hj+}(u), \qquad \forall h \neq j, \tag{IV.10}$$

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

Notons que $J_h(s)/Y_{h+}(s)$ est interprété comme étant 0 quand $Y_{h+}(s) = 0$. L'origine « naturelle » de cet estimateur provient de l'équation (IV.5) qui peut s'écrire symboliquement sous la forme

$$dN_{hj+}(t) = Y_{h+}(t)\alpha_{hj}(t)dt + dM_{hj+}(t),$$

où $dM_{hj+}(t)$ peut être considéré comme un bruit aléatoire.

Remarque 17 *Soit $T_{hj,1} < T_{hj,2} < \cdots$ les temps de saut successifs de N_{hj+}. N_{hj+} attribue une masse 1 à chacun de ces temps de saut et une masse 0 ailleurs. Il s'ensuit qu'on peut écrire $\hat{A}_{hj}(t)$ comme une simple somme*

$$\hat{A}_{hj}(t) = \sum_{\{k:T_{hj,k} \leqslant t\}} \frac{1}{Y_{h+}(T_{hj,k})}.$$

Ainsi $\hat{A}_{hj}$ est une fonction en escalier, croissante, continue à droite, de saut $1/Y_{h+}(T_{hj,k})$ au temps de saut $T_{hj,k}$ de N_{hj}.

Proposition 5 Un estimateur de la variance de $\hat{A}_{hj}(t)$ est

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

Proposition 6 Sous certaines conditions (pages 190-191 de Andersen et al. [1993]), $\hat{A}_{hj}(t)$ est un estimateur

– biaisé, tel que

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

– uniformément consistant,

– et asymptotiquement normal, tel que

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

avec U_{hj} martingales gaussiennes indépendantes telle que

$$\begin{cases} U_{hj}(0) = 0 \\ \mathbb{V}(U_{hj}(t)) = \int_0^t \frac{\alpha_{hj}(u)}{\sum_{j=1}^k p_{jh}(0,u)} du. \end{cases}$$

Le biais de cet estimateur est très faible puisqu'en pratique, la probabilité qu'à un instant t tous les individus ne soient plus à risque est proche de zéro. Les hypothèses sont discutées et les démonstrations sont données pages 179 et 190-199 de Andersen et al. [1993].

3.3 Estimation des probabilités de transition

Rappelons que la matrice des probabilités de transition est définie en fonction de la matrice des intensités cumulées par le produit intégral (IV.3),

$$\mathbf{P}(s,t) = \underset{u\in]s,t]}{\mathcal{P}} \left(\mathbf{Id} + d\mathbf{A}(u)\right) \qquad 0 < s \leqslant t,\ t,s \in \mathcal{T}.$$

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

Définition 9 L'estimateur de **Aalen-Johansen** de la matrice des probabilités de transition est défini par

$$\hat{\mathbf{P}}(s,t) = \underset{u\in]s,t]}{\mathcal{P}} \left(\mathbf{Id} + d\hat{\mathbf{A}}(u)\right), \qquad 0 < s \leqslant t,\ t,s \in \mathcal{T}, \tag{IV.11}$$

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

Remarque 18 *$p_{hj}(s,t)$ correspond à la probabilité de transiter dans l'état j à l'instant t sachant que le patient est dans l'état h à l'instant s. En pratique, on interprètera souvent la quantité $p_{hj}(0,t)$.*

Remarque 19 *En pratique, il y a un nombre fini de transitions. Soient $s < T_1 < ... < T_k < t$, les temps de transition entre deux états. L'estimateur de Aalen-Johansen devient un produit fini de matrice tel que*

$$\hat{\mathbf{P}}(s,t) = \prod_{l=1}^{k} \left(\mathbf{Id} + \Delta \hat{\mathbf{A}}(T_l) \right).$$

Dans le cas d'un modèle à trois états, où toutes les transitions entre états sont possibles,

$$\mathbf{I}+\Delta\hat{\mathbf{A}}(T_l) = \begin{pmatrix} 1 - \frac{\Delta N_{12+}(T_l)+\Delta N_{13+}(T_l)}{Y_{1+}(T_l)} & \frac{\Delta N_{12+}(T_l)}{Y_{1+}(T_l)} & \frac{\Delta N_{13+}(T_l)}{Y_{1+}(T_l)} \\ \frac{\Delta N_{21+}(T_l)}{Y_{2+}(T_l)} & 1 - \frac{\Delta N_{21+}(T_l)+\Delta N_{23+}(T_l)}{Y_{2+}(T_l)} & \frac{\Delta N_{23+}(T_l)}{Y_{2+}(T_l)} \\ \frac{\Delta N_{31+}(T_l)}{Y_{3+}(T_l)} & \frac{\Delta N_{32+}(T_l)}{Y_{3+}(T_l)} & 1 - \frac{\Delta N_{31+}(T_l)+\Delta N_{32+}(T_l)}{Y_{3+}(T_l)} \end{pmatrix},$$

où pour un processus $Z(.)$ cadlag, $\Delta Z(T_l) = Z(T_l) - Z(T_{l-1})$, pour $l = 2, ..., k$, et $\Delta Z(T_1) = Z(T_1)$.

Proposition 7 Un premier estimateur de la covariance de $\hat{p}_{hj}(s,t)$ et $\hat{p}_{mr}(s,t)$ est

$$\begin{aligned} \widehat{Cov}(\hat{p}_{hj}(s,t), \hat{p}_{mr}(s,t)) = \sum_{l=1}^{k} \sum_{g \neq l} \int_s^t & \hat{p}_{hg}(s,u)\, \hat{p}_{mg}(s,u) \{\hat{p}_{lj}(u,t) - \hat{p}_{gj}(u,t)\} \\ & \times \{\hat{p}_{lr}(u,t) - \hat{p}_{gr}(u,t)\} J_g(u) (Y_{g+}(u))^{-2} dN_{gl+}(u). \end{aligned}$$

L'estimateur de la variance de $\hat{p}_{hj}(s,t)$ est

$$\hat{\mathbb{V}}(\hat{p}_{hj}(s,t)) = \sum_{l=1}^{k} \sum_{g \neq l} \int_s^t (\hat{p}_{hg}(s,u))^2 \{\hat{p}_{lj}(u,t) - \hat{p}_{gj}(u,t)\}^2 J_g(u) (Y_{g+}(u))^{-2} dN_{gl+}(u).$$

Proposition 8 Un deuxième estimateur, de type Greenwood, de la covariance de $\hat{p}_{hj}(s,t)$ et $\hat{p}_{mr}(s,t)$ est

$$\begin{aligned} \widehat{Cov}(\hat{p}_{hj}(s,t), \hat{p}_{mr}(s,t)) = \sum_{l=1}^{k} \sum_{g \neq l} \int_s^t & \hat{p}_{hg}(s,u^-)\, \hat{p}_{mg}(s,u^-) \{\hat{p}_{lj}(u,t) - \hat{p}_{gj}(u,t)\} \\ & \times \{\hat{p}_{lr}(u,t) - \hat{p}_{gr}(u,t)\} J_g(u) (Y_{g+}(u) - 1) (Y_{g+}(u))^{-3} dN_{gl+}(u). \end{aligned}$$

L'estimateur correspondant de la variance de $\hat{p}_{hj}(s,t)$ est

$$\hat{\mathbb{V}}(\hat{p}_{hj}(s,t)) = \sum_{l=1}^{k} \sum_{g \neq l} \int_s^t (\hat{p}_{hg}(s,u^-))^2 \{\hat{p}_{lj}(u,t) - \hat{p}_{gj}(u,t)\}^2 J_g(u) \frac{(Y_{g+}(u) - 1)}{(Y_{g+}(u))^3} dN_{gl+}(u).$$

Aalen et Johansen (1978) utilisent la décomposition en martingales (IV.5) et le théorème de limite centrale pour martingales (*cf.* page 175) pour obtenir les propriétés asymptotiques de $\hat{\mathbf{P}}$ quand $n \to \infty$.

Proposition 9 Sous le même type de condition que pour l'estimateur de Nelson-Aalen (pages 317-319 de Andersen et al. [1993]), $\hat{p}_{hj}(s,t)$ est un estimateur

- biaisé, tel que
$$\mathbb{E}\left(\hat{\mathbf{P}}(s,t)\right) \geqslant \mathbf{P}(s,t),$$
- uniformément consistant,
- et asymptotiquement normal, tel que
$$\sqrt{n}\left(\hat{\mathbf{P}}(s,.) - \mathbf{P}(s,.)\right) \xrightarrow{\mathcal{L}} \int_s^{\cdot} \mathbf{P}(s,u)\, d\mathbf{U}(u)\mathbf{P}(u,.),$$
avec $\mathbf{U} = \{U_{hj}\}_{hj}$ où U_{hj} martingales gaussiennes indépendantes telle que
$$\begin{cases} U_{hj}(0) = 0 \\ \mathbb{V}(U_{hj}(t)) = \int_0^t \frac{\alpha_{hj}(u)}{\sum_{j=1}^k p_{jh}(0,u)} du. \end{cases}$$

Les démonstrations de ces propositions sont données pages 197 ; 289-290 et 317-321 de Andersen et al. [1993].

3.4 Test des intensités de transition

Cette méthode d'estimation ne prend pas en compte l'effet des covariables dans la modélisation. L'effet de certaines covariables peut cependant être testé en comparant les intensités de transition obtenues dans chacune des strates. Considérons par exemple, le cas de deux bras de traitements. Il est intéressant de comparer les intensités de transition dans chacune des strates afin de comparer les deux traitements.

Proposition 10 Soit un processus de comptage $\mathbf{N} = (N_1, N_2)$ avec un processus d'intensité $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ tel que $\lambda_h(t) = \alpha_h(t)Y_h(t)$. Soit l'hypothèse nulle,

$$H_0 : \alpha_1 = \alpha_2, \tag{IV.12}$$

Alors la statistisque de test $X^2(t)$ suit asymptotiquement une loi du χ^2 à un degré de liberté sous l'hypothèse (IV.12), où

$$X^2(t) = \frac{H^2(t)}{\sigma(t)},$$

avec,

$$\begin{aligned} H(t) &= \int_0^t L(u) d\hat{A}_1(u) - \int_0^t L(u) d\hat{A}_2(u), \\ \sigma(t) &= \int_0^t L^2(u) \left\{Y_1(u)Y_2(u)\right\}^{-1} d(N_1 + N_2)(u), \\ L(t) &= K(t)Y_1(t)Y_2(t) \left\{Y_1(u) + Y_2(u)\right\}^{-1}. \end{aligned}$$

Le test du **Log-rank** est obtenu avec

$$K(t) = \mathbb{1}_{\{Y_1(t)+Y_2(t)>0\}}. \tag{IV.13}$$

Le test de **Gehan-Wilcoxon** est obtenu avec

$$K(t) = Y_1(t) + Y_2(t). \tag{IV.14}$$

La proposition précédente est un cas particulier ($k = 2$) d'une classe de statistiques de test pour l'hypothèse suivante,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k,$$

(*cf.* page 345 Andersen et al. [1993]).

3.5 Cas particulier : données de survie

L'analyse des données de survie est l'étude de la survenue au cours du temps d'un événement précis pour un groupe d'individus.

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

Le modèle de survie peut être considéré comme un modèle de Markov non-homogène particulier comportant deux états avec une seule transition possible (Figure IV.1). Le processus est Markovien dans le sens où le passé du processus se résume à l'état présent.

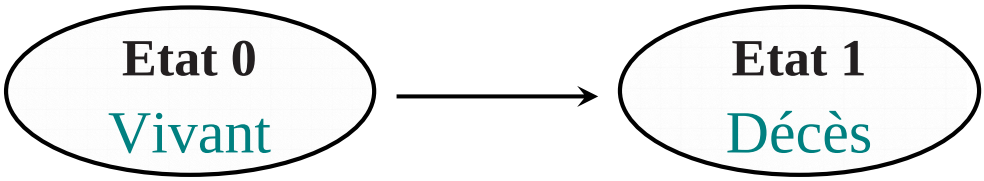


FIG. IV.1 – Modèle de survie à deux états : vivant et décès.

Le processus ponctuel de comptage associé est un processus dont les marques sont : état 0 (vivant), état 1 (décès). La matrice des probabilités de transition associée est définie par

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

et la matrice des intensités cumulées par

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

L'application de l'équation différentielle « forward » de Kolmogorov (IV.1) aux matrices $\mathbf{A}$ et $\mathbf{P}$ donnent l'équation suivante

$$\frac{\partial P_{00}(s,t)}{\partial t} = P_{00}\left(s,t^{-}\right) dA_{00}(t).$$

En intégrant par rapport à t, nous obtenons alors la probabilité de rester dans l'état 0 entre les instants s et t sous la forme d'une équation de type Volterra

$$\begin{aligned} P_{00}(s,t) &= \int_s^t \frac{\partial P_{00}(s,t)}{\partial t} dt, \\ &= \int_s^t P_{00}\left(s, v^-\right) dA_{00}(v). \end{aligned}$$

Or, comme $\mathbf{A}$ est une matrice d'intensité de transition d'un processus Markovien, $dA_{00}(v) = -dA_{01}(v)$. La probabilité $P_{00}(s,t)$ est de la forme

$$Z(s,t) = W(s,t) + \int_s^t Z\left(s, v^-\right) dX(v),$$

avec

$$\begin{aligned} Z(s,t) &= P_{00}(s,t), \\ W(s,t) &= 0, \\ X(v) &= -A_{01}(v). \end{aligned}$$

Par le théorème 3 page 163, il existe une unique solution de la forme

$$P_{00}(s,t) = \underset{u\in]s,t]}{\mathcal{P}} \{1 - dA_{01}(u)\}. \tag{IV.15}$$

$P_{00}(0,t)$ 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. Ainsi, on peut déduire de l'équation (IV.15) et de l'estimateur de Nelson-Aalen, l'estimateur bien connu Kaplan-Meier.

Proposition 11 Soient, pour tout $i = 1, ..., n$,
- T_i la date de survenue de l'événement,
- C_i la date de censure,
- $\tilde{T}_i = \min(T_i, C_i)$,
- $Y_i(t) = \mathbb{1}_{\left\{\tilde{T}_i \geqslant t\right\}}$, vaut 1 si l'individu est à risque l'instant t, 0 sinon.
- $N_i(t) = \mathbb{1}_{\{T_i \leqslant t\}}$, vaut 1 si l'individu est décédé à l'instant t, 0 sinon.

Enfin, définissons
- $Y_+(t) = \sum_{i=1}^n Y_i(t)$, le nombre total de survenues de l'événement à l'instant t,
- $N_+(t) = \sum_{i=1}^n N_i(t)$, le nombre total d'individus à risque juste avant l'instant t,
- $J(t) = \mathbb{1}_{\{Y_+(t)>0\}}$.

Alors l'estimateur de **Kaplan-Meier** de la fonction de survie est défini par (Kaplan et Meier [1958])

$$\hat{S}(t) = \underset{u\leqslant t}{\mathcal{P}} (1 - dA(u)),$$

où A est l'estimateur de Nelson-Aalen,

$$dA(u) = \frac{J(u)}{Y_+(u)} dN_+(u).$$

En pratique, le nombre d'évements est fini. Soient $s < T_1 < ... < T_k < t$, les temps de transition entre deux états. L'estimateur de Kaplan-Meier devient

$$\hat{S}(t) = \prod_{\{l;T_l \leqslant t\}} \left(1 - \frac{1}{Y_+(T_l)}\right).$$

Remarque 20 *La construction de l'estimateur Kaplan et Meier [1958], repose sur l'idée intuitive qu'être encore en vie après l'instant t, c'est être en vie juste avant t et ne pas mourir en t. Cette idée traduite en termes probabilistes mène à l'estimateur*

$$\begin{aligned} S(t) &= \mathbb{P}(T \geqslant t) \\ &= \mathbb{P}(T \geqslant t \mid T \geqslant t-1)\mathbb{P}(T \geqslant t-1) \\ &= \cdots \\ &= \mathbb{P}(T \geqslant t \mid T \geqslant t-1) \ \ldots \ \mathbb{P}(T \geqslant 1 \mid T \geqslant 0)\mathbb{P}(T \geqslant 0). \end{aligned}$$

4 Estimation semi-paramétrique

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

4.1 Définitions et notations

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

- $N_{hji}(t)$ qui compte le nombre de transitions de l'état h vers l'état j dans $[0, t]$ pour l'individu i;
- $Y_{hi}(t) = \mathbb{1}_{\{X_i(t-)=h\}}$, 1 si l'individu i est à risque dans l'état h juste avant le temps t, 0 sinon ;
- $\mathbf{Z}_i = (Z_{1i}, \ldots, Z_{pi})$, le vecteur de covariables de dimension p.
- $N_{hj+}(t) = \sum_{i=1}^{n} N_{hji}(t)$ et $Y_{h+}(t) = \sum_{i=1}^{n} Y_{hi}(t)$, les processus agrégés.
- $\boldsymbol{\beta}_{hj} = (\beta_{hj,1}, \ldots, \beta_{hj,p})$, le vecteur de dimension p des coefficients de régression ;
- $\mathbf{N} = \{N_{hji}, h, j \in S, h \neq j; i = 1, \ldots, n\}$, le processus de comptage multivarié associé aux n individus ;
- $\boldsymbol{\lambda} = \{\lambda_{hji}, h, j \in S, h \neq j; i = 1, \ldots, n\}$, le processus d'intensité par rapport à la filtration $\mathcal{F}_t$.

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

$$\lambda_{hji}(t) = Y_{hi}(t)\, \alpha_{hji}(t; \mathbf{Z}_i),$$

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

$$\alpha_{hji}(t;\mathbf{Z}_i) = \alpha_{hj0}(t)\exp(\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i), \qquad h,j \in S, h \neq j, i = 1,\ldots,n. \tag{IV.16}$$

où $\alpha_{hj0}(t)$ est l'intensité de transition de base associée à la transition de l'état h vers l'état j. Plus précisement, $\alpha_{hj0}(.)$ est la fonction de risque des sujets pour lesquels toutes les covariables explicatives sont nulles. Ce modèle est dit **semi-paramétrique** du fait de la présence, dans la définition des intensités de transition, d'une partie paramétrique (la partie de régression $\exp(\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i)$) et d'une partie non-paramétrique (le risque de base $\alpha_{hj0}(.)$).

De plus, le modèle (IV.16) est dit à **risques proportionnels** car, par définition, quels que soient deux individus (1 et 2), le rapport des intensités de transition ne varie pas au cours du temps

$$\frac{\alpha_{hj1}(t)}{\alpha_{hj2}(t)} = \exp(\boldsymbol{\beta}_{hj}^T(\mathbf{Z}_1 - \mathbf{Z}_2)). \tag{IV.17}$$

Les intensités de transition sont donc proportionnelles. Ceci est une conséquence du modèle, mais c'est aussi une hypothèse qu'il faudra vérifier.

Notons que tous les modèles à structure multiplicative, c'est-à-dire les modèles où les intensités de transition sont séparables en deux termes dont l'un dépend uniquement du temps et l'autre non (par exemple (IV.16)), ont cette propriété ; ce sont des modèles à risques proportionnels. Dans ces modèles, le rapport des intensités de transition représente un risque relatif à l'instant t des sujets de caractéristique $\mathbf{Z}_1$ par rapport aux sujets de caractéristique $\mathbf{Z}_2$.

Notons également que le modèle (IV.16) est **log-linéaire**, c'est à dire que le logarithme de l'intensité de transition est une fonction linéaire de $\mathbf{Z}_i$,

$$\log\alpha_{hji}(t) - \log\alpha_{hj0}(t) = \boldsymbol{\beta}_{hj}^T\mathbf{Z}_i.$$

Cette hypothèse est contraignante lorsqu'on utilise une variable explicative continue. En effet, l'hypothèse de log-linéarité suppose que le risque relatif est constant pour une augmentation d'une unité quelle que soit la valeur de la covariable explicative. C'est une hypothèse qu'il convient de vérifier ou tout au moins d'avoir à l'esprit quand on utilise ce modèle de régression. Par exemple, si l'on considère l'âge comme variable explicative continue et que l'on étudie une maladie qui touche essentiellement les personnes âgées, le modèle supposera un même risque relatif pour une augmentation de 1 an, que ce soit pour un âge de 30 ans ou pour un âge de 70 ans.

Cette relation log-linéaire est souvent utilisée dans la littérature (Cox [1972]) car elle permet d'avoir des intensités définies positives quelle que soit la valeur des coefficients de régression. De plus, les résultats obtenus sont bien connus des cliniciens et sont facilement interprétables. Cependant, d'autres choix de fonctions sont possibles (m tel que $m(Z) > 0$ et $m(0) = 1$).

4.2 Estimation des intensités de base

Supposons pour l'estimation que les fonctions $\alpha_{hj0}(.)$ sont positives et que $A_{hj0}(t) = \int_0^t \alpha_{hj0}(u)du < \infty$, $\forall h \neq j$.

Les résultats qui suivent reposent sur la partie théorie mathématique des processus de comptage présentée en annexe page 161. Rappelons que d'après le théorème 6 page 166 et par la définition du compensateur (IV.4), la **vraisemblance complète** associée à un processus non censuré $\mathbf{N}^*$ est

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

De plus, d'après la proposition 3 page 79, la vraisemblance associée à un processus censuré (censure à droite indépendante) a une forme identique à la vraisemblance complète. Cette vraisemblance d'un processus censuré $\mathbf{N}$ est appelée la **vraisemblance partielle.** Dans le cadre d'un modèle semi-paramétrique multiplicatif, la vraisemblance partielle s'écrit

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) = \prod_{t\leqslant\tau}\prod_{i=1}^{n}\prod_{h\neq j}\left\{\left(dA_{hj0}(t)Y_{hi}(t)\exp(\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i)\right)^{\Delta N_{hji}(t)}\right\} \\ \times \exp\left[-\sum_{h\neq j}\int_0^\tau S_{hj}^{(0)}(\boldsymbol{\beta},u)dA_{hj0}(u)\right] \qquad \text{(IV.18)}\end{aligned}$$

avec

$$S_{hj}^{(0)}(\boldsymbol{\beta},t) = \sum_{i=1}^{n}\exp(\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i)Y_{hi}(t).$$

En considérant $\boldsymbol{\beta}$ fixé, la maximisation de la vraisemblance (IV.18) par rapport à $\Delta A_{hj0}(.)$ conduit à

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

Proposition 12 Pour $\boldsymbol{\beta}$ fixé, $A_{hj0}(t) = \int_0^t \alpha_{hj0}(u)du$ est estimé par l'**estimateur de Breslow** (Breslow [1974]),

$$\hat{A}_{hj0}(t) = \int_0^t \frac{J_h(u)}{\sum_{i=1}^n \exp(\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i)Y_{hi}(t)}dN_{hj+}(u), \qquad \text{(IV.19)}$$

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

4.3 Estimation des coefficients de régression

En remplaçant, dans (IV.18), $A_{hj0}(t)$ par son estimation obtenue en (IV.19), la vraisemblance partielle devient,

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

avec

$$\mathcal{L}_{Cox}(\boldsymbol{\beta}) = \prod_t \prod_{i=1} \prod_{h\neq j} \left[\frac{Y_{hi}(t) \exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i)}{S_{hj}^{(0)}(\boldsymbol{\beta}, t)} \right]^{\Delta N_{hji}(t)}. \tag{IV.20}$$

La vraisemblance $\mathcal{L}(\boldsymbol{\beta})$ se décompose en une partie dépendante de $\boldsymbol{\beta}$ ($\mathcal{L}_{Cox}$) et une partie indépendante de $\boldsymbol{\beta}$. Par définition, $\mathcal{L}_{Cox}(\boldsymbol{\beta})$ est **la vraisemblance partielle de Cox.** Cette vraisemblance est introduite par Cox (1972) mais par une approche complètement différente. La vraisemblance partielle de Cox n'est pas une vraisemblance dans le sens statistique du terme, mais il est établi qu'elle peut être utilisée comme telle pour estimer les coefficients de régression.

Considérons la fonction de log-vraisemblance partielle de Cox,

$$\log \mathcal{L}_{Cox}(\boldsymbol{\beta}) = \sum_i \sum_{h\neq j} \int_0^\tau \left[\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i - \log S_{hj}^{(0)}(\boldsymbol{\beta}, t) \right] dN_{hji}(t).$$

Les vecteurs scores (dérivées de la Log-Vraisemblance par rapport à $\boldsymbol{\beta}_{hj}$) sont donnés par

$$\begin{aligned}
\mathbf{U}_{hj}(\boldsymbol{\beta}) &= \frac{\partial \log \mathcal{L}_{Cox}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{hj}} \\
&= \sum_i \int_0^\tau \left[\mathbf{Z}_i - \frac{S_{hj}^{(1)}(\boldsymbol{\beta}, t)}{S_{hj}^{(0)}(\boldsymbol{\beta}, t)} \right] dN_{hji}(t),
\end{aligned}$$

avec

$$S_{hj}^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial S_{hj}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}_{hj}} = \sum_{i=1}^n Y_{hi}(t) \mathbf{Z}_i \exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i).$$

Proposition 13 L'estimateur $\hat{\boldsymbol{\beta}}$ du maximum de la vraisemblance partielle de Cox vérifie

$$\mathbf{U}_{hj}(\hat{\boldsymbol{\beta}}) = 0.$$

De plus,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathcal{L}} N(0, n\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})),$$

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

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

avec

$$S_{hj}^{(2)}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_{hi}(t) \left(\mathbf{Z}_i\right)^2 \exp(\boldsymbol{\beta}_{hj}\mathbf{Z}_i).$$

L'inverse de la matrice d'information de Fisher étant égale à la matrice de variance-covariance, elle fournit une estimation de la variance de $\hat{\boldsymbol{\beta}}_{hj}$. En pratique, les estimations des coefficients de régression sont obtenues par maximisation de la log-vraisemblance à l'aide d'algorithmes itératifs, comme par exemple l'algorithme de quasi-Newton (Nocedal et Wright [1999]).

Remarque 21 *Les vraisemblances successivement utilisées sont résumées par le schéma suivant :*

Vraisemblance complète
(relative à un processus non censuré)
$L^*(\beta)$

↓

Vraisemblance « partielle »
(relative à un processus censuré)
$L(\beta)$

↓

Vraisemblance partielle de Cox
$L_{Cox}(\beta)$

TAB. IV.1 – Les vraisemblances successives.

4.4 Estimation des probabilités de transition

A partir des estimateurs précédents (du risque de base et des coefficients), la matrice des probabilités de transition s'obtient en suivant la démarche permettant d'obtenir l'estimateur de Aalen-Johansen. Considérons $\mathbf{Z}_0$, la valeur des covariables pour un individu.

Proposition 14 Un estimateur de la matrice des probabilités de transition est donné par le produit intégral

$$\hat{\mathbf{P}}\left(s,t \mid \mathbf{Z}_0\right) = \underset{u\in]s,t]}{\mathcal{P}} \left(\mathbf{Id} + d\hat{\mathbf{A}}\left(u \mid \mathbf{Z}_0\right)\right), \qquad s \leqslant t \leqslant \tau,$$

avec $\hat{\mathbf{A}} = \left\{\hat{A}_{hj}\right\}_{hj}$,

$$\begin{aligned} d\hat{A}_{hj}(t \mid \mathbf{Z}_0) &= d\hat{A}_{hj0}(t \mid \mathbf{Z}_0) \times \exp(\hat{\boldsymbol{\beta}}_{hj}^T \mathbf{Z}_0) \\ &= \frac{J_h(t) \times \exp(\hat{\boldsymbol{\beta}}_{hj}^T \mathbf{Z}_0)}{\sum_{i=1}^n \exp(\hat{\boldsymbol{\beta}}_{hj}^T \mathbf{Z}_0) Y_{hi}(t)} \times \Delta N_{hj+}(t), \qquad h \neq j, \end{aligned}$$

et

$$d\hat{A}_{hh}(t \mid \mathbf{Z}_0) = -\sum_{j\neq h} d\hat{A}_{hj}(t \mid \mathbf{Z}_0), \qquad h = 1, ..., s.$$

Proposition 15

Les propriétés asymptotiques de $\hat{\mathbf{P}}\left(s,t \mid \mathbf{Z}_0\right)$ s'obtiennent à partir de celles de $\hat{\boldsymbol{\beta}}$ et $\hat{A}_{hj0}(t)$ et de l'utilisation de la delta-méthode (Andersen et al. [1991]).

Proposition 16 Un estimateur de la covariance de $\hat{p}_{hj}\left(s,t \mid \mathbf{Z}_0\right)$ et $\hat{p}_{mr}\left(s,t \mid \mathbf{Z}_0\right)$ est

$$\widehat{Cov}(\hat{p}_{hj}\left(s,t \mid \mathbf{Z}_0\right), \hat{p}_{mr}\left(s,t \mid \mathbf{Z}_0\right)) = \widehat{Cov}_1 + \widehat{Cov}_2,$$

avec

$$\begin{aligned} \widehat{Cov}_1 = \sum_{g\neq l} \int_s^t & \hat{p}_{hg}\left(s,u \mid \mathbf{Z}_0\right) \hat{p}_{mg}\left(s,u \mid \mathbf{Z}_0\right) \left\{\hat{p}_{lj}\left(u,t \mid \mathbf{Z}_0\right) - \hat{p}_{gj}\left(u,t \mid \mathbf{Z}_0\right)\right\} \\ & \times \left\{\hat{p}_{lr}\left(u,t \mid \mathbf{Z}_0\right) - \hat{p}_{gr}\left(u,t \mid \mathbf{Z}_0\right)\right\} J_g(u) \left\{\exp(\hat{\boldsymbol{\beta}}_{gl}^T \mathbf{Z}_0)\right\}^2 \left\{S_{gl}^{(0)}(\hat{\boldsymbol{\beta}}, u)\right\}^{-2} dN_{gl+}(u), \end{aligned}$$

et

$$\begin{aligned} \widehat{Cov}_2 = & \int_s^t \sum_{g,l} \hat{p}_{hg}\left(s,u \mid \mathbf{Z}_0\right) dW_{gl}(u) \hat{p}_{lj}\left(u,t \mid \mathbf{Z}_0\right) \times \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}) \\ & \times \int_s^t \sum_{g,l} \hat{p}_{mg}\left(s,u \mid \mathbf{Z}_0\right) dW_{gl}(u) \hat{p}_{lr}\left(u,t \mid \mathbf{Z}_0\right), \end{aligned}$$

où $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$ est la matrice de variance-covariance de $\hat{\boldsymbol{\beta}}$ et pour tout $g \neq l$,

$$W_{gl}(t) = \exp(\hat{\boldsymbol{\beta}}_{gl}^T \mathbf{Z}_0) \int_0^t \left(\mathbf{Z}_0 - \frac{S_{hj}^{(1)}(\hat{\boldsymbol{\beta}}, t)}{S_{hj}^{(0)}(\hat{\boldsymbol{\beta}}, t)} \right) J_g(u) \left\{ S_{gl}^{(0)}(\hat{\boldsymbol{\beta}}, u) \right\}^{-1} dN_{gl+}(u),$$

et $W_{gg} = -\sum_{g \neq l} W_{gl}$.

4.5 Covariables dépendantes du temps

Dans ce qui précède, la méthode d'estimation est présentée pour des covariables indépendantes du temps afin de simplifier les écritures. Cependant, l'utilisation de la vraisemblance partielle et tous les résultats précédents restent vrais pour des covariables dépendantes du temps. Quand les covariables sont dépendantes du temps, le rapport des risques pour deux individus n'est plus indépendant du temps mais l'impact relatif de deux valeurs données d'une covariable (par exemple, un taux de cholestérol de 200 contre 250) est toujours résumé par un seul coefficient β_{hj}. Ce n'est pas « l'effet » de la variable qui varie avec le temps mais la variable elle-même.

L'utilisation de covariables dépendantes du temps est pertinente dans certaines situations. Par exemple, pour prendre en compte des informations qui sont recueillies au cours du suivi : type de traitement ou mesures répétées d'un dosage biologique. On pourrait aussi étudier l'effet du temps de séjour avant le prochain événement afin de vérifier l'hypothèse de Markov qui suppose que cette durée n'influence pas le processus d'évolution.

Notons cependant quelques différences avec l'utilisation de covariables dépendantes du temps.

- Afin de calculer la contribution à la vraisemblance partielle de Cox à chaque temps d'événement, il est nécessaire de connaître la valeur de ces covariables pour tous ces temps. Le plus souvent, ces valeurs ne sont pas disponibles. Ce problème de données manquantes peut se résoudre en supposant que la valeur de la covariable à l'instant t est la valeur enregistrée à la *dernière* consultation (avant t^-). Ce choix semble naturel, mais on pourrait aussi utiliser la valeur enregistrée lors de la consultation la plus proche de t.
- L'interprétation des résultats est moins aisée même si celle des coefficients de régression demeure identique. En effet, l'interprétation des probabilités de transition pour des covariables dépendantes du temps devient délicate car il y a un estimateur des probabilités de transition pour chaque « histoire » de covariable.
- Les temps de calcul peuvent singulièrement augmenter.

4.6 Tests des coefficients

Trois statistiques de test de l'hypothèse nulle « $H_0 : \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ » peuvent être déduites des résultats concernant la convergence asymptotique de $\hat{\boldsymbol{\beta}}$.

– Test du rapport de vraisemblance :
Ce test mesure les différences des valeurs prises par le logarithme de la vraisemblance en $\hat{\boldsymbol{\beta}}$ et $\boldsymbol{\beta}_0$,

$$2\left[\log \mathcal{L}_{Cox}(\hat{\boldsymbol{\beta}}) - \log \mathcal{L}_{Cox}(\boldsymbol{\beta}_0)\right] \rightsquigarrow \chi^2(p).$$

– Test de Wald :
Il mesure l'écart entre $\hat{\boldsymbol{\beta}}$ et $\boldsymbol{\beta}_0$, qui est nul en moyenne sous H_0 car $\hat{\boldsymbol{\beta}}$ est asymptotiquement sans biais,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightsquigarrow \chi^2(p).$$

– Test du score :
Ce test mesure la pente de la tangente en $\boldsymbol{\beta}_0$,

$$\mathbf{U}_{hj}(\boldsymbol{\beta}_0)^T \mathcal{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{U}_{hj}(\boldsymbol{\beta}_0) \rightsquigarrow \chi^2(p).$$

4.7 Test de l'hypothèse de proportionnalité des risques

La méthode d'estimation semi-paramétrique suppose que les risques sont à hasards proportionnels (IV.17). Les méthodes pour vérifier ce type d'hypothèses sont bien connues dans les modèles de survie et certaines peuvent être adaptées au modèle multi-états. Une alternative à l'hypothèse de proportionnalité des risques est d'utiliser un modèle avec des coefficients de régression dépendants du temps. On peut écrire $\beta_{hj}(t)$ comme une régression sur $g(t)$,

$$\beta_{hj}(t) = \beta_{hj} + \gamma_{hj} \times g(t).$$

où $g(t)$ est une fonction du temps. L'hypothèse nulle de proportionnalité des risques correspond alors à $\gamma_{hj} = 0$. Ainsi, pour tester l'hypothèse de proportionnalité pour la covariable Z_1, le modèle suivant peut être considéré

$$\alpha_{hj}(t; Z_1) = \alpha_{hj0}(t) \exp(\beta_{hj} Z_1 + \gamma_{hj} \times g(t) \times Z_1), \qquad h, j = 1, ..., s;\ h \neq j.$$

Si γ_{hj} est statistiquement différent de zéro, alors un coefficient de régression dépendant du temps sera mieux adapté et l'hypothèse de proportionnalité ne sera pas respectée. Au contraire, si $\gamma_{hj} = 0$, la transition sera bien à risques proportionnels pour la variable Z_1.

Différents choix sont possibles pour la fonction $g(.)$, par exemple $g(t) = \log(t)$ ou $g(t) = t$. Notons que cette méthode permet de tester uniquement un effet linéaire du coefficient avec une fonction du temps. L'hypothèse de proportionnalité des risques peut également être testée par des méthodes graphiques comme les résidus de Schoenfeld (Therneau et Grambsch [2000]).

4.8 Cas particulier : données de survie

Le modèle de survie est un modèle à deux états où une seule transition est possible, comme décrit page 87. La méthodologie présentée précédemment comprend ainsi le cas particulier des données de survie.

En considérant l'espace d'états $\{0,1\}$ (« vivant » : Etat 0 et « décès » : Etat 1), la vraisemblance partielle de Cox (IV.20) s'écrit

$$\mathcal{L}_{Cox}(\boldsymbol{\beta}) = \prod_t \prod_{i=1} \left[\frac{Y_{0i}(t) \exp(\boldsymbol{\beta}_{01}^T \mathbf{Z}_i)}{S_{01}^{(0)}(\boldsymbol{\beta}, t)} \right]^{\Delta N_{01i}(t)},$$

où $N_{01i}(t)$ vaut 1 si l'individu i passe de l'état 0 à l'état 1 au temps t et 0 sinon ; $\boldsymbol{\beta}_{01}$ représente le coefficient de régression associé à la transition vers le « décès ». Cette expression est bien la vraisemblance obtenue par Cox pour des données de survie. De même, l'estimateur (IV.19) des intensités cumulées de base généralise l'estimateur de Breslow introduit en 1974 pour des données de survie,

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

où $\hat{A}_{010}(t)$ représente l'intensité cumulée de base associée au « décès ».

Le modèle de Cox à risques proportionnels qui est couramment utilisé représente ainsi un cas particulier de la méthodologie présentée précédemment.

Remarque 22 *Le concept de vraisemblance partielle est introduit par Cox en 1972. La vraisemblance est scindée en deux parties de manière à conserver uniquement la partie de la vraisemblance qui concerne les coefficients $\boldsymbol{\beta}$ que l'on cherche à estimer. Soient*

$$T_{(1)} < \cdots < T_{(m)},$$

les différents temps de décès observés chez les sujets $(1), \ldots, (m)$. La probabilité conditionnelle que le sujet (i) décède en $T_{(i)}$ sachant qu'il est à risque au temps $T_{(i)}$ et qu'il n'y a qu'un seul décès en $T_{(i)}$ parmi les individus à risque en $T_{(i)}$ est égale à :

$$\begin{aligned} V_i &= \frac{\alpha_0(T_{(i)}) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{(i)})}{\sum_{l \in R_{(i)}} \alpha_0(T_{(i)}) \exp(\boldsymbol{\beta}^T \mathbf{Z}_{(l)})} \\ &= \frac{\exp(\boldsymbol{\beta}^T \mathbf{Z}_{(i)})}{\sum_{l \in R_{(i)}} \exp(\boldsymbol{\beta}^T \mathbf{Z}_{(l)})}. \end{aligned}$$

où $R_{(i)}$ représente l'ensemble des personnes à risque en $T_{(i)}$. La vraisemblance partielle de Cox correspond au produit de ces probabilités conditionnelles calculées à chaque temps de décès,

$$\mathcal{L}_{Cox} = \prod_{i=1}^m V_i.$$

Cette quantité ne dépend pas de la fonction de risque de base α_0. Ainsi, seul l'ordre et non la valeur des temps d'évement est important.

Remarque 23 *La méthodologie présentée dans ce chapitre peut être adaptée au modèle semi-Markovien (Huber-Carol et Pons [2004], Chang et al. [2000], Andersen et al. [1993], Gill [1980]). La vraisemblance dans le cas semi-Markovien a exactement la même forme que dans le cas Markovien. L'estimateur des intensités cumulées est la même fonction des données dans chacun des cas et, de plus, les distributions pour les grands échantillons ont la même forme.*

5 Application à l'asthme

Les modèles de Markov non-homogènes supposent que la durée du suivi influence l'évolution de la maladie (les intensités de transition). Dans le cas de l'asthme, le modèle homogène semble trop restrictif (*cf.* chapitre II). En effet, les patients sont traités et éduqués et les intensités de transition en début de suivi peuvent ainsi être différentes de celles après trois ans de suivi.

La méthodologie précédente est appliquée à la base de données présentée au chapitre I par l'intermédiaire d'un modèle à trois états de contrôle (Figure IV.2). Les données observées sont censurées à droite de manière aléatoire (car l'inclusion dans l'étude est aléatoire) à cause de l'arrêt de l'étude (à la date point). Dans notre cas, la date de point correspond à la dernière consultation enregistrée dans la base de données. Ainsi, il est supposé, qu'après sa dernière consultation, le patient reste à risque jusqu'à la date de point.

Notons également que les temps d'événements sont supposés distincts. Cette hypothèse découle de la définition d'un processus de comptage multivarié (*cf.* page 74). Ainsi, les quelques temps ex-æquo présents dans la base de données sont légèrement modifiés en ajoutant ou en retranchant un ou deux jours.

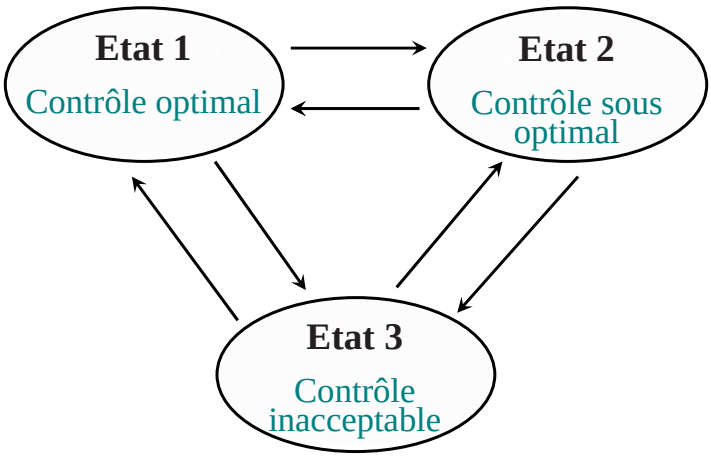


FIG. IV.2 – Modèle à trois états de contrôle pour l'asthme.

Les résultats présentés concernent l'indice de masse corporelle (IMC). Dans la méthode d'estimation non-paramétrique, il est nécessaire de stratifier la base de données pour observer

l'effet de la covariable. Ainsi, dans tout ce qui suit, l'IMC sera considéré comme une variable indépendante du temps et sa valeur sera fixée à la première consultation. La variable IMC sera

- codée 0 si l'IMC à la première consultation est inférieur à 25,
- codée 1 si l'IMC à la première consultation est supérieur ou égal à 25.

Les objectifs consistent à

- estimer les probabilités de transition du modèle par les méthodes non-paramétrique et semi-paramétrique,
- observer et tester l'effet du surpoids,
- comparer les estimations des probabilités de transition obtenues avec un modèle homogène et un modèle non-homogène.

5.1 Estimation non-paramétrique

Dans un premier temps, l'estimateur de **Aalen-Johansen** (Equation (IV.11)) est utilisé pour estimer les probabilités de transition du modèle. Cet estimateur ne prend pas en compte de covariable dans sa construction. Ainsi, il est nécessaire d'appliquer les estimateurs dans des sous-groupes afin de pouvoir ensuite comparer les estimations.

La base est stratifiée en deux groupes suivant l'IMC :

- 260 patients ont un IMC < 25 à la première consultation,
- 146 patients ont un IMC $\geqslant 25$ à la première consultation.

Les figures IV.3 **(a)** et **(b)** présentent les estimateurs de **Nelson-Aalen** (Equation (IV.10)) des intensités cumulées $1 \to 3$ et $3 \to 1$ dans les deux groupes liés à l'IMC. L'intensité cumulée de l'état optimal vers l'état inacceptable ($1 \to 3$) est plus importante dans la strate des patients en surpoids. Inversement, l'intensité cumulée de l'état inacceptable vers l'état optimal ($3 \to 1$) est plus faible dans cette strate.

Les estimateurs de Aalen-Johansen des probabilités de transition $1 \to 3$ et $3 \to 1$ dans chacun des groupes sont représentés sur les figures IV.4 **(a)** et **(b).** Comme pour les intensités cumulées, les estimations des probabilités de transition montrent de manière graphique les effets négatifs du surpoids sur l'évolution de l'asthme.

Afin de confirmer statistiquement l'effet du surpoids, les tests du Log-rank et de Gehan-Wilcoxon (Equations (IV.13) et (IV.14)) peuvent être utilisés pour comparer les intensités de transition obtenues dans chaque groupe. Ces tests permettent de tester l'égalité des intensités de transition de l'état h vers l'état j obtenues dans les deux groupes :

$$H_0 : \alpha_{hj}^0 = \alpha_{hj}^1$$

où α_{hj}^0 est l'intensité associée à la strate IMC < 25 et α_{hj}^1 est l'intensité associée à la strate IMC $\geqslant 25$. Les résultats obtenus avec les deux statistiques de test sont présentés dans le tableau IV.2. Il apparaît que les différences graphiques précédemment observées

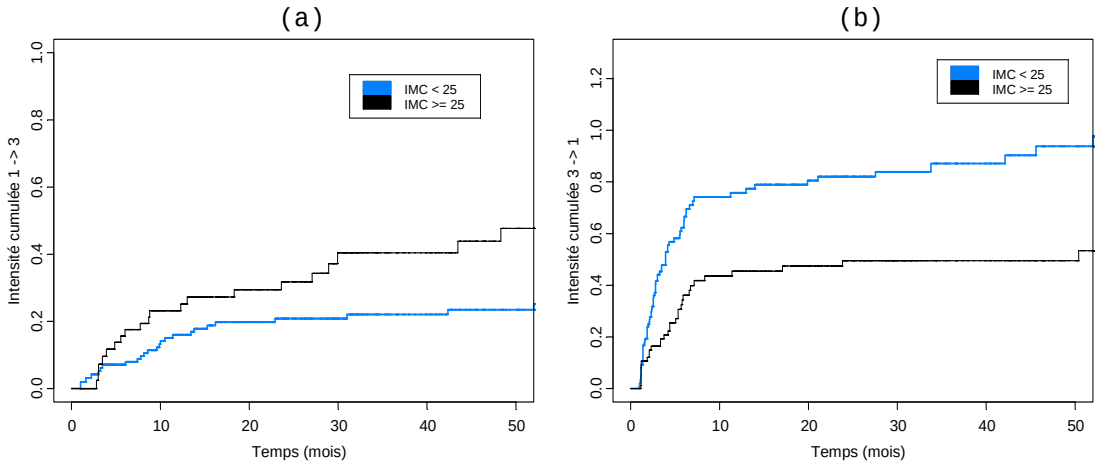


FIG. IV.3 – Estimation non-paramétrique (Nelson-Aalen) des intensités cumulées dans les strates IMC < 25 et IMC $\geqslant 25$: **(a)** transition de l'état optimal vers l'état inacceptable $(1 \to 3)$; **(b)** transition de l'état inacceptable vers l'état optimal $(3 \to 1)$.

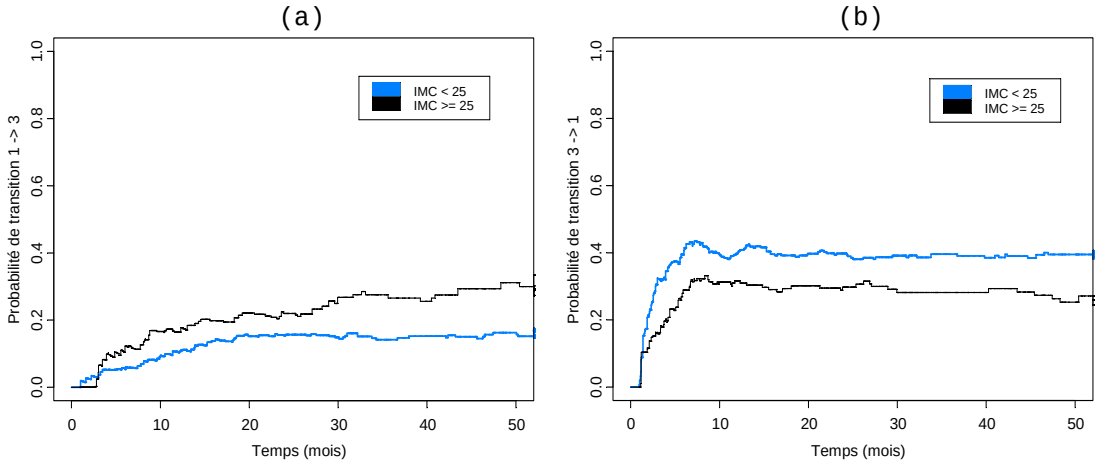


FIG. IV.4 – Estimation non-paramétrique (Aalen-Johansen) des probabilités de transition dans les strates IMC < 25 et IMC $\geqslant 25$: **(a)** transition de l'état optimal vers l'état inacceptable $(1 \to 3)$; **(b)** transition de l'état inacceptable vers l'état optimal $(3 \to 1)$.

sont confirmées par ces résultats. Les deux tests rejettent l'hypothèse d'égalité des intensités pour les transitions de l'état optimal vers l'état inacceptable ($1 \to 3$) et de l'état inacceptable vers l'état optimal ($3 \to 1$). Pour les autres intensités, les tests ne rejettent pas l'égalité des intensités de transition.

	Test	
Hypothèse H_0	$p-value$ Log-rank	$p-value$ Gehan-Wilcoxon
$\alpha_{12}^0 = \alpha_{12}^1$	0.29	0.19
$\alpha_{13}^0 = \alpha_{13}^1$	0.03	0.02
$\alpha_{21}^0 = \alpha_{21}^1$	0.88	0.98
$\alpha_{23}^0 = \alpha_{23}^1$	0.41	0.61
$\alpha_{31}^0 = \alpha_{31}^1$	< 0.01	< 0.01
$\alpha_{32}^0 = \alpha_{32}^1$	0.22	0.22

TAB. IV.2 – Test du Log-rank et de Gehan-Wilcoxon pour comparer les intensités de transition dans les strates IMC < 25 (codée 0) et IMC $\geqslant 25$ (codée 1).

Les résultats graphiques (Figures IV.3 et IV.4) et les tests (Log-rank et Gehan-Wilcoxon) permettent de mettre en évidence l'effet du surpoids sur les intensités de transition. L'intensité associée à la transition $1 \to 3$ est statistiquement plus grande pour les patients en surpoids à l'inclusion. Inversement, l'intensité associée à la transition $3 \to 1$ est plus faible pour ces mêmes patients. Les autres intensités ne montrent pas de différences significatives

5.2 Estimation semi-paramétrique

Un **modèle semi-paramétrique** à risques proportionnels (Equation (IV.16)) est ajusté pour prendre en compte l'effet des covariables (et pour quantifier cet effet). Le modèle à trois états de la figure IV.2 est utilisé.

Nous considérons deux modèles prenant en compte l'effet de l'IMC,

$$\begin{aligned} \text{Modèle } A: &\quad \alpha_{hj}(t) = \alpha_{hj0}(t)\exp(\beta_{hj} \times Z_1) \\ \text{Modèle } B: &\quad \alpha_{hj}(t) = \alpha_{hj0}(t)\exp(\beta_{hj} Z_1 + \gamma_{hj} Z_2(t)) \end{aligned}$$

où Z_1 est la variable associée à l'IMC à la première consultation,

$$Z_1 = \begin{cases} 0 & \text{si IMC} < 25 \\ 1 & \text{sinon,} \end{cases}$$

et $Z_2(t)$ est une covariable artificielle permettant de tester l'hypothèse de proportionnalité des risques (*cf.* page 96),

$$Z_2(t) = Z_1 \times \log(t/1000)$$

où t est le temps en jours.

Le modèle A permet de quantifier l'effet de l'IMC par l'intermédiaire de β_{hj}. Si β_{hj} est statistiquement différent de zéro alors le surpoids influence de manière significative la transition de l'état h vers l'état j.

Le modèle B comprend la variable Z_1 et une interaction de cette variable avec le temps. Ainsi, si le coefficient γ_{hj} est différent de zéro, alors l'effet de Z_1 n'est pas constant au cours du temps et l'hypothèse de proportionnalité n'est pas respectée pour la transition de h vers j.

Type de transition	Covariable	Modèle A $\hat{\beta}$	Modèle A $(e.c)^1$	Modèle A $(p)^2$	Modèle B $\hat{\beta}$	Modèle B $(e.c)^1$	Modèle B $(p)^2$
optimal vers	Z_1	-0.248	(0.236)	(0.28)	-0.143	(0.386)	(0.54)
sous-optimal $(1 \to 2)$	$Z_2(t)$				0.086	(0.240)	(0.55)
optimal vers	Z_1	0.655	(0.303)	(0.03)	0.863	(0.509)	(<0.01)
inacceptable $(1 \to 3)$	$Z_2(t)$				0.176	(0.325)	(0.36)
sous-optimal vers	Z_1	-0.030	(0.203)	(0.88)	-0.197	(0.442)	(0.33)
optimal $(2 \to 1)$	$Z_2(t)$				-0.089	(0.215)	(0.37)
sous-optimal vers	Z_1	0.204	(0.249)	(0.42)	0.426	(0.320)	(0.10)
inacceptable $(2 \to 3)$	$Z_2(t)$				0.281	(0.259)	(0.15)
inacceptable vers	Z_1	-0.595	(0.220)	(<0.01)	-0.479	(0.515)	(0.02)
optimal $(3 \to 1)$	$Z_2(t)$				0.047	(0.210)	(0.57)
inacceptable vers	Z_1	-0.234	(0.191)	(0.21)	-0.260	(0.296)	(0.16)
sous-optimal $(3 \to 2)$	$Z_2(t)$				-0.021	(0.152)	(0.83)

[1] estimations des écarts-types.
[2] p avec le test du LRT pour $H_0 : \beta_{ij} = 0$.

Tab. IV.3 – Estimation semi-paramétrique des coefficients de régression associés aux covariables : Z_1 = IMC (0 si IMC < 25, 1 si IMC $\geqslant 25$) et $Z_2(t) = Z_1 \times \log(t/1000)$ où t est le temps en jours.

Le tableau IV.3 présente l'estimation des coefficients de régression dans les modèles A et B. La fonction *optim*() du logiciel R qui utilise l'algorithme de quasi-Newton (Nocedal et Wright [1999]) a permis d'obtenir les estimateurs du maximum de vraisemblance. Dans les deux modèles, l'hypothèse de nullité des coefficients de régression est rejetée pour les transitions $1 \to 3$ et $3 \to 1$. Ainsi, dans le modèle A, les intensités de la transition $1 \to 3$ sont multipliées par 1.9 (exp(0.655)) pour les patients en surpoids ; les intensités de la transition $3 \to 1$ sont multipliées par 0.55 (exp(-0.595)) pour les patients en surpoids. Dans le modèle B, aucun coefficient associé à la variable artificielle Z_2 n'est statistiquement différent de zéro. Par conséquent, l'hypothèse de proportionnalité des risques, qui n'est jamais rejetée, ne semble pas trop contraignante. Notons également que le coefficient associé à l'IMC pour la transition $3 \to 1$ (modèle B) reste significatif après ajustement (avec l'interaction entre le temps et la variable).

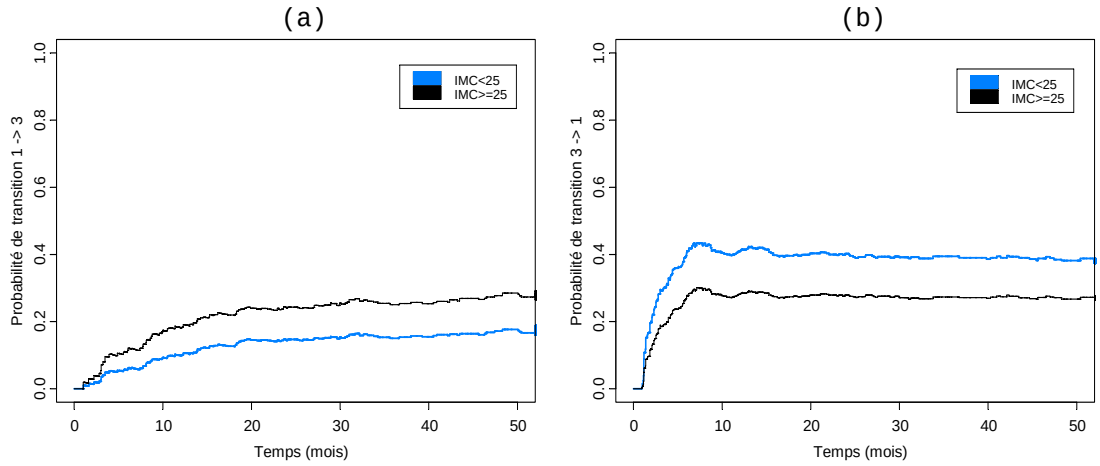


FIG. IV.5 – Estimation semi-paramétrique des probabilités de transition avec l'IMC en covariable : IMC < 25 et IMC $\geqslant 25$: **(a)** transition de l'état optimal vers l'état inacceptable $(1 \rightarrow 3)$; **(b)** transition de l'état inacceptable vers l'état optimal $(3 \rightarrow 1)$.

Les probabilités de transition $1 \rightarrow 3$ et $3 \rightarrow 1$ du modèle A sont représentées, pour chaque valeur de la covariable, sur les figures IV.5 **(a)** et **(b)**. Ces courbes permettent d'observer graphiquement l'effet négatif du surpoids sur les probabilités de transition. Notons que les estimations des probabilités de transition obtenues avec le modèle semi-paramétrique (Figure IV.5) sont proches de celles obtenues en stratifiant la base (Figure IV.4). Les deux modèles mettent en évidence les mêmes effets du surpoids sur l'évolution de la maladie. Le modèle semi-paramétrique est attractif car il permet de quantifier l'effet de la covariable.

5.3 Comparaison des modèles de Markov homogène et non-homogène

Dans cette partie, les résultats obtenus avec un modèle non-homogène sont comparés à ceux obtenus avec un modèle homogène.

Premièrement, un modèle homogène sans covariable (*cf.* chapitre II) et un modèle non-homogène sans covariable (Equation (IV.9)) sont ajustés. Les figures IV.6 **(a)** et **(b)** représentent les estimations des probabilités de transition $1 \rightarrow 3$ et $3 \rightarrow 1$ dans les deux modèles. Les courbes lisses correspondent aux estimations dans le cas homogène (paramétrique) ; les courbes en escalier correspondent aux estimations dans le cas non-homogène (non-paramétrique). Les deux estimations des probabilités de transition $3 \rightarrow 1$ (IV.6 **(b)**) sont proches alors que celles associées à la transition $1 \rightarrow 3$ (IV.6 **(a)**) ne convergent pas vers la même valeur et n'ont pas la même forme (convergence plus lente dans le cas non-homogène).

Ensuite, la covariable IMC à la première consultation est étudiée dans un modèle homogène et dans un modèle non-homogène (IV.16). Les coefficients de régression et les probabi-

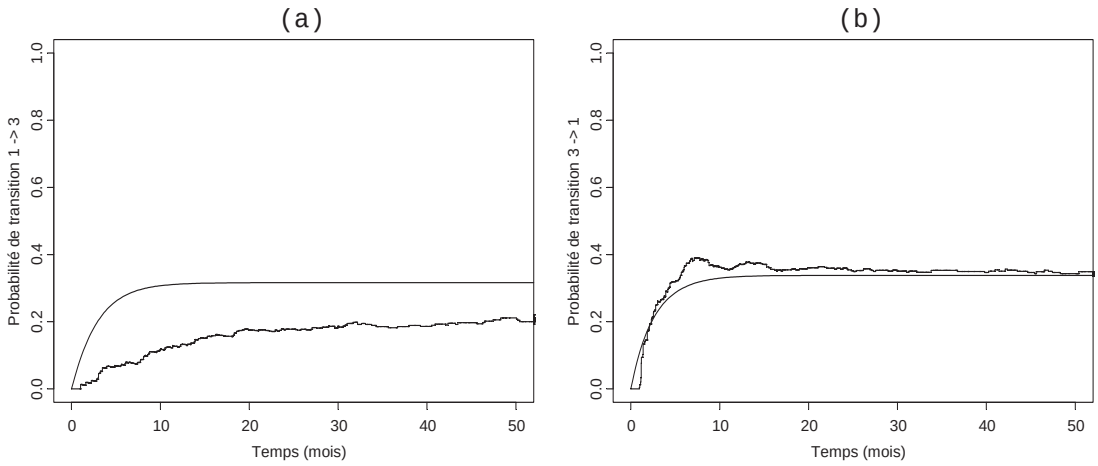


FIG. IV.6 – Estimation des probabilités de transition dans un modèle homogène (courbe lisse) et dans un modèle non-homogène (courbe en escalier) ; **(a)** transition de l'état optimal vers l'état inacceptable ($1 \rightarrow 3$) ; **(b)** transition de l'état inacceptable vers l'état optimal ($3 \rightarrow 1$).

lités de rejet de l'hypothèse de nullité des coefficients sont présentés dans le tableau IV.4. De manière générale, les estimations vont dans le même sens, même si les estimations du modèle non-homogène ont tendance à être plus petites (en valeur absolue). Notons cependant des différences significatives entre les estimations des coefficients associés à la transition $1 \rightarrow 3$ et $2 \rightarrow 1$. Plusieurs coefficients sont significatifs avec une méthode mais pas avec l'autre. Finalement, le coefficient associé à la transition $3 \rightarrow 1$ est le seul qui est statistiquement différent de zéro avec les deux modélisations.

6 Discussion

Ce chapitre présente des méthodes d'estimation dans le cadre d'un modèle de Markov non-homogène. La théorie des processus de comptage permet d'obtenir des estimateurs non-paramétrique et semi-paramétrique des probabilités de transition. Cette méthodologie pour modèle de Markov non-homogène est ensuite appliquée à une base de données sur l'asthme où l'hypothèse d'homogénéité semble trop restrictive (*cf.* chapitre II). Notons également qu'une méthode de programmation de ces estimateurs est présentée en annexe page 177.

6.1 Application

L'impact du surpoids a tout d'abord été observé en comparant les intensités de transition obtenues dans différentes strates. Ensuite, un modèle à risques proportionnels a été ajusté pour étudier l'effet de la covariable associée au surpoids. Les deux méthodes mettent en évidence les mêmes résultats : les patients en surpoids ont un risque plus important de transiter de l'état optimal vers l'état inacceptable, et également un risque plus faible de subir la transition inverse. Ces résultats sont similaires, en particulier parce que l'hypothèse de

Type de transition	Modèle homogène			Modèle non-homogène		
	$\hat{\beta}$	$(e.c)^1$	$(p)^2$	$\hat{\beta}$	$(e.c)^1$	$(p)^2$
optimal vers sous-optimal ($1 \rightarrow 2$)	-0.409	(0.383)	(0.02)	-0.248	(0.236)	(0.28)
optimal vers inacceptable ($1 \rightarrow 3$)	-0.122	(0.269)	(0.71)	0.655	(0.303)	(0.03)
sub-optimal vers optimal ($2 \rightarrow 1$)	0.542	(0.364)	(<0.01)	-0.030	(0.203)	(0.88)
sous-optimal vers inacceptable ($2 \rightarrow 3$)	0.041	(0.378)	(0.87)	0.204	(0.249)	(0.42)
inacceptable vers optimal ($3 \rightarrow 1$)	-1.170	(0.370)	(<0.01)	-0.595	(0.220)	(<0.01)
inacceptable vers sous-optimal ($3 \rightarrow 2$)	-0.561	(0.293)	(<0.01)	-0.234	(0.191)	(0.21)

[1] estimations des écarts-types.
[2] p-values avec le test du LRT pour $H_0 : \beta_{ij} = 0$.

TAB. IV.4 – Estimations des coefficients de régression associés à l'IMC dans un modèle homogène et dans un modèle non-homogène.

proportionnalité des risques est respectée dans le modèle semi-paramétrique. Ainsi, quand l'hypothèse de proportionnalité est respectée, le modèle semi-paramétrique est plus intéressant. En effet, ce modèle utilise toute la base de données pour l'estimation et les résultats sur l'effet de la covariable (coefficients de régression) sont facilement interprétables en terme de risque relatif.

Les résultats du modèle non-homogène ont ensuite été comparés à ceux du modèle homogène. Certains résultats sont proches, en particulier l'effet négatif du surpoids sur la transition de l'état inacceptable vers l'état optimal qui est significatif dans les deux méthodes. D'autres résultats sont différents : les probabilités de transition de l'état optimal vers l'état inacceptable ne convergent pas vers la même valeur et le coefficient associé à cette transition n'est pas significatif dans le modèle homogène.

Les différences de résultats s'expliquent en partie par la restriction induite par l'hypothèse d'homogénéité du temps. Ces différences peuvent aussi s'expliquer par la différence entre les méthodes d'estimation. En effet, le modèle homogène suppose (par la construction de la vraisemblance) que des événements peuvent avoir lieu entre deux consultations alors que l'estimation semi-paramétrique suppose que l'individu reste dans le même état entre deux consultations. Le modèle semi-paramétrique fait une hypothèse plus forte, à savoir que l'information est disponible en continu (ce qui est faux en pratique). De plus, il ajuste des données censurées à droite et suppose que l'état du patient ne varie pas entre la dernière consultation et la fin de l'étude. Le modèle de Markov homogène, lui, ne prend pas en compte la durée ou l'état de santé entre la dernière consultation et l'arrêt de l'étude.

Finalement, d'un point de vue clinique, les modèles homogènes et non-homogènes mettent en évidence que le surpoids diminue les chances de revenir à un état optimal. Ce résultat va-

lide une fois de plus le fait que le surpoids a un effet négatif sur l'évolution de l'asthme. Dans notre cas, où l'hypothèse d'homogénéité semble trop contraignante, les résultats obtenus avec le modèle non-homogène sont jugés plus fiables et sont conservés pour l'interprétation clinique.

Le modèle homogène et le modèle non-homogène n'étant pas emboîtés, il est difficile de comparer ces deux modèles. Le modèle homogène est facile à mettre en oeuvre mais l'hypothèse d'homogénéité est souvent trop restrictive. Le modèle non-homogène est plus complexe et par conséquent plus flexible. Cependant, les estimations peuvent parfois être difficilement interprétables. Ainsi, pour choisir la méthode la mieux adaptée, certains compromis doivent être faits en fonction de la base de données.

6.2 Méthodes

L'utilisation des modèles de Markov comporte certaines limites. La définition et le choix optimal des états de santé ne sont toujours aisés et l'hypothèse de Markov est parfois trop restrictive dans certaines applications. Des covariables artificielles, comme le temps de séjour avant de transiter ou le nombre de visites dans l'état présent, peuvent être utilisées pour tester cette hypothèse Markovienne. L'hypothèse de proportionnalité des risques induite par l'introduction de covariables dans les modèles doit aussi être vérifiée (en introduisant une covariable artificielle, par exemple).

Dans l'analyse des modèles multi-états et plus généralement dans les études de suivi, il est supposé que les changements d'états interviennent exactement au moment de la visite à l'hôpital. Cette approximation peut être plus ou moins contraignante. Dans notre cas, certains patients consultent en fonction de leurs besoins de santé et l'hypothèse n'est pas trop forte. Cependant, cette hypothèse devient moins réaliste pour les visites de routine fixées tous les trois mois. Une alternative consiste à utiliser des méthodes d'estimation pour données censurées par intervalles (Commenges [2002], Joly et al. [2002]). Ces méthodes considèrent que le moment de la transition intervient entre les deux visites à l'hôpital.

Les estimateurs présentés dans ce chapitre reposent sur une hypothèse de censure à droite indépendante. Cela suppose que la censure n'apporte aucune information sur le processus d'événement. Dans de nombreuses applications, cette hypothèse est une source de biais. Dans le cas de l'asthme, il semble qu'un patient qui se porte bien a tendance à ne pas venir en consultation. Ainsi, on peut penser que la censure apporte une information de bon diagnostic qui n'est pas prise en compte. Certaines méthodes, en particulier dans le cadre du modèle de Cox, s'intéressent à l'estimation de la survie sous une hypothèse de censure dépendante de l'événement.

6.3 Perspectives

Le modèle semi-paramétrique repose sur une structure multiplicative, c'est-à-dire que les intensités de transition sont le produit d'une fonction de risque et d'une fonction des

coefficients de régression. Cependant, on pourrait utiliser des modèles avec une structure additive, c'est-à-dire

$$\lambda_{hj}(t) = \lambda_{hj0}(t) + \boldsymbol{\beta}^T \mathbf{Z}.$$

Dans ce chapitre, des méthodes d'estimation non-paramétrique ont été discutées dans le cadre des modèles à intensité multiplicative. Cependant, des méthodes d'estimation paramétrique peuvent également être utilisées dans ces modèles en considérant

$$\lambda_{hj}(t;\boldsymbol{\theta}) = Y_h(t)\alpha_{hj}(t;\boldsymbol{\theta}),$$

où $\alpha_{hj}(t;\boldsymbol{\theta})$ est une fonction spécifiée par le vecteur de paramètres $\boldsymbol{\theta}$. La méthode du maximum de vraisemblance peut être utilisée pour estimer le vecteur de coefficients.

Les modèles de fragilité (frailty models) permettent de prendre en compte l'effet de variables omises dans la modélisation quand, par exemple,

- ces variables ne sont pas observées,
- les effets de ces variables sont déjà bien connus,
- il n'est pas certain que ces variables influencent les intensités.

Les modèles à intensité multiplicative peuvent être étendus à des modèles de fragilité (Huber-Carol et Vonta [2004], Andersen et al. [1993], Nielsen et al. [1992]). Dans ce cas, les intensité du processus de comptage peuvent dépendre en partie d'une variable aléatoire non observée qui agit de manière multiplicative sur les intensités, c'est-à-dire

$$\lambda_{hj}(t) = ZY_h(t)\alpha_{hj}(t).$$

Ce problème statistique peut être vu comme un problème de données manquantes. Les données complètes mais non observables sont les processus Z, $\mathbf{N}$, et $\mathbf{Y}$. Les données disponibles correspondent aux processus $\mathbf{N}$ et $\mathbf{Y}$. La variable de fragilité Z peut, par exemple, suivre une loi gamma et l'algorithme EM peut être utilisé pour résoudre ce problème de données incomplètes.

Chapitre V

Prise en compte de la censure informative - Méthode IPCW

L'objectif de ce chapitre est de présenter une méthode qui permette de prendre en compte l'information contenue dans le phénomène de censure. En effet, les méthodes classiques d'analyse des modèles de survie et de modèles multi-états supposent une indépendance entre le processus d'événement et le processus de censure. Cependant, cette hypothèse est souvent trop stricte et les estimations sont biaisées. Dans le cas des modèles de survie, la méthode IPCW (Inverse Probability Censoring Weight) introduit par Robins [1993], prend en compte la dépendance entre la censure et l'évènement. Le principe de la méthode est d'étudier le risque de censure en fonction de covariables afin de déduire des poids spécifiques à chaque individu. Les poids sont ensuite intégrés dans les estimateurs classiques pour modifier le nombre de personnes à risque pour prendre en compte les personnes censurées et ainsi tenir compte du phénomène de censure informative.

Ce chapitre se décomposera en trois parties. Une première partie présente la méthode IPCW pour modèles de survie. Cette partie reprend essentiellement les résultats d'un article (Saint-Pierre et al. [2005b]) soumis dans la revue *Biometrical Journal*. Elle est inspirée des travaux de Castelli [2004] et de Robins et Finkelstein [2000]. Dans une deuxième partie, nous proposons une généralisation de la méthode IPCW à certains modèles multi-états Markovien. En effet, il existe peu de méthodes pour traiter la censure informative dans de tels modèles bien que le phénomène de censure pose les mêmes difficultés que dans les études de survie. Dans une troisième partie nous présentons l'application des méthodes IPCW pour données de survie et pour modèles multi-états à la base de données sur l'asthme.

1 Méthode IPCW pour les modèles de survie

1.1 Introduction

Les données de survie ou plus généralement des données liées au temps sont souvent présentes dans les études épidémiologiques. Les analyses de ces données sont alors réalisées afin de mieux connaître la pathologie et les effets des différentes covariables. Dans ces études,

on cherche à expliquer l'apparition d'un événement : le décès, une rechute, une cicatrisation, un rejet de greffe, ou un autre état spécifique à la maladie étudiée. La particularité des données de survie repose sur le fait que ces données sont très souvent incomplètes à cause du phénomène de censure. En effet, un événement n'est jamais observé sur un temps infini et de plus, dans les bases de données observationnelles, il est impossible de contrôler la présence aux visites ce qui génère des individus perdus de vue. Les méthodes statistiques pour ces données de survie doivent alors tenir compte des données censurées. Cette problématique est essentielle car on se trouve en situation naturelle d'utilisation des ressources du système de santé.

Dans les analyses de survie traditionnelles (Kaplan-Meier, modèle de Cox, modèles paramétriques) (Hill et al. [1996]), et dans la quasi-totalité des études, la censure est considérée comme indépendante du processus étudié. Cette censure est dite non informative, car quand elle se produit, elle n'apporte aucune information sur l'événement étudié. Dans ces analyses, l'information potentielle apportée par les patients perdus de vue est ignorée. Cette hypothèse d'indépendance entre la censure et l'événement est évidemment une hypothèse forte qui induit un biais dans l'estimation de la survie. De plus, cette hypothèse n'est que rarement vérifiée dans la réalité. En effet, prenons l'exemple d'une pathologie lourde comme le VIH : les patients perdus de vue sont généralement les patients les plus malades. Dans le cas de l'asthme, il semble au contraire que les patients perdus de vue soient des patients bien contrôlés qui ne ressentent pas le besoin de venir consulter. Dans ces deux exemples, le processus de censure est clairement lié à l'événement étudié.

Les méthodes permettant de prendre en compte cette information apportée par la censure sont fondées sur l'hypothèse de dépendance entre la censure et l'événement et visent à réduire le biais dans l'estimation de la survie. La méthode Inverse Probability Censoring Weighted (IPCW) est souvent citée dans la littérature (Rotnitzky et Robins [2003], Scharfstein et al. [2001], Scharfstein et al. [1999], Robins [1993], Robins et Finkelstein [2000]). Elle adapte les méthodes traditionnelles d'analyse de données de survie par pondération des estimateurs. D'autres méthodes pour prendre en compte cette dépendance dans l'analyse de données répétées sont présentes dans la littérature. Par exemple, Huang et Wolfe [2002], Liu et al. [2004] proposent des modèles de fragilité pour prendre en compte la corrélation entre la censure et l'événement dans le cas de données en grappe, c'est-à-dire pour des données regroupées en cluster. Minini et Chavance [2004] proposent une approche visant à multiplier le risque de décès par un paramètre après la survenue de la censure. Robins et Finkelstein [2000] présentent le G-algorithme qui utilise l'intégration de Monte-Carlo pour estimer la courbe de survie, cependant cette méthode d'estimation est en général moins performante que la méthode IPCW. La méthode IPCW a été généralisée au cadre des modèles GEE (équations d'estimation généralisées) par Matsuyama [2003]. Ce modèle GEE pondéré permet de modéliser une variable à mesures répétées dans le temps sous l'hypothèse de dépendance de la censure.

Dans le cas de l'asthme, il y a un nombre important de perdus de vue et la censure semble informative. Dans ce contexte de censure dépendante, l'utilisation des méthodes traditionnelles conduirait à introduire un biais dans l'estimation de la fonction de survie. Une alternative est l'utilisation de la méthode IPCW qui ajuste le modèle de Cox pour censure dépendante en utilisant les données collectées sur les facteurs de risques d'événement et de censure. La particularité de la méthode est d'intégrer dans les estimations traditionnelles des poids dépendants de la probabilité d'être non censuré au temps t. Cette partie sur la

méthode IPCW pour données de survie est inspirée des travaux de et .

1.2 Mécanisme de censure

Dans les bases de données observationnelles, différents types de censures sont rencontrés : la censure à gauche, la censure à droite ou encore la censure par intervalle.

La censure par intervalle se produit quand on n'observe pas les temps de changement d'état : la seule information disponible est que le changement d'état se produit entre deux consultations. Dans le cas d'observations intermittentes, les données sont censurées par intervalle car l'état du patient est observé uniquement au moment de la consultation. Les méthodes présentées dans cette thèse ne prennent pas en compte cette censure. Elles supposent que les changements d'état se produisent exactement aux moments des consultations. Dans le cas de l'asthme, les données observées sont censurées par intervalle. Pour appliquer les méthodes présentées, nous supposons que les changements d'état se produisent aux moments des visites à l'hôpital. Cette hypothèse est restrictive mais semble acceptable dans notre cas car les patients doivent venir consulter quand ils ressentent un changement dans leurs états de santé. Cependant, afin d'être plus réaliste dans la modélisation d'observations discrètes, on devrait considérer des méthodes d'estimation permettant d'ajuster une censure par intervalle (Commenges [2002], Joly et al. [2002]).

Dans cette section, nous considérons le phénomène de censure à droite qui est le plus courant dans les études de survie. Il existe différents mécanismes de censure à droite : un individu est considéré comme censuré s'il est perdu de vue (déménagement, arrêt du suivi pour diverses raisons), exclu vivant (si l'événement se produit après la date de point) ou décédé pour cause de risques compétitifs (décès non imputable à la maladie comme par exemple un accident de la route). Dans la pluspart des études, la censure engendrée par les exclus vivants et décédés pour causes de risques compétitifs peut le plus souvent être considérée comme non informative. Par contre, la censure par perdus de vue est, dans la majorité des cas, informative de l'événement étudié et se manifeste de différentes façons.

En effet, les perdus de vue sont les patients dont on ne connait pas l'état à la date de point. Ils représentent une perte d'information et sont source de biais. Le mécanisme de censure devient dépendant quand un patient n'est plus suivi volontairement par le médecin car il sait pertinemment que le décès est proche ou encore si le patient est non compliant c'est à dire qu'il a arrêté son traitement et ne vient plus aux consultations. Ce phénomène de non compliance est courant pour les patients atteints de VIH par exemple. En effet, le traitement est souvent lourd avec des effets secondaires multiples et provoque ainsi des arrêts du suivi. De plus il a été observé que ce sont les patients les plus malades qui arrêtent leurs traitements soit parce qu'ils ne supportent plus la thérapie soit parce qu'ils n'ont plus d'espoir. Dans le cas de l'asthme au contraire, les patients non compliants semblent être des patients qui se portent bien et ne ressentent pas le besoin d'être suivi. Dans ces exemples, la censure devient dépendante de l'événement puisqu'elle apporte une information sur l'événement étudié. Dans les essais cliniques, afin d'étudier la différence entre deux traitements en terme de survie, on est amené à considérer l'arrêt ou le changement de traitement après une certaine date comme une censure. Un résultat thérapeutique insuffisant, des effets secondaires trop gênants, un

traitement mal toléré mais aussi un état satisfaisant du patient peuvent être la cause de ces censures. Ce mécanisme de censure devient souvent dépendant car l'arrêt ou le changement de traitement informe généralement sur l'état du patient.

Ainsi, un grand nombre d'individus des bases observationnelles est censuré et une part importante de ces censures sont a priori informatives. Pour cette raison les méthodes traditionnelles occultant cette information sont souvent biaisées. Il est donc important d'avoir des méthodes plus proches de la réalité pour améliorer les estimations de la survie.

1.3 Hypothèses

Robins et Rotnitzky [1992] et Robins [1993] ont montré que si l'on dispose de tous les facteurs de risques dépendants du temps pour l'événement, alors on peut corriger la dépendance entre la censure et l'événement étudié en remplaçant les estimateurs de Kaplan-Meier et de la vraisemblance partielle de Cox par les versions corrigées de la méthode IPCW.

Soit $\mathbf{V}(t)$ les covariables qui prédisent l'événement au temps t et soit $\bar{\mathbf{V}}(t) = \{\mathbf{V}(x); 0 \leqslant x \leqslant t\}$ l'histoire de ces covariables. L'hypothèse fondamentale de la méthode est que le risque de censure au temps t ne dépend plus du possible temps d'événement non observé T, *i.e*,

$$\lambda_C(t \mid \bar{\mathbf{V}}(t), T, T > t) = \lambda_C(t \mid \bar{\mathbf{V}}(t), T > t). \tag{V.1}$$

Cette hypothèse signifie que la connaissance temporelle des covariables apportent suffisamment d'informations pour que l'on puisse se passer de celle de T. En effet, la variable aléatoire T n'apporte qu'une information d'événement déjà prédite par les covariables (qui prédisent cet événement). Cette hypothèse peut être considérée comme l'hypothèse en « miroir » nécessaire pour étudier le risque de l'événement. En pratique, on ne s'attend pas à ce que cette hypothèse soit vraie, en effet, si un important facteur de risque de l'événement (et de la censure) n'est pas inclu dans $\mathbf{V}(t)$ alors l'hypothèse (V.1) se rapproche de l'indépendance entre l'événement et la censure ce qui amoindrit l'intérêt de la méthode. Il est donc important pour l'utilisation d'une telle méthode de disposer d'un nombre suffisant de variables dans $\mathbf{V}(t)$ pour que (V.1) soit possible.

Soit $\lambda_T(t \mid .) = \lim_{h \to 0} \frac{1}{h} \Pr(t \leqslant T \leqslant t + h \mid T > t, .)$, alors, l'hypothèse de censure indépendante nécessaire à l'application de Kaplan-Meier et de la vraisemblance partielle est définie par

$$\lambda_T(t \mid C > t) = \lambda_T(t). \tag{V.2}$$

où C est le temps de censure. Cette hypothèse de censure indépendante sera fausse si $\bar{\mathbf{V}}(t)$ prédit le risque de censure au temps t, *i.e* si

$$\lambda_C(t \mid \bar{\mathbf{V}}(t), T > t) \neq \lambda_C(t \mid T > t). \tag{V.3}$$

En effet, par hypothèse, $\bar{\mathbf{V}}(t)$ prédit l'événement, donc s'il détermine aussi la censure, censure et événement seront dépendants par l'intermédiaire des covariables. Pour que la censure soit considérée comme non informative, il faut que

$$\lambda_C(t \mid \bar{\mathbf{V}}(t), T > t) = \lambda_C(t \mid T > t). \tag{V.4}$$

Il n'existe pas de méthode (test empirique) pour tester la validité de l'hypothèse (V.1), par contre on peut aisément vérifier la validité de (V.4) en étudiant le risque de censure

$\lambda_C(t \mid \bar{\mathbf{V}}(t), T > t)$ par un modèle de Cox. Si certains coefficients de régression associés aux covariables sont statistiquement différents de zéro alors, (V.4) et donc l'hypothèse de censure indépendante seront fausses. Dans ce cas, les estimations de Kaplan-Meier et de la vraisemblance partielle de Cox sont biaisées. On pourra alors utiliser les versions IPCW de ces estimations pour corriger le biais dû à la dépendance de la censure.

1.4 Réduction du nombre de covariables

Dans un premier temps, on veut réduire la dimension de $\bar{\mathbf{V}}(t)$ qui comprend un nombre important de variables. Un modèle de Cox est ajusté pour l'événement étudié afin de sélectionner uniquement les facteurs qui sont significatifs :

$$\lambda_T(t \mid \bar{\mathbf{V}}(t)) = \gamma_0(t) \exp(\boldsymbol{\psi}^T \mathbf{V}(t)),$$

où $\mathbf{V}(t)$ sont les valeurs les plus récentes des covariables, $\boldsymbol{\psi}$ sont les coefficients de régression associés et $\gamma_0(t)$ est le risque de base. On peut noter que ce modèle suppose que conditionnellement aux valeurs récentes, les valeurs passées de $\mathbf{V}(.)$ ne prédisent pas l'événement en t (*i.e* $\bar{\mathbf{V}}(t)$ est résumé par $\mathbf{V}(t)$). En gardant uniquement les facteurs pour lesquels les coefficients de régression sont statistiquement différents de zéro, on obtient la version réduite $\bar{\mathbf{V}}^*(t)$ de $\bar{\mathbf{V}}(t)$. Afin d'utiliser la version réduite de $\bar{\mathbf{V}}(t)$ pour la modélisation, il faut supposer que les risques de censure sachant $\bar{\mathbf{V}}(t)$ et $\bar{\mathbf{V}}^*(t)$ sont égaux, *i.e*,

$$\lambda_T(t \mid \bar{\mathbf{V}}(t)) = \lambda_T(t \mid \bar{\mathbf{V}}^*(t)). \tag{V.5}$$

Une fois la version réduite $\bar{\mathbf{V}}^*(t)$ obtenue, on peut utiliser la méthode IPCW avec $\bar{\mathbf{V}}^*(t)$, cependant, il est nécessaire que l'hypothèse (V.1) soit vérifiée pour $\bar{\mathbf{V}}^*(t)$, *i.e*

$$\lambda_C(t \mid \bar{\mathbf{V}}^*(t), T, T > t) = \lambda_C(t \mid \bar{\mathbf{V}}^*(t), T > t). \tag{V.6}$$

Robins (1986) a montré que (V.1) et (V.5) n'impliquaient pas (V.6). Par contre si (V.1) et (V.5) sont vérifiées et que les variables éliminées $\mathbf{V}^+(u)$ ne prédisent pas les futures valeurs des facteurs $\mathbf{V}^*(u)$ *i.e*

$$f(\mathbf{V}^*(u) \mid \bar{\mathbf{V}}^+(u^-), \bar{\mathbf{V}}^*(u^-), T \geqslant u, C \geqslant u) = f(\mathbf{V}^*(u) \mid \bar{\mathbf{V}}^*(u^-), T \geqslant u, C \geqslant u), \tag{V.7}$$

alors (V.6) sera vérifiée. L'hypothèse (V.7) implique que $\bar{\mathbf{V}}^*(u^-)$ (*i.e* $\mathbf{V}^+(x); 0 \leqslant x \leqslant u^-$) ne prédit pas le processus $\mathbf{V}^*(u)$ sachant l'historique de $\mathbf{V}^*(u^-)$.

Une autre possibilité pour que l'hypothèse (V.6) soit vérifiée est d'utiliser le concept de variables Coarsened At Random (CAR) introduit par Heitjan et Rubin [1991]. En effet si $(\bar{\mathbf{V}}^*(T), T)$ sont CAR *i.e*

$$\lambda_C(t \mid \bar{\mathbf{V}}^*(T), T, T > t) = \lambda_C(t \mid \bar{\mathbf{V}}^*(t), T > t), \tag{V.8}$$

alors (V.6) sera vérifiée. On peut noter que (V.8) diffère de (V.6) car $\bar{\mathbf{V}}^*(T)$ remplace $\bar{\mathbf{V}}^*(t)$. L'hypothèse (V.8) implique (V.6) mais la réciproque est fausse. En fait, l'hypothèse implique que les valeurs des covariables ne varient plus entre le temps t où on veut calculer le risque de censure et le temps T où se serait produit l'évènement si le sujet n'avait pas été censuré (*i.e* $\bar{\mathbf{V}}^*(t)$ apporte le même information que $\bar{\mathbf{V}}^*(T)$). La notion de Coarsening At Random a été

introduite pour les cas où on n'observe pas la valeur exacte des données mais seulement un ensemble qui contient la vraie valeur. Cette définition a l'avantage de couvrir de nombreux problèmes aux données incomplètes qui ont émergé en biomédecine, incluant les arrondis, les censures et les données manquantes.

1.5 Etude du risque de censure

1.5.1 Notations et estimation

L'estimation IPCW repose sur une pondération des estimateurs de Kaplan-Meier et de la vraisemblance partielle. Les poids permettent de modifier le nombre de personnes à risque et le nombre de personnes qui subissent l'événement de manière à prendre en compte les patients censurés. Ces poids sont spécifiques à chaque individu et sont construits à partir de la probabilité que l'individu soit censuré. L'estimation du risque de censure permet d'obtenir ces pondérations et d'étudier la dépendance entre l'événement et la censure. Si (V.6) est vérifiée, on peut étudier le risque de censure en fonction des facteurs de risque $\mathbf{V}^*(t)$ qui prédisent l'événement. Le modèle pour le risque de censure peut s'écrire de la façon suivante :

$$\lambda_C(t \mid \bar{\mathbf{V}}^*(t), T > t) = \lambda_0(t) \exp(\boldsymbol{\alpha}^T \mathbf{V}^*(t)). \tag{V.9}$$

On pose pour chaque individu $(k = 1, ..., n)$:

- $X_k = min(T_k, C_k)$: temps d'événement T_k ou temps de censure C_k pour l'individu k : celui qui se produit en premier.
- $Y_k(u) = \mathbb{1}_{\{X_k \geqslant u\}}$: vaut 1 si l'individu k est à risque (n'a pas subi l'événement et n'est pas censuré), 0 sinon.
- $\tau_k = \mathbb{1}_{\{T_k = X_k\}}$: vaut 1 si on observe l'événement et 0 si l'individu k est censuré.
- $\mathbf{V}_k^*(u)$ les covariables au temps u associés à l'individu k.
- $(\bar{\mathbf{V}}_k^*(T_k), T_k)$ sont CAR, ainsi (V.6) est vérifiée.

Un estimateur de $\boldsymbol{\alpha}$ s'obtient par maximisation de la vraisemblance partielle suivante :

$$V = \prod_{j=1}^{n} \left(\frac{\exp(\boldsymbol{\alpha}^T \mathbf{V}_k^*(X_j))}{\sum_{k=1}^{n} Y_k(X_j) \exp(\boldsymbol{\alpha}^T \mathbf{V}_k^*(X_j))} \right)^{1-\tau_j}$$

Soit $\hat{\alpha}$ l'estimateur de α par maximisation de la vraisemblance précédente. On peut ensuite estimer le risque de base par un estimateur du type Kaplan-Meier, pour chaque temps :

$$\hat{\lambda}_0(X_j) = \frac{1 - \tau_j}{\sum_{k=1}^{n} Y_k(X_j) \exp(\hat{\boldsymbol{\alpha}}^T \mathbf{V}_k^*(X_j))}, \qquad j = 1, ..., n.$$

1.5.2 Extension possible

Dans le modèle décrit précédemment, tous les patients ne subissant pas l'événement (*i.e* tous les patients censurés) sont utilisés pour estimer le risque de censure. Autrement dit, aucune différence n'est faite entre les patients perdus de vue et les patients exclus vivants (censurés à la date de point). Pourtant, il semble évident qu'il y ait une différence entre ces

deux phénomènes de censure. En effet, la censure engendrée par les patients perdus de vue est potentiellement informative alors que la censure engendrée par le gel de la base n'apporte a priori aucune information sur la survenue de l'événement étudié.

La méthodologie présentée précédemment peut aisément être adaptée pour étudier uniquement le risque de la censure générée par les patients perdus de vue. Pour cela, il est nécessaire de distinguer les deux types de censure. Considérons pour chaque individu, $k = 1, ..., n$,

$$\gamma_k = \begin{cases} 1 & \text{si l'individu est perdu de vue} \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance partielle s'écrit alors,

$$V = \prod_{j=1}^{n} \left(\frac{\exp(\boldsymbol{\alpha}^T \mathbf{V}_k^*(X_j))}{\sum_{k=1}^{n} Y_k(X_j) \exp(\boldsymbol{\alpha}^T \mathbf{V}_k^*(X_j))} \right)^{\gamma_j}$$

et le risque de censure de base est estimé par :

$$\hat{\lambda}_0(X_j) = \frac{\gamma_j}{\sum_{k=1}^{n} Y_k(X_j) \exp(\hat{\boldsymbol{\alpha}}^T \mathbf{V}_k^*(X_j))}, \qquad j = 1, ..., n.$$

De cette manière, le risque de censure est estimé en supposant a priori que la censure générée par les exclus vivants est non informative. On estime un risque de transiter vers un état de censure par perdu de vue ou encore un « risque de censure informative ». De cette façon, les pondérations calculées par la suite modifieront les contributions des individus dans les estimations, en prenant compte uniquement les personnes perdues de vue.

1.5.3 Estimation de la survie de la censure et calcul des poids

Soit $K_i(t) = \Pr(C_i > t)$, la survie de la censure pour l'individu i. Un estimateur de $K_i(t)$ est donné par le produit limite :

$$\begin{aligned} \hat{K}_i(t) &= \prod_{\{j; X_j < t, \tau_j = 0\}} [1 - \hat{\lambda}_C(X_j \mid \bar{\mathbf{V}}_i^*(X_j))] \\ &= \prod_{\{j; X_j < t, \tau_j = 0\}} \left[1 - \hat{\lambda}_0(X_j) \times \exp(\hat{\boldsymbol{\alpha}}^T \mathbf{V}_i^*(X_j))\right] \\ &= \prod_{\{j; X_j < t\}} \left[1 - \frac{(1 - \tau_j) \times \exp(\hat{\boldsymbol{\alpha}}^T \mathbf{V}_i^*(X_j))}{\sum_{k=1}^{n} Y_k(X_j) \exp(\hat{\boldsymbol{\alpha}}^T \mathbf{V}_i^*(X_j))}\right]. \end{aligned}$$

Soit $\hat{K}^0(t)$ l'estimateur de la survie de la censure dans un modèle sans covariable, c'est-à-dire l'estimateur classique de Kaplan-Meier,

$$\hat{K}^0(t) = \prod_{\{j; X_j < t\}} \left[1 - \frac{(1 - \tau_j)}{\sum_{k=1}^{n} Y_k(X_j)}\right]$$

Les poids spécifiques à chaque individu vont être déterminés à partir du calcul de ces probabilités :

$$\hat{W}_i(t) = \frac{\hat{K}^0(t)}{\hat{K}_i(t)}. \tag{V.10}$$

Notons que $\hat{K}^0(t)$ est identique pour tous les individus. Par contre, les individus se différenciant par l'historique de leurs covariables, $\hat{K}_i(t)$ sera différent d'un individu à l'autre. De plus, si $\hat{\boldsymbol{\alpha}}$ est le vecteur nul, c'est-à-dire si l'historique de $\bar{\mathbf{V}}^*(t)$ ne prédit pas le risque de censure en t (censure indépendante) alors $\hat{K}_i(t) = \hat{K}^0(t)$ et ainsi pour tout t, les poids convergent vers 1. En revanche, si la censure est dépendante, les poids ne convergeront pas vers 1. Si $\hat{W}_i(t) > 1$ alors la probabilité de rester non censuré jusqu'au temps t sera plus importante dans un modèle sans covariable ce qui signifie que les covariables associées à l'individu i augmentent la probabilité d'être censuré pour l'individu i. Si au contraire $\hat{W}_i(t) < 1$, alors les covariables de l'individu i diminuent la probabilité d'être censuré avant t. Ces poids vont ainsi permettre de modifier l'estimation du risque d'événement en prenant en compte l'effet de la censure sur chaque individu.

1.6 Estimation IPCW

1.6.1 Estimation IPCW de la survie

L'estimateur IPCW de la survie est construit en modifiant l'estimateur de Kaplan-Meier par les poids calculés précédemment. L'estimation de la probabilité de rester vivant jusqu'au temps t est donnée par

$$\begin{aligned}\hat{S}_T(t) &= \prod_{\{j;X_j<t\}} \left[1 - \frac{\tau_j \times \hat{W}_j(X_j))}{\sum_{k=1}^n Y_k(X_j)\hat{W}_k(X_j)}\right] \\ &= \prod_{\{j;X_j<t\}} \left[1 - \frac{\tau_j/\hat{K}_j(X_j))}{\sum_{k=1}^n Y_k(X_j)/\hat{K}_k(X_j)}\right].\end{aligned} \tag{V.11}$$

Sous les hypothèses (V.6) et (V.9), Robins [1993] montre que $\hat{S}_T(t)$ est un estimateur consistant de $S_T(t) = Pr(T > t)$. La consistance vient principalement du fait que $\tau_j/\hat{K}_j(X_j)$ estime le nombre de sujets qui devraient subir l'événement au temps X_j en absence de censure et que $\sum_{k=1}^n Y_k(X_j)/\hat{K}_k(X_j)$ estime le nombre de sujets qui devraient être à risque au temps X_j en absence de censure. Ainsi le ratio de ces deux quantités estime le risque de décès en absence de censure. Il s'ensuit que $\hat{S}_T(t)$ estime la probabilité rester vivant jusqu'au temps t en absence de censure.

Le principe de la méthode est d'augmenter le poids d'un individu qui a subi l'événement à une date t pour considérer les censures qui ont lieu avant cette date t et qui sont expliquées par les covariables avant t. Pour chaque individu k qui est à risque au temps X_j (*i.e.* $Y_k(X_j) = 1$) et dont la probabilité d'être non censuré après X_j est $\hat{K}_k(X_j) = 0.25$, il devrait y avoir, afin de corriger la censure, en moyenne trois autres sujets similaires en pronostics qui ont été censurés avant X_j et qui devraient, comme l'individu k, avoir survécu au moins jusqu'en X_j. Un poids de $4 = 1/0.25$ est alors attribué à l'individu k dans l'estimation du nombre de sujets à risque au temps X_j. En augmentant le poids du sujet k, on prend en compte les sujets censurés avant X_j du fait de l'interdépendance entre l'événement étudié et la censure. De manière similaire un individu qui subit l'événement en X_j avec un poids $\hat{K}_j(X_j) = 0.25$ doit avoir trois images similaires en pronostics qui devraient avoir le même temps d'événement.

1.6.2 Version IPCW du score de la vraisemblance partielle

La méthode IPCW permet aussi d'ajuster un modèle de Cox avec covariables dépendantes du temps. Elle fournit ainsi des estimations des coefficients de régression sous l'hypothèse de censure dépendante. Considérons le modèle suivant

$$\lambda_T(t \mid \mathbf{Z}(.), C > t) = \lambda_0(t) \times e^{\boldsymbol{\beta}^T \mathbf{Z}(t)}, \tag{V.12}$$

où $\lambda_0(.)$ est le risque de base et $\boldsymbol{\beta}$ les coefficients de régression associés à $\mathbf{Z}(.)$. Par exemple, dans Robins et Finkelstein [2000], le modèle précédent est utilisé pour comparer la survie IPCW dans deux bras de traitements, avec $Z = 0$ représente l'appartenance au bras A et $Z = 1$ l'appartenance au bras B. On peut ainsi interpréter le coefficient $\boldsymbol{\beta}$ en terme de risques relatifs : si $\boldsymbol{\beta}$ est significativement positif, le traitement A aura un effet bénéfique sur la survie comparé au traitement B (risque de décès accéléré pour le traitement B).

Afin d'estimer $\boldsymbol{\beta}$ dans un contexte de censure dépendante, Robins (1993) a montré que le score IPCW de la vraisemblance partielle $U(\boldsymbol{\beta})$ se différenciait du score ordinaire de la vraisemblance partielle de Cox seulement par le fait que la contribution du sujet k au temps X_j est pondérée par le poids $\hat{W}_k(X_j)$, *i.e*,

$$U(\boldsymbol{\beta}) = \sum_{j=1}^{n} \tau_j \hat{W}_j(X_j)) \times \left[\mathbf{Z}_j(X_j) - \frac{\sum_{k=1}^{n} Y_k(X_j) \hat{W}_k(X_j) \mathbf{Z}_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)}}{\sum_{k=1}^{n} Y_k(X_j) \hat{W}_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)}} \right]. \tag{V.13}$$

En effet, la vraisemblance partielle est :

$$V = \prod_{j=1}^{n} \left(\frac{Y_j(X_j) e^{\boldsymbol{\beta}_j^T \mathbf{Z}(X_j)}}{\sum_{k=1}^{n} Y_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)}} \right)^{\tau_j},$$

$$\log V = \sum_{i=1}^{n} \tau_j \left[\log(Y_j(X_j)) + \boldsymbol{\beta}_j^T \mathbf{Z}(X_j) - \log \left(\sum_{k=1}^{n} Y_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)} \right) \right],$$

et la dérivée de la log-vraisemblance est donnée par :

$$\frac{\partial \log V}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{n} \tau_j \times \left[\mathbf{Z}_j(X_j) - \frac{\sum_{k=1}^{n} Y_k(X_j) \mathbf{Z}_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)}}{\sum_{k=1}^{n} Y_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)}} \right].$$

Soit

$$U^0(\boldsymbol{\beta}) = \frac{\partial \log V}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{n} \left[\tau_j \frac{\hat{K}^0(X_j)}{\hat{K}^0(X_j)} \right] \left(\mathbf{Z}_j(X_j) - E_j^0(\beta) \right),$$

avec,

$$E_j^0(\boldsymbol{\beta}) = \frac{\sum_{k=1}^{n} Y_k(X_j) \mathbf{Z}_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)} \frac{\hat{K}^0(X_j)}{\hat{K}^0(X_j)}}{\sum_{k=1}^{n} Y_k(X_j) e^{\boldsymbol{\beta}^T \mathbf{Z}_k(X_j)} \frac{\hat{K}^0(X_j)}{\hat{K}^0(X_j)}}.$$

De la même façon que pour $E_j^0(\boldsymbol{\beta})$, $E_j(\boldsymbol{\beta})$ est défini en remplaçant $\hat{K}^0(X_j)$ au dénominateur par $\hat{K}_k(X_j)$:

$$E_j(\boldsymbol{\beta}) = \frac{\sum_{k=1}^{n} Y_k(X_j) \mathbf{Z}_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)} \frac{\hat{K}^0(X_j)}{\hat{K}_k(X_j)}}{\sum_{k=1}^{n} Y_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)} \frac{\hat{K}^0(X_j)}{\hat{K}_k(X_j)}},$$

et

$$U(\boldsymbol{\beta}) = \sum_{j=1}^{n} \left[\tau_j \frac{\hat{K}^0(X_j)}{\hat{K}_j(X_j)} \right] (\mathbf{Z}_j(X_j) - E_j(\boldsymbol{\beta}))$$
$$= \sum_{j=1}^{n} \left[\tau_j \hat{W}_j(X_j) \right] \left(\mathbf{Z}_j(X_j) - \frac{\sum_{k=1}^{n} Y_k(X_j) \mathbf{Z}_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)} \hat{W}_k(X_j)}{\sum_{k=1}^{n} Y_k(X_j) e^{\boldsymbol{\beta}_k^T \mathbf{Z}(X_j)} \hat{W}_k(X_j)} \right).$$

$U(\boldsymbol{\beta})$ est appelée fonction score modifiée.

Sous l'hypothèse (V.6) et (V.9), Robins [1993] montre que la solution $\hat{\boldsymbol{\beta}}$ de l'équation score IPCW $U(\boldsymbol{\beta}) = 0$ est un estimateur consistant du paramètre $\boldsymbol{\beta}$ du modèle de Cox. On peut noter que si les poids $\hat{W}_k(X_j)$ sont remplacés par $\frac{1}{\hat{K}_j(X_j)}$, la solution de $U(\boldsymbol{\beta}) = 0$ est encore consistante et asymptotiquement normale, cependant cet estimateur est moins efficace.

1.6.3 Ecarts-types

D'après les théorèmes de convergence asymptotique de Robins [1993], sous les conditions (V.6), (V.9) et (V.12), la variance asymptotique de $\hat{\boldsymbol{\beta}}$ est estimée de manière consistante par $\hat{I}^{-1}\hat{\Sigma}(\hat{\boldsymbol{\beta}}) \left(\hat{I}^{-1}\right)^T$, avec $\hat{\Sigma}(\hat{\boldsymbol{\beta}}) = \Omega_1(\hat{\boldsymbol{\beta}}) - \Omega_2(\hat{\boldsymbol{\beta}}) - \Omega_3(\hat{\boldsymbol{\beta}})$ et $\hat{I} = \frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$.

– $\Omega_1(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{A}_i(\boldsymbol{\beta}, 0) \right\}^2$, où

$$\hat{A}_i(\boldsymbol{\beta}, x) = \int_x^{\infty} d\hat{M}_{T_i}(u, \boldsymbol{\beta}) \hat{W}_i(u) \times \left[Z - \frac{\sum_{j=1}^{n} Y_j(u) \hat{W}_j(u) Z_j e^{\boldsymbol{\beta}_j^T \mathbf{Z}(u)}}{\sum_{j=1}^{n} Y_j(u) \hat{W}_j(u) e^{\boldsymbol{\beta}_j^T \mathbf{Z}(u)}} \right],$$

$$\hat{M}_{T_i}(u, \boldsymbol{\beta}) = N_{T_i}(u) - \int_0^u d\hat{\Lambda}_0(t, \beta) Y_i(t) e^{\boldsymbol{\beta}_i^T \mathbf{Z}(t)},$$
$$N_{T_i} = \mathbb{1}_{\{X_i \leqslant u, \tau_i = 1\}},$$
$$d\hat{\Lambda}_0(t, \boldsymbol{\beta}) = \frac{\sum_{j=1}^{n} \hat{W}_j(t) dN_{T_j}}{\sum_{j=1}^{n} Y_j(t) \hat{W}_j(t) e^{\boldsymbol{\beta}_j^T \mathbf{Z}(t)}}.$$

– $\Omega_2(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^{\infty} dN_{C_i}(u) \left\{ G(\hat{A}(\boldsymbol{\beta}, u) \right\}^2 \right]$. Sachant le model V.12, pour tous $\mathbf{H}(u) = (\mathbf{H}_j(u))_{j=1,\ldots,n}$, où les $\mathbf{H}_j(u)$ peuvent être des vecteurs de fonctions, G est définie par,

$$G\{\mathbf{H}(u)\} = \frac{\sum_{j=1}^{n} \mathbf{H}_j(u) Y_j(u) e^{\hat{\boldsymbol{\alpha}}^T \mathbf{V}_i^*(u)}}{\sum_{j=1}^{n} Y_j(u) e^{\hat{\boldsymbol{\alpha}}^T \mathbf{V}_i^*(u)}},$$
$$N_{C_i} = \mathbb{1}_{\{X_i \leqslant u, \tau_i = 0\}}.$$

– $\Omega_3(\boldsymbol{\beta}) = \Phi\{\hat{A}(\boldsymbol{\beta},u),\mathbf{V}^*(u)\}\left[\Phi\{\mathbf{V}^*(u),\mathbf{V}^*(u)\}\right]^{-1}\Phi\{\mathbf{V}^*(u),\hat{A}(\boldsymbol{\beta},u)\}$, où pour tous $\mathbf{H_1}(u)$ et $\mathbf{H_2}(u)$ des vecteurs de fonctions,

$$\Phi\{\mathbf{H_1}(u),\mathbf{H_2}(u)\} = \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^\infty dN_{C_i}(u)\left(G\{\mathbf{H_1}(u)\left[\mathbf{H_2}(u)\right]^t\} - G\{\mathbf{H_1}(u)\}\left[G\{\mathbf{H_2}(u)\}\right]^t\right)\right].$$

De même, à partir des théorèmes de convergence de Robins [1993], sous les conditions (V.6) et (V.9), la variance asymptotique de la fonction de risques cumulés $\hat{\Lambda}_T(t) = -\ln(\hat{S}_T(t))$, est estimé de manière consistante par $\hat{\Sigma}^*(t) = \Omega_1^*(t) - \Omega_2^*(t) - \Omega_3^*(t)$.

– $\Omega_1^*(t) = \frac{1}{n}\sum_{i=1}^n\left\{\hat{A}_i^*(t,0)\right\}^2$, où

$$\begin{aligned}
\hat{A}_i^*(t,x) &= \int_x^t \frac{d\hat{M}_{T_i}^*(u)\hat{W}_i(u)}{\sum_{j=1}^n Y_j(u)\hat{W}_j(u)},\\
\hat{M}_{T_i}^*(u) &= N_{T_i}(u) - \int_0^u d\hat{\Lambda}(t)Y_i(t),\\
N_{T_i} &= \mathbb{1}_{\{X_i\leqslant u,\tau_i=1\}},\\
d\hat{\Lambda}(t) &= \frac{\sum_{j=1}^n \hat{W}_j(t)dN_{T_j}}{\sum_{j=1}^n Y_j(t)\hat{W}_j(t)}.
\end{aligned}$$

– $\Omega_2^*(t) = \frac{1}{n}\sum_{i=1}^n\left[\int_0^\infty dN_{C_i}(u)\left\{G(\hat{A}^*(t,u)\right\}^2\right]$. Avec la fonction G définit comme précedemment, $N_{C_i} = \mathbb{1}_{\{X_i\leqslant u,\tau_i=0\}}$ et $\hat{A}^* = (\hat{A}_i^*)_{i=1,\dots,n}$.

– $\Omega_3^*(t) = \Phi\{\hat{A}^*(t,u),\mathbf{V}^*(u)\}\left[\Phi\{\mathbf{V}^*(u),\mathbf{V}^*(u)\}\right]^{-1}\Phi\{\mathbf{V}^*(u),\hat{A}^*(t,u)\}$, où Φ est défini précédemment.

D'après les théorèmes précedents, la variance des coefficients de régression est

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \hat{I}^{-1}\hat{\Sigma}(\hat{\boldsymbol{\beta}})\left(\hat{I}^{-1}\right)^T \tag{V.14}$$

et les intervalles de confiance asymptotiques de Wald à 95% pour $\hat{\beta}$ et $\hat{\Lambda}_T(t)$ sont donnés par

$$\begin{aligned}
IC_{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} \pm 1.96\left\{\hat{I}^{-1}\hat{\Sigma}(\hat{\boldsymbol{\beta}})\left(\hat{I}^{-1}\right)^T\right\}^{1/2}\\
IC_\Lambda(t) &= \hat{\Lambda}_T(t) \pm 1.96\left\{\hat{\Sigma}^*(t)\right\}^{1/2}
\end{aligned}$$

avec $\hat{I} = \frac{\partial U(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ et $\hat{\Sigma}(\hat{\boldsymbol{\beta}})$ et $\hat{\Sigma}^*(t)$ sont les estimateurs des variances asymptotiques décrits précédemment.

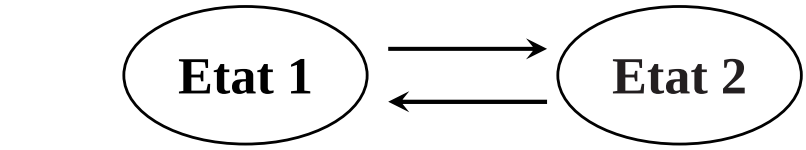
2 Méthode IPCW adaptée au modèle Markovien

2.1 Introduction

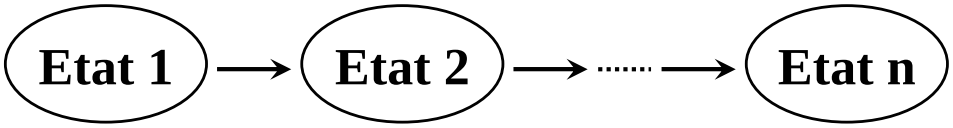
Dans la plupart des études de survie, la censure est supposée indépendante de l'événement étudié même si cela ne correspond pas toujours à la réalité clinique. Cette hypothèse d'indépendance permet d'utiliser les méthodes d'analyse classiques. Cependant, ces estimations sont souvent biaisées du fait de la dépendance entre la censure et l'événement. La méthode IPCW (Inverse Probability Censoring Weight) présentée dans la section précédente, permet de prendre en compte cette dépendance. Cependant, la méthode IPCW est développée pour des données de survie (Rotnitzky et Robins [2003], Scharfstein et al. [2001]) ce qui correspond à un modèle à deux états où il n'y a pas de retour possible (vivant-décès). Pourtant, la censure par perdus de vue dans les modèles multi-états pose les mêmes difficultés et entraine souvent un biais dans les estimations des intensités de transition. L'adaptation de la méthode IPCW aux modèles multi-états permettrait de prendre en compte l'information issue de la censure dans l'étude de ces modèles. La méthode d'estimation semi-paramétrique pour modèles markoviens présentée au chapitre précédent suppose que la censure est indépendante du processus d'événement. Cependant, elle peut être modifiée de manière à prendre en compte un phénomène de censure informative.

L'objet de cette partie est de généraliser la méthode IPCW pour modèles de survie aux modèles Markoviens à deux états avec retour (Figure V.1 **(a)**) et aux modèles Markoviens progressifs (Figure (V.1 **(b)**). La généralisation se restreint pour le moment à ces modèles : en effet, l'adaptation aux modèles non progressifs à plus de deux états entraîne des difficultés supplémentaires. Prenons l'exemple d'un modèle à trois états : « sain », « malade » et « décès » (Figure V.1 **(c)**). La difficulté vient du fait que la pondération associée au phénomène de censure à partir de l'état 'malade' sera identique pour l'estimation de la transition « malade » vers « décès » et pour l'estimation de la transition « malade » vers « sain » : cela suppose que le phénomène de censure informative influence de la même manière l'estimation des deux transitions. Dans les modèles progressifs (Figure V.1 **(b)**) où une seule transition est possible cette difficulté n'apparait pas. Suite à la présentation de la méthodologie IPCW au modèle à deux états avec retour, nous discuterons les perspectives d'adaptation de la méthodologie aux modèles à plus de 2 états.

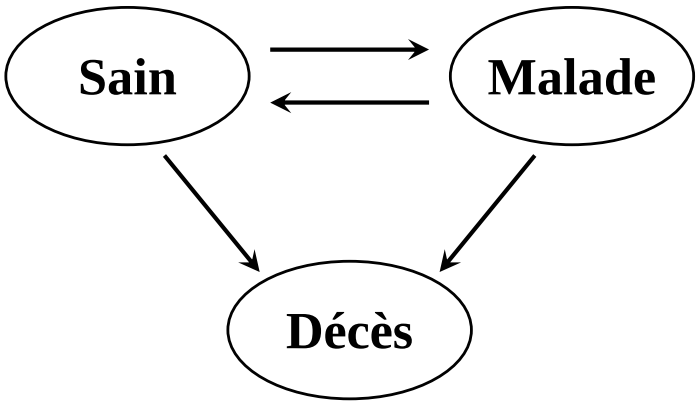
Dans le cas de données de survie, la méthode IPCW adapte l'estimation de Kaplan-Meier et de la vraisemblance partielle à l'hypothèse de censure dépendante. La généralisation de la méthode IPCW aux modèles multi-états s'obtient en adaptant la méthode d'estimation semi-paramétrique pour les modèles multi-états Markoviens présentée au chapitre IV. L'estimation semi-paramétrique (Andersen et al. [1991]) qui est une extension de la méthode non-paramétrique de Aalen et Johansen [1978] à un modèle de régression de Cox, repose sur la théorie des processus de comptage et de la vraisemblance partielle. Ainsi, le principe de pondération des estimations pour prendre en compte le phénomène de censure informative peut être utilisé. L'adaptation de la méthode IPCW se sépare en deux parties : dans un premier temps, la méthode d'estimation semi-paramétrique pour modèle Markovien est utilisée pour estimer les risques de censure à partir de chaque état, dans un deuxième temps, la méthode d'estimation semi-paramétrique est pondérée afin d'estimer les probabilités de transition sous l'hypothèse de censure dépendante.



(a)



(b)



(c)

FIG. V.1 – **(a)** Modèle à deux états avec retour. **(b)** Modèle progressif. **(c)** Modèle à trois états « sain », « malade », « décès ».

2.2 Modèle avec un état de censure

2.2.1 Introduction

Les méthodes d'estimation pour modèles Markoviens qui prennent en compte le phénomène de censure telle que l'estimation non-paramétrique de Aalen-Johansen ou l'estimation semi-paramétrique ne traitent pas du cas de la censure dépendante (Andersen et al. [1993]). Afin de définir la notion de censure indépendante, considérons un modèle de Markov à deux états avec retour, les intensités de transition de l'état h vers l'état j sont

$$\lambda_{hj}(t \mid .) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(X(t + \Delta t) = j \mid X(t) = h, .), \qquad h, j = 1, 2,\ h \neq j.$$

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

$$\lambda_{hj}(t \mid C_h > t) = \lambda_{hj}(t), \qquad h, j = 1, 2,\ h \neq j,$$

où C_h correspond au temps de censure pour un individu dans l'état h. Cette hypothèse sera vérifiée si et seulement si les facteurs qui prédisent le processus d'événement ne prédisent pas le processus de censure. En effet, si on suppose que les covariables déterminent le processus d'événement alors si elles déterminent aussi la censure : le processus d'événement et la censure seront dépendants par l'intermédiaire des covariables. L'objectif est alors d'étudier les risques de censure afin de montrer un lien entre la censure et événement. On considère pour cela un modèle à 2 états de santé (états transients) avec un état absorbant correspondant à ce que nous appellerons l'état « censure ». Le modèle considéré peut être schématisé par la figure V.2 où Etat 1 et Etat 2 représentent les états de santé et Etat C correspond à l'état de censure.

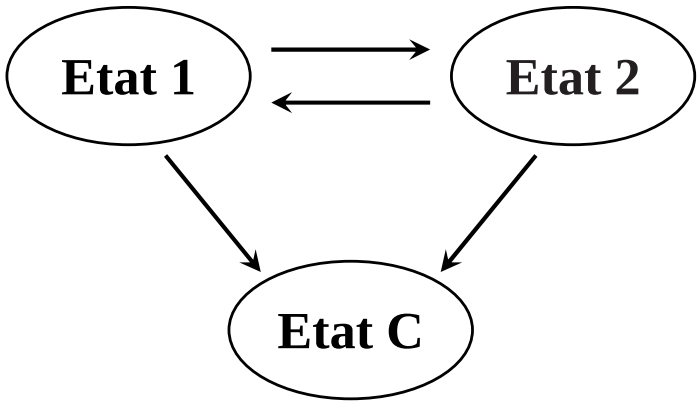


FIG. V.2 – Modèle à deux états de santé et un état absorbant représentant la censure (Etat C).

2.2.2 Hypothèses

L'utilisation de la méthode IPCW pour données de survie nécessitait certaines hypothèses comme le fait de disposer de tous les facteurs de risques de l'événement. Afin d'adapter la méthode à un modèle à deux états, des hypothèses du même ordre doivent être satisfaites.

Soit $\mathbf{V}(t)$ les covariables qui prédisent le processus d'événement au temps t et soit $\bar{\mathbf{V}}(t) = \{\mathbf{V}(x); 0 \leqslant x \leqslant t\}$ l'histoire de ces covariables. L'hypothèse fondamentale de la méthode est que le risque de censure à partir de l'état h au temps t ne dépend plus du possible temps d'événement non observé T_h, où l'événement est une transition à partir de l'état h, ce qui peut aussi s'écrire :

$$\lambda_{hC}(t \mid \bar{\mathbf{V}}(t), T_h, T_h > t) = \lambda_{hC}(t \mid \bar{\mathbf{V}}(t), T_h > t), \qquad h = \{1, 2\}. \tag{V.15}$$

Dans la suite, le concept de variable Coarsened At Random (CAR) introduit précedemment (Heitjan et Rubin [1991]), sera utilisé pour s'assurer de l'hypothèse (V.15). En effet, si on suppose que $(\mathbf{V}(T_h), T_h)$ sont CAR, c'est-à-dire

$$\lambda_{hC}(t \mid \bar{\mathbf{V}}(T_h), T_h, T_h > t) = \lambda_{hC}(t \mid \bar{\mathbf{V}}(t), T_h > t), \qquad h = \{1, 2\}, \tag{V.16}$$

alors l'hypothèse (V.15) sera vérifiée. Cette hypothèse entraîne que les covariables prédisent « parfaitement » le processus d'événement. Par conséquent, si le risque de censure dépend de ces mêmes covariables, c'est-à-dire,

$$\lambda_{hC}(t \mid \bar{\mathbf{V}}(t), T_h > t) \neq \lambda_{hC}(t \mid T_h > t) \qquad h = \{1, 2\},$$

alors la censure (à partir de l'état h) sera dépendante du processus d'événement (par l'intermédiaire des covariables). Cette dépendance peut être testée en vérifiant si, dans ce modèle, certains coefficients de régression associés aux covariables sont statistiquement différents de zéro : si tel est le cas, l'hypothèse de censure indépendante sera fausse. Ainsi, le recours à la méthode d'estimation IPCW pour corriger les biais dus à cette dépendance sera justifié.

2.2.3 Estimations des risques de censure

Le but est d'estimer les risques de censure à partir des deux états de santé. On rappelle que les risques de censure dépendent des mêmes covariables qui prédisent les transitions entre les deux états de santé. Sous les hypothèses présentées précédemment, on peut utiliser l'estimation semi-paramétrique pour estimer les risques de censure dans le modèle (V.2).

Pour chaque individu, $i = 1, ..., n$, on observe un processus de Markov $\{X_i(t), t \in \mathcal{T}\}$ à espace d'états $\{1, 2, C\}$, $\mathcal{T} = [0, \tau]$, $0 < \tau \leqslant +\infty$. Les intensités de transition de l'état h vers l'état j sont définies de la façon suivante :

$$\lambda_{hj}(t \mid \bar{\mathbf{V}}(t)) = \lambda^0_{hj}(t) \exp\left(\boldsymbol{\alpha}^T_{hj} \mathbf{V}(t)\right), \qquad (h, j) \in S,$$

avec $S = \{(1, 2), (2, 1), (1, C), (2, C)\}$, $\lambda^0_{hj}(t)$ l'intensité de transition de base associée à la transition de l'état h vers l'état j, $\boldsymbol{\alpha}_{hj}$ le vecteur des coefficients de régression associé à la transition de h vers j, $\mathbf{V}_i(t)$ le vecteur des covariables dépendantes du temps associées à l'individu i et $\bar{\mathbf{V}}_i(t) = \{\mathbf{V}_i(u); 0 \leqslant u \leqslant t\}$ l'histoire de ces covariables. On suppose que $(\mathbf{V}(T_{hi}), T_{hi})$ sont CAR (V.16), où T_{hi} sont des temps de transition à partir de l'état h pour l'individu i. Pour chaque individu i, on peut associer un processus de comptage $N_{hji}(t) = \#\{\text{transitions observées de } h \to j \text{ dans } [0, t] \text{ pour l'individu } i\}$, pour tous $(h, j) \in S$. De plus, on définit $Y_{hi}(t) = \mathbb{1}_{\{X_i(t-)=h\}}$ qui vaut 1 si l'individu i est à risque dans l'état h $(h = 1, 2)$ au temps t^- et 0 sinon.

Les estimateurs des coefficients de régression s'obtiennent en maximisant la vraisemblance partielle de Cox suivante

$$V = \prod_{t\in\mathcal{T}}\prod_{i=1}^{n}\prod_{(h,j)\in S}\left(\frac{\exp\left(\boldsymbol{\alpha}_{hj}^{T}\mathbf{V}_i(t)\right)}{\sum_{k=1}^{n}Y_{hk}(t)\exp\left(\boldsymbol{\alpha}_{hj}^{T}\mathbf{V}_k(t)\right)}\right)^{\Delta N_{hji}(t)}$$

De plus, le risque cumulé de base $A_{hj}^{0}(t) = \int_0^t \lambda_{hj}^{0}(u)du$ peut être estimé par un estimateur du type Nelson-Aalen

$$\hat{A}_{hj}^{0}(t \mid \hat{\boldsymbol{\alpha}}) = \int_0^t \frac{J_h(u)}{\sum_{i=1}^{n}Y_{hi}(u)\exp(\hat{\boldsymbol{\alpha}}_{hj}^{T}\mathbf{V}_i(u))}dN_{hj.}(u), \qquad (h,j)\in S.$$

où $\hat{\boldsymbol{\alpha}}_{hj}$ est l'estimateur du vecteur des coefficients de régression, $N_{hj.}(t) = \sum_{i=1}^{n}N_{hji}(t)$ compte le nombre de transitions observées de l'état h vers l'état j dans l'intervalle $[0,t]$ dans toute la population, $Y_h(t) = \sum_{i=1}^{n}Y_{hi}(t)$ compte le nombre de personnes à risque de subir une transition à partir de l'état h juste avant le temps t et $J_h(t) = \mathbb{1}_{\{Y_h(t)>0\}}$, $(h = 1, 2)$.

2.2.4 Estimation des probabilités de censure

Un estimateur de la matrice $\mathbf{P} = \{p_{hj}\}$ des probabilités de transition est donné par le produit intégral :

$$\hat{\mathbf{P}}(t,t') = \prod_{]t,t']}(\mathbf{I} + d\hat{\mathbf{A}}(u)),$$

où $\hat{\mathbf{A}} = \{\hat{A}_{hj}\}$ est l'estimateur de la matrice des intensités de transition et $\mathbf{I}$ la matrice identité $(s\times s)$. En pratique, on observe un nombre fini d'événements : soient $0 < T_1 < \cdots < T_l < \tau$ les temps d'événements. Ainsi $\hat{\mathbf{A}}$ est une fonction en escalier et on peut écrire

$$\hat{\mathbf{P}}(t,t' \mid \bar{\mathbf{V}}_i(t')) = \prod_{\{k;T_k\in]t,t']\}}(\mathbf{I} + \Delta\hat{\mathbf{A}}(T_k \mid \hat{\boldsymbol{\alpha}}, \bar{\mathbf{V}}_i(T_k))),$$

avec

$$\Delta\hat{A}_{hj}(T_k \mid \hat{\boldsymbol{\alpha}}, \bar{\mathbf{V}}_i(T_k))) = \frac{J_h(T_k)\times\exp(\hat{\boldsymbol{\alpha}}_{hj}^{T}\mathbf{V}_i(T_k))}{\sum_{i=1}^{n}Y_{hi}(T_k)\exp(\hat{\boldsymbol{\alpha}}_{hj}^{T}\mathbf{V}_i(T_k))}\times(N_{hj.}(T_k) - N_{hj.}(T_{k-1}))\,,$$

pour tous $(h,j)\in S$. De plus, on a $\Delta\hat{A}_{C1} = \Delta\hat{A}_{C2} = 0$ et $\Delta\hat{A}_{hh} = -\sum_{j\neq h}\Delta\hat{A}_{hj}$. On obtient ainsi une estimation de la matrice des probabilités de transition en fonction des covariables qui va permettre de définir les poids associés à chaque individu.

2.2.5 Calcul des poids

Dans la méthode IPCW pour modèles de survie, le poids pour l'individu i au temps t est défini à partir de la probabilité $K_i(t)$, qui est la probabilité que le sujet i soit non censuré jusqu'au temps t sachant $\bar{\mathbf{V}}_i(.)$ (« survie de la censure » $= P(C_i > t \mid \bar{\mathbf{V}}_i(.))$). Par contre, dans l'estimation d'un modèle à deux états, il y a plusieurs probabilités d'être non censuré au temps t : une pour chaque temps de départ dans $[0, t[$. Soit $1 - \hat{p}_{hC}(s, t \mid \bar{\mathbf{V}}_i(.))$, la probabilité que le sujet i soit non censuré au temps t sachant qu'il était dans l'état h au temps s ($h = 1, 2$). Pour pouvoir utiliser la méthode IPCW, il faut sélectionner une seule probabilité pour chaque temps t. Un choix naturel est la probabilité d'être non censuré au temps t sachant qu'il était dans l'état h au temps 0 : $1 - \hat{p}_{hC}(0, t \mid \bar{\mathbf{V}}_i(.))$. Ce choix est arbitraire mais il semble cependant le mieux adapté.

Soit $\hat{K}_{hi}(t) = 1 - \hat{p}_{hC}(0, t \mid \bar{\mathbf{V}}_i(.)) =$ la probabilité pour l'individu i de ne pas être censuré au temps t sachant qu'il est dans l'état h au temps 0. Soit $\hat{K}_h^0(t) = 1 - \hat{p}_{hC}(0, t)$ la même probabilité dans un modèle sans covariable. Cette probabilité ne dépend pas de l'indice des individu car elle est identique pour tous les individus. Les poids pour chaque individu sont définis comme le rapport de ces deux probabilités :

$$\hat{W}_{hi}(t) = \frac{\hat{K}_h^0(t)}{\hat{K}_{hi}(t)}. \tag{V.17}$$

Les poids sont égaux à un si les covariables ne prédisent pas le risque de censure, ce qui correspond à une censure qui n'est pas dépendante du processus d'événement (censure non informative). Et inversement, si les poids diffèrent de un, la censure sera dépendante.

2.3 Extension possible

Il est possible de considérer un modèle mieux adapté aux observations cliniques en prenant en compte l'information a priori sur les différents phénomènes de censure. En effet, le modèle précédent ne fait aucune différence entre les patients perdus de vue et les patients exclus vivants (censurés à la date de point). Pourtant, il semble évident qu'il y ait une différence entre ces deux phénomènes de censure. En effet, la censure engendrée par les patients perdus de vue est potentiellement informative alors que la censure engendrée par le gel de la base n'apporte a priori aucune information sur la survenue de l'événement étudié.

La méthodologie présentée précédemment peut aisément être adaptée pour étudier uniquement le risque de la censure générée par les patients perdus de vue. Pour cela, on peut considérer le modèle de la figure V.3 qui comprend deux états de censure : un pour la censure par perdu de vue (état C_1) et un pour la censure engendrée par l'arrêt de l'étude (état C_2).

L'ensemble des transitions possibles est alors $S = \{(1,2), (2,1), (1,C_1), (2,C_1), (1,C_2), (2,C_2)\}$. En considérant les processus de comptage $N_{hji}(.)$ correspondant pour tout $(h,j) \in S$, les estimations des risques de transition s'obtiennent comme précédemment. Ainsi, en reprenant la méthodologie précédente, en remplaçant $N_{hC}(.)$ par $N_{hC_1}(.)$, le risque de censure étudié est un risque de censure par perdu de vue ou encore un « risque de censure informative ». Ce risque est estimé en supposant a priori que la censure générée par les exclus vivants est non informative. De cette façon, les pondérations calculées modifieront les contributions

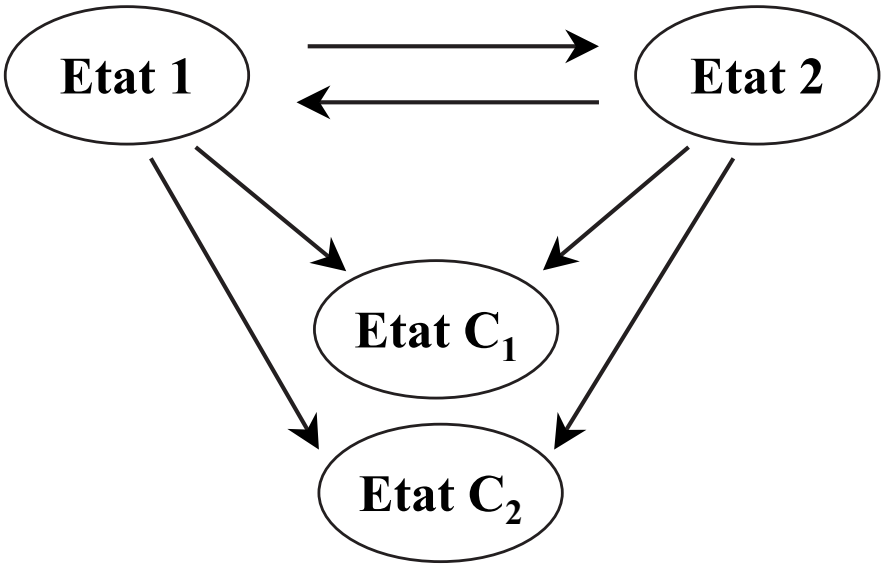


FIG. V.3 – Modèle à deux états de santé et à deux états absorbants représentant la censure. L'état C_1 représente la censure par perdu de vue et l'état C_2 représente la censure engendrée par les exclus vivants.

des individus dans les estimations, en prenant compte uniquement les personnes perdues de vue. Ce modèle semble mieux adapté dans le sens où il utilise l'information a priori sur le type de censure.

2.4 Modèle non-paramétrique IPCW

La nature dépendante de la censure entraîne que l'estimation semi-paramétrique dans les modèles Markoviens est biaisée et n'est plus justifiée. Cependant, afin de corriger le biais engendré par la dépendance, on peut modifier l'estimateur « classique » en utilisant les pondérations calculées précédemment. La méthode IPCW consiste en quelque sorte à modifier le nombre de personnes à risque et le nombre d'événements dans les estimateurs de façon à prendre en compte les personnes qui ne sont plus dans l'étude pour cause de censure. Afin de décrire la méthode d'estimation IPCW pour des modèles Markoviens, on considèrera dans ce qui suit le modèle à deux états de santé de la figure V.4.

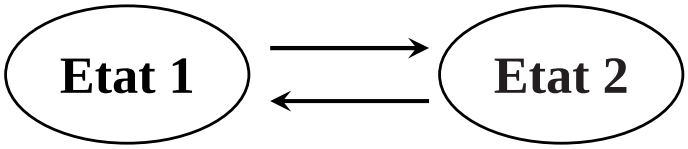


FIG. V.4 – Modèle à deux états de santé avec retour.

Afin d'adapter la méthode IPCW au modèle à deux états avec retour, il faut supposer

que les temps d'événement sont tous différents. Cette hypothèse nécessaire pour utiliser les méthodes d'estimation non-paramétriques (*cf.* chapitre IV page 74), permet d'avoir un seul poids à prendre en compte à chaque temps d'événement. Notons qu'il est également nécessaire de faire cette modification des temps pour l'estimation des risques de censure et le calcul des poids dans la section précédente.

Soient $0 < T_1 < \cdots < T_l < \tau$ les temps d'événement (tous différents), on peut créer un vecteur $\hat{W}_h$ $(h = 1, 2)$ tel que pour $k = 1, ..., l$:

$$\begin{aligned}\hat{W}_h(T_k) &= \hat{W}_{hi}(T_k) \text{ si l'individu } i \text{ subit un événement au temps } T_k \\ &= 0 \text{ si aucun individu ne subit d'événement,}\end{aligned}$$

où l'événement est une transition à partir de l'état h. On considère un modèle sans covariable où les intensités de transition sont

$$\lambda_{hj}(t) = \lambda^0_{hj}(t), \qquad (h, j) \in \{(1, 2), (2, 1)\}.$$

L'estimateur IPCW de $A_{hj}(t) = \displaystyle\int_0^t \lambda^0_{hj}(u)du$ s'obtient par pondération de l'estimateur de Nelson-Aalen suivant

$$\hat{A}_{hj}(t) = \int_0^t \frac{J_h(u)}{\sum_{i=1}^n Y_{hi}(u)} dN_{hj.}(u).$$

En suivant la démarche de la construction de l'estimateur IPCW pour études de survie, on obtient l'estimateur IPCW des intensités cumulées $A_{hj}(t)$:

$$\hat{A}_{hj}(t \mid \hat{W}_h) = \int_0^t \frac{J_h(u)\hat{W}_h(u)}{\sum_{i=1}^n Y_{hi}(u)\hat{W}_{hi}(u)} dN_{hj.}(u).$$

On peut ainsi déduire un estimateur IPCW de la matrice des probabilités de transition $\mathbf{P} = \{p_{hj}\}$:

$$\hat{\mathbf{P}}(t, t') = \prod_{u \in]t, t']} (\mathbf{I} + d\hat{\mathbf{A}}(u \mid \hat{W}_h(u))).$$

Remarque 24 *On peut noter que la méthode d'estimation IPCW reste stable dans le cas où la censure n'est pas dépendante du processus d'événement. En effet, les poids sont égaux à un et on retrouve les estimations sans pondération (estimation semi-paramétrique « classique »).*

Remarque 25 *Dans le calcul des poids, il a été nécessaire de choisir une des probabilités de censure au temps t. Le choix arbitraire mais naturel a été $\hat{p}_{hC}(0, t)$ qui correspond à la probabilité d'être censuré au temps t sachant un état h au temps 0. Par conséquent, pour que les résultats soient les plus corrects possibles, il faut interpréter les probabilités de transition $\hat{p}_{hj}(0, t)$. Cela ne pose pas vraiment de problème en pratique, en effet, ce sont souvent ces probabilités qui sont interprétées.*

2.5 Modèle semi-paramétrique IPCW

Nous avons vu dans la section précédente, comment obtenir l'estimation de la survie IPCW dans un modèle sans covariable. La méthode IPCW permet aussi d'ajuster un modèle semi-paramétrique de Cox et d'estimer les coefficients de régression dans un contexte de censure dépendante. On peut considèrer le modèle suivant

$$\lambda_{hj}(t) = \lambda^0_{hj}(t) \times \exp(\boldsymbol{\beta}_{hj}\mathbf{Z}(t)), \qquad (h,j) \in \{(1,2),(2,1)\}.$$

2.5.1 Estimation des coefficients de régression

Afin d'estimer les coefficients de régression $\boldsymbol{\beta}_{hj}$ tout en prenant compte de la censure dépendante, il faut modifier la fonction score de la vraisemblance partielle. En s'inspirant de la méthode IPCW pour la survie, on peut obtenir une fonction score modifiée qui se différencie par le fait que la contribution de l'individu i pour la transition à partir de l'état h au temps T_k est pondéré par $\hat{W}_{hi}(T_k)$. Un estimateur de $\boldsymbol{\beta}_{hj}$, $(h,j) \in \{(1,2),(2,1)\}$ est ainsi obtenu en annulant la fonction score IPCW suivante :

$$U_{hj}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau \left[\mathbf{Z}_i(t) - \frac{\sum_{i=1}^n Y_{hi}(t) Z_i(t) e^{\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t)} \hat{W}_{hi}(t)}{\sum_{i=1}^n Y_{hi}(t) e^{\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t)} \hat{W}_{hi}(t)} \right] \times \hat{W}_{hi}(t) dN_{hji}(t). \qquad \text{(V.18)}$$

On note que si les poids sont égaux à un, on retrouve la fonction score de la vraisemblance partielle de la méthode d'estimation traditionnelle.

Preuve *La construction de la fonction score IPCW pour le modèle à deux états suit la même démarche de construction que le score IPCW pour la survie. La vraisemblance partielle est donnée par l'expression suivante :*

$$V = \prod_t \prod_{i=1}^n \prod_{h,j} \left(\frac{\exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t))}{\sum_{i=1}^n Y_{hi}(t) \exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t))} \right)^{\Delta N_{hji}(t)},$$

La log-vraisemblance est :

$$LogV = \sum_{h,j} \sum_{i=1}^n \int_0^\tau \left(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t) - \log \left(\sum_{i=1}^n Y_{hi}(t) \exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t)) \right) \right) dN_{hji}(t).$$

La dérivée de la log-vraisemblance par rapport à β_{hj} est donnée par :

$$\frac{\partial LogV}{\partial \boldsymbol{\beta}_{hj}} = \sum_{i=1}^n \int_0^\tau \left[\mathbf{Z}_i(t) - \frac{\sum_{i=1}^n Y_{hi}(t) \mathbf{Z}_i(t) \exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t))}{\sum_{i=1}^n Y_{hi}(t) \exp(\boldsymbol{\beta}_{hj}^T \mathbf{Z}_i(t))} \right] dN_{hji}(t),$$

On peut ensuite définir $U^0_{hj}(\boldsymbol{\beta})$:

$$\begin{aligned} U^0_{hj}(\boldsymbol{\beta}) &= \frac{\partial LogV}{\partial \boldsymbol{\beta}_{hj}} \\ &= \sum_{i=1}^n \int_0^\tau \left[\mathbf{Z}_i(t) - E^0_{hj}(\boldsymbol{\beta}, t) \right] \times \frac{\hat{K}^0_h(t)}{\hat{K}^0_h(t)} dN_{hji}(t), \end{aligned}$$

avec

$$E^0_{hj}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^n Y_{hi}(t)\mathbf{Z}_i(t)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i(t)}\frac{\hat{K}^0_h(t)}{\hat{K}^0_h(t)}}{\sum_{i=1}^n Y_{hi}(t)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i(t)}\frac{\hat{K}^0_h(t)}{\hat{K}^0_h(t)}}.$$

De manière similaire, on peut définir $E^i_{hj}(\boldsymbol{\beta})$ *et* $U_{hj}(\boldsymbol{\beta})$ *en remplaçant au dénominateur* $\hat{K}^0_h(t)$ *par* $\hat{K}_{hi}(t)$ *:*

$$E^i_{hj}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^n Y_{hi}(t)\mathbf{Z}_i(t)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i(t)}\frac{\hat{K}^0_h(t)}{\hat{K}_{hi}(t)}}{\sum_{i=1}^n Y_{hi}(t)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i(t)}\frac{\hat{K}^0_h(t)}{\hat{K}_{hi}(t)}},$$

$$U_{hj}(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \left[\mathbf{Z}_i(t) - E^i_{hj}(\boldsymbol{\beta},t)\right] \times \frac{\hat{K}^0_h(t)}{\hat{K}_{hi}(t)} dN_{hji}(t).$$

La fonction $U_{hj}(\boldsymbol{\beta},t)$ *est appelée fonction score modifié. Avec* $\hat{W}_{hi}(t) = \frac{\hat{K}^0_h(t)}{\hat{K}_{hi}(t)}$ *on a*

$$U_{hj}(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\tau \left[\mathbf{Z}_i(t) - \frac{\sum_{i=1}^n Y_{hi}(t)Z_i(t)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i(t)}\hat{W}_{hi}(t)}{\sum_{i=1}^n Y_{hi}(t)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_i(t)}\hat{W}_{hi}(t)}\right] \times \hat{W}_{hi}(t)dN_{hji}(t).$$

Un estimateur de $\boldsymbol{\beta}_{hj}$ *est la solution de l'équation* $U_{hj}(\boldsymbol{\beta}) = 0$*.*

■

En pratique, le nombre d'événements est fini et la fonction $N_{hji}(t)$ est une fonction en escalier. Soit $0 < T_1 < \cdots < T_l < \tau$ les temps d'événements, $U_{hj}(\boldsymbol{\beta})$ peut ainsi s'écrire :

$$U_{hj}(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{\{k;T_k\in[0,\tau[\}} \left[\mathbf{Z}_i(T_k) - \frac{\sum_{i=1}^n Y_{hi}(T_k)Z_i(T_k)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_{i(T_k)}}\hat{W}_{hi}(T_k)}{\sum_{i=1}^n Y_{hi}(T_k)e^{\boldsymbol{\beta}_{hj}^T\mathbf{Z}_{i(T_k)}}\hat{W}_{hi}(T_k)}\right] \times \hat{W}_{hi}(T_k) \times \left(N_{hji}(T_k) - N_{hji}(T_{k-1})\right).$$

2.5.2 Estimation des probabilités de transition

Une estimation des probabilités de transition en fonction des covariables s'obtient à partir de l'estimateur des intensités cumulées. En suivant une démarche identique au cas non-paramétrique, $A_{hj}(t) = \int_0^t \lambda_{hj}(u)du$ est estimé par :

$$\hat{A}_{hj}(t \mid \hat{W}_h(.), \mathbf{Z}_i(.)) = \int_0^t \frac{J_h(u)\hat{W}_h(u)}{\sum_{i=1}^n Y_{hi}(u)\hat{W}_{hi}(u)\exp(\hat{\boldsymbol{\beta}}_{hj}^T\mathbf{Z}_i(u))} \exp(\hat{\boldsymbol{\beta}}_{hj}^T\mathbf{Z}_i(u))dN_{hj.}(u),$$

où $\hat{\boldsymbol{\beta}}_{hj}$ est l'estimateur défini dans la section précédente. Soit $0 < T_1 < \cdots < T_l < \tau$ les temps d'événement, on peut calculer les valeurs de l'estimateur $\Delta\hat{A}_{hj}$ pour chaque temps

T_k qui sont données par l'écriture suivante :

$$\Delta \hat{A}_{hj}(T_k \mid \hat{W}_h(T_k), \mathbf{Z}_i(T_k)) = \frac{J_h(T_k)\hat{W}_h(T_k)}{\sum_{i=1}^{n} Y_{hi}(T_k)\hat{W}_{hi}(T_k)\exp(\hat{\boldsymbol{\beta}}_{hj}^T \mathbf{Z}_i(T_k))} \times \exp(\hat{\boldsymbol{\beta}}_{hj}^T \mathbf{Z}_i(T_k)) \times (N_{hj.}(T_k) - N_{hj.}(T_{k-1}))$$

A l'aide du produit intégral, on déduit un estimateur de la matrice des probabilités de transition $\mathbf{P} = \{p_{hj}\}$ pour les valeurs de la covariable $\mathbf{Z}(.)$:

$$\hat{\mathbf{P}}(t, t' \mid \hat{W}_h(.), \mathbf{Z}(.)) = \prod_{T_k \in]t,t']} (\mathbf{I} + \Delta \hat{\mathbf{A}}(T_k \mid \hat{W}_h(T_k), \mathbf{Z}(T_k))).$$

3 Application à l'asthme

Dans cette section, on s'intéresse à l'application de la méthode IPCW à la base de données de patients asthmatiques. Dans cette base les patients perdus de vue sont très nombreux et après discussion avec les pneumologues, il semble que le phénomène de censure soit informatif. En effet, pour cette maladie, il semble que les perdus de vue soient essentiellement des patients qui ont un asthme bien contrôlé et qui ne ressentent pas le besoin d'être suivis régulièrement. L'objectif est de comparer les méthodes traditionnelles avec les versions IPCW afin d'observer l'impact de l'hypothèse de censure dépendante sur la modélisation. Ces résultats seront ensuite discutés et interprétés d'un point de vue clinique.

Dans une première partie, la méthode IPCW sera appliquée à des données de survie sur l'asthme et dans une deuxième partie l'extension de la méthode IPCW sera utilisée dans le cas d'un modèle de Markov à deux états.

3.1 Application à des données de survie

L'objet de cette partie est de présenter l'application de la méthode IPCW à des données de survie sur l'asthme.

3.1.1 Définition du modèle

La base de données sur l'asthme présentée au chapitre I (*cf.* page 9) est utilisée afin d'étudier des données de survie. Dans le cas de l'asthme, l'événement d'intérêt pour les cliniciens est « le passage dans un état de contrôle inacceptable ». Le modèle est représenté par la figure V.5.

Afin d'obtenir des données de survie à partir de la base, nous avons sélectionné pour chaque patient, la « séquence » la plus longue d'au moins deux consultations consécutives commençant par un état de contrôle acceptable (contrôle optimal et sous-optimal) et se terminant soit par la fin du suivi soit par un état de contrôle inacceptable. Après cette sélection, la base comprend 334 patients et un total de 777 consultations. Les quelques temps d'événements ex-æquos ont été légèrement modifiés. Trois types de suivi sont rencontrés :

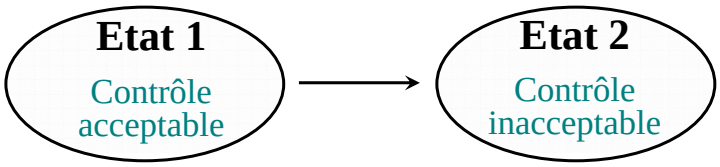


FIG. V.5 – Modèle de survie pour l'asthme.

- Les patients qui subissent l'événement, c'est-à-dire des patients qui passent dans l'état de contrôle inacceptable ($24,8\%$; $n = 83$).
- Les patients considérés comme perdus de vues ($68,3\%$; $n = 228$). Ces patients ne subissent pas l'événement et n'ont pas eu de consultation 3 mois avant la date de fin d'étude.
- Les patients considérés comme vivants à la date de point (6.9%; $n = 23$). Ces patients ne subissent pas l'événement et sont allés consulter 3 mois avant la date de fin d'étude.

Pour modéliser le risque de passer dans un état inacceptable, nous utiliserons les covariables suivantes.

- L'indice de masse corporelle à chaque consultation : codée 0 si IMC < 25, 1 sinon.
- La sévérité à chaque consultation : codée 0 si le patient est non sévère, 1 sinon.
- La dose de corticoïdes inhalés à chaque consultation : codée 0 si la dose est inférieure à 500 μg, 1 sinon.
- La dose de corticoïdes oraux à chaque consultation : codée 0 si la dose est égale à 0 mg, 1 sinon.
- Le nombre d'exacerbations à chaque consultation : codée 0 si le patient n'a aucune exacerbations entre deux consultations, 1 sinon.
- La dose cumulée de corticoïdes oraux pendant l'année avant l'inclusion : codée 0 si le patient a une dose cumulée inférieure ou égale à 2 grammes, 1 sinon.

Cette étude se déroulera en plusieurs étapes. Tout d'abord, les facteurs prédictifs de l'événement d'intérêt (passage dans l'état l'inacceptable) seront sélectionnés à l'aide d'un modèle de Cox. Cette étape permettra de choisir les variables à utiliser dans la modélisation du risque de censure.

La seconde étape consistera à estimer le risque ainsi que la survie de la censure associée. Ceci permettra de calculer les pondérations, d'observer l'impact des covariables sur le phénomène de censure et de mettre en évidence la dépendance entre censure et événement.

Puis, la survie de l'événement « passage dans l'état inacceptable » sera estimée par la méthode IPCW et sera comparée aux résultats obtenus par Kaplan-Meier. Pour finir, l'impact des différentes covariables sur la survie pourra être étudié en estimant les coefficients de regression à l'aide de la fonction score modifiée. Nous comparerons ces résultats avec ceux obtenus par un modèle de Cox classique.

3.1.2 Modèle de Cox pour l'événement

Dans un premier temps, l'événement « passage dans l'état inacceptable » est étudié à l'aide d'un modèle de Cox afin de déterminer les facteurs de risque de l'événement. On

considére dans un premier temps un modèle univarié pour chacune des covariables,

$$\lambda_T(t) = \lambda_0(t) \exp(\beta^T V(t)).$$

Le tableau V.1 présente les résultats des estimations des coefficients de régression, les écarts-types et les p-value du test de Wald pour le test de $H_0 : \beta = 0$.

Covariable	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	0.389	(0.220)	(0.08)
Sévérité	0.589	(0.223)	(<0.01)
Exacerbations	-0.312	(0.267)	(0.24)
Corticoïdes Inhalés	0.431	(0.162)	(<0.01)
Corticoïdes Oraux	0.437	(0.129)	(<0.01)
Antécédents Corticoïdes	0.759	(0.248)	(<0.01)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

Tab. V.1 – Modèle de Cox univarié pour l'événement.

Afin de définir un modèle qui ajuste au mieux le risque d'événement, toutes les covariables significativement différentes de 0 (sévérité de l'asthme, corticoïdes inhalés, oraux, antécédents de traitements par corticoïdes oraux) sont incluses dans une analyse multivariée. Les estimations pour les modèles à trois et quatre covariables sont données dans le tableau V.2. Dans le modéle avec les quatre covariables, l'effet de la sévérité n'est plus significatif ce qui n'est pas vraiment surprenant de part la définition de la sévérité de l'asthme. Finalement le modèle avec trois covariables sera retenu pour modéliser le risque de passer dans un état de contrôle inacceptable.

Covariable	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
Sévérité	0.024	(0.289)	(0.93)		——	
Corticoïdes Inhalés	0.317	(0.169)	(0.06)	0.320	(0.167)	(0.05)
Corticoïdes Oraux	0.347	(0.137)	(0.01)	0.350	(0.132)	(<0.01)
Antécédents Corticoïdes	0.563	(0.304)	(0.06)	0.577	(0.252)	(0.02)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

Tab. V.2 – Modèle de Cox multivarié pour l'événement.

3.1.3 Risque et survie de la censure

La méthode nécessite d'étudier le phénomène de censure en fonction des covariables qui prédisent le processus d'événement. En effet, si les covariables qui prédisent l'événement

prédisent aussi la censure, alors la censure et l'événement seront dépendants par l'intermédiaire de ces covariables. Un modèle de Cox est alors utilisé pour étudier le risque censure avec les covariables qui prédisent l'événement dans l'analyse multivariée. A partir de ce modèle et des estimations de la survie de la censure, les poids nécessaires à l'utilisation de la méthode IPCW seront estimés et interprétés. Dans ce qui suit, aucune distinction n'a été faite entre les différents types de censure, tous les individus qui ne subissent pas l'événement sont considérés comme censurés (méthode IPCW de Robins [1993]).

Le tableau V.3 présente les résultats des estimations des coefficients de régression, les écarts-types et les p-value avec le test de Wald. Parmi les trois covariables qui influencent le risque d'événement, les corticoïdes oraux et les antécédents de corticoïdes oraux influencent aussi de manière significative le risque de censure. Ces deux covariables ont un effet significatif sur le risque de censure et sur le risque d'événement. Ainsi il y a une dépendance entre le processus de censure et le processus d'événement par l'intermédiaire de ces variables. Ce résultat montre bien l'utilité de la méthode IPCW pour essayer de réduire les biais (dus à cette dépendance) présents dans les méthodes d'estimations traditionnelles. Si aucune des variables n'était significative à la fois dans la modélisation du risque de censure et de l'événement, l'utilisation de la méthode IPCW n'aurait pas été justifiée (car les résultats auraient été ceux de la méthode de Kaplan-Meier).

Covariable	$\hat{\beta}$	$(ec)^1$	$(p)^2$
Corticoïdes Inhalés	-0.094	(0.070)	(0.18)
Corticoïdes Oraux	-0.670	(0.192)	(<0.01)
Antécédents Corticoïdes	-0.524	(0.256)	(0.04)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

Tab. V.3 – Modèle de Cox multivarié pour la censure.

En ce qui concerne l'effet de ces covariables, les patients avec un traitement par corticoïdes oraux et les patients avec des antécédents de corticoïdes oraux ont un risque plus faible d'être censuré. Autrement dit, les facteurs qui augmentent le risque de passage dans l'état inacceptable diminuent le risque de censure. De plus, il est particulièrement intéressant d'étudier l'impact des covariables sur le risque de censure. En effet, dans l'analyse univariée, il ressort que mis à part les deux covariables précédentes, un asthme sévère et la présence d'exacerbations diminue de manière significative le risque de censure. Tous ces résultats sur le risque de censure tendent à corroborer l'intuition clinique selon laquelle les patients qui se portent bien ont plus de chance d'être censurés.

Afin d'observer de manière graphique l'impact des covariables qui prédisent l'événement sur le processus de censure, on peut tracer les risques de censure (Figure V.6) dans les trois cas suivants :

– Le risque de censure sans covariable. L'estimation de ce risque correspond au rapport du nombre de personnes qui subissent la censure et du nombre de personnes à risque.

– Le risque de censure pour un individu théorique qui aurait ses covariables égales à 0 en tout temps t est le suivant : $\lambda_C(t \mid \bar{\mathbf{V}}^*(t), T > t)$ où $\bar{\mathbf{V}}^*(t)$ est le vecteur nul de dimension (3×1). Les valeurs des covariables restent constantes au cours du suivi.
– Le risque de censure pour un individu théorique qui aurait ses covariables égales à 1 en tout temps t est le suivant : $\lambda_C(t \mid \bar{\mathbf{V}}^*(t), T > t)$ où $\bar{\mathbf{V}}^*(t)$ est le vecteur unité de dimension (3×1).

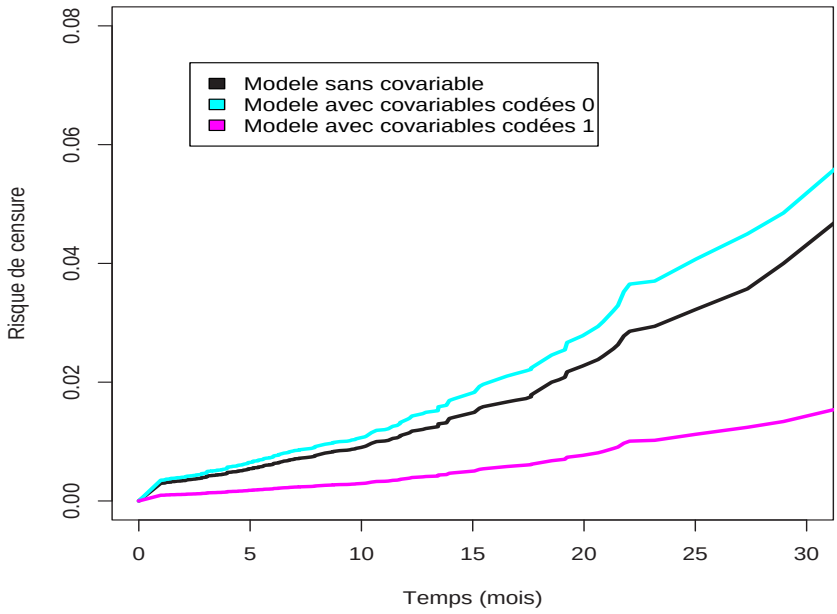


FIG. V.6 – Estimations du risque de censure dans un modèle sans covariable, un modèle avec covariables codées 0 et un modèle avec covariables codées 1.

Graphiquement, il apparaît également que ce sont les patients qui vont le mieux (*i.e* n'ont pas d'antécédents de corticoïdes et n'ont pas de traitement corticoïdes oraux ou inhalés) qui sont le plus à risque de censure. A l'inverse, les patients avec les covariables codées par 1 ont un risque de censure plus faible. En gardant le même raisonnement, nous avons tracé sur la figure V.7 la probabilité de rester non censuré jusqu'au temps t (ou survie de la censure), pour les deux mêmes individus théoriques et le modèle sans covariable. Notons que le modèle sans covariable correspond à l'estimation de la survie de la censure par l'estimateur de Kaplan-Meier. Une fois encore, la probabilité de rester non censuré jusqu'au temps t est la plus élevée chez les individus qui prennent un traitement et qui avaient des antécédents (modèle avec covariables codées 1).

Les figures V.6 et V.7 montrent des écarts importants dans les estimations suivant les valeurs des covariables. Ainsi ce sont les patients les moins à risque de subir l'événement « passage dans l'état inacceptable » qui ont le plus de chance d'être censuré.

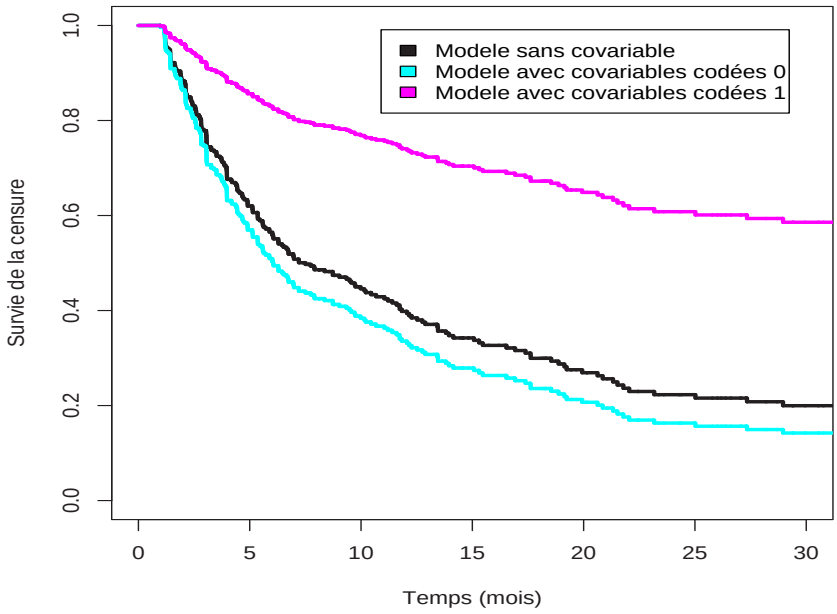


FIG. V.7 – Estimations de la survie de censure dans un modèle sans covariable, un modèle avec covariables codées 0 et un modèle avec covariables codées 1.

3.1.4 Estimation de la survie

A partir des estimations de la survie de la censure par l'estimateur de Kaplan-Meier (où la survie est identique pour tous les individus, $\hat{K}^0(t)$) et par un modèle semi-paramétrique de Cox (la survie est spécifique à chaque individu puisqu'elle prend en compte les covariables, $\hat{K}_i(t)$), il est possible de définir les poids spécifiques à chaque individu :

$$\hat{W}_i(t) = \frac{\hat{K}^0(t)}{\hat{K}_i(t)}$$

Si la censure est indépendante $\hat{K}_i(.)$ va converger vers $\hat{K}^0(.)$ et les poids des individus convergeront vers 1. Dans notre cas, si par exemple, le patient a toutes ses covariables codées par 0 alors la survie (de la censure) sera inférieure à celle estimée par Kaplan-Meier (Figure V.7). Le poids associé sera supérieur à 1 et par conséquent la contribution de cet individu sera augmenter. Sur la figure V.8, l'estimation IPCW de la survie de l'événement « passage dans l'état inacceptable » est comparée à l'estimation de Kaplan-Meier.

La survie IPCW est supérieure à l'estimation de Kaplan-Meier. Ce résultat signifie que l'information issue de la censure améliore la survie. Ce résultat est cohérent puisque les cliniciens estiment que les patients les plus censurés sont ceux qui se portent le mieux. Les pondérations modifient les estimations en prenant en compte l'information des patients censurés qui souvent se portent bien. Ainsi, la survie est améliorée avec la méthode IPCW.

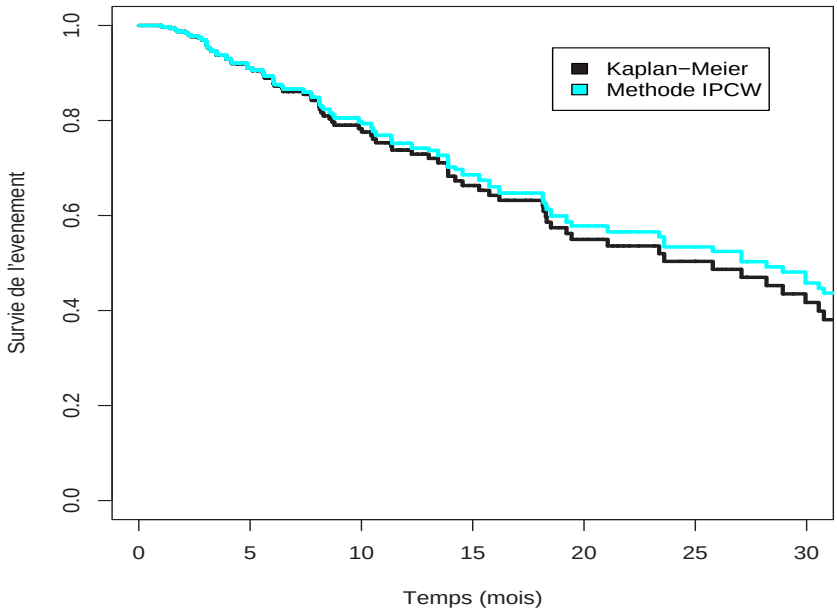


FIG. V.8 – Estimations de la survie de l'événement (transition dans l'état inacceptable) par Kaplan-Meier et par la méthode IPCW.

3.1.5 Méthode IPCW avec covariables

La méthode IPCW permet aussi d'estimer les effets des covariables sur le risque d'événement en prenant en compte la censure dépendante. Pour cela, il faut résoudre la version pondérée de la fonction score de la vraisemblance partielle (équation (V.13)). Les résultats des estimations du paramètres β dans $\lambda_T(t) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{Z}(t))$, par un modèle de Cox et par la fonction score modifiée sont présentés dans le tableau V.4. Les écarts-types des estimations IPCW donnés dans le tableau sont obtenus par la formule (V.14).

Covariable	Risque d'événement					
	Modèle de Cox univarié			Méthode IPCW		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	0.389	(0.220)	(0.08)	0.355	(0.221)	(0.11)
Sévérité	0.589	(0.223)	(<0.01)	0.697	(0.225)	(<0.01)
Exacerbations	-0.312	(0.267)	(0.24)	-0.158	(0.263)	(0.55)
Corticoïdes Inhalés	0.431	(0.162)	(<0.01)	0.893	(0.316)	(<0.01)
Corticoïdes Oraux	0.437	(0.129)	(<0.01)	0.880	(0.249)	(<0.01)
Antécédents Corticoïdes	0.759	(0.248)	(<0.01)	0.839	(0.256)	(<0.01)

[1] estimations des écarts-types.

[2] p avec le test de Wald pour $H_0 : \beta = 0$.

TAB. V.4 – Estimation des coefficients de régression pour la survie de l'événement par le modèle de Cox et par la méthode IPCW.

L'estimation de ces coefficients de régression par la méthode IPCW sont assez proches de ceux obtenus avec un modéle de Cox. Les facteurs de risque de passage dans l'état de contrôle inacceptable sont les mêmes dans les deux méthodes d'estimation. La méthode IPCW augmente de manière significative l'effet des corticoïdes oraux et inhalés sur le risque d'événement. Ceci s'explique certainement par le fait que ces covariables prédisent la censure et qu'elles sont utilisées dans le calcul des pondérations. Les résultats peuvent être aussi discutés en terme de risque relatif. Par exemple, le risque de passer dans l'état 2 est supérieur pour les individus dont l'asthme est sévère par rapport aux non sévères. Autrement dit avec la méthode IPCW, les patients sévères ont deux fois plus de chance de passer dans l'état de contrôle inacceptable. L'effet de l'IMC est légèrement atténué et n'est plus significatif avec un seuil de 10 %.

3.1.6 Extension

Dans ce qui précède, tous les patients n'ayant pas subi l'événement sont considérés comme censurés et sont pris en compte dans l'estimation du risque de censure. Ainsi aucune différence n'est faite entre les phénomènes de censure. Cependant, il semble raisonnable de penser que la censure générée par l'arrêt de l'étude n'apporte aucune information sur la survenue de l'événement d'intérêt.

Pour cela, il est intéressant de distinguer deux types de censure à savoir la censure potentiellement informative (perdus de vue) et non informative (exclus vivants). Ainsi il semble naturel de considérer uniquement les perdus de vue dans l'estimation du risque de censure. Les pondérations calculées de la sorte modifieront les contributions des individus (dans les estimations) en prenant en compte uniquement les personnes perdues de vue.

Cette conception de la censure a été mise en oeuvre dans le cas de l'asthme afin de comparer les nouveaux résultats avec les précédents. Ils ne sont pas présentés car ils ne sont pas significativement différents. En effet, les coefficients de régression associés au risque de censure ainsi que les estimations de la survie de la censure sont très proches. De même, les estimations de la survie IPCW dans les deux cas sont quasiment superposables. Les faibles différences observées entre ces résultats s'expliquent en partie par la faible proportion de patients exclus vivants (6.9%, $n = 23$).

3.2 Application à un modèle de Markov à deux états

3.2.1 Modèle avec deux états de contrôle

Dans un premier temps, on considère un modèle à deux états de contrôle représenté par la figure V.9.

Deux types de suivi sont rencontrés.

– Les patients considérés comme perdus de vues ($90,8\%$; $n = 369$). Ce sont les patients qui n'ont pas eu de consultation 3 mois avant la date de point. A la dernière consultation,

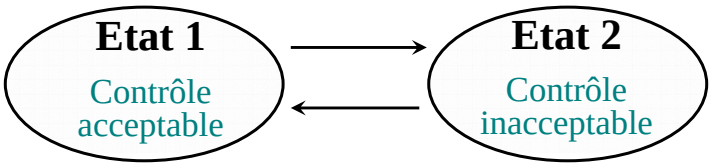


FIG. V.9 – Modèle à deux états de contrôle pour l'asthme.

  - 256 patients étaient dans l'état acceptable,
  - 113 patients étaient dans l'état inacceptable.
- Les patients considérés comme vivants à la date de point (9.2%; $n = 37$). Ces patients sont allés consulter au moins 3 mois avant la date de fin d'étude. A la dernière consultation,
  - 27 patients étaient dans l'état acceptable,
  - 10 patients étaient dans l'état inacceptable.

L'objectif de cette étape est de sélectionner les covariables qui influencent l'évolution du contrôle de l'asthme dans ce modèle. Un modèle univarié est ajusté pour chaque covariable. Le tableau V.5 donne les estimations des coefficients de régression, les écarts-types et les p-value avec le test de Wald pour tester si les coefficients sont statistiquement différents de zéro.

Covariable	Transition $1 \rightarrow 2$			Transition $2 \rightarrow 1$		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	0.238	(0.189)	(0.21)	-0.356	(0.132)	(<0.01)
Sévérité	1.013	(0.196)	(<0.01)	-0.180	(0.170)	(0.29)
Exacerbations	-0.149	(0.223)	(0.50)	0.198	(0.183)	(0.28)
Corticoïdes Inhalés	0.899	(0.288)	(<0.01)	-0.244	(0.136)	(0.07)
Corticoïdes Oraux	0.903	(0.219)	(<0.01)	-0.249	(0.360)	(0.49)
Antécédents Corticoïdes	1.117	(0.205)	(<0.01)	-0.320	(0.225)	(0.16)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

TAB. V.5 – Modèle semi-paramétrique univarié pour les risques de transition entre états de santé.

La sévérité de l'asthme, le traitement par corticoïdes oraux et inhalés et les antécédents de corticoïdes oraux augmentent le risque de passer d'un état acceptable vers un état inacceptable. La seule covariable qui influence de manière significative le risque de passer de l'état inacceptable vers un état acceptable, est l'IMC : un patient en surpoids à moins de chance de passer dans un état acceptable (comparé à un patient avec un IMC < 25). A partir de cette analyse univariée, on peut définir un modèle multivarié qui inclut les cinq covariables qui ont un impact significatif sur l'évolution de l'asthme. Le tableau V.6 donne les résultats des estimations et des tests dans ce modèle multivarié.

Covariable	Transition $1 \rightarrow 2$			Transition $2 \rightarrow 1$		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	0.068	(0.193)	(0.72)	-0.291	(0.143)	(0.04)
Sévérité	0.451	(0.245)	(0.07)	-0.032	(0.159)	(0.84)
Corticoïdes Inhalés	0.515	(0.302)	(0.09)	-0.103	(0.152)	(0.49)
Corticoïdes Oraux	0.617	(0.230)	(<0.01)	-0.065	(0.198)	(0.74)
Antécédents Corticoïdes	0.696	(0.244)	(<0.01)	-0.190	(0.190)	(0.32)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

TAB. V.6 – Modèle semi-paramétrique multivarié pour les risques de transition entre états de santé.

Les effets qui sont significatifs dans l'analyse univariée le sont encore dans l'analyse multivariée même s'ils sont atténués par l'ajustement. Ce modèle sera retenu pour modéliser l'évolution de la maladie.

3.2.2 Modèle avec état de censure

Après avoir sélectionné les covariables qui ajustent le mieux l'évolution de l'asthme dans un modèle à deux états de santé, on va vérifier que ces covariables influencent aussi les processus de censure à partir de l'état acceptable et inacceptable. Afin d'étudier les risques de censure, on considère le modèle à trois états de la figure V.2. Ce modèle comprend deux états de contrôle (acceptable et inacceptable) et un état de « censure » dans lequel l'individu peut transiter à partir de l'état acceptable et de l'état inacceptable. L'intérêt de ce modèle est de permettre une estimation des risques de censure et d'observer si les covariables ont un impact sur ces risques. Le tableau V.7 donne les estimations, dans un modèle multivarié, des coefficients de régression associés aux transitions des états de contrôle vers l'état de censure.

Covariable	Transition $1 \rightarrow C$			Transition $2 \rightarrow C$		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	0.038	(0.134)	(0.78)	-0.135	(0.202)	(0.50)
Sévérité	-0.116	(0.193)	(0.55)	0.154	(0.224)	(0.49)
Corticoïdes Inhalés	-0.035	(0.139)	(0.80)	0.829	(0.264)	(<0.01)
Corticoïdes Oraux	-1.292	(0.365)	(<0.01)	-0.228	(0.265)	(0.39)
Antécédents Corticoïdes	-0.111	(0.249)	(0.66)	-0.586	(0.269)	(0.03)

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

TAB. V.7 – Modèle semi-paramétrique multivarié pour les risques de censure.

Ces résultats montrent l'effet des covariables sur les risques de censure. A partir de l'état acceptable, les patients avec un traitement par corticoïdes oraux ont un risque de censure

diminué. De même, lors de l'analyse univariée avec différentes covariables, il ressort que la présence d'exacerbations diminue le risque de censure à partir de l'état acceptable. Ces résultats corroborent les intuitions cliniques et les résultats obtenus avec la méthode IPCW pour données de survie. A propos du risque de censure, à partir de l'état inacceptable, il semble que les antécédents de corticoïdes oraux diminue ce risque alors que le traitement par corticoïdes inhalés augmente ce risque. Ce résultat semble surprenant et contradictoire avec le fait que les patients les plus atteints soit les moins censurés. Cependant, il peut s'expliquer par le fait que le phénomène de censure à partir d'un état acceptable et à partir d'un état inacceptable sont de nature différente. Un patient qui est soigné et qui se trouve dans un état de contrôle inacceptable peut commencer à douter du médecin et ainsi quitter l'étude pour se faire soigner ailleurs. Ce phénomène est courant dans le cas du VIH, par exemple, où les patients les plus malades abandonnent le suivi.

Ensuite, les résultats de ce tableau montrent que parmi les coefficients qui ont un effet significatif sur la transition $1 \rightarrow 2$, seul l'effet des corticoïdes oraux reste significatif pour la transition de l'état acceptable vers l'état « censure ». L'IMC était la seule variable à avoir un effet sur la transition $2 \rightarrow 1$, cependant elle ne modifie pas de manière significative le risque de censure à partir de l'état 2. Ainsi la seule covariable qui influence de manière significative le processus d'évolution de la maladie et le processus de censure est le traitement par corticoïdes oraux. Le processus d'événement étant dépendant de la censure par l'intermédiaire de cette variable, on peut appliquer la méthode IPCW pour réduire les biais engendrés par cette dépendance. Même si la dépendance n'est pas flagrante (une seule variable prédit les deux processus), la méthode IPCW peut être utilisée. En effet, si les covariables qui prédisent l'évolution de la maladie ne prédisent pas la censure alors les pondérations vont tendre vers 1 et on retombera sur les estimations traditionnelles.

Afin d'étudier les phénomènes de censure à partir de l'état acceptable et de l'état inacceptable, il est possible de tracer les probabilités d'être censurés au temps t sachant l'état de départ. La figure V.10, montre les probabilités de censure dans un modèle sans covariable. Un patient dans l'état acceptable au temps 0 a plus de chance d'être censuré au temps t qu'un patient dans un état inacceptable. Ce résultat est en accord avec les résultats précédents : en effet, un patient qui se porte bien a tendance à arrêter le suivi, alors qu'un patient dans un état inacceptable revient plus facilement en consultation.

3.2.3 Estimations des probabilités de transition

Le calcul des probabilités de censure dans le modèle multivarié (avec état de censure), permet de définir des pondérations spécifiques à chaque individu. On définit ainsi des poids associés à l'état acceptable et à l'état inacceptable. Ces poids sont ensuite intégrés dans la méthode d'estimation semi-paramétrique pour modèle Markovien afin de modifier les contributions des individus. Sur les figures V.11 **(a)** et V.11 **(b)**, les estimations des probabilités de transition (**(a)** probabilité de transition $1 \rightarrow 2$, **(b)** probabilité de transition $2 \rightarrow 1$) par la méthode IPCW sont comparées à celles obtenues par la méthode d'estimation semi-paramétrique pour modèle Markovien.

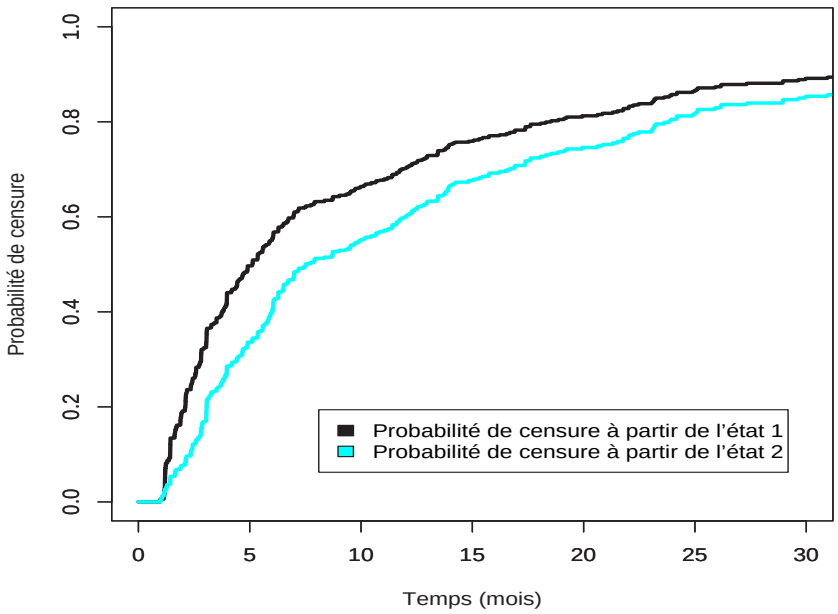


FIG. V.10 – Probabilités de censure à partir des états de contrôle acceptable et inacceptable.

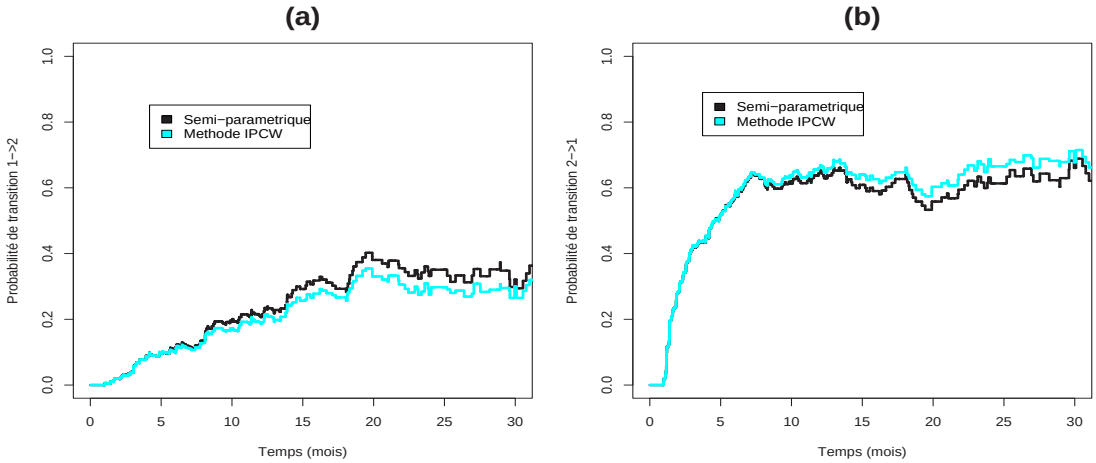


FIG. V.11 – Estimations des probabilités de transition par la méthode semi-paramétrique et par la méthode IPCW. **(a)** Probabilité de transition de l'état de contrôle acceptable vers l'état inacceptable ($1 \rightarrow 2$). **(b)** Probabilité de transition de l'état de contrôle inacceptable vers l'état acceptable ($2 \rightarrow 1$).

Comme pour l'étude de l'événement passage dans l'état inacceptable (modèle de survie), l'estimation de la probabilité de passer d'un état acceptable vers un état inacceptable par la méthode IPCW est plus faible que l'estimation par la méthode semi-paramétrique (Figure V.11 **(a)**). D'autre part, la probabilité de passer de l'état inacceptable vers l'état acceptable (Figure V.11 **(b)**) est plus grande avec la méthode IPCW. Ainsi, dans ce modèle à deux états avec retour possible, le fait de prendre en compte les patients censurés conduit à de meilleurs résultats d'un point de vue clinique : les estimations IPCW améliorent la probabilité de retour dans un état stable et diminuent la probabilité de passer dans un état instable.

3.2.4 Méthode IPCW avec covariables

Dans un deuxième temps, la méthode IPCW permet d'estimer les effets des covariables sur les risques de transitions entre les états de contrôle en intégrant les pondérations calculées précédemment. On peut ainsi obtenir des estimations des coefficients de régression en prenant en compte l'information issue des patients censurés. Ces estimations sont obtenues par résolution des équations modifiées du score de la vraisemblance partielle (équation (V.18)). Les résultats sont présentés dans les tableaux V.8 et V.9. Les estimations avec la vraisemblance partielle de Cox et avec la version modifiée de la fonction score sont proches et peu de différences significatives sont à noter. Les écarts-types et les probabilités de rejet ne sont pas présentés car il n'y a pas encore de résultats de convergence asymptotique.

Covariable	Transition $1 \rightarrow 2$					
	Modèle (univarié) semi-paramétrique			Méthode IPCW (univarié)		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	0.238	(0.189)	(0.21)	0.203		
Sévérité	1.013	(0.196)	(<0.01)	1.060		
Exacerbations	-0.149	(0.223)	(0.50)	-0.032		
Corticoïdes Inhalés	0.899	(0.288)	(<0.01)	0.787		
Corticoïdes Oraux	0.903	(0.219)	(<0.01)	1.072		
VEMS	0.693	(0.288)	(0.02)	0.607		
Antécédents Corticoïdes	1.117	(0.205)	(<0.01)	1.081		

[1] estimations des écarts-types.

[2] p avec le test de Wald pour $H_0 : \beta = 0$.

TAB. V.8 – Estimation des coefficients de régression pour la transition $1 \rightarrow 2$ par le modèle semi-paramétrique et par la méthode IPCW.

3.2.5 Extension

De manière similaire à l'analyse de données de survie, il est intéressant de différencier la censure générée par l'arrêt de l'étude (pouvant être supposée non informative) et celle

Covariable	Transition 2 → 1					
	Modèle (univarié) semi-paramétrique			Méthode IPCW (univarié)		
	$\hat{\beta}$	$(ec)^1$	$(p)^2$	$\hat{\beta}$	$(ec)^1$	$(p)^2$
IMC	-0.356	(0.132)	(<0.01)	-0.337		
Sévérité	-0.180	(0.170)	(0.29)	-0.228		
Exacerbations	0.198	(0.183)	(0.28)	0.197		
Corticoïdes Inhalés	-0.244	(0.136)	(0.07)	-0.250		
Corticoïdes Oraux	-0.249	(0.360)	(0.49)	-0.261		
VEMS	0.084	(0.136)	(0.54)	0.117		
Antécédents Corticoïdes	-0.320	(0.225)	(0.16)	-0.254		

[1] estimations des écarts-types.
[2] p avec le test de Wald pour $H_0 : \beta = 0$.

TAB. V.9 – Estimation des coefficients de régression pour la transition 2 → 1 par le modèle semi-paramétrique et par la méthode IPCW.

engendrée par les perdus de vue (potentiellement informative). Ainsi, seul les patients perdus de vue peuvent être considérés dans l'estimation du risque de censure.

Avec cette conception de la censure, les résultats obtenus dans le cas de l'asthme sont très proches des résultats précedents. Les faibles différences observées s'expliquent en partie par la faible proportion de patients exclus vivants (27 à partir de l'état 1 et 10 à partir de l'état 2).

4 Discussion

L'objectif de ce travail était de modéliser l'évolution d'une maladie dans le cas où la censure est dépendante de cette évolution. Dans un premier temps, la méthode est décrite dans le cadre d'analyse de survie. La méthode permet d'étudier les risques de censure et fournit ainsi des informations sur les facteurs expliquant ce phénomène. L'extension proposée de la méthode permet de différencier a priori la censure par perdus de vue et par exclus vivants afin d'utiliser uniquement l'information issue de la censure potentiellement informative. Dans un deuxième temps, la méthode est généralisée à certains modèles multi-états afin de prendre en compte la dépendance entre l'évolution de la maladie et la censure dans ces modèles. Enfin, les méthodes présentées sont appliquées à la base de données sur l'asthme. Notons que la programmation de ces méthodes suit une démarche identique à celle de l'estimation semi-paramétrique (*cf.* chapitre IV) présentée en annexe page 177.

4.1 Application

L'implémentation des méthodes a donné de résultats forts intéressants. Dans le cadre de données de survie, nous sommes en mesure d'expliquer la censure par différentes covariables : la prise ou non de traitement de corticoïdes inhalés ou oraux, les antécédents de traitement

mais aussi la présence d'exacerbations et la sévérité de l'asthme. Il s'avère que ce sont souvent les patients qui vont le mieux qui sont perdus de vue ce qui est cohérent avec l'intuition des cliniciens. Le deuxième point important à noter est que certaines covariables prédisent à la fois la censure et l'événement modélisé ce qui met en évidence la dépendance entre la censure et l'événement dans le cas de l'asthme. Le fait de prendre en compte l'information issue des patients censurés, nous a permis d'améliorer la survie des patients asthmatiques. Certes, cette différence entre les deux estimations n'est pas majeure mais elle va dans le sens des cliniciens. En effet, si ce sont les patients les plus censurés qui vont le mieux il est logique que la survie soit améliorée en prenant en compte cette information. Cependant, dans la population, il y a aussi des individus qui sont censurés parce qu'ils vont mal ou parce qu'ils ne sont pas satisfaits du résultat thérapeutique, et changent de médecin (même si ce n'est pas la majorité des cas). Dans ce cas là, les poids associés à ces individus sont inférieurs à un. De ce fait, pour chaque temps, il y a un mélange de poids inférieurs et supérieurs à un, ce qui entraîne un effet barycentrique dans l'estimation de la survie. Ceci pourrait expliquer en partie que la différence entre les deux méthodes ne soit pas plus importante.

Dans le cadre d'un modèle à deux états de contrôle avec retour possible, les résultats confirment ceux obtenus avec un modèle de survie. En effet, la censure est expliquée par plusieurs covariables et elle est dépendante du processus d'événement par l'intermédiaire de certaines de ces covariables. Notons également que la probabilité de censure à partir de l'état acceptable est plus importante que celle à partir de l'état inacceptable ce qui est logique d'un point de vue clinique. Les résultats sur les probabilités de transition confirment l'intuition clinique selon laquelle un patient qui se portent bien a plus tendance à être censuré : l'estimation IPCW améliore la probabilité de retour vers un état stable et diminue la probabilité de passage vers un état instable.

Dans le cadre des modèles de survie ou dans le cadre des modèles multi-états, nous avons réalisé deux modélisations des risques de censure. Tout d'abord, il est considéré que toute la population qui n'a pas subi d'événement est censurée, que ce soit par perdu de vue ou par exclus vivants (Robins et Finkelstein [2000]). Cependant, nous disposons d'une information que nous occultons dans ce cas précis. En effet, il est raisonnable de penser que la censure due au gel de la base est engendrée de manière aléatoire et n'apporte aucune information sur l'événement étudié. Ceci motive notre choix de modéliser un « risque de censure informative » pour utiliser au maximum l'information a priori dont nous disposons. En réalisant les estimations selon les deux hypothèses présentées, nous avons constaté qu'une faible différence existait entre les résultats des probabilités de transition. Ceci s'explique notamment par le fait que la part des exclus vivants représente moins de 10% de l'échantillon, ainsi, prendre en compte ou non cette information ne change guère les résultats.

4.2 Méthodes

L'originalité de cette méthode repose sur la modélisation des risques de censure, ce qui habituellement n'est pas traité dans la littérature. Cette étape permet de déterminer les facteurs prédictifs de la censure et ainsi de mieux comprendre le mécanisme de censure dans la pathologie étudiée. De plus, cette étape permet juger de la dépendance entre la censure et l'événement en déterminant les covariables qui influencent à la fois le risque d'événement et de censure. La méthode IPCW est fondée sur l'étude de la survie de la

censure afin d'obtenir les pondérations spécifiques à chaque individu. Ceci permet de prendre en compte l'information issue des perdus de vue et d'améliorer l'estimation de la survie de l'événement. Cette méthode a également un grand avantage puisque les estimateurs IPCW convergent vers les estimateurs traditionnels (Kaplan-Meier et estimateur non-paramétrique dans les modèles multi-états). En effet, les variables prédisant l'événement n'ont pas un impact significatif sur le risque de censure, par conséquent, les poids convergent vers un ($\hat{K}^0(t) \simeq \hat{K}_i(t)$) et ainsi, l'estimateur IPCW converge vers l'estimateur traditionnel. Ceci est un atout pour la méthode puisqu'il n'est pas nécessaire de faire d'hypothèses a priori sur la relation entre la censure et l'événement. La méthode est particulièrement adaptée pour réduire les biais quand la censure semble informative et qu'une part importante des patients sont des perdus de vue.

Cependant, comme dans toutes études des limites existent. En particulier, l'hypothèse (V.1) est contraignante puisqu'elle suppose que la connaissance temporelle des covariables apporte suffisamment d'informations pour que l'on puisse se passer de celle apportée par le temps d'événement. Cette hypothèse est d'autant plus vraie qu'on dispose d'un maximum de facteurs de risque de l'événement. Il est donc important de prendre en compte les facteurs essentiels afin de minimiser la contrainte imposer par l'hypothèse.

4.3 Perspectives

Afin de poursuivre le développement de la méthode IPCW pour modèle de Markov à deux états, il faudra étudier les propriétés de convergence asymptotique en s'inspirant des travaux de Robins [1993] et Andersen et al. [1993]. Ces résultats permettront d'obtenir une estimation de la variance des coefficients de régression. Cette variance est utile pour tester si certains coefficients sont statistiquement différents de zéro et pour obtenir des intervalles de confiance.

Il serait également intéressant de généraliser la méthode à toutes les formes de modèles multi-états. En effet, dans la méthode présentée, les poids sont identiques pour toutes les transitions à partir d'un même état. Cela ne pose pas de problèmes dans le cas des modèles à deux états et des modèles progressifs car une seule transition est possible à partir d'un même état. Si plusieurs transitions sont possibles, les poids modifieront les estimations de manière identique. Ainsi, dans l'exemple d'un modèle à trois états : « sain », « malade » et « décès » (Figure V.1 **(c)**), la censure aura un impact identique sur la transition « malade » vers « décès » et sur la transition « malade » vers « sain ». Dans le cas où plusieurs transitions sont possibles, on pourrait, par exemple essayer de construire des poids spécifiques à chaque transition (par l'intermédiaire des covariables par exemple). Ceci permettrait de prendre en compte l'effet de la censure informative en fonction du type d'événement.

Conclusion Générale

Décrire l'évolution des phénomènes dans le temps est d'un intérêt capital en épidémiologie. Par exemple, les maladies chroniques, l'impact d'un traitement ou encore la qualité et le coût du suivi doivent être étudiés de manière dynamique. L'étude statistique de ces données de cohorte est devenue essentielle pour une meilleure compréhension des maladies et une amélioration du suivi. Les modèles multi-états de type Markovien répondent à cette problématique et constituent un outil important pour l'analyse de ces données.

5 Récapitulatif de la thèse

Dans un premier temps, nous avons rappelé la méthodologie relative au modèle de Markov homogène (Saint-Pierre et al. [2003]). Ce modèle est le moins « complexe » des modèles de type Markovien car il suppose que les intensités de transition sont constantes dans le temps. Cette hypothèse d'homogénéité simplifie la méthodologie statistique et la programmation des méthodes d'estimation. Cependant, elle impose une contrainte qui est souvent trop forte dans de nombreuses applications.

Dans un second temps, nous avons étudié deux méthodes d'estimation (paramétrique et non-paramétrique) des intensités de transition dans un modèle semi-Markovien homogène. En effet, le modèle semi-Markovien propose une alternative quand le temps écoulé dans un état de santé semble être un facteur important de l'évolution de la maladie. Le modèle semi-Markovien homogène suppose lui que les intensités de transition dépendent de la durée écoulée dans un état.

La théorie des processus de comptage est ensuite présentée afin d'introduire des méthodes d'estimation (non-paramétrique et semi-paramétrique) dans le cadre d'un modèle de Markov non-homogène (Saint-Pierre et al. [2005c]). Dans ce modèle, les intensités de transition dépendent de la durée du suivi (temps depuis l'inclusion dans l'étude). La méthodologie des processus de comptage fournit un cadre rigoureux qui permet notamment de généraliser, aux modèles Markoviens, les estimateurs traditionnels des modèles de survie.

Les méthodes d'estimation présentées dans le cadre du modèle Markov non-homogène (et du modèle semi-Markovien) supposent que le mécanisme de censure n'apporte aucune information sur l'évolution de la maladie. Cette hypothèse étant rarement vérifiée en pratique, nous présentons une méthode d'estimation permettant de prendre en compte une censure informative dans l'étude de la survie (Saint-Pierre et al. [2005b]). En s'inspirant des méthodes d'estimation dans un modèle de Markov non-homogène, nous avons étendu

cette méthodologie au cas des modèles progressifs et des modèles Markoviens à deux états réversibles où la censure pose les mêmes difficultés.

Nous présentons en annexe un complément concernant la théorie statistique relative aux processus de comptage. Ce complément permet d'approfondir certains résultats certes complexes, mais indispensables à la bonne compréhension de cette théorie. Les annexes fournissent également un « guide » visant à faciliter la programmation des estimateurs basés sur les processus de comptage (Saint-Pierre et al. [2004]). Ainsi, les méthodes non-paramétriques et semi-paramétriques dans le cadre d'un modèle de Markov non-homogène (*cf.* chapitre IV) et les méthodes IPCW (*cf.* chapitre V) peuvent être implémentées en suivant la démarche décrite dans cette aide (*cf.* annexe page 177).

6 Résultats cliniques sur l'asthme

Dans les différents modèles étudiés, chaque méthode d'estimation est appliquée à une cohorte de patients asthmatiques. Nous avons particulièrement observé l'impact de l'indice de masse corporelle sur l'évolution de l'asthme. Dans tous les modèles utilisés (Markov homogène, semi-Markov homogène et Markov non-homogène), le surpoids diminue significativement l'intensité de transition d'un état de contrôle inacceptable vers un état de contrôle optimal. De plus, l'utilisation d'un modèle de Markov homogène à deux états de contrôle a permis d'obtenir des estimations ajustées sur la sévérité et la corticothérapie orale. Ainsi, il semble que l'effet négatif du surpoids sur la transition de l'état inacceptable vers l'état acceptable soit indépendant du traitement et de la sévérité. Ce résultat qui fait l'objet d'une publication dans une revue médicale (Saint-Pierre et al. [2005a]) corrobore et renforce les conclusions de plusieurs travaux visant à montrer un lien entre l'asthme et le surpoids (même si une preuve définitive est toujours manquante). Ainsi, l'effet négatif du surpoids devrait être pris en compte dans l'élaboration de rapports et consignes sur le suivi des asthmatiques. En particulier, il semble important que l'asthmatique se maintienne à un poids normal (IMC < 25) afin de ne pas diminuer ses chances de retourner dans un état stable. De plus, il faudrait faire le maximum pour éviter qu'un patient en surpoids transite vers un état de contrôle inacceptable.

La prise en compte de la censure informative dans le cas de l'asthme fournit également des résultats intéressants. En effet, les méthodes prenant en compte l'information contenue dans la censure permettent de diminuer l'estimation de la probabilité de transition vers un état de contrôle inacceptable et d'augmenter la probabilité de revenir à un contrôle acceptable. Ces résultats mettent en évidence l'intuition des médecins selon laquelle les patients censurés sont souvent des patients bien contrôlés.

7 Choix du modèle

Le choix du modèle dépend essentiellement des échelles de temps qui influencent l'évolution du phénomène. Le modèle de Markov homogène est adapté quand aucune échelle de temps ne semble influencer l'évolution de la maladie. Le modèle de Markov non-homogène et le modèle semi-Markovien homogène sont comparables en terme de flexibilité mais accordent

de l'importance à des échelles de temps différentes. Le modèle semi-Markovien homogène sera adapté pour prendre en compte la durée écoulée dans un état alors que le modèle de Markov non-homogène sera préférable pour la durée du suivi, l'âge ou le temps calendaire. Un modèle semi-Markovien non-homogène permettra de prendre en compte deux échelles de temps dans la modélisation.

Cependant, le rôle des échelles de temps est parfois mal connu. Dans ce cas, le modèle de Markov homogène est très utile. Ce modèle permet de tester l'hypothèse d'homogénéité et l'hypothèse Markovienne par l'intermédiaire de certaines covariables artificielles (respectivement, homogénéité par périodes, temps de séjour). Cette étape permet ainsi de choisir entre les différents modèles. Cependant, plusieurs modèles peuvent être appropriés. Dans ce cas, il devient difficile de mesurer l'apport de chaque modèle. En effet, les modèles ne sont pas emboîtés et les vraisemblances des modèles sont de formes différentes. Il est alors intéressant de comparer et de croiser les résultats afin d'approfondir les interprétations.

8 Discussion des biais

Les modèles multi-états constituent un outil performant pour l'analyse de données répétées en particulier lorsque les temps d'observation sont quelconques. Les résultats obtenus sont facilement interprétables d'un point de vue clinique. Les probabilités de transition permettent de bien comprendre l'évolution de la maladie. De plus, l'introduction de covariables dans ces modèles permet de mesurer l'impact des différents facteurs de risque.

Cependant, il faut rester très prudent quant à l'interprétation des résultats obtenus avec de telles analyses. En effet, les hypothèses nécessaires à la modélisation peuvent être sources de biais.

- L'hypothèse de Markov résume l'historique de l'individu au dernier état visité (processus sans mémoire).
- Dans toutes les méthodes présentées, il est supposé que la transition entre les états se produit au moment de la consultation. Les méthodes pour ajuster une censure par intervalles sont une alternative quand cette hypothèse semble trop forte.
- Les méthodes d'estimation dans le modèle de Markov non-homogène et dans le modèle semi-Markovien homogène supposent qu'il n'y a pas de changement d'état non observé entre deux consultations consécutives. Cependant, quand aucune information sur le patient n'est disponible entre deux consultations consécutives (patients suivis par intermittence) et que le modèle comprend des états réversibles, il est possible que certains changements d'état ne soient pas observés. Pour éviter ce problème, on suppose que les observations sont suffisamment rapprochées pour que l'on puisse considérer en pratique que tous les changements d'état sont observés (ainsi les durées de séjours sont disponibles). Dans le cas où cette hypothèse n'est pas vérifiée, l'analyse peut être biaisée.
- Aucune information n'est disponible sur l'état du patient après sa dernière consultation à cause du phénomène de censure à droite. Le modèle de Markov non-homogène et le modèle semi-Markovien font intervenir la durée entre la dernière consultation et la date de fin d'étude. Cette durée varie en fonction du choix de la date de fin d'étude et peut influencer les estimations. Ces modèles considèrent que l'état du patient ne

change pas entre la dernière consultation et l'arrêt de l'étude. Cette hypothèse semble d'autant plus vraie que la durée entre la dernière consultation et la fin de l'étude est courte. Le modèle de Markov homogène ne prend pas en compte ce qui se passe après la dernière consultation.
- L'introduction de covariables par l'intermédiaire d'un modèle à risques proportionnels implique une hypothèse de proportionnalité.
- L'utilisation de covariables dépendantes du temps nécessite de supposer que la valeur de la covariable ne change pas entre deux consultations (indépendamment du temps entre deux consultations).
- Les méthodes d'estimation dans ce type de modèles font souvent une hypothèse de censure non-informative. Cependant, dans de nombreuses applications, il est difficile de supposer que la censure n'apporte aucune information sur le phénomène étudié. Dans ce cas, les méthodes prenant en compte une censure informative permettent de mieux comprendre le mécanisme de censure et d'avoir des estimations corrigées.

Notons également que de nombreux biais sont inhérents aux bases observationnelles et sont très difficiles à éliminer. Parmi les biais les plus courants dans ce type d'étude, on rencontre :
- Le biais de confusion dû aux corrélations entre les covariables et qui n'est que partiellement réglé par l'ajustement.
- Le biais de sélection dû au fait que ce n'est pas n'importe quelle population qui prend un médicament et de ce fait la population traitée est particulière.
- Le biais de compliance, lui est d'une autre forme : en général, les patients qui prennent bien les traitements adhèrent au système de soins, ont une meilleure hygiène de vie et une meilleure santé (comparativement aux individus non compliants).

9 Perspectives

L'introduction d'un effet aléatoire dans l'écriture des intensités de transition peut être utile pour réduire certains biais. Ces modèles de fragilité (frailty) permettent de prendre en compte le lien entre éléments de la population. Ils ajustent les estimations sur l'effet des variables qui ne sont pas incluses dans le modèle ou qui ne sont pas mesurées. L'utilisation des techniques bayesiennes fournit un outil attractif pour ajuster ce type de modèles.

Il semble également important de développer les méthodes permettant de prendre en compte une censure informative. En effet, le phénomène de censure est toujours présent dans les données de cohorte, et, dans bien des cas, il apporte une information sur l'événement étudié. La prise en compte de cette information est importante pour tenter de réduire une partie des biais.

La théorie des processus de comptage et des martingales fournit un cadre formel à de nombreuses problématiques complexes. L'utilisation de cette théorie semble essentielle pour le développement théorique des méthodes liées aux modèles multi-états.

Il nous semble également important de continuer à développer des programmes et des algorithmes afin de faciliter l'utilisation des modèles multi-états. Il serait ainsi intéressant de mettre en place des logiciels permettant d'ajuster certaines méthodes d'estimation pour ce type de modèle.

D'un point de vue clinique, il semble important d'approfondir l'analyse de cette base de données et l'étude de l'évolution de l'asthme. En particulier, l'impact de l'indice de masse corporelle mérite d'être mieux connu. Nous nous sommes limités à l'étude du surpoids (IMC $\geqslant 25$) car la base ne comporte pas suffisamment de patients obèses (IMC $\geqslant 30$). La mise en place d'une nouvelle cohorte pourrait permettre, entre autre, d'étudier plus en détail l'effet de l'indice de masse corporelle en considérant une variable à plusieurs modalités : poids trop faible, poids normal, surpoids, obésité.

L'application des méthodes statistiques pour données de cohorte a permis de mettre en évidence certaines limites liées à la base de données. D'une part, le nombre de covariables pouvant être inclus simultanément dans la modélisation est limité par le nombre d'observations de la base de données. Ainsi, un nombre important de covariables présentes dans la base ne sont pas utilisées. Il serait préférable de mesurer les variables les plus pertinentes afin de simplifier la construction de la base et ainsi permettre d'inclure plus de patients dans l'étude. D'autre part, la base ne contient aucune information sur les patients qui arrêtent le suivi. Il serait intéressant de faire quelques efforts pour disposer de renseignements sur ces patients afin de mieux comprendre le phénomène de censure. Cette information pourrait aider à réduire certains biais présents dans les études de cohorte.

Bibliographie

Aalen O. O. (1978). *Nonparametric inference for a family of counting processes.* The Annals of Statistics, vol. 6. pages 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, vol. 5. pages 141–150.

Aguirre-Hernandez R. et Farewell V. T. (2002). *A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models.* Stat Med, vol. 21, n°13. pages 1899–1911.

Akerman M. J., Calacanis C. M. et Madsen M. K. (2004). *Relationship between asthma severity and obesity.* J Asthma, vol. 41. pages 521–526.

Alioum A. et Commenges D. (2001). *MKVPCI : a computer program for Markov models with piecewise constant intensities and covariates.* Comput Methods Programs Biomed, vol. 64, n°2. pages 109–119.

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

Andersen P. K., Borgan Ø., Gill R. D. et Keiding N. (1993). *Statistical models based on counting processes.* Springer-Verlag.

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, vol. 18. pages 153–167.

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

Boudemaghe T. et Daures J.-P. (2000). *Modeling asthma evolution by a multi-state model.* Rev Epidemiol Sante Publique, vol. 48, n°3. pages 249–255.

Boulet L. P., Becker A. et Berube D. (1999), *Canadian asthma consensus report.* Rapport technique Canadian Asthma Consensus Group CMAJ n°161.

Cardot H., Ferraty F. et Sarda P. (1999). *Linear Functional Model.* Statistics and Probability Letters, vol. 45. pages 11–22.

Castelli C. (2004). *Prise en compte de la censure informative dans le cadre d'analyse de survie et application à l'asthme*. Mémoire de DEA sous la direction du professeur JP. Daurès, Université Montpellier II.

Chang I. C., Hsiung C. A. et Wu S. (2000). *Estimation in a proportionnal hazard model for semi-Markov counting processes*. Statistica Sinica, vol. 10. pages 1257–1266.

Chen E. B. et Cook R. J. (2003). *Regression modeling with recurrent events and time-dependent interval-censored marker data*. Lifetime Data Analysis, vol. 9. pages 275–291.

Chouaid C., Vergnenegre A., Vandewalle V., Liebaert F. et Khelifa A. (2004). *The costs of asthma in France : an economic analysis by a Markov model*. Rev Mal Respir, vol. 21. pages 493–499.

Cockcroft D. W. et Swystum V. A. (1996). *Asthma control versus asthma severity*. J Allergy Clin Immunol, vol. 98. pages 1016–1018.

Colvert R. E. et Boardman T. J. (1999). *Estimation in the piece-wise constant hazard rate model*. Comm. Statist.-Theor. Meth., vol. A5. pages 1013–1029.

Combescure C., Chanez P., Saint-Pierre P., Daurès J.-P., Proudhon H. et Godard P. (2003). *Assessment of variations in control of asthma over time*. Eur Respir J, vol. 22, n°2. pages 298–304.

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

Commenges D. (2002). *Inference for multi-state models from interval-censored data*. Statistical Methods in Medical Research, vol. 11, n°2. pages 167–182.

Cook R. J. (1999). *A mixed model for two-state Markov processes under panel observation*. Biometrics, vol. 55, n°3. pages 915–920.

Cook R. J., Kalbfleisch J. D. et Yi G. Y. (2002). *A generalized mover-stayer model for panel data*. Biostatistics, vol. 3, n°3. pages 407–420.

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

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

de Stavola B. L. (1988). *Testing departures from time homogeneity in multistate Markov processes*. Applied Statistics, vol. 37. pages 242–250.

Diggle P., Liang K. Y. et Zeger S. L. (1994). *Analysis of longitudinal*. Oxford University Press.

Duchateau L., Janssen P., Kezic I. et Fortpied C. (2003). *Evolution of recurrent asthma event rate over time in frailty models*. Journal of the Royal Statistical Society Series C, vol. 52. pages 355–363.

Flegal K. M., Caroll M. D., Ogden C. L. et Johnson C. L. (2002). *Prevalence and trends in obesity among US adults, 1999-2000.* JAMA, vol. 288. pages 1723–1727.

Foucher Y. (2004). *Modèles semi-Markoviens - Place dans la gestion des maladies chroniques.* Mémoire de DEA sous la direction du professeur JP. Daurès, Université Montpellier II.

Foucher Y., Mathieu E., Saint-Pierre P., Durand J. et Daurès J.-P. (Octobre 2004). A semi-markov model based on generalized weibull distribution with an illustration for hiv disease, soumis dans dans Biometrical Journal. Octobre 2004.

Gentleman R. C., Lawless J. F., Lindsey J. C. et Yan P. (1994). *Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease.* Stat Med, vol. 13, n°8. pages 805–821.

Gill R. D. (1980). *Nonparametric estimation based on censored observations of a Markov renewal process.* Z. Wahrscheinlichkeitstheorie verw. Gebiete, vol. 53. pages 97–116.

Godard P., Clark T. J. et Busse W. W. (1998). *Clinical assessment of patients.* Eur Respir J, vol. 26. pages 2S–5S.

Heitjan D. F. et Rubin D. B. (1991). *Ignorability and coarse data.* The annals of statistics, vol. 19. pages 2244–2253.

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

Hill C., Com-Nougué C., Kramar A., Moreau T., O'Quiegley J. et Senoussi R. (1996). *Analyse statistique des données de survie.* Médecine-Science Flammarion, seconde éd.

Hougaard P. (1995). *Frailty models for survival data.* Lifetime Data Analysis, vol. 1. pages 255–273.

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

Hougaard P. (2000). *Analysis of multivariate survival data.* Springer – Statistics for biology and health.

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.* Stat Med, vol. 21. pages 3369–3382.

Huang X. et Wolfe A. R. (2002). *A fraitly model for non informative censoring.* Biometrics, vol. 58. pages 510–520.

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

Huber-Carol C. et Vonta I. (2004). *Fraitly models for arbitrarily censored and truncated data*. Lifetime Data Analysis, vol. 10. pages 369–388.

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

Jain S. (1986). *Markov chain model and its application*. Comput Biomed Res, vol. 19. pages 374–378.

Janssen J. (1986). *In semi-markov models*. Plenum Press.

Janssen J. et Limnios N. (1999). *Semi-markov models and applications*. Kluwer Academic Publishers.

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

Joly P., Commenges D., Helmer C. et Letenneur L. (2002). *A Penalized Likelihood Approach for an Illness-Death Model With Interval-Censored Data : Application to Age-Specific Incidence of Dementia*. Biostatistics, vol. 3. pages 433–443.

Juniper E. F., O'Byrne P. M., Guyatt G. H., Ferrie P. J. et King D. R. (1999). *Development and validation of a questionnaire to measure asthma control*. Eur Respir J, vol. 14. pages 902–907.

Kalbfleisch J. D. et Lawless J. F. (1985). *The analysis of panel data under a Markov assumption*. Journal of the American Statistical Association, vol. 80. pages 863–871.

Kaplan E. L. et Meier P. (1958). *Non-parametric estimation from incomplete observations*. J. Amer. Statist. Assoc., vol. 53. pages 457–481.

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

Korn E. L. et Whittemore A. S. (1979). *Methods for analyzing panel studies of acute health effects of air pollution*. Biometrics, vol. 35, n°4. pages 795–802.

Kousignian I., Autran B., Chouquet C., Calvez V. et Gomard E. e. a. (2003). *Markov modelling of changes in HIV-specific cytotoxic T-lymphocyte responses with time in untreated HIV-1 infected patients*. Stat Med, vol. 22. pages 1675–1690.

Liang K. Y. et Zeger S. L. (1986). *Longitudinal analysis using generalized linear models*. Biometrika, vol. 73. pages 13–22.

Limnios N. (1997). *Dependability analysis of semi-Markov systems*. Reliability Engineering and System Safety, vol. 55. pages 203–207.

Lindsay J. C. et Ryan L. M. (1993). *A three-state multiplicative model for rodent tumorigenicity experiments*. Applied Statistics, vol. 42. pages 283–300.

Little R. J. A. (1995). *Modelling the drop-out mechanism in repeated-measures studies*. Journal of the American Statistical Association, vol. 90. pages 1112–1121.

Liu L., Wolfe A. R. et Huang X. (2004). *Shared fraitly models for recurrent events and a terminal event.* Biometrics, vol. 60. pages 747–756.

Longini I. M. J., 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.* Stat Med, vol. 8, n°7. pages 831–843.

Marshall G. et Jones R. H. (Sep 1995). *Multi-state models and diabetic retinopathy.* Stat Med, vol. 14, n°18. pages 1975–1983.

Masoli M., Fabian D., Holt S. et Beasley R. (2004). *The global burden of asthma : executive summary of the GINA Dissemination Committee report.* Allergy, vol. 59. pages 469–478.

Matsuyama Y. (2003). *Sensitivity analysis for the estimation of rates of change with non-ignorable drop-out : an application to a randomized clinical trial of the vitamin D3.* Statistics in Medicine, vol. 22. pages 811–827.

Minini P. et Chavance M. (2004). *Sensitive analysis of longitudinal normal data with drop-outs.* Statistics in Medicine, vol. 23. pages 1039–1054.

National Institutes of Health (1997), *Expert Panel Report 2 : Guidelines for the diagnosis and management of asthma.* Rapport technique NIH 97-4051.

Nielsen G. G., Gill R. D., Andersen P. K. et Sørensen T. I. A. (1992). *A counting process approach to maximum likelihood estimation in frailty models.* Scandinavian Journal of Statistics, vol. 8. pages 25–43.

Nocedal J. et Wright S. J. (1999). *Numerical optimization.* Springer.

Ouhbi B. et Limnios N. (1999). *Nonparametric Estimation for Semi-Markov Processes Based on its Hazard Rate Functions.* Statistical Inference for Stochastic Processes, vol. 2, n°2. pages 151–173.

Paes A. T. et de Lima A. (2004). *A SAS macro for estimating transition probabilities in semiparametric models for recurrent events.* Computer Methods and Programs in Biomedicine, vol. 75. pages 59–65.

Paltiel A. D., Fuhlbrigge A. L., Kitch B. T., Liljas B. et Weiss S. T. (2001). *Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma : results from the asthma policy model.* J Allergy Clin Immunol, vol. 108. pages 39–46.

Papadopoulou A. A. et Vassiliou P. C. (1994). *Asymptotic behaviour of non homogeneous semi-Markov systems.* Linear Algebra and Its Applications, vol. 210. pages 153–198.

Perez-Ocon R. et Ruiz-Castro J. E. (1999). *Semi-markov models and applications*, chapter 14, pages 229–238. Kluwer Academic Publishers.

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

Pinheiro J. C. et Bates D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Price M. J. et Briggs A. H. (2002). *Development of an economic model to assess the cost effectiveness of asthma management strategies*. Pharmacoeconomics, vol. 20. pages 183–194.

Redline S., Tager I. B., Segal M. et Gold D. (1989). *The relationship between longitudinal change in pulmonary function and nonspecific airway responsiveness in children and young adults*. Am Rev Respir Dis, vol. 140. pages 179–184.

Robins J. M. (1993). *Information recovery and bias adjustement in proportionnal hazards regression analysis of randomized trials using surrogate markers*. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, Berlin. Springer. pages 24–33.

Robins J. M. et Finkelstein D. M. (2000). *Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests*. Biometrics, vol. 56, n°3. pages 779–788.

Robins J. M. et Rotnitzky A. (1992). *Recovery of information and adjustement for dependent censoring using surrogate markers*. AIDS Epidemiology-Methodological Issues. pages 297–331.

Rotnitzky A. et Robins J. M. (2003). *Inverse probability weighted estimation in survival analysis*. The Encyclopedia of Biostatistics.

Saint-Pierre P. (2001). *Modèle linéaire généralisé pour variables fonctionnelles*. Mémoire de DEA sous la direction du professeur P. Sarda, Université Toulouse III.

Saint-Pierre P., Bourdin A., Chanez P., Daurès J.-P. et Godard P. (Février 2005a). Overweighted asthmatics are more difficult to control, soumis dans Chest. Février 2005a.

Saint-Pierre P., Castelli C. et Daurès J.-P. (Février 2005b). Informative censoring in survival analysis and application to asthma, soumis dans Biometrical Journal. Février 2005b.

Saint-Pierre P., Combescure C., Daurès J.-P. et Godard P. (2003). *The analysis of asthma control under a Markov assumption with use of covariates*. Statistics in Medicine, vol. 22. pages 3755–3770.

Saint-Pierre P., Daurès J.-P. et Godard P. (Septembre 2004). Implementation of semi-parametric estimation in multi-state models, soumis dans Computer Methods and Programs in Biomedicine. Septembre 2004.

Saint-Pierre P., Daurès J.-P. et Godard P. (2005c). *The analysis of asthma using homogeneous and non-homogeneous markov models*. A paraitre dans Far East Journal of Theorical Statistics.

Satten G. A. (1999). *Estimating the extent of tracking in interval-censored chain-of-events data*. Biometrics, vol. 55, n°4. pages 1228–1231.

Scharfstein D. O., Robins J. M., Eddings W. et Rotnitzky A. (2001). *Inference in randomized studies with informative censoring and discrete time-to-event endpoints.* Biometrics, vol. 57. pages 404–413.

Scharfstein D. O., Rotnitzky A. et Robins J. M. (1999). *Adjusting for non ignorable drop-out using semiparametric nonresponse models.* Journal of the American Statistical Association, vol. 94, n°448. pages 1096–1146.

Self S. et Liang K.-Y. (1987). *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions.* Journal of the American Statistical Association, vol. 82. pages 605–610.

Singpurwalla N. D. et Wong M. Y. (1983). *Estimation of the failure rate : A survey of nonparametric methods. Part I : Non-bayesian methods.* Commun. Statist.-Theor.Meth., vol. 12, n°5. pages 559–588.

Sternberg M. R. et Satten G. A. (1999). *Fitting Semi-Markov Models to Interval-Censored Data with Unknown Initiation Times.* Biometrics, vol. 55. pages 507–513.

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

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

Ware J. H., Lipsitz S. et Speizer F. E. (1988). *Issues in the analysis of repeated categorical outcomes.* Stat Med, vol. 7, n°1-2. pages 95–107.

Weisberg S., McCann D., Desai M., Rosenbaum M., Leibel R. et Ferrante A. (2003). *Obesity is associated with macrophage accumulation in adipose tissue.* J Clin Invest, vol. 112. pages 1796–1808.

Weiss S. T. et Shore S. (2004). *Obesity and asthma : directions for research.* Am J Respir Crit Care Med, vol. 169. pages 963–968.

Annexe A

Théorie statistique

1 Processus aléatoires et intégrales stochastiques

Définition 10 Soit un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ où $\mathbb{P}$ est la mesure de probabilité sur $(\Omega, \mathcal{A})$. Un **processus aléatoire**, ou encore une **fonction aléatoire réelle** (f.a.r.) est une fonction à deux variables

$$\begin{array}{rcl} X : \mathcal{T} \times \Omega & \longrightarrow & \mathbb{R} \\ (t, \omega) & \longmapsto & X(t, \omega) \end{array}$$

où t représente le temps et ω le hasard. Pour tout $t \geqslant 0$, la fonction $X_t : \omega \longmapsto X(t, \omega)$ est une variable aléatoire réelle appelée **coordonnée** à l'instant t. Pour tout $\omega \in \Omega$, la **trajectoire** est la fonction $t \longmapsto (X(t, \omega), t \geqslant 0)$. Si, pour presque tout ω, la trajectoire est continue alors $X(t, \omega)$ est une f.a.r. à trajectoire continue.

Définition 11 Soit un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, $t \in \mathcal{T}$. Une **filtration** est une famille de tribus $\{\mathcal{F}_t : t \in \mathcal{T}\}$, telle que

$$\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{A}, \qquad \forall s \leqslant t.$$

Définition 12 Soient un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et $\mathcal{F}_t$ une filtration. Un processus $X = X(t, \omega)$ est dit **$\mathcal{F}_t$-adapté** si $\forall t$, X_t est $\mathcal{F}_t$-mesurable.

Définition 13 Soit un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. Soit $(M_t)_t$, $t \in \mathcal{T}$ un processus réel défini sur Ω. Soit $(\mathcal{F}_t)_t$ une filtration sur Ω.

$(M_t)_t$ est une **$\mathcal{F}_t$-martingale** si

(i) $\forall t$, M_t est $\mathcal{F}_t$-adaptée et $M_t \in \mathcal{L}^1$ (i.e. est intégrable),

(ii) $\forall s$, t, $0 \leqslant s \leqslant t$, $\mathbb{E}(M_t \mid \mathcal{F}_s) = M_s$ *p.s.*

Définition 14 Soit un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ et une filtration $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{R}^+}$. Un **$\mathcal{F}$-processus croissant** A est un processus adapté à $\mathcal{F}$ à valeurs réelles satisfaisant la propriété suivante : pour tout $\omega \in \Omega$, les $t \longmapsto A_t(\omega)$ sont croissantes, continues à droites et nulles en 0.

Définition 15 La tribu P de $\mathcal{T} \times \Omega$ engendrée par les ensembles $]s, t] \times \Gamma$, où $0 < s < t$ et $\Gamma \in \mathcal{F}_s$, est **la tribu des ensembles prévisibles**. Une v.a. $C(t, \omega)$ définie sur $(\mathcal{T} \times \Omega, \mathbb{P})$ est un processus **prévisible**.

Proposition 17 Soient A un processus croissant et C un processus prévisible. Alors, pour tout t,

$$\int_0^t C(s)dA(s) = \int_0^t C(s)A(ds), \tag{A.1}$$

est une variable aléatoire.

2 Produit intégral (ou infini)

Définition 16 Soit $\mathbf{X}(t)$, $t \in \mathcal{T}$, une matrice $p \times p$ de processus cadlag, nul en 0, et à variation bornée. On obtient une mesure additive en posant

$$\mathbf{X}(]s, t]) = \mathbf{X}(t) - \mathbf{X}(s).$$

Soit une partition $t_0 = s < t_1 < ... < t_n = t$. Son pas est

$$|\delta| = \sup_i |t_i - t_{i-1}|,$$

On appelle **produit intégral** (ou **produit infini**)

$$\underset{]s,t]}{\mathcal{P}} (\mathbf{Id} + d\mathbf{X}) = \lim_{|\delta|\to 0} \prod_{i=1}^{n} [\mathbf{Id} + \mathbf{X}(]t_{i-1}, t_i])]$$

où $\mathbf{I}$ représente la matrice identité $p \times p$.

Théorème 2 Soit $\mathbf{A}$ une matrice de fonctions de dimension $k \times k$ correspondant à une mesure d'intensité (matrice des intensités cumulées).

Alors la matrice,

$$\mathbf{P}(s, t) = \underset{u\in]s,t]}{\mathcal{P}} (\mathbf{Id} + \mathbf{A}(du)), \qquad s \leqslant t,\ t, s \in \mathcal{T},$$

est la matrice de probabilité de transition d'un processus de Markov à espace d'états fini $\{1, \ldots, k\}$.

Le processus peut être construit de la façon suivante

– Sachant que le processus est dans h au temps t_0, il reste dans l'état h pour une durée avec une intensité cumulée

$$-\left(A_{hh}(t)\right) - A_{hh}(t_0)), \qquad t_0 \leqslant t \leqslant \inf\left\{u \geqslant t_0 : \Delta A_{hh}(u) = -1\right\}.$$

– Sachant qu'il quitte h au temps t, il transite vers l'état j $(j \neq h)$ avec une probabilité

$$-\frac{dA_{hj}(t)}{dA_{hh}(t)}.$$

Dans le cas scalaire, où $p = 1$, si X est continu, le produit intégral devient l'exponentielle

$$\underset{]0,t]}{\mathcal{P}}(1 + dX) = \exp(X(t)).$$

Notons que le produit intégral est introduit par Volterra (1887) comme l'unique solution de certaines équations intégrales.

Théorème 3 Soient $\mathbf{Z}$, $\mathbf{W}$, 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 $[0, t]$

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

3 Processus de comptage

Soient $(\mathcal{T}, \mathfrak{T})$ un espace mesurable, $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et $\mathcal{F}_t$ une filtration,

Définition 17 Soit la fonction $\mu(.)$ définie par

$$\begin{array}{rcl}\mu : \mathcal{T} & \longrightarrow & \mathbb{N} \\ t & \longmapsto & \mu(t)\end{array}$$

est une **mesure de comptage** sur $\mathcal{T}$ si, pour chaque $B \subset \mathfrak{T}$

$$\mu(B) \in \mathbb{N}, \text{ et } \mu(B) < \infty \; p.s.$$

Définition 18 Un **processus aléatoire de comptage** $N(.)$ sur $\mathcal{T}$ est une fonction aléatoire

$$\begin{array}{rcl} N : \mathcal{T} \times \Omega & \longrightarrow & \mathbb{N} \\ (t, \omega) & \longmapsto & N(t, \omega), \end{array}$$

cadlag (continue à droite avec une limite à gauche), $\mathcal{F}_t$-adaptée, nulle en zéro, croissante et ayant des sauts d'amplitude 1.

Proposition 18 Soit $N(.)$ un processus de comptage. Il existe un processus $\Lambda(.)$ $\mathcal{F}_t$-prévisible, croissant, continu à droite et nul en zéro tel que

$$M(t) = N(t) - \Lambda(t),\ t \in \mathcal{T} \tag{A.2}$$

soit une martingale.

$\Lambda(.)$ s'appelle le compensateur de $N(.)$, ou encore son processus d'intensité cumulée, ou encore sa mesure d'intensité.

Ce résultat est l'application de la **décomposition de Doob** à la sous-martingale locale $N(.)$.

Proposition 19 Soient N un processus comptage de dimension 1 et Λ son compensateur.

Si N est absolument continu, alors N possède une intensité λ, i.e. il existe un processus prévisible λ tel que $\forall\ A \in \mathcal{T}$

$$\Lambda(A) = \int_A \lambda(u)\, du\ .$$

L'intensité λ est définie par $\forall\ u \in \mathcal{T}$

$$\lambda(u) = \lim_{\varepsilon \to 0} P\left(N(u+\varepsilon) - N(u) \geqslant 1 \mid \mathcal{F}_u\right),$$

où $\mathcal{F}_u$ est la **filtration naturelle** engendrée par les $N(s)$ pour $s \leqslant u$, c'est-à-dire l'ensemble des événements observables à l'instant u,

$$\mathcal{F}_u = \sigma\left\{N(s) : s \leqslant u\right\}.$$

Définition 19 Un processus de comptage k-dimensionnel $\mathbf{N} = (N_1, N_2, ..., N_k)$ est appelé **processus de comptage multivarié** si chacune de ses composantes est un processus de comptage univarié et s'il ne peut y avoir simultanéité des sauts de deux (ou plus) de ses composantes.

4 Processus ponctuel marqué

A chaque processus de comptage multivarié $\mathbf{N}$, on peut associer une séquence des *temps du saut* et une séquence des *marques du saut* renseignant sur le temps et la nature de chaque événement. Comme il ne peut y avoir simultanéité des sauts de deux composantes de $\mathbf{N}$, la somme des composantes est un processus de comptage. En définissant,

$$N_{\cdot} = \sum_{h=1}^{k} N_h,$$

Considérons la séquence des *temps du saut*

$$0 < T_1 \leqslant T_2 \leqslant T_3 \leqslant \cdots$$

et les variables aléatoires représentant les *marques du saut*

$$J_1, J_2, \ldots,$$

à valeurs dans $\{1, 2, \ldots, k\} \cup \{0\}$ tels que pour $n \leqslant N_{\cdot}(\tau)$, $T_n \in \mathcal{T}$, $J_n \neq 0$, $T_n > T_{n-1}$,

$$N_{\cdot}(T_n) = n \text{ et } \Delta N_{J_n}(T_n) = 1.$$

Pour $n > N_{\cdot}(\tau)$, on fixe $J_n = 0$ et $T_n = \tau$ pour que tous soient bien définis. Les variables T_n sont des temps d'arrêts et de plus, J_n est $\mathcal{F}_{T_n}$-mesurable.

Ces notations sont essentiellement celles d'un processus ponctuel marqué où $\mathbf{N}$ est considéré comme une mesure de comptage sur l'espace produit : *espace de temps* $\times$ *espace de marques*. Ainsi, chaque événement d'un processus ponctuel marqué est caractérisé par le moment de son apparition et par une marque fournissant une description de cet événement. De manière générale, nous observons, sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, des instants aléatoires successifs $\{T_n\}_{n\in\mathbb{N}}$,

$$\begin{array}{rcl} T_n : \Omega & \longrightarrow & \mathcal{T} \\ \omega & \longmapsto & T_n(\omega) \end{array}$$

A chaque instant T_n, nous faisons une observation J_n qui est une fonction mesurable de $(\Omega, \mathcal{A}, \mathbb{P})$ dans un espace mesurable $(\mathcal{K}, \mathfrak{K})$ appelé espace de marques ou espace d'états

$$\begin{array}{rcl} J_n : \Omega & \longrightarrow & \mathcal{K} \\ \omega & \longmapsto & J_n(\omega). \end{array}$$

Définition 20 Un **processus ponctuel marqué** est une suite, finie ou dénombrable, de variables aléatoires $\{Y_n\}_{n\in\mathbb{N}}$ de $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans l'espace produit $(\mathcal{T} \times \mathcal{K}, \mathfrak{T} \otimes \mathfrak{K})$

$$Y_n = (T_n, J_n) \in \mathcal{T} \times \mathcal{K}. \tag{A.3}$$

La mesure de comptage associée est $\forall\ A \times L \subset \mathfrak{T} \otimes \mathfrak{K}$

$$\mu(\omega, A \times L) = \mu_L^{\omega}(A) = \sum_{n\in\mathbb{N}} \mathbb{1}_{[T_n(\omega)\in A,\ J_n(\omega)\in L]},$$

avec

$$\mu_L^{\omega}(A) \in \mathbb{N} \text{ et } \mu_L^{\omega}(A) < \infty\ p.s.$$

Cette approche permet de prendre en compte différents types d'événements qui varient de manière continue, par exemple, la mesure d'un markeur à chaque consultation. Dans le cas particulier d'un processus de comptage multivarié, l'espace de marques est fini, $\mathcal{K} = \{1, ..., k\}$.

5 Vraisemblance d'un processus de comptage

Théorème 4 (Jacod) Soit un processus de comptage multivarié $\mathbf{N} = (N_1, \ldots, N_k)$ sur $[0, \tau]$. Soient $\mathbb{P}$ et $\tilde{\mathbb{P}}$ deux mesures de probabilité sur deux espaces tels que N ait pour compensateur dans chacun d'eux respectivement Λ et $\tilde{\Lambda}$. On suppose que $\tilde{\mathbb{P}}$ est absolument continue par rapport à $\mathbb{P}$ *i.e.* $\tilde{\mathbb{P}} << \mathbb{P}$. Alors

$$\tilde{\Lambda}_h << \Lambda_h \qquad \text{pour tout } h,\ \mathbb{P} - p.s.$$

et

$$\begin{aligned}\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} &= \left.\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}\right|_{\mathcal{F}_0} \frac{\underset{t\in[0,\tau]}{\mathcal{P}} \left\{\prod_h d\tilde{\Lambda}_h(t)^{\Delta N_h(t)}(1 - d\tilde{\Lambda}_\cdot(t))^{1-\Delta N_\cdot(t)}\right\}}{\underset{t\in[0,\tau]}{\mathcal{P}} \left\{\prod_h d\Lambda_h(t)^{\Delta N_h(t)}(1 - d\Lambda_\cdot(t))^{1-\Delta N_\cdot(t)}\right\}} \\ &= \left.\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}\right|_{\mathcal{F}_0} \underset{t\in[0,\tau]}{\mathcal{P}} \prod_h \left(\frac{d\tilde{\Lambda}_h}{d\Lambda_h}(t)\right)^{\Delta N_h(t)} \frac{\mathcal{P}_{t\in[0,\tau]:\Delta N_\cdot(t)\neq 1}(1 - d\tilde{\Lambda}_\cdot(t))}{\mathcal{P}_{t\in[0,\tau]:\Delta N_\cdot(t)\neq 1}(1 - d\Lambda_\cdot(t))}.\end{aligned}$$

Théorème 5 Si Λ et $\tilde{\Lambda}$ sont sont presque sûrement continues, alors

$$\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \left.\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}\right|_{\mathcal{F}_0} \frac{\underset{t\in[0,\tau]}{\mathcal{P}} \prod_h d\tilde{\Lambda}_h(t)^{\Delta N_h(t)} \exp\left[-\tilde{\Lambda}_\cdot(\tau)\right]}{\underset{t\in[0,\tau]}{\mathcal{P}} \prod_h d\Lambda_h(t)^{\Delta N_h(t)} \exp\left[-\Lambda_\cdot(\tau)\right]}$$

et si Λ et $\tilde{\Lambda}$ sont presque sûrement absolument continues, alors

$$\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}} = \left.\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}\right|_{\mathcal{F}_0} \frac{\underset{t\in[0,\tau]}{\mathcal{P}} \prod_h \tilde{\lambda}_h(t)^{\Delta N_h(t)} \exp\left[-\tilde{\Lambda}_\cdot(\tau)\right]}{\underset{t\in[0,\tau]}{\mathcal{P}} \prod_h \lambda_h(t)^{\Delta N_h(t)} \exp\left[-\Lambda_\cdot(\tau)\right]}.$$

Théorème 6 La vraisemblance associée au processus $\mathbf{N}$ est

$$d\mathbb{P} = d\mathbb{P}|_{\mathcal{F}_0} \underset{t\in[0,\tau]}{\mathcal{P}} \left(\prod_h d\Lambda_h(t)^{\Delta N_h(t)}(1 - d\Lambda_\cdot(t))^{1-\Delta N_\cdot(t)}\right).$$

Si Λ est sont presque sûrement continue, alors

$$d\mathbb{P} = d\mathbb{P}|_{\mathcal{F}_0} \underset{t\in[0,\tau]}{\mathcal{P}} \left(\prod_h d\Lambda_h(t)^{\Delta N_h(t)} \exp\left[-\Lambda_\cdot(\tau)\right]\right).$$

Si Λ est sont presque sûrement absolument continue, alors

$$d\mathbb{P} = d\mathbb{P}|_{\mathcal{F}_0} \underset{t\in[0,\tau]}{\mathcal{P}} \left(\prod_h \lambda_h(t)^{\Delta N_h(t)} \exp\left[-\Lambda_{\cdot}(\tau)\right] \right).$$

6 Vraisemblance partielle

D'après la section précédente, la vraisemblance pour un processus de comptage $\mathbf{N}(t)$ par rapport à la filtration naturelle $\mathcal{F}_t = \sigma\{\mathbf{N}(s) : s \leqslant t\}$ vaut

$$d\mathbb{P} = \underset{t\in[0,\tau]}{\mathcal{P}} \prod_h d\Lambda_h(t)^{dN_h(t)} \left[1 - d\Lambda_{\cdot}(t)\right]^{1-dN_{\cdot}(t)}. \tag{A.4}$$

Pour introduire la notion de vraisemblance partielle, il est utile de présenter la vraisemblance à l'aide des processus ponctuels marqués (A.3). La vraisemblance d'un processus $\mathbf{N}(t)$ par rapport à la filtration naturelle $\mathcal{F}_t$, s'écrit

$$\underset{t\in[0,\tau]}{\mathcal{P}} \prod_{x\in\mathcal{K}} \underbrace{\mathbf{\Lambda}(dt,dx)^{\mathbf{N}(dt,dx)}}_{=(A)} \underbrace{\left[1 - \mathbf{\Lambda}(dt,\mathcal{K})\right]^{1-\mathbf{N}(dt,\mathcal{K})}}_{=(B)} \tag{A.5}$$

où

- $\mathbf{N}$ est le processus ponctuel marqué dans $(\mathcal{K}, \mathfrak{K})$,
- $\mathbf{\Lambda}$ est le compensateur de $\mathbf{N}$,
- $\mathbf{N}$ et $\mathbf{\Lambda}$ sont des mesures aléatoires sur $(\mathcal{T}\times\mathcal{K}, \mathfrak{T}\otimes\mathfrak{K})$.

Cette formule peut s'interpréter de la manière suivante :

- (A) représente le fait, conditionnellement à ce qui s'est passé avant t, l'événement de marque dx se produit dans l'intervalle de temps $[t, t+dt]$ (*i.e.* $\mathbf{N}(dt,dx) = 1$) avec la probabilité $\mathbf{\Lambda}(dt,dx)$.
- (B) représente le fait que, conditionnellement à ce qui s'est passé avant t, aucun événement ne s'est produit dans l'intervalle de temps $[t, t+dt]$ (*i.e.* $\mathbf{N}(dt,d\mathcal{K}) = 0$) avec la probabilité $1 - \mathbf{\Lambda}(dt,d\mathcal{K})$.

La vraisemblance peut être reformulée sous forme de plusieurs facteurs conditionnels afin de ne conserver que les termes qui dépendent des paramètres d'intérêt. Ce qui revient à supprimer les termes qui dépendent d'une manière complexe ou inconnue des paramètres de nuisance. La vraisemblance partielle ainsi obtenue sera en fait la vraisemblance complète pour le paramètre d'intérêt.

Considérons $\emptyset$ la « marque vide » qui représente l'absence d'événement, $\emptyset \notin \mathcal{K}$. Notons de plus $\overline{\mathcal{K}} = \mathcal{K}\cup\{\emptyset\}$. Soit G un autre espace de marques tel que $\emptyset \notin G$ et $\overline{G} = G\cup\{\emptyset\}$. Soit $g : \overline{\mathcal{K}} \to \overline{G}$ une application mesurable telle que $g(\emptyset) = \emptyset$. Considérons enfin, $\mathbf{N}^g$ le processus ponctuel marqué, appelé processus réduit, d'espace de marques G, défini par

$$\mathbf{N}^g(]0,t]\times A) = \mathbf{N}\left(]0,t]\times g^{-1}(A)\right).$$

Ainsi, lorsque $\mathbf{N}$ est défini par les points (T_n, J_n) tels que $g(J_n) \neq \emptyset$, ceux du processus ponctuel marqué $\mathbf{N}^g$ sont définis par $(T_n, g(J_n))$. Le compensateur $\mathbf{\Lambda}^g$ du processus marqué $\mathbf{N}^g$ est défini par

$$\mathbf{\Lambda}^g(]0,t] \times A) = \mathbf{\Lambda}\left(]0,t] \times g^{-1}(A)\right).$$

La distribution conditionnelle des événements dans l'intervalle de temps $[t, t+dt[$ sachant $\mathcal{F}_{t-}$ peut être reformulée à l'aide du processus ponctuel $\mathbf{N}^g$ de la manière suivante :

(1) avec la probabilité $\mathbf{\Lambda}^g(dt, dy)$, le processus réduit $\mathbf{N}^g$ a un événement dans $[t, t+dt[\times dy$ (*i.e.* $\mathbf{N}^g(dt, dy) = 1$) ou avec la probabilité $1 - \mathbf{\Lambda}^g(dt, G)$, le processus réduit $\mathbf{N}^g$ n'a aucun événement dans $[t, t+dt[$ (*i.e.* $\mathbf{N}^g(dt, G) = 0$)

(2) étant donné que le processus réduit $\mathbf{N}^g$ a un événement dans $[t, t+dt[\times dy$ *i.e.* $\mathbf{N}^g(dt, dy) = 1$, le processus $\mathbf{N}$ a un événement dans $[t, t+dt[\times dx$, pour x tel que $g(x) = y$, avec la probabilité conditionnelle $\mathbf{\Lambda}(dt, dx) / \mathbf{\Lambda}^g(dt, dy)$.

(3) sachant que le processus réduit $\mathbf{N}^g$ n'a pas d'événement dans $[t, t+dt[$ *i.e.* $\mathbf{N}^g(dt, G) = 0$, le processus $\mathbf{N}$ a soit

- **a.** un événement dans $[t, t+dt[\times dx$ pour x tel que $g(x) = \emptyset$, *i.e.* $\mathbf{N}(dt, dx) = 1$ avec la probabilité conditionnelle $\mathbf{\Lambda}(dt, dx) / [1 - \mathbf{\Lambda}^g(dt, dy)]$,
- **b.** aucun événement dans $[t, t+dt[$ *i.e.* $\mathbf{N}(dt, \mathcal{K}) = 0$ avec une probabilité conditionnelle complémentaire égale à $1 - \mathbf{\Lambda}\left(dt, g^{-1}(\emptyset)\right) / [1 - \mathbf{\Lambda}^g(dt, dy)]$.

La vraisemblance (A.5) s'exprime comme le produit intégral des distributions conditionnelles des événements dans $[t, t+dt[$ sachant $\mathcal{F}_{t-}$ pour tout $t \in [0, \tau]$. Combinant les possibilités dans le même ordre que celui de leur description, la vraisemblance (A.5) peut se réécrire sous la forme :

$$\begin{aligned} d\mathbb{P} \;=\; & \underset{t\in[0,\tau]}{\mathcal{P}} \left[\left\{ \prod_{y\in G} \mathbf{\Lambda}^g(dt, dy)^{\mathbf{N}^g(dt,dy)} \left[1 - \mathbf{\Lambda}^g(dt, G)\right]^{1-\mathbf{N}^g(dt,G)} \right\} \right. \\ & \times \prod_{y\in G} \left\{ \prod_{x:g(x)=y} \left(\frac{\mathbf{\Lambda}(dt, dx)}{\mathbf{\Lambda}^g(dt, dy)} \right)^{\mathbf{N}(dt,dx)} \right\}^{\mathbf{N}^g(dt,dy)} \\ & \times \left\{ \prod_{x:g(x)=\emptyset} \left(\frac{\mathbf{\Lambda}(dt, dx)}{1 - \mathbf{\Lambda}^g(dt, G)} \right)^{\mathbf{N}(dt,dx)} \right. \\ & \left. \left. \times \left(1 - \frac{\mathbf{\Lambda}\left(dt, g^{-1}(\emptyset)\right)}{1 - \mathbf{\Lambda}^g(dt, G)} \right)^{1-\mathbf{N}(dt,\mathcal{K})} \right\}^{1-\mathbf{N}^g(dt,G)} \right] \end{aligned} \tag{A.6}$$

L'interprétation intuitive et probabilitiste de (A.6) est la suivante :

(1) la première ligne de (A.6) a la même forme que la vraisemblance basée sur $\mathbf{N}^g$ dans le cas d'un processus adapté à sa filtration naturelle et peut donc être interprétée comme la vraisemblance partielle basée sur $\mathbf{N}^g$ en ignorant le reste de l'information basée sur $\mathbf{N}$.

(2) la deuxième ligne donne la contribution $\mathbf{\Lambda}(dt, dx) / \mathbf{\Lambda}^g(dt, dy)$ pour t, x et $y = g(x)$, qui correspond aux événements (en nombre fini) communs à $\mathbf{N}$ et $\mathbf{N}^g$. Puisque $\mathbf{\Lambda}^g$ est une « marginalisation » de $\mathbf{\Lambda}$, on peut « désintégrer » $\mathbf{\Lambda}$ [restreint à $\mathcal{F} \times \mathcal{K} \backslash g^{-1}(\emptyset)$] en un produit de l'image $\mathbf{\Lambda}^g$ de $\mathbf{\Lambda}$ et d'une mesure de probabilité de transition $\mathbf{\Lambda}(dx|t, y)$ sur $\{x : g(x) = y\}$. Ainsi, on peut écrire

$$\mathbf{\Lambda}(dt, dx) = \mathbf{\Lambda}^g(dt, dy)\, \mathbf{\Lambda}(dx|t, y),$$

dans ce sens que l'intégrale suivant t et x du membre de gauche est égale à la triple intégrale suivant t, y et $x = g(y)$ du membre de droite. Intuitivement, pour $y = g(x)$, nous écrivons la probabilité d'avoir une marque dans dx dans la période dt, étant donné le passé, comme la probabilité d'avoir une marque dans dy dans la période dt connaissant le passé multiplié par la probabilité d'avoir une marque dans dx étant donnée une marque réduite y au temps t.

(3) la troisième ligne n'intervient qu'un nombre fini de fois ; cela suggère que $\mathbf{\Lambda}(dt, dx) / [1 - \mathbf{\Lambda}^g(dt, G)]$ peut être interprété mathématiquement comme $\mathbf{\Lambda}(dt, dx) \times [1 - \mathbf{\Lambda}^g(\{t\} \times G)]^{-1}$.

(4) la quatrième ligne intervient uniquement lorsque N n'a pas d'événement dans $[t, t + dt[$. Notons que

$$\begin{aligned} 1 - \frac{\mathbf{\Lambda}\left(dt, g^{-1}(\emptyset)\right)}{1 - \mathbf{\Lambda}^g(dt, G)} &= \frac{1 - \mathbf{\Lambda}^g(dt, G) - \mathbf{\Lambda}\left(dt, g^{-1}(\emptyset)\right)}{1 - \mathbf{\Lambda}^g(dt, G)} \\ &= \frac{1 - \mathbf{\Lambda}\left(dt, g^{-1}(G)\right) - \mathbf{\Lambda}\left(dt, g^{-1}(\emptyset)\right)}{1 - \mathbf{\Lambda}^g(dt, G)} \\ &= \frac{1 - \mathbf{\Lambda}(dt, \mathcal{K})}{1 - \mathbf{\Lambda}^g(dt, G)}. \end{aligned}$$

Avec ces transformations, par le fait que

$$\prod_{y \in G} \prod_{x : g(x) = y} \sim \prod_{x : g(x) \neq \emptyset},$$

que dans (2), $\mathbf{N}^g(dt, dy) = 1$, et dans (3) et (4), $\mathbf{N}^g(dt, dy) = 0$, l'expression de la vraisemblance devient

$$\begin{aligned} d\mathbb{P} = \underset{t \in [0,\tau]}{\mathcal{P}} \Bigg[& \left\{ \prod_{y \in G}^{g} \mathbf{\Lambda}(dt, dy)^{\mathbf{N}^g(dt,dy)} \left[1 - \mathbf{\Lambda}^g(dt, G)\right]^{1 - \mathbf{N}^g(dt,G)} \right\} \\ & \times \left\{ \prod_{x : g(x) \neq \emptyset} \mathbf{\Lambda}(dx|t, g(x))^{\mathbf{N}(dt,dx)} \right\} \times \left\{ \prod_{x : g(x) = \emptyset} \left(\frac{\mathbf{\Lambda}(dt, dx)}{1 - \mathbf{\Lambda}^g(\{t\} \times G)} \right)^{\mathbf{N}(dt,dx)} \right\} \\ & \times \left\{ \left(\frac{1 - \mathbf{\Lambda}(dt, \mathcal{K})}{1 - \mathbf{\Lambda}^g(dt, G)} \right)^{1 - \mathbf{N}(dt,\mathcal{K})} \right\} \Bigg]. \end{aligned} \tag{A.7}$$

La première ligne de la formule (A.7) est la vraisemblance partielle basée sur $\mathbf{N}^g$ et le produit des autres lignes forme la vraisemblance partielle basée sur le reste de $\mathbf{N}$.

Le cas particulier où $\mathcal{K}$ et G sont dénombrables et où $\mathbf{\Lambda}$ est absolument continue sur $\mathcal{F} \times \mathcal{K}$ par rapport à la mesure de Lebesgue que multiplie la mesure de comptage est

intéressant. Dans ce cas, $\mathbf{N}$ est un processus de comptage dans le sens usuel (c-à-d avec un nombre dénombrable de composantes) et $\mathbf{N}^g$ est une agrégation de $\mathbf{N}$. Soient λ_x l'intensité du processus $\mathbf{N}$ pour $x \in \mathcal{K}$ et λ_y^g celle de $\mathbf{N}^g$ pour $y \in G$ avec

$$\lambda_y^g(t) = \sum_{x:g(x)=y} \lambda_x(t).$$

Pour l'ensemble fini $\{x : g(x) = y\}$, la mesure de transition

$$\begin{aligned}\Lambda(dx|t, g(x)) &= \frac{\mathbf{\Lambda}(dt, dx)}{\mathbf{\Lambda}^g(dt, dy)}, \\ &= \frac{\lambda_x(t)}{\lambda_y^g(t)},\end{aligned}$$

est une mesure de probabilité car $\mathbf{\Lambda}$ est absolument continue.

La partie atomique $1 - \mathbf{\Lambda}^g(\{t\} \times G)$ disparaît et la factorisation (A.7) devient :

$$\begin{aligned}d\mathbb{P} \propto & \underset{t\in[0,\tau]}{\mathcal{P}} \left[\prod_{y\in G} \lambda_y^g(t)^{N_y^g(dt)} \left[1 - \sum_{y\in G} \lambda_y^g(t)\, dt \right]^{1-\sum_{y\in G} N_y^g(dt)} \right. \\ & \times \prod_{x:g(x)\neq\emptyset} \frac{\lambda_x(t)}{\lambda_{g(x)}^g(t)}^{N_x(dt)} \times \prod_{x:g(x)=\emptyset} \lambda_x(t)^{N_x(dt)} \left[\frac{1 - \sum_{x\in\mathcal{K}} \lambda_x(t)\, dt}{1 - \sum_{y\in G} \lambda_y^g(t)\, dt} \right]^{1-\sum_{x\in\mathcal{K}} N_x(dt)} \\ & \propto \prod_{t,y} \lambda_y^g(t)^{N_y^g(dt)} \exp\left(- \int_0^\tau \sum_{y\in G} \lambda_y^g(t)\, dt \right) \times \prod_{t,x:g(x)\neq\emptyset} \frac{\lambda_x(t)}{\lambda_{g(x)}^g(t)}^{N_x(dt)} \\ & \times \prod_{t,x:g(x)=\emptyset} \lambda_x(t)^{N_x(dt)} \exp\left(- \int_0^\tau \left(\sum_{x\in\mathcal{K}} \lambda_x(t) - \sum_{y\in G} \lambda_y^g(t) \right) dt \right).\end{aligned}$$

On démontre facilement que cette quantité est réellement une factorisation de la vraisemblance totale :

$$d\mathbb{P} \propto \underset{t\in[0,\tau]}{\mathcal{P}} \left[\prod_{x\in\mathcal{K}} \lambda_x(t)^{N_x(dt)} \left(1 - \sum_{x\in\mathcal{K}} \lambda_x(t)\, dt \right)^{1-\sum_{x\in\mathcal{K}} N_x(dt)} \right].$$

7 Processus de comptage et censure à droite

7.1 Notations et définitions

Considérons,

– les conditions initiales, $\mathbf{X}_0 = (X_{i0}, i = 1, ..., n)$ tel que $X_i(0) = X_{i0}$.

– $\mathbf{N}(t)$ le processus de comptage multivarié non censuré par rapport à $\mathcal{F}_t = \sigma(\mathbf{X}_0) \vee \sigma(\mathbf{N}(u), 0 \leqslant u \leqslant t)$ est

$$\mathbf{N}(t) = (N_{hi}(t), i = 1, ..., n; h = 1, ..., k),$$

où

$$N_{hi}(t) = \mathbb{1}_{\{T_{hi} \leqslant t\}}$$

avec T_{hi} le temps d'apparition de l'événement h pour l'individu i.

– $Y_{hi}(t)$ qui vaut 1 si l'individu i est à risque pour l'événement h au temps t, 0 sinon.

– $\mathbf{C}(t)$ le processus de censure à droite,

$$\mathbf{C}(t) = (C_{hi}(t), i = 1, ..., n; h = 1, ..., k),$$

avec

$$C_{hi}(t) = I_{\{t \leqslant U_{hi}\}},$$

où U_{hi} est le temps de censure pour l'individu i à risque pour l'événement h.

– La filtration $\mathcal{G}_t$ telle que

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(\mathbf{C}(u), 0 \leqslant u \leqslant t),$$

– $\mathbf{N}^c(.)$ le processus de comptage censuré à droite qui représente la partie observable de $\mathbf{N}(.)$:

$$\mathbf{N}^c(t) = (N^c_{hi}(t), i = 1, ..., n; h = 1, ..., k),$$

avec

$$N^c_{hi}(t) = \int_0^t C_{hi}(s)\, dN_{hi}(s),$$

– $Y^c_{hi}(t)$ tel que,

$$Y^c_{hi}(t) = C_{hi}(t)\, Y_{hi}(t), \qquad h = 1, \ldots, k.$$

Afin d'obtenir la vraisemblance pour des données censurées, considérons la famille P de mesures de probabilité sur $(\Omega, \mathcal{A})$,

$$P = \{\mathbb{P}_{\theta\phi} : (\theta, \phi) \in \Theta \times \Phi\}.$$

Le paramètre θ doit être interprété comme le paramètre d'intérêt définissant les intensités de transition alors que ϕ est un paramètre de nuisance. Les ensembles Θ et Φ peuvent être des espaces de dimension finie (modèle paramétrique) ou des espaces de fonctions (modèle non-paramétrique). Supposons que la distribution de $\mathbf{X}_0$ dépend des paramètres θ et ϕ.

7.2 Censure à droite indépendante

Définition 21 Soit $\mathbf{N}(.)$ un processus de comptage multivarié de compensateur $\mathbf{\Lambda}(.)$ par rapport à la filtration $\mathcal{F}_t$ et la probabilité $\mathbb{P}_{\theta\phi}$. Soit $\mathbf{C}(.)$ un processus de censure à droite prévisible par rapport à $\mathcal{G}_t \supseteq \mathcal{F}_t$.

Alors, la censure à droite générée par $\mathbf{C}(.)$ est indépendante si le compensateur de $\mathbf{N}(.)$ par rapport à $\mathcal{G}_t$ est aussi $\mathbf{\Lambda}(.)$.

Proposition 20 Sous l'hypothèse de censure à droite indépendante, le compensateur de $\mathbf{N}^c(.)$ par rapport à $(\mathbb{P}_{\theta\phi}, \mathcal{G}_t)$ est

$$\Lambda^c_{hi}(t,\theta) = \int_0^t C_{hi}(s)\, d\Lambda_{hi}(s,\theta).$$

En effet, sous l'hypothèse de censure à droite indépendante, chaque $N_{hi}(t)$ a une décomposition (A.2)

$$N_{hi}(t) = \Lambda_{hi}(t,\theta) + M_{hi}(t),$$

avec M_{hi} une martingale locale de carré intégrable par rapport à $\mathcal{G}_t$.

D'après (A.1),

$$\begin{aligned} N^c_{hi}(t) &= \int_0^t C_{hi}(s)\, dN_{hi}(s), \\ &= \int_0^t C_{hi}(s)\, d\Lambda_{hi}(s,\theta) + \int_0^t C_{hi}(s)\, dM_{hi}(s), \\ &= \Lambda^c_{hi}(t,\theta) + M^c_{hi}(t). \end{aligned}$$

Puisque M_{hi} est une martingale locale de carré intégrable, $C_{hi}(.)$ est un processus prévisible par rapport à $\mathcal{G}_t$ et C_{hi} est bornée alors $\int_0^t C_{hi}(s)\, dM_{hi}(s)$ est une martingale locale de carré intégrable et $\Lambda^c_{hi}(t,\theta)$ est un compensateur de $N^c_{hi}(t)$ (la décomposition est unique).

Proposition 21 Sous l'hypothèse de censure à droite indépendante, si le processus $\mathbf{N}(.)$ a une intensité multiplicative par rapport à $\mathcal{G}_t$ alors $\mathbf{N}^c(.)$ a aussi une intensité multiplicative par rapport à $\mathcal{G}_t$.

En effet, si

$$\lambda_{hi}(t,\theta) = \alpha_{hi}(t,\theta)\, Y_{hi}(t),$$

est l'intensité de N par rapport à $\mathcal{G}_t$ alors

$$\begin{aligned} d\Lambda^c_{hi}(t,\theta) &= C_{hi}(t)\, d\Lambda_{hi}(t,\theta), \\ &= C_{hi}(t)\, \alpha_{hi}(t,\theta)\, Y_{hi}(t), \\ &= \alpha_{hi}(t,\theta)\, Y^c_{hi}(t), \\ &= \lambda^c_{hi}(t,\theta). \end{aligned}$$

N^c_{hi} a la même intensité individuelle $\alpha_{hi}(t,\theta)$ que le processus non censuré N_{hi}.

Proposition 22 Sous l'hypothèse de censure à droite indépendante, si le processus $\mathbf{N}$ a une intensité multiplicative par rapport à $\mathcal{G}_t$ alors $\mathbf{N}^c$ a aussi une intensité multiplicative par rapport à $\mathcal{F}_t^c$, avec

$$\mathcal{F}_t^c = \sigma\left(\mathbf{X}_0, \mathbf{N}^c\left(u\right), Y^c\left(u\right), 0 \leqslant u \leqslant t\right),$$

et

$$\mathbf{N}^c\left(u\right) = \left(N_{hi}^c\left(u\right), h = 1, ..., k; i = 1, ..., n\right),$$
$$Y^c\left(u\right) = \left(Y_{hi}^c\left(u\right), h = 1, ..., k; i = 1, ..., n\right),$$

7.3 Vraisemblance sous censure indépendante

Soit $\mathbf{N}^c\left(t\right)$ le processus de comptage censuré à droite, avec censure indépendante. Les données disponibles au temps t comprennent X_0, $\mathbf{N}^c\left(t\right), 0 \leqslant u \leqslant t$ et les temps de censure $U_i \leqslant t$ pour les individus n'ayant pas de temps d'absorption avant U_i ce qui peut, dans le cadre d'un modèle multiplicatif, être formalisé par

$$\mathcal{F}_t^c = \sigma\left(\mathbf{X}_0, \mathbf{N}^c\left(u\right), Y^c\left(u\right), 0 \leqslant u \leqslant t\right)$$

La vraisemblance sous censure indépendante est construite en suivant la démarche de la vraisemblance partielle présentée page 167. Les observations sont représentées par un processus ponctuel marqué $\mathbf{N}^*$ de marques de la forme $x = (y, u)$, avec

- y le type d'événement pour un individu (la censure n'est pas un événement), $y = \emptyset$ correspond à l'absence d'événement (« marque vide »),
- u le sous-ensemble de $\{1, ..., n\}$ (ensemble des individus) des individus censurés.

Le couple $(\emptyset, \emptyset)$ est impossible car si personne n'est censuré, c'est que forcément un événement survient. Le processus de comptage observé $\mathbf{N}^c$ (correspondant au processus réduit $\mathbf{N}^g$ de la page 167) est obtenu par agrégation de $\mathbf{N}^*$:

$$\mathbf{N}_y^c = \sum_{x:g(x)=y} \mathbf{N}_x^*.$$

La fonction g définie par

$$g(y, u) = y,$$

permet, via la processus $\mathbf{N}^c\left(.\right)$, de décomposer la vraisemblance en plusieurs facteurs conditionnels.

La vraisemblance $L^*\left(\theta, \phi\right)$ pour $\mathbf{N}^*$ par rapport à $\mathcal{F}_t^c$ est

$$L_\tau^*\left(\theta, \phi\right) = L_0\left(\theta, \phi\right) \underset{t\in[0,\tau]}{\mathcal{P}} \mathbb{P}_{\theta\phi}\left(d\mathbf{N}^*\left(t\right) | \mathcal{F}_{t^-}^c\right).$$

La construction de la vraisemblance partielle page 167 est appliquée du processus réduit $\mathbf{N}^c$,

$$\begin{aligned} L_\tau^*\left(\theta, \phi\right) &= \underbrace{\underset{t\in[0,\tau]}{\mathcal{P}} \mathbb{P}_{\theta\phi}\left(d\mathbf{N}^c\left(t\right) | \mathcal{F}_{t^-}^c\right)}_{=L_\tau^c(\theta)} \\ &\times \underbrace{L_0\left(\theta, \phi\right) \underset{t\in[0,\tau]}{\mathcal{P}} \mathbb{P}_{\theta\phi}\left(d\mathbf{N}^*\left(t\right) | d\mathbf{N}^c\left(t\right), \mathcal{F}_{t^-}^c\right)}_{=L"(\theta,\phi)}, \end{aligned}$$

où $L_\tau^c(\theta)$ correspond à la vraisemblance partielle basée sur $\mathbf{N}^c$

$$L_\tau^c(\theta) = \underset{t\in[0,\tau]}{\mathcal{P}} \prod_{i=1}^{n}\prod_{h=1}^{k} \left[d\Lambda_{hi}^c(t,\theta)\right]^{\Delta N_{hi}^c(t)} \left[1 - d\Lambda_{\cdot\cdot}^c(t,\theta)\right]^{1-\Delta N_{\cdot\cdot}^c(t)}.$$

et $L^{"}(\theta,\phi)$ correspond à la contribution de $\mathbf{X}_0$ et des événements liés au processus N^* dans $[t, t+dt[$ sachant $\mathbf{N}^c$ et $\mathcal{F}_{t-}$.

La vraisemblance partielle avec censure a la même forme que la vraisemblance (partielle) sans censure. Le fait que la forme de la vraisemblance (partielle) soit préservée par la censure indépendante implique que les propriétés de martingales restent les mêmes. Le processus de score $\dfrac{\partial}{\partial\theta}\log\left[L_\tau^c(\theta)\right]$ est une $\mathcal{F}_t^c$-martingale de la même manière que $\dfrac{\partial}{\partial\theta}\log\left[L_\tau(\theta)\right]$ est $\mathcal{F}_t$-martingale dans le modèle sans censure. Ainsi, lorsque la censure est indépendante, l'inférence et la théorie asymptotique sont toujours applicables de la même manière que pour des données non censurées (par la structure de martingale).

7.4 Censure à droite non informative

Soit $\mathbf{N}(.)$ un processus de comptage multivarié,

$$\mathbf{N}(t) = \left(N_{hi}(t), i = 1,...,n; h = 1,...,k\right),$$

sujet à une censure à droite indépendante par le processus $\mathbf{C}(t)$. Les observations sont considérées comme un processus ponctuel marqué $\mathbf{N}^*$, d'après ce qui précède la vraisemblance pour $\mathbf{N}^*$ par rapport à $\mathcal{F}_t^c$ est

$$L_\tau^*(\theta,\phi) = L_\tau''(\phi,\theta)\, L_\tau^c(\theta),$$

où $L_\tau^c(\theta)$ est la vraisemblance partielle basée sur le processus de comptage observé $\mathbf{N}^c$. Afin de déterminer si $L_\tau^c(\theta)$ est la vraisemblance totale pour θ pour un $\phi \in \Phi$ fixé, on introduit la notion de censure non informative.

Définition 22 Si, pour chaque $\phi \in \Phi$ fixé, la vraisemblance $L_\tau^*(\theta,\phi)$ peut s'écrire

$$L_\tau^*(\theta,\phi) = L_\tau''(\phi)\, L_\tau^c(\theta)$$

alors la censure à droite indépendant $C(.)$ est non informative pour θ.

La censure à droite non informative peut être interprétée de la manière suivante : si pour chaque $\phi \in \Phi$ et $t \in \mathcal{T}$, l'intensité conditionnelle des individus censurés au temps t sachant le passé juste avant t et sachant un possible événement d'interêts au temps t ne dépend pas de θ, alors la censure est non informative pour θ.

En d'autres termes, si $L_\tau^c(\theta)$ est la vraisemblance complète pour le paramètre θ, alors la censure est non informative.

8 Théorème de la limite centrale

Théorème 7 (Robolledo) Si M_n est une suite de martingales, et si

(i) $\langle M_n \rangle_t$ converge en probabilité vers v_t déterministe,

(ii) $\forall \varepsilon, \exists M_{n,\varepsilon}$ suite de martingales telles qu'aucune différence $M_n - M_{n,\varepsilon}$ n'ait une amplitude supérieure à ε,

alors $M_n(t)$ a une limite $M(t)$ de processus croissant v_t, et $M(t)$ est un processus gaussien :

$$\frac{M_n(t)}{v_t} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Annexe B

Programmation de l'estimateur semi-paramétrique

Cette section fournit des explications détaillées pour une programmation de l'estimateur semi-paramétrique dans un modèle de Markov non-homogène (*cf.* chapitre IV page 89). Ce travail a été soumis, en septembre 2004, dans la revue *Computers Methods and Programs in Biomédicine* (Saint-Pierre et al. [2004]).

Les notations et la méthodologie sont celles du Chapitre IV page 71. La méthode de programmation a été testée sur le jeu de données utilisé dans Andersen et al. [1991] et Andersen et al. [1993]. Notons que nos programmes (en language *R*) correspondant à ces explications dans le cas particulier d'un modèle à trois états avec toutes les transitions possibles, sont disponibles sur internet ($stpierre31@free.fr$). Notons également que tous les programmes utilisés dans cette thèse (en langage *S-Plus* ou *R*) sont disponibles auprès de l'auteur.

La programmation est décrite dans le cas d'un modèle semi-paramétrique avec une seule covariable. Cependant, les explications sont facilement généralisables à l'implémentation d'un modèle avec plusieurs covariables. La méthodologie programmée sera brièvemment rappelée en début de section dans le cas particulier d'une seule covariable. Considérons les intensités de transition associées à chaque individu i,

$$\alpha_{hji}(t) = \alpha_{hj0}(t)\exp(\beta_{hj}Z_i(t)) \qquad , i = 1, ..., n,$$

où $Z_i(t)$ et β_{hj} sont des réels.

L'objectif est de décrire la programmation de :
- l'estimation les coefficents β_{hj} de manière paramétrique en maximisant la vraisemblance partielle ;
- l'estimation des variances des coefficients de régression ;
- l'estimation de $A_{hj0}(t) = \int_0^t \alpha_{hj0}(u)du$ de manière non-paramétrique par l'estimateur de Nelson-Aalen ;
- l'estimation de la matrice des probabilités de transition $\mathbf{P}(s,t)$.

Remarque 26

- *Les estimatueurs non-paramétriques (cf. chapitre IV page 81) sont obtenus en fixant* $\beta_{hj} = 0$.
- *Le méthode fonctionne pour des variables qualitatives et quantitatives. Cependant, l'utilisation de variables quantitatives est délicate car il y a un estimateur semi-paramétrique de la matrice des probabilités de transition pour chaque valeur de la covariable.*
- *La méthode d'estimation reste valable pour des covariables dépendantes du temps. L'estimation des coefficients de régression fournit des résultats interprétables ; par contre, l'interprétation des probabilités de transition pour des covariables dépendantes du temps est délicate : il y a un estimateur pour chaque "histoire" de covariable.*
- *La programmation des estimateurs IPCW (cf. chapitre V page 109) suit la même démarche que l'estimation semi-paramétrique. Il faut calculer les pondérations et les inclures dans les estimateurs semi-paramétriques.*

1 Données requises pour l'estimation

1.1 Base d'entrée

Les données nécessaires pour l'utilisation du programme se présentent sous forme d'une matrice avec 6 colonnes (dans le cas d'un modèle avec une covariable). Le nombre de lignes est égal au nombre total de consultations présentes dans la base de données. L'information pour un patient est donnée par au moins 2 lignes dans la base de données (1 ligne pour chaque consultation). Les colonnes de la matrice sont :

- ID : identifiant du sujet, chaque identifiant est répété plusieur fois ;
- TEMPS0 : représente le temps d'entrée dans ETAT0. Pour la première ligne de chaque patient, TEMPS0 vaut 0 ;
- ETAT0 : représente l'état de l'individu pendant l'intervalle de temps]TEMPS0 ;TEMPS1] ;
- TEMPS1 : représente le temps de transition et peut être un temps d'entrée dans ETAT1 ou un temps de censure ;
- ETAT1 : représente l'état d'arrivée d'une transition de ETAT0. Pour la dernière ligne associée à un patient, ETAT1 est soit un état absorbant soit un état de censure ;
- COVARIABLE : représente la valeur de la covariable pendant l'intervalle de temps]TEMPS0 ;TEMPS1].

1.2 Quantités à calculer

Certaines quantités doivent être calculées afin d'obtenir les estimations.

- *Temps* : un vecteur donnant tous les temps de transition différents, classés par ordre croissant . Vecteur de longueur l.
- *Patient* : un vecteur donnant les identifiants des patients (tous différents). Vecteur de longueur n.

- $MatriceYh$: une matrice renseignant sur l'état des individus pour chaque temps. Matrice de dimension $l \times n$, pour $i = 1, ..., n$, et $k = 1, ..., l$, $MatriceYh[k, i] = 1$ si l'individu i est à risque dans l'état h au $Temps[k]$, 0 sinon. La colonne i de $MatriceYh$ correspond à la quantité « $Y_{hi}(t)$ ».
- Yh : une matrice renseignant sur le nombre total d'individus à risque dans l'état h pour chaque temps. Le vecteur Yh (longueur l) correspond à la quantité « $Y_{h+}(t)$ ».
- Jh : un vecteur de longueur l correspondant à la quantité « $J_h(t)$ ».
- $MatriceNhj$: une matrice renseignant sur les transitions observées de l'état h vers l'état j. Matrice de dimension $l \times n$, pour $i = 1, ..., n$, et $k = 1, ..., l$, $MatriceNhj[k, i] = 1$ si l'individu i subit une transition de l'état h vers l'état j à $Temps[k]$, 0 sinon. La colonne i de $MatriceNhj$ correspond à la quantité « $\Delta N_{hji}(t)$ ».
- Nhj : un vecteur renseignant sur le nombre total de transitions observées de l'état h vers l'état j à chaque temps. Le vecteur Nhj (longueur l) correspond à la quantité « $\Delta N_{hj+}(t)$ ».
- $MatriceCov$: un vecteur renseignant sur la valeur de la covariable à chaque temps. Matrice de dimension $l \times n$, pour $i = 1, ..., n$, et $k = 1, ..., l$, $MatriceCov[k, i]$ donne la valeur de la covariable pour l'individu i à l'instant $Temps[k]$. La colonne i de $MatriceCov$ correspond à la quantité « $Z_i(t)$ ».

1.3 Programmation des quantités

- $MatriceYh$, $MatriceNhj$ et $MatriceCov$ peuvent être programmées dans la même boucle :
  Initialiser les matrices à 0.
  Faire une double boucle pour $i = 1, ..., n$, et $k = 1, ..., l$.
  Sélectionner les lignes de la matrice d'entrée correspondant à l'individu i. Ensuite :
  - pour $MatriceYh$: sélectionner la ligne pour laquelle TEMPS0 $< Temps[k] \leqslant$ TEMPS1. A partir de ce vecteur, si l'élément ETAT0 $= h$ alors $MatriceYh[k, i] = 1$ sinon $MatriceYh[k, i] = 0$. Si ce vecteur n'existe pas, l'individu n'est pas à risque pour ce temps.
  - pour $MatriceNhj$: sélectionner la ligne pour laquelle TEMPS1 $= Temps[k]$. A partir de ce vecteur, si les éléments ETAT0 $= h$ et ETAT1 $= j$ alors $MatriceNhj[k, i] = 1$ sinon $MatriceNhj[k, i] = 0$. Si ce vecteur n'existe pas, il n'y a pas de transition pour ce temps.
  - pour $MatriceCov$: sélectionner la ligne pour laquelle TEMPS0 $< Temps[k] \leqslant$ TEMPS1. A partir de ce vecteur, $MatriceCov[k, i] =$ COVARIABLE. Quand ce vecteur n'existe pas, la valeur de la covariable n'est pas nécessaire.
- Yh et Nhj sont obtenus en faisant la somme des colonnes de $MatriceYh$ et $MatriceNhj$, respectivement.

2 Fonction de la log-vraisemblance partielle

Afin d'obtenir l'estimation des coefficients de régression, il faut maximiser le logarithme de la vraisemblance partielle. Pour un modèle avec une seule covariable :

$$LogVrais = \sum_{h,j=1,...,s} \left[\underbrace{\sum_{i=1}^{n} \int_0^\tau \beta_{hj} Z_i(t) dN_{hji}(t)}_{=F_1} - \underbrace{\int_0^\tau \log(S_{hj}^{(0)}(\beta,t)) dN_{hj+}(t)}_{=F_2} \right],$$

avec

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

Définissons *betahj* un nombre réel. Afin de maximiser la vraisemblance, l'utilisateur doit programmer une fonction qui reçoit en entrée un vecteur de paramètres et renvoie en sortie la valeur de la log-vraisemblance.

1. Calcul de $S_{hj}^{(0)}(\beta,t) = \sum_{i=1}^{n} Y_{hi}(t) \exp(\beta_{hj} Z_i(t))$:
   - Considérer $Shj^{(0)}$ un vecteur de longueur l (la valeur de $Shj^{(0)}$ sera renseignée pour tous les temps de transition).
   - Définir $MatriceCalcul$ une matrice avec l lignes (nombre de temps de transition différents) et n colonnes (nombre d'individus).
     $for\ (i = 1, ..., n)$,
     $MatriceCalcul[,i] = MatriceYh[,i] \times \exp(betahj \times MatriceCov[,i])$
     Pour calculer $Shj^{(0)}$, il faut calculer la somme des colonnes de $MatriceCalcul$ (permet de calculer$\sum_{i=1}^{n}$).

2. Calcul de $F_2 = \int_0^\tau \log(S_{hj}^{(0)}(\beta,t)) dN_{hj}(t)$:
   - Considérer F_2 un réel.
   - Soit $VecteurCalcul$ un vecteur de longueur l,
     $VecteurCalcul = \log(Shj^{(0)}) \times Nhj$,
     Le processus de comptage $N_{hj.}(t)$ est une fonction en escalier, ainsi, $\int_0^\tau$ est une somme finie des contributions à chaque temps. Pour obtenir F_2, il faut calculer la somme de chaque coordonnée de $VecteurCalcul$.

3. Calcul de $F_1 = \sum_{i=1}^{n} \int_0^\tau \beta_{hj} Z_i(t) dN_{hji}(t)$:
   - Considérer F_1 un réel.
   - Définir $MatriceCalcul$ une matrice avec l lignes et n colonnes,
     $for\ (i = 1, ..., n)$,
     $MatriceCalcul[,i] = MatriceNhj[,i] \times \exp(betahj \times MatriceCov[,i])$
     Pour calculer F_1 il faut calculer la somme de chaque élément de la matrice (pour le calcul de $\sum_{i=1}^{n}$ et $\int_0^\tau$)

4. Minimisation de $-LogVrais$:
   - La contribution à la vraisemblance pour *betahj* est la somme de F_1 et F_2. La vraisemblance totale est la somme de toutes les contributions pour $h, j = 1, ..., s$, $h \neq j$.
   - Les paramètres initiaux sont fixés à 0 (aucun effet de la covariable).

– La fonction "*nlminb*" de *Splus* ou la fonction "*optim*" de R peuvent être utilisées pour minimiser $-LogVrais$. La fonction "*optim*" de R fournit la matrice hessienne et permet ainsi d'obtenir les écart-types. Cependant, les écart-types peuvent également être programmés ($cf.$ section suivante).

Remarque 27 *Dans ce cas simple, la log-vraisemblance peut être maximisée séparément (car la log-vraisemblance est une somme de plusieurs termes où chaque terme comprend un seul paramètre à estimer). Ainsi, il suffit de calculer F_1 et F_2 pour un couple h, j fixé, et de maximiser la somme pour obtenir la valeur du coefficient β_{hj}.*

3 Ecart-types des coefficients de régression

3.1 Méthodologie pour une covariable

$$LogVrais = \sum_{h,j=1,\ldots,s} \sum_{i=1}^{n} \int_0^\tau \left[\beta_{hj} Z_i(t) - \log(S_{hj}^{(0)}(\beta,t))\right] dN_{hji}(t).$$

$$Hessienne = \begin{pmatrix} \frac{\partial^2 LogVrais}{\partial\beta_{11}\partial\beta_{11}} & \cdots & \frac{\partial^2 LogVrais}{\partial\beta_{11}\partial\beta_{ss}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 LogVrais}{\partial\beta_{ss}\partial\beta_{11}} & \cdots & \frac{\partial^2 LogVrais}{\partial\beta_{ss}\partial\beta_{ss}} \end{pmatrix}.$$

Dans la dérivée de la log-vraisemblance par rapport à β_{hj}, la somme $\sum_{h,j=1,\ldots,s}$ disparaît car seul le terme avec β_{hj} est différent de 0.

$$\frac{\partial LogVrais}{\partial\beta_{hj}} = \sum_{i=1}^{n} \int_0^\tau \left[Z_i(t) - \frac{S_{hj}^{(1)}(\beta,t)}{S_{hj}^{(0)}(\beta,t)} \right] dN_{hji}(t),$$

avec

$$\begin{aligned} S_{hj}^{(1)}(\beta,t) &= \frac{\partial S_{hj}^{(0)}(\beta,t)}{\partial\beta_{hj}} \\ &= \sum_{i=1}^{n} Y_{hi}(t) Z_i(t) \exp(\beta_{hj} Z_i(t)). \end{aligned}$$

Les dérivées secondes sont,

$$\frac{\partial^2 LogVrais}{\partial\beta_{hj}\partial\beta_{uv}} = 0,$$

$$\frac{\partial^2 LogVrais}{\partial\beta_{hj}\partial\beta_{hj}} = \sum_{i=1}^{n} \int_0^\tau - \left[\frac{S_{hj}^{(2)}(\beta,t) S_{hj}^{(0)}(\beta,t) - \left(S_{hj}^{(1)}(\beta,t)\right)^2}{\left(S_{hj}^{(0)}(\beta,t)\right)^2} \right] dN_{hji}(t),$$

avec

$$S_{hj}^{(2)}(\beta,t) = \frac{\partial^2 S_{hj}^{(0)}(\beta,t)}{\partial\beta_{hj}\partial\beta_{hj}} = \frac{\partial S_{hj}^{(1)}(\beta,t)}{\partial\beta_{hj}}$$
$$= \sum_{i=1}^{n} Y_{hi}(t)\,(Z_i(t))^2 \exp(\beta_{hj}Z_i(t)).$$

La dérivée seconde de la log-vraisemblance par rapport à β_{hj} peut aussi s'écrire comme suit (car le terme sous l'intégrale ne dépend pas de i) :

$$\frac{\partial^2 LogVrais}{\partial\beta_{hj}\partial\beta_{hj}} = -\int_0^\tau \left[\frac{S_{hj}^{(2)}(\beta,t)S_{hj}^{(0)}(\beta,t) - \left(S_{hj}^{(1)}(\beta,t)\right)^2}{\left(S_{hj}^{(0)}(\beta,t)\right)^2}\right] dN_{hj}(t).$$

La matrice de variance-covariance est égale à moins l'inverse de la matrice hessienne. Comme les termes qui ne sont pas sur la diagonale sont nuls, la variance des béta est donnée par (inverse d'une matrice diagonale) :

$$Var(\beta_{hj}) = -\frac{1}{\frac{\partial^2 LogVrais}{\partial\beta_{hj}\partial\beta_{hj}}}.$$

3.2 Programmation pour une covariable

Soit $\widehat{betahj}$ l'estimateur du maximum de vraisemblance de $betahj$.

1. Calcul de $S_{hj}^{(0)}(\hat{\beta},t)$, $S_{hj}^{(1)}(\hat{\beta},t)$, $S_{hj}^{(2)}(\hat{\beta},t)$:

$$S_{hj}^{(0)}(\hat{\beta},t) = \sum_{i=1}^{n} Y_{hi}(t)\exp(\hat{\beta}_{hj}Z_i(t)),$$
$$S_{hj}^{(1)}(\hat{\beta},t) = \sum_{i=1}^{n} Y_{hi}(t)Z_i(t)\exp(\hat{\beta}_{hj}Z_i(t)),$$
$$S_{hj}^{(2)}(\hat{\beta},t) = \sum_{i=1}^{n} Y_{hi}(t)\,(Z_i(t))^2\exp(\hat{\beta}_{hj}Z_i(t)).$$

Soit $MatriceCalcul$ une matrice avec l lignes et n colonnes
$Shj^{(0)}$, $Shj^{(1)}$, $Shj^{(2)}$ sont des vecteurs de longueur l.

– $\underline{Shj^{(0)}:}$
$for\ (i = 1, ..., n)$
$MatriceCalcul[,i] = MatriceYh[,i] \times \exp(\widehat{betahj} \times MatriceCov[,i])$
$Shj^{(0)}$ est la somme des colonnes de $MatriceCalcul$.

– $\underline{Shj^{(1)}:}$
$for\ (i = 1, ..., n)$
$MatriceCalcul[,i] = MatriceYh[,i] \times MatriceCov[,i] \times \exp(\widehat{betahj} \times MatriceCov[,i])$
$Shj^{(1)}$ est la somme des colonnes de $MatriceCalcul$.

– $\underline{Shj^{(2)}}$:
$for\ (i = 1, ..., n)$,
$MatriceCalcul[, i] = MatriceYh[, i] \times (MatriceCovariable[, i])^2 \times \exp(\widehat{betahj} \times MatriceCovariable[, i])$
$Shj^{(2)}$ est la somme des colonnes de $MatriceCalcul$.

2. Calcul de $\frac{\partial^2 LogVrais}{\partial\hat{\beta}_{hj}\partial\hat{\beta}_{hj}}$:
Soit $VecteurCalcul$ un vecteur de longueur l,
$VecteurCalcul = -\frac{\left[Shj^{(2)}Shj^{(0)} - \left(Shj^{(1)}\right)^2\right]}{(Shj^{(0)})^2} \times Nhj$, avec la convention $0/0 = 0$.
$\frac{\partial^2 LogVrais}{\partial\hat{\beta}_{hj}\partial\hat{\beta}_{hj}}$ est la somme des coordonnées du vecteur $VecteurCalcul$.

3. Calcul des écart type :
Soit $sd(\hat{\beta}_{hj})$ l'écart type de $\widehat{betahj}$
$sd(\hat{\beta}_{hj}) = \sqrt{-\frac{1}{\frac{\partial^2 LogVrais}{\partial\beta_{hj}\partial\beta_{hj}}}}$

3.3 Méthodologie pour deux covariables

Dans le cas de plusieurs covariables, la matrice hessienne n'est plus une matrice diagonale. Il faut calculer toutes les composantes de la matrice hessienne afin de l'inverser pour obtenir la matrice de variance-covariance.

$$LogVrais = \sum_{h,j=1,...,s}\sum_{i=1}^{n}\int_0^\tau \left[(\beta_{hj}Z_i^1(t) + \alpha_{hj}Z_i^2(t)) - \log(S_{hj}(\beta, t))\right] dN_{hji}(t),$$

$$S_{hj}^{(0)}(t) = \sum_{i=1}^{n} Y_{hi}(t)\exp(\beta_{hj}Z_i^1(t) + \alpha_{hj}Z_i^2(t)),$$

Dans la dérivée de la log-vraisemblance par rapport à β_{hj} ou α_{hj} la somme $\sum_{h,j=1,...,s}$ disparait.

$$\frac{\partial LogVrais}{\partial\beta_{hj}} = \sum_{i=1}^{n}\int_0^\tau \left[Z_i^1(t) - \frac{\partial_{\beta_{hj}}S_{hj}^{(0)}(t)}{S_{hj}^{(0)}(t)}\right] dN_{hji}(t),$$

$$\frac{\partial LogVrais}{\partial\alpha_{hj}} = \sum_{i=1}^{n}\int_0^\tau \left[Z_i^2(t) - \frac{\partial_{\alpha_{hj}}S_{hj}^{(0)}(t)}{S_{hj}^{(0)}(t)}\right] dN_{hji}(t),$$

avec

$$\partial_{\beta_{hj}}S_{hj}^{(0)}(t) = \frac{\partial S_{hj}^{(0)}(\beta, t)}{\partial\beta_{hj}} = \sum_{i=1}^{n} Y_{hi}(t)Z_i^1(t)\exp(\beta_{hj}Z_i^1(t) + \alpha_{hj}Z_i^2(t)),$$

$$\partial_{\alpha_{hj}} S_{hj}^{(0)}(t) = \frac{\partial S_{hj}^{(0)}(\beta, t)}{\partial \alpha_{hj}} = \sum_{i=1}^{n} Y_{hi}(t) Z_i^2(t) \exp(\beta_{hj} Z_i^1(t) + \alpha_{hj} Z_i^2(t)).$$

Les termes différents de zéro dans la matrice hessienne sont $\forall h, j = 1, ..., s$

$$\frac{\partial^2 LogVrais}{\partial \beta_{hj} \partial \beta_{hj}} = \sum_{i=1}^{n} \int_0^{\tau} - \left[\frac{\partial^2_{\beta_{hj}\beta_{hj}} S_{hj}^{(0)}(t) S_{hj}^{(0)}(t) - \left(\partial_{\beta_{hj}} S_{hj}^{(0)}(t)\right)^2}{\left(S_{hj}^{(0)}(t)\right)^2} \right] dN_{hji}(t),$$

$$\frac{\partial^2 LogVrais}{\partial \alpha_{hj} \partial \alpha_{hj}} = \sum_{i=1}^{n} \int_0^{\tau} - \left[\frac{\partial^2_{\alpha_{hj}\alpha_{hj}} S_{hj}^{(0)}(t) S_{hj}^{(0)}(t) - \left(\partial_{\alpha_{hj}} S_{hj}^{(0)}(t)\right)^2}{\left(S_{hj}^{(0)}(t)\right)^2} \right] dN_{hji}(t),$$

$$\begin{aligned} \frac{\partial^2 LogVrais}{\partial \beta_{hj} \partial \alpha_{hj}} &= \frac{\partial^2 LogVrais}{\partial \alpha_{hj} \partial \beta_{hj}} \\ &= \sum_{i=1}^{n} \int_0^{\tau} - \left[\frac{\partial^2_{\alpha_{hj}\beta_{hj}} S_{hj}^{(0)}(t) S_{hj}^{(0)}(t) - \partial_{\alpha_{hj}} S_{hj}^{(0)}(t) \partial_{\beta_{hj}} S_{hj}^{(0)}(t)}{\left(S_{hj}^{(0)}(t)\right)^2} \right] dN_{hji}(t). \end{aligned}$$

Avec

$$\partial^2_{\beta_{hj}\beta_{hj}} S_{hj}^{(0)}(t) = \sum_{i=1}^{n} Y_{hi}(t) \left(Z_i^1(t)\right)^2 \exp(\beta_{hj} Z_i^1(t) + \alpha_{hj} Z_i^2(t)),$$

$$\partial^2_{\alpha_{hj}\alpha_{hj}} S_{hj}^{(0)}(t) = \sum_{i=1}^{n} Y_{hi}(t) \left(Z_i^2(t)\right)^2 \exp(\beta_{hj} Z_i^1(t) + \alpha_{hj} Z_i^2(t)),$$

$$\partial^2_{\beta_{hj}\alpha_{hj}} S_{hj}^{(0)}(t) == \partial^2_{\alpha_{hj}\beta_{hj}} S_{hj}^{(0)}(t) = \sum_{i=1}^{n} Y_{hi}(t) Z_i^1(t) Z_i^2(t) \exp(\beta_{hj} Z_i^1(t) + \alpha_{hj} Z_i^2(t)).$$

La matrice de variance-covariance est égale à moins l'inverse de la matrice hessienne.

3.4 Programmation pour deux covariables

La programmation des écart-types pour un modèle avec deux covariables se fait en suivant la démarche de la programmation pour une covariable.

4 Estimateur de Nelson-Aalen

L'estimateur de Nelson-Aalen du risque de base est,

$$\hat{A}_{hj0}(t;\hat{\beta}_{hj}) = \int_0^t \frac{J_h(u)}{\sum_{i=1}^n Y_{hi}(u)\exp(\hat{\beta}_{hj}Z_i(u))} dN_{hj+}(u), \qquad h \neq j,$$

- Calculer le vecteur $Shj^{(0)}(\widehat{betahj})$ de longueur l comme précédemment (valeur de $Shj^{(0)}$ pour les valeurs de $betahj$ qui maximise la vraisemblance partielle).
- Soit $VecteurCalcul$ un vecteur de longueur l,
  $VecteurCalcul = \frac{Jh}{Shj^{(0)}(\widehat{betahj})} \times Nhj$, avec la convention $0/0 = 0$.
- Programmer le vecteur $Ahj0$ de longueur l,
  $for\ (k = 1, ..., l)$,
  $Ahj0[k] = \sum_{u=1}^k VecteurCalcul[u]$ (la somme des k premières coordonnées du vecteur $VecteurCalcul$ permet de calculer $\int_0^t$).

5 Estimateur des probabilités de transition

La matrice des probabilités de transition est définie par

$$\begin{aligned}
\hat{\mathbf{P}}(s,t \mid Z_i(t)) &= \prod_{]s,t]} (\mathbf{I} + d\hat{\mathbf{A}}(u \mid Z_i(t))), \\
d\hat{A}_{hj}(t \mid Z_i(t)) &= \frac{J_h(t)}{S_{hj}^{(0)}(\hat{\beta}_{hj}, t)} \times \Delta N_{hj.}(t) \times \exp(\hat{\beta}_{hj}Z_i(t)), \qquad h \neq j, \\
d\hat{A}_{hh}(t \mid Z_i(t)) &= -\sum_{j \neq h} d\hat{A}_{hj}(t \mid Z_i(t)), \qquad h = 1, ..., s,
\end{aligned}$$

avec

$$\mathbf{I} + d\hat{\mathbf{A}} = \begin{pmatrix} 1 - \sum_{j\neq 1} d\hat{A}_{1j} & d\hat{A}_{12} & \cdots & d\hat{A}_{1s} \\ d\hat{A}_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & d\hat{A}_{s-1s} \\ d\hat{A}_{s1} & \cdots & d\hat{A}_{ss-1} & 1 - \sum_{j\neq s} d\hat{A}_{sj} \end{pmatrix}.$$

- calcul de $d\hat{\mathbf{A}}(u \mid Z_i(t))$ (associé à l'individu i) :
  Définir un tableau dA de dimension $s \times s \times l$ (s nombre d'états et l nombre de temps d'événements) et calculer :
  $dA[h, j,] = \frac{Jh}{Shj^{(0)}(\widehat{betahj})} \times Nhj \times \exp(\widehat{betahj} \times MatriceCov[, i])\ \forall h \neq j$,
  $dA[h, h,] = -\sum_{j\neq h} dA[h, j,]$.

– calcul de $\hat{\mathbf{P}}(0,t \mid Z_i(t))$:
Définir un tableau P de dimension $s \times s \times l$, la $(s \times s)$-matrice identité I et calculer :
$P[,,1] = I + dA[,,1]$,
$P[,,k] = P[,,k-1]\% \times \%(I + dA[,,k])$, $k = 2, ..., l$, où $\% \times \%$ est la multiplication matricielle.

Remarque 28 *La probabilité* $\mathbf{P}(0,t)$ *est souvent utilisée pour l'interprétation des résultats.* $p_{hj}(0,t)$ *représente le probabilité d'être dans l'état* h *à l'instant* 0 *et dans l'état* j *à l'instant* t.

6 Test de l'hypothèse de proportionnalité des risques

L'hypothèse de proportionnalité des risques peut être testée par l'intermédiaire d'une covariable artificielle en suivant la démarche présentée au chapitre IV page 96. La méthode d'implémentation présentée précédemment doit être adaptée au cas de deux covariables afin de tester l'effet de la covariable artificielle. La programmation de la covariable artificielle dépendante du temps suit la démarche permettant la programmation de $MatriceCov$.

Annexe C

Définition des états de contrôle

L'état de contrôle est défini à partir de plusieurs variables de la base de données. Chaque variable présentée ci-dessous permet d'attribuer un état de santé à un patient à chaque consultation. L'état de santé finalement choisi sera le maximum des états obtenus avec les critères pris en compte. Cet état sera appelé état de contrôle. L'état **optimal** est symbolisé par l'**état 1**, l'état **sous-optimal** par l'**état 2** et l'état **inacceptable** par l'**état 3**.

Définition des exacerbations

Les variables se rapportant aux exacerbations sont les variables *E-Type* (nature des épisodes) et *E-corticoïde* (consommation ou non de corticoïdes).

Les modalités de la variable *E-Type* sont :
- aucun
- instabilité
- attaque
- crise_grave_hôpital
- crise_grave_maison
- crise_grave_réa

Les modalités de la variable *E-corticoïde* sont :
- FALSE
- TRUE

Les patients sont classés dans les états de la manière suivante :

- si *E-Type* est codée « crise_grave_$\cdots$ » et *E-corticoïde* est codée « TRUE », alors ce patient est classé dans l'**état 3** ;
- si *E-Type* est codée « crise_grave_$\cdots$ » et *E-corticoïde* est codée « FALSE », alors ce patient est classé dans l'**état 2** ;
- Dans tous les autres cas, le patient est classé dans l'**état 1**.

Consommation de beta2 associée à la présence de symptômes

Les variables *entree_beta2_spray_dose* et *entree_beta2_spray_unite* renseignent sur la quantité de béta2 consommée. La présence de symptôme est renseignée par la variable *crise_frequence_recent*.

La variable *entree_beta2_spray_dose* est quantitative (elle varie de 0 à 30).
Les modalités de la variable *entree_beta2_spray_unite* sont :
- bouffées/semaine
- bouffées/jour

Les modalités de la variable *crise_frequence_recent* sont :
- absente
- une_par_mois
- moins_une_par_semaine
- une_par_semaine
- plusieurs_par_semaine
- une_par_jour
- plusieurs_par_jour

Les patients sont classés dans les états de la manière suivante :

- si la quantité de béta2 consommée est supérieure (ou égale) à 1 bouffée par jour et si la fréquence des crises est supérieure (ou égale) à 1 par jour, alors le patient est classé dans l'**état 3** ;
- si le patient ne vérifie pas les conditions pour être classé dans l'état 3 et si la quantité de béta2 consommée est supérieure (ou égale) à 1 bouffée par semaine et si la fréquence des crise est supérieure (ou égale) à 1 par semaine, alors le patient est classé dans l'**état 2** ;
- dans tous les autres cas (la quantité de béta2 consommée est inférieure à 1 bouffée par semaine ou si la fréquence des crise est inférieure (ou égale) à 1 par semaine), le patient est classé dans l'**état 1**.

VEMS mesuré (Volume Maximal Expiré en une Seconde)

La variable correspondante à ce critère est : *VEMS_basal_rapport*. Elle indique le rapport du VEMS mesuré sur le VEMS théorique. Une baisse de plus de 30% du VEMS mesuré par rapport au VEMS maximal mesuré chez un patient amène à classer ce patient dans l'état **3**, une chute comprise entre 10 et 30% dans l'état **2** , et sinon le patient est classé dans l'état **1**.

Autrement dit :

- $VEMS\_basal\_rapport < 70\%$ correspond à l'**état 3** ;
- $70\% < VEMS\_basal\_rapport < 90\%$ correspond à l'**état 2** ;
- $VEMS\_basal\_rapport > 90\%$ correspond à l'**état 1**.

Dyspnée

Les variables correspondant à ce critère sont les variables *Dyspnee_effort_recent* et *Dyspnee_intercritique_recent*.

Les modalités de la variable *Dyspnee_effort_recent* sont :

– absent
– effort_violent
– pas_rapide
– propre_pas
– moindre_effort

Les modalités de la variable *Dyspnee_intercritique_recent* sont :
– TRUE
– FALSE

Les patients sont classés dans les états de la manière suivante :

- si la variable *Dyspnee_intercritique_recent* est codée « TRUE », alors le patient est classé dans l'**état 3** ;
- lorsque la variable *Dyspnee_intercritique_recent* est codée « FALSE » :
  - si la variable *Dyspnee_effort_recent* est codée « moindre_effort » ou « propre_pas », alors le patient est classé dans l'**état 3** ;
  - si la variable *Dyspnee_effort_recent* est codée « pas rapide », alors le patient est classé dans l'**état 2** ;
  - si la variable *Dyspnee_effort_recent* est codée « effort_violent » ou « absent », alors le patient est classé dans l'**état 1**.

Résumé

Dans de nombreux domaines, décrire l'évolution des phénomènes dans le temps est d'un intérêt capital, en particulier pour aborder les problématiques de la prédiction et de la recherche de facteurs causaux. En épidémiologie, on dispose de données de cohorte qui renseignent sur un groupe de patients suivis dans le temps. Les modèles multi-états de type Markovien proposent un outil intéressant qui permet d'étudier l'évolution d'un patient à travers les différents stades d'une maladie. Dans ce manuscript, nous rappelons tout d'abord la méthodologie 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 étudions un modèle semi-Markovien homogène qui suppose que les intensités de transition dépendent du temps écoulé dans un état de santé. La théorie des processus de comptage est ensuite présentée afin d'introduire des méthodes d'estimations non-paramétriques dans le cadre d'un modèle de Markov non-homogène. Dans ce modèle, les intensités de transition dépendent du temps depuis l'inclusion dans l'étude. Les méthodes d'estimation supposent que le mécanisme de censure n'apporte aucune information sur l'évolution de la maladie. Cette hypothèse étant rarement vérifiée en pratique, nous proposons une méthode d'estimation permettant de prendre en compte une censure informative. Nous présentons également une méthode de programmation visant à faciliter la mise en oeuvre des estimateurs basés sur les processus de comptage. Toute ces méthodes sont appliquées afin d'étudier une base de données de patients asthmatiques. L'objectif est d'aider les cliniciens à mieux comprendre l'évolution de la maladie. Les résultats permettent de mettre en évidence l'impact négatif du surpoids sur l'évolution de l'asthme.

Mots-clés : Asthme, modèles multi-états, processus de Markov homogène et non-homogène, processus semi-Markovien, processus de comptage, censure informative.

Abstract

In many fields, describe the evolution of the phenomena in time is of a capital interest, in particular to approach the problems of prediction and research for causal factors. In epidemiology, one has cohort data which inform about a group of patients followed in time. The multi-states Markov type models propose an interesting tool which makes it possible to study the evolution of a patient through the various stages of a disease. First of all, we recall the methodology relating to the homogeneous Markov model. This model is the least complex, it supposes that the intensities of transition between the states are constant in time. In the second time, we study a homogeneous semi-Markov model which supposes that the transition intensities depend on the time spend in the health state. The theory of counting processes is then presented in order to introduce a non-parametric estimation methods within the framework of a non-homogeneous Markov model. In this model, the transition intensities depend on time since inclusion in the study. Most of the estimation methods suppose that the mechanism of censoring does not bring any information on the evolution of the disease. This assumption seldom being checked in practice, we propose an estimation method allowing to take into account an informative censoring. We also present a programmation guideline aiming to facilitate the implementation of the estimators based on counting processes. These methods are applied in order to study a data base of asthmatic patients. The objective is to help the clinicians with a better understanding of the disease evolution. The results make it possible to highlight the negative impact of overweight on the asthma evolution.

Keyword : Asthma, multi-states models, homogeneous and non-homogeneous Markov process, semi-Markov process, counting process, informative censoring.