Contribution à la statistique des processus : modélisation et applications

Anne Gegout-Petit

UNIVERSITÉ BORDEAUX I

ECOLE DOCTORALE MATHÉMATIQUE ET INFORMATIQUE

MÉMOIRE

pour obtenir

L'HABILITATION À DIRIGER DES RECHERCHES

Mention : Mathématiques appliquées

Présenté par
Anne Gégout-Petit

Contribution à la statistique des processus : modélisation et applications

préparé à l'Institut Mathématiques de Bordeaux
et à INRIA Bordeaux Sud-Ouest, Equipe CQFD

19 novembre 2012

Jury :

Rapporteurs :	Odd Aalen	Université d'Oslo
	Fabienne Comte	Université Paris Descartes
	Jean-François Delmas	Ecole Nationale des Ponts et Chaussées
Présidente :	Sylvie Méléard	Ecole Polytechnique
Examinateurs :	Daniel Commenges	INSERM
	François Dufour	Institut Polytechnique de Bordeaux
	Etienne Pardoux	Aix-Marseille Université.

Remerciements

Je tiens tout d'abord à remercier chaleureusement Odd Aalen, Fabienne Comte et Jean-François Delmas qui, en cette période chargée de dossiers à rédiger, ont accepté de prendre le temps d'évaluer ce document de synthèse. Je suis bien évidemment très honorée de recevoir des avis positifs de personnalités scientifiques aussi prestigieuses.

Sylvie Méléard a été témoin de tous les moments clés de ma carrière y compris ceux qui ne m'ont pas menée tout droit vers cette habilitation. J'ai toujours apprécié sa lecture profondément humaine et sensible des situations et je suis particulièrement touchée de sa présence à Bordeaux pour cette nouvelle étape. Daniel Commenges m'a accueillie dans son équipe de biostatistique de l'INSERM alors que j'étais PRAG à l'Institut de Santé Publique de Bordeaux et a su créer un environnement propice à la reprise de mes activités de recherche, je l'en remercie. Je suis aussi tout particulièrement reconnaissante à François Dufour dont j'apprécie chaque jour les qualités professionnelles et humaines et qui m'a accordé sa confiance en me proposant de faire partie du projet INRIA CQFD, me permettant ainsi de bénéficier d'un environnement scientifique des plus dynamisants. L'aura et l'énergie d'Etienne Pardoux accompagnent ses étudiants longtemps après la thèse et c'est un très grand plaisir pour moi qu'il participe à ce jury.

Cette habilitation est aussi celle de mes coauteur-e-s : Benoîte de Saporta avec laquelle je confronte idées, humeurs et eaux chaudes presque chaque jour et bien sûr Laurence Marsalle qui rétablit l'ordre alphabétique entre femmes et hommes et complète efficacement notre trio ; Bernard Bercu qui a fait bifurquer mes recherches vers les BAR (j'explore toujours la branche !), Daniel Commenges, François Dufour et Jérôme Saracco. J'aime confronter les disciplines avec Marie Touzet et Lucia Guérin. Romain Azaïs et Camille Baysse sont de biens charmants cobayes sur lesquels j'ai exercé la direction de recherche avant cette habilitation.

Cette habilitation est aussi celle des personnes qui m'ont encouragée à la préparer, parmi elles mon mari Michel est sûrement le premier, il y a aussi beaucoup d'autres personnes dont Avner Bar-Hen, Gérard Biau, Thierry Colin, Michel Langlais, ...

Je n'aurais pu réaliser ce travail sans un environnement professionnel porteur et épanouissant : celui de l'ISPED et son équipe pédagogique des plus stimulantes : Marthe-Aline, Pierre, Valérie, Alioum notamment ; de l'UFR Sciences et Modélisation où règne un doux climat de "résistance" pour continuer à exercer au mieux notre mission de service public auprès des étudiants : Frédérique, Manue, Brigitte, Olivier, Vincent, Pierrick... et celui de mes collègues de l'IMB et d'INRIA (plusieurs ont déjà été cités) : Héloïse, Lisl, Cécile avec une mention toute spéciale à Marie Chavent et Jérôme Saracco. Ces derniers m'ont embarqué dans l'aventure JdS 2009 et SFdS qui a permis de fructueuses et enrichissantes collaborations avec Gérard Biau et Jean-Michel Poggi notamment. L'inclassable (et presque toujours formidable !) Ingrid Rochel m'assiste efficacement et assure l'ambiance dans ces différentes structures. J'aurais aimé dire plus et mieux sur chacune des personnes citées ici ainsi qu'à beaucoup d'autres personnes encore. A tous, je vous redis merci.

Last but not least, mes trois grands enfants ont su me dire qu'ils tenaient à ce que je persévère dans la voie scientifique, je leur dédie ce modeste mémoire.

Table des matières

Introduction

Mes travaux de thèse de doctorat qui relèvent du calcul stochastique ont donné lieu à la publication de deux articles dans des revues internationales : le premier porte sur l'existence et l'unicité de la solution d'une Equation Différentielle Rétrograde Réfléchie dans un convexe (1). Le second, dans le prolongement des travaux de Jean Picard (114), propose un filtre approché et établit ses propriétés de convergence dans le cadre d'un processus partiellement observé (fonction d'observation dépendant d'une seule des composantes du processus à observer) (2). Suite à mes différentes affectation sur des postes du secondaire, mon activité de recherche s'est interrompue. Mon affectation en septembre 2001 sur un poste de professeur agrégé détaché à l'Institut de Santé Publique et du Développement (ISPED) de l'Université Bordeaux 2, m'a placée dans un cadre propice à la reprise progressive d'activités de recherche grâce mon intégration de l'équipe INSERM de Biostatistique. Le présent document décrit donc les travaux que j'ai menés depuis l'année 2003 et jusqu'à ce jour.

C'est l'étude ou l'utilisation des processus stochastiques qui unit l'ensemble de ces travaux. Même si certains des résultats sont purement théoriques, l'application potentielle était présente ou imaginée lors de leur élaboration. C'est pourquoi les processus à sauts constituent une partie de l'objet d'étude dont certains très élaborés comme les Processus Markoviens Déterministes par Morceaux (PDMP) ; ces processus sont en effet souvent utilisés pour modéliser des problématique d'histoire de vie ou de survie en biostatistique ou l'état de systèmes complexes dans des problématiques de sûreté de fonctionnement. Une seconde thématique est consacrée à l'estimation de certains paramètres des processus de bifurcation adaptés à la modélisation de la division cellulaire binaire, soit la modélisation de la stricte descendance de la cellule, soit la modélisation d'une valeur quantitative attachée à chacune de ces cellules et modélisée par un processus autorégressif adapté à l'arbre binaire. Enfin, puisque nous sommes "dans la vraie vie" et que dans celle-ci l'observation des phénomènes est souvent partielle ou bruitée, les notions d'observation partielle ou de données non observées ou manquantes sont la plupart du temps prises en compte ou étudiées spécifiquement.

Notre présentation est divisée en quatre chapitres. Le premier présente les problématiques étudiées lorsque j'étais à l'équipe INSERM de biostatistique et traite de l'utilisation des processus pour la modélisation des modèles d'histoire de vie et de survie. Le second traite de nos travaux sur les processus de bifurcation avec application à la division cellulaire. Les travaux réalisés dans le cadre de l'équipe INRIA CQFD, souvent liés aux PDMP et parfois dans le cadre de collaborations industrielles et/ou de projet ANR font l'objet du chapitre suivant. Enfin, au chapitre quatre, nous présentons des collaborations avec des collègues d'autres disciplines qui correspondent à du soutien méthodologique en statistique.

Chapitre 1

Les processus pour l'étude des modèles d'histoire de vie

Ce chapitre présente les travaux réalisés en collaboration avec Daniel Commenges (plus Pierre Joly et Benoît Liquet pour l'article (4)) lorsque j'étais dans l'équipe INSERM de biostatistique E0338 . Cette équipe développe des modèles appliqués en épidémiologie notamment par les autres équipes INSERM de santé publique, spécialistes par exemple de l'épidémiologie du VIH ou de la maladie d'Alzheimer. Mon insertion en tant que probabiliste a été possible par une approche théorique des modèles étudiés. D'un point de vue général, la thématique de recherche est « l'étude mathématique des modèles de survie et d'histoire de vie dans le cadre de données censurées ». Nous avons travaillé sur trois thèmes qui se recoupent : d'une part la modélisation du mécanisme menant aux données manquantes et son lien avec l'inférence sur le processus d'intérêt. Cette problématique, présentée à la section 1.1, est présente dans les trois articles publiés (3), (4) et (7). D'autre part nous avons justifié les écritures de vraisemblance, souvent heuristiques dans les articles de biostatistique, de modèle multi-états partiellement observé. C'est l'objet de la section 1.2 correspondant à l'article (3). Enfin, à la section 1.3.1 nous présentons les notions d'indépendance locale entre processus et le lien avec la causalité correspondant aux deux articles publiés (7) et (8).

1.1 Vraisemblance et mécanisme d'observation

Supposons que l'on veuille faire de l'inférence à partir de données "partiellement observées" ; ce terme ne s'entend pas ici comme le "partiellement observé" du titre de ma thèse mais tente de traduire le "Coarsened Observations" du titre de l'article (3). Le processus d'intérêt n'est pas observé pour tout $t \in \mathbb{R}$ et peut par exemple être censuré à droite (ce qui est standard dans des études de modèles de survie ou d'histoire de vie), ou observé à des dates ponctuelles (visites dans un établissement de santé, ou examens réguliers des études de cohortes) ce qui correspond à une censure par intervalle. On peut bien sûr varier les schémas d'observations en combinant observations ponctuelles et sur des intervalles. Un exemple concret d'observation mêlant observations ponctuelles et par intervalle est donné à la section 5.2. de (3) au sujet d'un modèle à cinq états "Démence-Institutions-Décès".

Pour faire une inférence correcte à partir des données observées, il convient de mo-

déliser le mécanisme d'observation et d'en vérifier le lien avec les données d'intérêt. Soit $\boldsymbol{X} = (X_1, \dots X_p)$ le processus d'intérêt à p composantes étudié sur un intervalle $\mathcal{I} = [0, C]$: $X_j = (X_{jt})_{t\in\mathcal{I}}$ pour $1 \leq j \leq p$. Un mécanisme d'observation déterministe de $\boldsymbol{X}$ est résumé par la donnée de $\boldsymbol{r} = (r_1, \dots, r_p)$ tel que pour tout j et $t \in \mathcal{I}$, $r_j(t) = 1$ si et seulement si $X_j(t)$ est observé à l'instant t. La tribu $\mathcal{X}$ engendrée par $\boldsymbol{X}$ et la tribu $\mathcal{X}^r$ des observations de $\boldsymbol{X}$ sont alors définies par

$$\mathcal{X} = \sigma\{X_{jt}, t \in \mathcal{I}, 1 \leq j \leq p\} \qquad \mathcal{X}^r = \sigma\{r_j(t)X_{jt}, t \in \mathcal{I}, 1 \leq j \leq p\}.$$

Bien sûr, un mécanisme d'observation n'est pas forcément déterministe et dans ce cas, on le notera $\boldsymbol{R} = (R_1, \dots R_p)$. Pour en étudier la σ-algèbre notamment dans le cas d'observations ponctuelles, il peut être nécessaire de représenter $\boldsymbol{R}$ lui-même par des processus ponctuels (voir pour cela la section 6.1. de l'article (51)). De même, les observations ponctuelles du processus d'intérêt peuvent nécessiter une représentation par un processus marqué. Cependant nous gardons la définition empirique suivante de la filtration et de la σ-algèbre observée sur $\mathcal{I} = [0, C]$:

$$\mathcal{O}_t = \{R_j(s), R_j(s)X_{js}, s \leq t, 1 \leq j \leq p\} \qquad \mathcal{O} = \{R_j(t), R_j(t)N_{jt}, t \in \mathcal{I}, 1 \leq j \leq p\}.$$

Si le mécanisme qui mène aux données manquantes est déterministe, alors $\mathcal{O} = \mathcal{X}^r \subset \mathcal{X}$ et la vraisemblance des observations est donnée simplement par $\mathcal{L}^\theta_\mathcal{O} = \mathbb{E}_0[\mathcal{L}^\theta_\mathcal{X}|\mathcal{X}^r]$. En revanche, si $\boldsymbol{R}$ est aléatoire, sans hypothèse sur les liens entre les processus $\boldsymbol{R}$ et $\boldsymbol{X}$, l'inférence sur $\boldsymbol{X}$ à partir de la tribu $\mathcal{O}$ n'est pas forcément la même que l'inférence que l'on ferait avec les observations de $\boldsymbol{X}$ si $\boldsymbol{R}$ était déterministe. En effet, les informations apportées par $\mathcal{O}$ sur le processus $\boldsymbol{X}$ ne se réduisent pas forcément à $\mathcal{X}^r$ sur l'événement $(\boldsymbol{R} = \boldsymbol{r})$. En effet d'une part, on n'a pas forcément $\mathcal{O} \subset \mathcal{X}$ et de plus $\boldsymbol{R}$ peut être informatif sur le processus d'intérêt.

Soyons plus précis. Un modèle pour $(\boldsymbol{X}, \boldsymbol{R})$ est une famille de mesures $\{\mathbb{P}_{\theta\psi}\}_{(\theta,\psi)\in\Theta\times\Psi}$ sur un espace mesurable $(\Omega, \mathcal{F})$. On pose $\mathcal{F} = \mathcal{R} \vee \mathcal{X}$. La vraisemblance à considérer pour une probabilité de référence $\mathbb{P}_0$ donnée, est $\mathcal{L}^{\mathbb{P}_{\theta\psi}/\mathbb{P}_0}_\mathcal{F}$ dont il va falloir prendre l'espérance conditionnelle sachant $\mathcal{O}$. Seul θ, qui donne des informations sur la loi de $\boldsymbol{X}$, nous intéresse pour l'inférence. Aussi il faut "détricoter" θ et ψ d'une part et les informations venant de $\boldsymbol{X}$ et de $\boldsymbol{R}$ dans $\mathcal{F}$ d'autre part. Plusieurs hypothèses sont données dans (4) à ce sujet. Premièrement, le paramètre ψ doit être un paramètre de nuisance, c'est-à-dire que la restriction de $\mathbb{P}_{\theta\psi}$ à $\mathcal{X}$ ne dépend pas de ψ. Ensuite on doit avoir une condition de "non-information" sur le mécanisme d'observation qui se traduit par le fait que la loi conditionnelle sachant $\mathcal{X}$, $\mathbb{P}_{\theta\psi}(.|\mathcal{X})$ ne dépend pas de θ. Enfin, $\mathcal{L}^{\mathbb{P}_1/\mathbb{P}_0}_{\mathcal{R}/\mathcal{X}}$ doit être $\mathcal{O}$-mesurable pour toute probabilité $\mathbb{P}_0$ et $\mathbb{P}_1$ dans une famille de probabilité $\mathcal{Q}$ contenant $\{\mathbb{P}_{\theta\psi}\}_{(\theta,\psi)\in\Theta\times\Psi}$. On dit alors que le mécanisme d'observation est CAR pour Coarsenong at Random. Sous ces hypothèses, en utilisant la décomposition $\mathcal{L}^{\mathbb{P}_{\theta\psi}/\mathbb{P}_0}_\mathcal{F} = \mathcal{L}^{\mathbb{P}_{\theta\psi}/\mathbb{P}_0}_{\mathcal{R}/\mathcal{X}}\mathcal{L}^{\mathbb{P}_{\theta\psi}/\mathbb{P}_0}_\mathcal{X}$, on montre que pour de l'inférence sur θ, il suffit de considérer $\mathbb{E}_{\mathbb{P}_0}[\mathcal{L}^{\mathbb{P}_{\theta/\mathcal{X}}/\mathbb{P}_{0/\mathcal{X}}}_\mathcal{X}|\mathcal{O}]$. Sous ces conditions, on peut presque reprendre les notations naïves du début de section et considérer $\mathcal{L}^\theta_\mathcal{O} = \mathbb{E}_0[\mathcal{L}^\theta_\mathcal{X}|\mathcal{O}]$.

Ce résultat étant établi, sans condition supplémentaire, comme $\mathcal{O} \nsubseteq \mathcal{X}$, le calcul de l'espérance conditionnelle n'est pas facile. C'est pourquoi, il est intéressant d'étudier des

conditions dites d'ignorabilité, c'est-à-dire que l'inférence connaissant $\mathcal{O}$ est la même que l'inférence sachant $\mathcal{X}^r$ sur l'événement $(\boldsymbol{R} = \boldsymbol{r})$. Autrement dit, on procède comme si $\boldsymbol{R}$ avait été déterministe. Bien sûr cette notion n'est soigneusement définie que si $(\boldsymbol{R} = \boldsymbol{r})$ est de mesure non nulle (car les vraisemblances sont définies à un ensemble négligeable près). On peut trouver des conditions d'ignorabilité dans l'article (51). Ces conditions ne nécessitent pas l'indépendance des deux processus $\boldsymbol{R}$ et $\boldsymbol{X}$ mais de manière heuristique on peut les exprimer par "le processus $\boldsymbol{R}$ ne dépend pas des parties non-observées de $\boldsymbol{X}$". On aura ignorabilité par exemple si le processus $\boldsymbol{R}$ est $\mathcal{O}_t$ prévisible. Ce cas est un cas particulier de la condition CAR(DYN) énoncée dans la proposition 1.1.1. Nous le verrons, sa formulation est proche de la définition de l'indépendance locale que nous verrons à la section 1.3.2. Indépendamment de ce résultat, nous montrons au paragraphe 4.5. de (3) que la censure par le décès est un mécanisme ignorable.

Proposition 1.1.1 *Supposons que le processus d'observation $\boldsymbol{R}$ est représenté par un processus marqué $\boldsymbol{Y}$, alors on dit que $(\boldsymbol{X}, \boldsymbol{R})$ satisfait la condition CAR(DYN) si le compensateur de $\boldsymbol{Y}$ est le même dans les deux filtrations $\mathcal{O}_t$ et $\mathcal{F}_t^* = \mathcal{X} \vee \mathcal{O}_t$.*

Sous la condition CAR(DYN), le mécanisme d'observation est ignorable.

1.1.1 Applications aux critères de choix de modèle

Nous avons utilisé la notion de Coarsening At Random pour établir un critère de choix de modèle pour l'inférence d'un modèle multi-états (voir section 1.2.1 pour une définition) dans un cadre partiellement observé. C'est l'objet de (4). En effet, il peut être utile d'avoir un ou des critères pour choisir par exemple entre un modèle multi-états markovien ou semi-markovien, pour choisir la dépendance par rapport aux covariables de manière multiplicative ou additive (101), ou pour choisir des paramètres de lissage, etc ... Pour cela, nous utilisons le critère EKL pour Expected Kullback Leibler et une approximation de EKL par un critère noté LCV pour Likelihood-based Cross Validation. Les définitions précises de ces critères sont disponibles dans (4). Nous ne sommes pas, comme nos co-auteurs pour cet article B. Liquet et D. Commenges spécialistes de choix de modèles, aussi nous ne passerons pas de temps sur ce sujet, sauf pour dire qu'une étude soigneuse du mécanisme de "coarsening" est nécessaire pour remplacer la tribu $\mathcal{X}$ par la tribu $\mathcal{O}$ dans les vraisemblances et dans les critères EKL et LCV. C'est notre contribution à ce travail.

Dans (4) nous généralisons aussi cette notion de CAR au cas où l'étude du processus d'intérêt $\boldsymbol{X}$ tient compte de l'information d'un processus de variables aléatoires explicatives $\boldsymbol{Z} = (\boldsymbol{Z}_t)_{t \in \mathcal{I}}$ complètement observé. Nous supposons que seule la loi de $\boldsymbol{X}$ sachant $\boldsymbol{Z}$ a de l'intérêt pour le statisticien (pas la loi marginale de $\boldsymbol{Z}$). Le modèle a alors trois paramètres : γ qui paramétre la loi de $\boldsymbol{Z}$, ψ la loi de $\boldsymbol{R}$ sachant $\boldsymbol{X}$ et $\boldsymbol{Z}$, et θ la loi de $\boldsymbol{X}$ sachant $\boldsymbol{Z}$. Des conditions de type CAR sont données dans ce cas et elles impliquent que seule la maximisation de $\mathbb{E}_{\mathbb{P}_0}[\mathcal{L}_{\mathcal{X}|\mathcal{Z}}^{\mathbb{P}_{\theta/(\mathcal{X},\mathcal{Z})}/\mathbb{P}_{0/(\mathcal{X},\mathcal{Z})}}|\mathcal{O}]$ est utile pour l'inférence sur θ. On peut alors considérer les critères EKL et LCV correspondants.

1.2 Vraisemblance et modèle multi-états partiellement observés

Le travail présenté ici, en collaboration avec Daniel Commenges, est la justification des écritures de fonction de vraisemblance dans le cas d'un processus ponctuel ou d'un processus multi-états partiellement observé correspondant à l'article (3).

1.2.1 Processus multi-états et processus ponctuels

Les modèles multi-états sont utilisés depuis longtemps par les bio-statisticiens notamment pour leur applications en biologie et épidémiologie. Un processus multi-états est un processus en temps continu, continu à droite prenant un nombre fini de valeurs, c'est donc un processus de Markov en temps continu comme défini par exemple dans (72)[Chapitre 7]. C'est à la fin des années 1970 que les modèles de Markov homogènes ont laissé la place à des modèles non-homogènes notamment dans Fleming (74), Aalen et Johansen (33), Lagakagos (98). Le modèle le plus étudié est bien sûr le célèbre modèle "illness-death" (34; 96). De nos jours, l'estimation des modèles multi-états est toujours étudiée d'un point de vue théorique comme par exemple dans (102; 48; 125; 97) et les applications en épidémiologie sont multiples (46; 124; 94).

Dans (33) et (35, Section IV.4), les auteurs utilisent les processus de comptage des transitions pour utiliser les techniques usuelles d'inférence des processus ponctuels (l'estimateur de Nelson-Aalen !) pour faire de l'inférence sur les intensités de transition du processus multi-états étudiés. C'est ce que nous voulons faire ici pour écrire la vraisemblance d'un processus multi-états. En effet dans les applications biomédicales, la vraisemblance est souvent écrite de manière heuristique et nous voulons justifier rigoureusement les écritures de vraisemblance en faisant le lien entre les processus multi-états et les processus ponctuels puis en formalisant le mécanisme qui a mené aux données observées qui sont souvent incomplètes.

Soit (X_t) un processus multi-états à valeurs dans $\{0, 1, \ldots, K-1\}$ sur un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$. On appelle $(\boldsymbol{A}_t)$ la famille des Q-matrices associée à (X_t), $\boldsymbol{A}_t = (\alpha_{hj}(t))$ vérifiant les hypothèses habituelles : si $h \neq j$, $\alpha_{hj}(t) \geq 0$ et $\alpha_{hh}(t) = -\sum_{j \neq h} \alpha_{hj}(t)$ et liée aux probabilités de transition par les équations de Kolmogorov. Si le modèle n'est pas Markovien, ces intensité de transition sont en fait aléatoires et prévisibles dans la filtration considérée, on a alors $\alpha_{hj}(t, \mathcal{F}_t^-)$. Nous l'avons dit plus haut le processus d'intérêt n'est pas toujours complètement observé et la tribu $\mathcal{F}_t$ n'est pas forcément la filtration naturelle du processus d'intérêt. Nous supposons que le processus multi-états est irréversible (i.e. $\alpha_{hj}(t) = 0$ si $h > j$). Dans ce cas, nous donnons une représentation du processus (X_t) par un processus ponctuel multivarié $\boldsymbol{N} = (N_1, \ldots, N_p)$ où chacun des N_k est un processus à 1 saut. Cette représentation est différente de celle de (33), nous ne la détaillons pas ici mais elle permet d'exprimer les intensités λ_k des N_k en fonction des intensités de transition α_{hj} de (X_t). C'est l'objet du Théorème 1 de (3). On a alors que les filtrations engendrées par (X_t) et $(\boldsymbol{N}_t)$ sont les mêmes, ce qui va nous permettre d'écrire la vraisemblance de (X_t) à partir des formules de vraisemblance de $(\boldsymbol{N}_t)$.

1.2.2 Vraisemblance

Jacod en 75 (91) a donné la formule de changement de probabilité pour des processus ponctuels marqués très généraux dans une filtration naturelle du processus $\boldsymbol{N}$ de la forme $\mathcal{F}_t = \mathcal{F}_0 \vee \mathcal{N}_t$ pour $t \in \mathcal{I}$. Aalen (33), en se basant sur des résultats de Jacod et Mémin (92), en a déduit une formule simple du rapport de vraisemblance de processus ponctuels multivariés à un saut dans le cas d'une intensité absolument continue en prenant pour probabilité de référence, la probabilité $\mathbb{P}_0$ qui rend les N_j indépendants et d'intensité 1. A l'aide de cette formule, si on suppose que $\mathcal{I} = [0, C]$, on peut alors écrire facilement la vraisemblance d'une probabilité $\mathbb{P}_\theta \ll \mathbb{P}_0$ relativement à $\mathbb{P}_0$. Si dans cette vraisemblance, on enlève les termes qui ne dépendent pas de θ, on a par abus de notation :

$$\mathcal{L}^\theta_{\mathcal{F}_C} = f^\theta_C(\tilde{T}_1, \ldots, \tilde{T}_p) \prod_{j=1}^{p} e^{\tilde{T}_j} \quad \text{p.s.}$$

avec $\tilde{T}_j = \min(T_j, C)$, $\delta_j = \mathbb{1}_{\{\tilde{T}_j < C\}}$ et

$$f^\theta_C(s_1, \ldots, s_p) = \prod_{j=1}^{p} \left[\lambda^\theta_j(\tilde{T}_j; \tilde{T}_l \wedge \tilde{T}_j, l = 1, \ldots, p)\right]^{\delta_j} \exp[-\Lambda^\theta_\cdot(C, \tilde{T}_l \wedge C, l = 1, \ldots, p)].$$

Ici, l'écriture est donnée dans le cadre non-markovien et la dépendance au passé est explicitement décrite au travers de l'écriture de λ^θ en fonction des sauts passés.

Supposons maintenant que le processus X_t n'est que partiellement observé et que les conditions d'ignorabilité sont satisfaites. La probabilité $\mathbb{P}_0$ qui rend les temps T_j indépendants est particulièrement propice au calcul de l'espérance conditionnelle $\mathcal{L}^\theta_{\mathcal{O}} = \mathbb{E}_0[\mathcal{L}^\theta_{\mathcal{F}_C} | \mathcal{N}^r_C]$. Nous donnons l'écriture de cette vraisemblance lorsque une seule des composantes de $\boldsymbol{N}$ est observée par intervalle et les autres sont complètement observées. C'est l'objet du lemme suivant.

Lemme 1.2.1 *Pour N_1 observé aux temps discrets $v_0, \ldots, v_m$ et N_k $(2 \leq k \leq p)$ observés en temps continu, la vraisemblance de $\boldsymbol{N}$ est donnée par*

$$\begin{aligned}\mathcal{L}^\theta_{\mathcal{O}} \quad = \quad & \sum_{l=1}^{m} \mathbb{1}_{\{v_{l-1} < T_1 \leq v_l\}} \prod_{j=2}^{p} \frac{g(\Gamma)}{e^{-v_{l-1}} - e^{v_l}} \int_{v_{l-1}}^{v_l} f^\theta_C(s, \Gamma) ds \\ & + \mathbb{1}_{\{T_1 > v_m\}} g(\Gamma) e^{v_m} \left[\int_{v_m}^{C} f^\theta_C(s, \Gamma) ds + f^\theta_C(C, \Gamma)\right]\end{aligned}$$

avec $\Gamma = (\tilde{T}_1, \ldots, \tilde{T}_p)$, $g(\Gamma) = \prod_{j=2}^{p} e^{\tilde{T}_j}$ et $f^\theta_C(t, \Gamma)$ une abbréviation pour $f^\theta_C(t, \tilde{T}_2, \ldots, \tilde{T}_p)$.

Nous donnons au Lemme 2 de (3) la formule obtenue lorsque les deux premières composantes sont observées par intervalle et l'on comprend rapidement que la formule se généralise mais qu'elle est de plus en plus difficile à écrire. C'est pourquoi nous montrons que le calcul de l'espérance conditionnelle dans le schéma d'observation entier n'est pas utile. En effet, le lemme de localisation de Kallenberg (95)[Lemma 6.2] nous permet de remplacer la tribu $\mathcal{O} = \mathcal{N}^r_C$ du conditionnement par une tribu plus simple. En effet avec les données en main, on connait exactement :

1. les composantes du processus $\boldsymbol{N}$ qui ont été observées exactement,
2. celles qui ont été observées par intervalle et l'intervalle correspondant,
3. celles qui ont été censurées à droite et la date de censure correspondante.

Ainsi pour les composantes de type 1, on peut faire comme si elles étaient partout observées, pour celles de type 2, on peut oublier les autres intervalles d'observation possibles et ainsi de suite. On a donc un événement A appelé "pseudo-atome" et qui s'est réalisé dans les données qui nous servent à faire l'inférence et une tribu $\tilde{\mathcal{O}} \subset \mathcal{O}$ telle que $A \cap \tilde{\mathcal{O}} = A \cap \mathcal{O}$. Le lemme de localisation nous permet de dire que $\mathcal{L}^{\theta}_{\mathcal{O}} = \mathcal{L}^{\theta}_{\tilde{\mathcal{O}}}$ p.s. sur A. La définition précise d'un pseudo-atome est donnée dans la définition 3 de (3) et le calcul de la vraisemblance sur un pseudo atome est donné dans le Théorème 2 du même article. Ils ne sont pas détaillés ici. Ce théorème permet de donner une expression simple de la vraisemblance et facilement utilisable en pratique. Et on retrouve les formules empiriques, écrites par les biostatisticiens.

1.3 Indépendance locale entre processus

Nous abordons ici les notions d'indépendance locale ou "d'influence" entre processus avec en filigrane une interprétation causale possible dans un modèle dynamique. Ces travaux ont été publiés dans les articles (7) et (8).

1.3.1 Contexte

La notion de causalité n'est pas l'apanage des statisticiens, elle est aussi centrale dans d'autres sciences et préoccupe les philosophes des sciences (45) (119). Même si les scientifiques sont d'accord pour reconnaître qu'une corrélation ou un lien statistique n'implique pas forcément de relation de cause à effet, les scientifiques qui utilisent la statistique et les statisticiens eux mêmes se demandent de quelle manière la modélisation peut aider à découvrir des liens de causalité entre les phénomènes.

Parmi les modèles qui permettent de décrire d'éventuels liens causaux, les modèles graphiques sont particulièrement adaptés. Utilisés et développés par Wright (129; 130) dès les années 20, ils sont l'objet d'un regain d'intérêt récent notamment par Dawid et Didelez (56; 61; 65; 66; 67) et aussi par Pearl dans sa monographie (112). Nous renvoyons le lecteur à Aalen *et al.* (31)[Chapitre 9] pour une revue complète des différentes approches de la causalité en statistique, notamment pour l'approche contrefactuelle et celle des modèles marginaux. Nous ne les aborderons pas ici sauf pour signaler les travaux récents et proches de cette thématiques de Chambaz et van der Laan (47) qui traitent de la mesure de l'importance d'une variable d'exposition. En revanche, les auteurs de (31) développent l'approche des caractéristiques locales dans les modèles dynamiques et celle de Granger et Schweder. C'est dans ce cadre que nous plaçons nos travaux. C'est à la fin des années 70 et dans la littérature économétrique que Granger (77) a introduit la notion de causalité entre deux séries temporelles : de manière heuristique on peut dire que X n'influence pas Y selon Granger si la prédiction de Y sachant tous les "prédicteurs" n'est pas meilleure que celle basée sur tous les prédicteurs privés de X. Schweder (120) définit lui la notion d'indépendance locale entre deux composantes d'un processus de Markov multivarié en donnant des conditions sur les intensités de transition correspondantes. Ces travaux placent la causalité dans le cadre des

processus stochastiques et prennent en compte le déroulement du temps (ce n'était pas le cas dans l'approche contrefactuelle et les premiers modèles graphiques) et l'évidente vérité " la cause doit précéder les effets". Ces notions ont été développées pour des processus stochastiques plus généraux notamment pour les processus ponctuels dans (29; 70; 32; 75) grâce à la décomposition de Doob-Meyer de ces processus. A noter aussi, le travail récent de Røysland (118) qui place la notion d'indépendance locale de Didelez dans un cadre biostatistique et donne une caractérisation de mesure d'un essai randomisé en terme de martingale. Si le modèle est observationnel, il est éventuellement possible de se ramener à un modèle randomisé par changement de probabilité (Girsanov).

Suite à la lecture des articles (29; 32) et à la rencontre avec V. Didelez de University College London, spécialiste des modèles graphiques et qui a étudié des liens d'indépendance locale dans une famille de processus ponctuels marqués (66; 67), nous avons orienté notre recherche sur les critères d'indépendances locales dans une famille de semi-martingales, les modèles graphiques associés et tenter de formaliser les notions de causalité dans ce cadre.

1.3.2 Indépendance locale

Nous donnons ici la notion d'indépendance locale pour une classe de semi-martingales très générale qui est celle proposée dans (8) qui généralise la classe pour laquelle nous avions déjà proposé une définition dans (7)[Section 2.2].

Soit un espace de probabilité filtré $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ et un processus stochastique multivarié $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ à valeurs dans l'espace de Skorohod $D(\Re^m)$ des fonctions càdlàg de $\mathbb{R}_+$ dans $\mathbb{R}^m$. On a donc $\boldsymbol{X} = (X_j, j = 1, \ldots, m)$ avec $X_j = (X_{jt})_{t\geq 0}$. On appelle $(\mathcal{X}_t) = (\mathcal{X}_t)_{t\geq 0}$ la filtration satisfaisant les conditions usuelles, engendrée par le processus $\boldsymbol{X}$ i.e. pour tout $t \geq 0$, $\mathcal{X}_t = \sigma\{\boldsymbol{X}_u, 0 \leq u \leq t\}$. De manière similaire on définit $(\mathcal{X}_{jt})$ la filtration associée à la composante X_j de $\boldsymbol{X}$. On définit $\mathcal{F}_t = \mathcal{H} \vee \mathcal{X}_t$; $\mathcal{H}$ peut contenir de l'information connue au temps $t = 0$ (covariables non dépendantes du temps par exemple) en plus de la valeur initiale de $\boldsymbol{X}$. On note aussi $\mathcal{F}_{-jt} = \mathcal{H} \vee \mathcal{X}_{-jt}$ avec $\mathcal{X}_{-jt} = \vee_{l\neq j}\mathcal{X}_{lt}$ de sorte que la filtration $(\mathcal{F}_{-jt})$ contient l'information de $(\mathcal{F}_t)$ moins celle propre à $(\mathcal{X}_{jt})$.

On suppose que $\boldsymbol{X}$ appartient à la classe des semi-martingales spéciales dans la filtration $(\mathcal{F}_t)$. Les caractéristiques de la semi-martingale $\boldsymbol{X}$ sous la probabilité P sont notées (B, C, ν), la partie martingale de X_j est notée M_j et la partie continue de cette dernière est notée M_j^c. On note aussi (B^j, C^j, ν^j) les caractéristiques de la semi-martingale X_j sous P. Une définition précise des caractéristiques (B, C, ν) est disponible dans (93)[SectionII.2]. Pour se représenter le rôle de chacune des trois caractéristiques, si $\mathcal{X}$ était à accroissements indépendants, B serait sa dérive (son drift !), C la variance de sa partie gaussienne (son crochet oblique !) et ν sa mesure de Levy (ou compensateur de la composante de saut).

Nous faisons les hypothèses **(H-M)** et **(H-MC)** sur les semi-martingales étudiées.

(H-M) Pour tout $j \neq k$, M_j et M_k sont des martingales de carré intégrable orthogonales.

Sous **(H-M)**, les parties sauts des martingales M_j et M_k sont orthogonales, il n'y a donc pas de sauts simultanés. De plus, la caractéristique C de $\boldsymbol{X}$ (le crochet de la partie continue de la martingale) est une matrice diagonale. En effet, par définition, $C_{ij} =< M_i^c, M_j^c >= 0$ pour tout $1 \leq i, j \leq m$; on note $C^j = C_{jj}$.

(H-C) C^j est déterministe pour tout j.

Ces deux conditions pourraient être résumées en une seule, cependant elles sont de nature très différentes. En faisant l'hypothèse **(H-M)**, nous supposons que deux composantes X_j et X_k ne peuvent pas refléter un même phénomène stochastique à travers leur partie martingale. Cette association entre les deux composantes ne serait de toute façon pas causale. Pötter et Blossfeld (116) disent que les deux composantes, dans ce cas, sont "autonomes". Dans (31) un exemple concret exclu par cette hypothèse est le déclenchement de deux phénomènes d'irritation (yeux et nez par exemple) dû à la même allergie. **(H-C)**, la deuxième hypothèse est, elle, plus technique et pour l'instant nous n'avons pas pu la contourner. Elle est liée au caractère mesurable du crochet de la partie continue de la martingale d'un processus dans la filtration propre de ce processus. L'influence d'une composante sur une autre par l'intermédiaire de ce crochet ne peut donc pas être captée par des notions de mesurabilité comme nous le faisons pour les autres caractéristiques dans la définition 1.3.1.

On appelle $\boldsymbol{\mathcal{D}'}$ la classe des semi-martingales spéciales qui vérifient **(H-A)** et **(H-C)**. La classe $\boldsymbol{\mathcal{D}'}$ est stable par un changement de probabilité absolument continu (C ne change pas avec la proba). De plus $\boldsymbol{\mathcal{D}'}$ est une large classe, elle inclut par exemple les mesures aléatoires, les processus ponctuels marqués, les diffusions et les diffusions avec sauts.

Nous introduisons maintenant les notions d'indépendance locale et d'influence entre composantes d'un processus de $\boldsymbol{\mathcal{D}'}$.

Définition 1.3.1 (Weak conditional local independence (WCLI)) *Soit $\boldsymbol{X}$ dans la classe $\boldsymbol{\mathcal{D}'}$. On dit que X_k est WCLI de X_j dans $\boldsymbol{X}$ pour $t \in [r,s]$, si et seulement si les caractéristiques B^k et ν^k sont telles que $B_{kt} - B_{kr}$ et $\nu_{kt} - \nu_{kr}$ sont $(\mathcal{F}_{-jt})$-prévisibles sur $[r,s]$. De manière équivalente, on peut dire que X_k a le même triplet de caractéristique (B^k, C^k, ν^k) dans les filtrations $(\mathcal{F}_t)$ et $(\mathcal{F}_{-jt})$ sur l'intervalle $[r,s]$.*

Remarquons que l'hypothèse CAR(DYN) que nous avons énoncée à la Proposition 1.1.1 s'énonce par une condition de mesurabilité sur le compensateur d'un processus marqué autrement dit sur sa caractéristique locale (les deux autres sont dégénérées). Cependant elle n'est pas traduisible en une condition d'indépendance locale car l'information de $\mathcal{X}$ dans $\mathcal{F}_t^* = \mathcal{X} \vee \mathcal{O}_t$ n'est pas dynamique mais donnée en 0. Les caractéristiques des semi-martingales dépendent de la probabilité sous-laquelle elles sont considérées, c'est pourquoi la notion d'indépendance locale en dépend aussi. On peut donc imaginer des changements de probabilité qui font disparaître des dépendances ou qui rendent un facteur localement indépendant de tous les autres facteurs : ceci permet de mesurer leur influence sur un événement d'intérêt, c'est cette possibilité qu'envisage Røysland dans (118).

En ce qui concerne les changements de probabilité, nous énonçons ici un autre critère d'indépendance locale en fonction de l'existence d'un rapport de vraisemblance "ne concernant que" la loi d'une composante et vérifiant une condition de mesurabilité. Nous donnons cette définition ci-après. Les hypothèses sont discutées dans (8) où il est aussi montré l'équivalence entre les deux notions WCLI et LWCLI pour une certaine classe de semi-martingales (Proposition 1).

Définition 1.3.2 *[Likelihood-based weak conditional local independence (LWCLI)]*
Soit $\boldsymbol{X}$ dans la classe $\boldsymbol{\mathcal{D}'}$. On suppose l'existence d'une probabilité $\mathbb{P}_0$ telle que

(i) $\mathbb{P} \ll \mathbb{P}_0$,

(ii) les caractéristiques des semi-martingales X_i avec $i \neq k$ sont les mêmes sous les probabilités $\mathbb{P}$ et $\mathbb{P}_0$,

(iii) les $\mathbb{P}_0$-caractéristiques (B_0^k, C_0^k, ν_0^k) de la semi-martingale X_k sont déterministes.

On dit que X_k est LWCLI de X_j dans $\boldsymbol{X}$ sur $[0,t]$ si et seulement si le processus de vraisemblance $Z_t^{P/P_0} = \mathcal{L}_{\mathcal{F}_t}^{P/P_0}$ est $(\mathcal{F}_{-jt})$-mesurable sur $[0,t]$.

La notion s'étend facilement sur un intervalle $[r,s]$ en considérant le processus $\frac{Z_t^{P/P_0}}{Z_r^{P/P_0}}$.

Les conditions d'indépendance étant posées, on peut définir par contraposée des notions d'influence.

Définition 1.3.3 (Influence directe) *Si X_k n'est pas WCLI de X_j dans $\boldsymbol{X}$, on dit que X_j influence directement X_k dans $\boldsymbol{X}$ et l'on note $X_j \longrightarrow_{\boldsymbol{X}} X_k$.*

Définition 1.3.4 (WCLI influence pour une groupe de composantes) *Soit A, B des sous-ensemble de $(1,\ldots,m)$. On dit que $X_A \longrightarrow_{\boldsymbol{X}} X_B$ s'il existe $j \in A$ et $k \in B$ tel que $X_j \longrightarrow_{\boldsymbol{X}} X_k$.*

Nous pouvons définir maintenant une notion plus forte d'indépendance locale entre processus.

Définition 1.3.5 (Strong conditional local independence (SCLI)) *X_k est SCLI de X_j dans $\boldsymbol{X}$ si et seulement si $X_j \not\longrightarrow_{\boldsymbol{X}} X_k$ et il n'existe pas de $X_D \in \boldsymbol{X}$ tel que $X_j \longrightarrow_{\boldsymbol{X}} X_D$ et $X_D \longrightarrow_{\boldsymbol{X}} X_k$. Dans ce cas, nous noterons $X_j \not\longrightarrow_{\boldsymbol{X}} X_k$.*

Définition 1.3.6 (Influence indirecte) *Si X_k n'est pas SCLI de X_j, X_j influence (au moins indirectement) X_k dans $\boldsymbol{X}$ et nous notons $X_j \rightarrow\!\!\rightarrow_{\boldsymbol{X}} X_k$.*

Si de plus $X_j \rightarrow\!\!\rightarrow_{\boldsymbol{X}} X_k$ et $X_j \not\longrightarrow_{\boldsymbol{X}} X_k$, on dit que l'influence est indirecte.

1.3.3 Représentation par un graphe

Nous serons très concis dans cette section mais comme l'a fait Didelez dans (66), à l'aide des définitions de la section précédente, il est facile et naturel de représenter les influences entre les composantes d'un processus dans un graphe orienté. Les sommets du graphe sont les composantes X_j de $\boldsymbol{X}$ et il y a une flèche de X_j vers X_k si et seulement si $X_j \longrightarrow_{\boldsymbol{X}} X_k$. On peut alors définir la notion de chemin entre deux composantes et l'existence d'un tel chemin signifie $X_j \rightarrow\!\!\rightarrow_{\boldsymbol{X}} X_k$. Nous donnons à la figure 1.1 deux exemples de graphes emboîtés représentants les liens entre quatre variables d'intérêt $(X_1,\ldots,X_4)$ auxquelles on ajoute des facteurs. En ajoutant des facteurs, certaines influences directes disparaissent et deviennent indirectes. Cette remarque nous conduit à considérer des systèmes emboités. C'est ce que nous faisons dans la prochaine section. Dans (7), nous présentons un modèle dynamique d'infection par le VIH et les graphes correspondants.

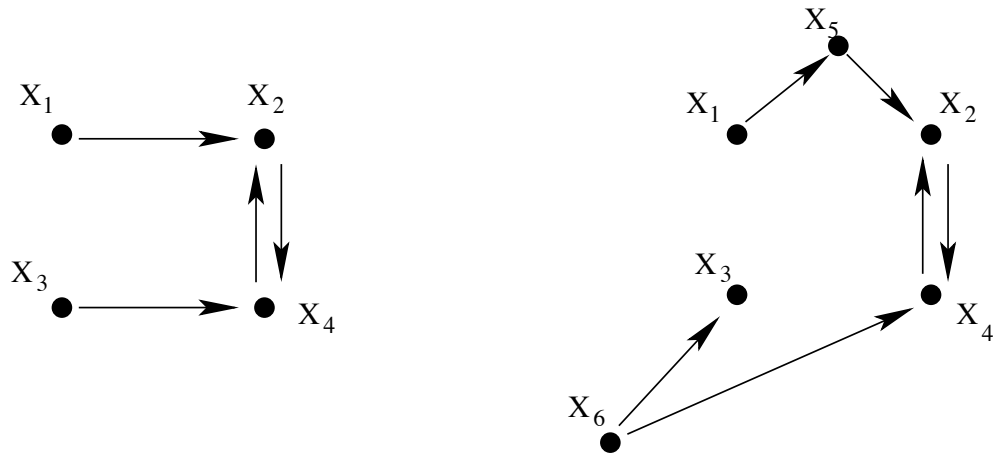


FIGURE 1.1 – Exemple de deux graphes emboîtés décrivant le même système physique.

1.3.4 Interprétation causale

Nous l'avons dit, tenter de définir un lien causal nous emmène vers la philosophie et vers des chemins au sol plus mouvant que celui de l'univers des mathématiques ou les assertions sont soit justes, soit fausses.... C'est l'objet de la discussion de la section 3. de (7). Pour tenter de définir un lien causal, nous y introduisons la notion de système dont l'état et les attributs à l'origine sont décrits par une tribu $\mathcal{A}$ et l'état au cours du temps est décrit par une filtration $(\mathcal{X}_t)$. Notre formation (ancienne certes !) à la physique, nous a toujours donné en filigrane un système régi par des lois physiques, mécaniques ou au moins issues de la nature (on parle bien de modèles mécanistes en biostatistique !). Par abus de notation, nous identifions le système à un processus stochastique et les lois mécanistes nous invitent à imaginer un système d'équations différentielles stochastiques et/ou de processus ponctuels d'intensité interprétables.

Dans ce système, on doit distinguer les événements ou processus d'intérêt des événements auxiliaires qui sont liés aux premiers. On comprend qu'un système trop pauvre ne prenant pas en compte tous les facteurs liés au processus d'intérêt va faire apparaître des liens injustifiés (ou de manière un peu moins probable, faire disparaître des liens réels). Les cas connus du paradoxe de Simpson en sont des exemples. Cependant il paraît illusoire de faire apparaître un niveau d'information trop pointu. On comprend cependant l'intérêt de définir des systèmes emboîtés. Un système $\mathcal{S}^{m'}$ est emboité dans $\mathcal{S}^m$ si $\mathcal{F}_t^{m'} \subset \mathcal{F}_t^m$ pour tout $t : \mathcal{S}^{m'}$ peut être plus riche que $\mathcal{S}^m$ soit du point de vue de ses attributs ($\mathcal{A}^{m'} \subset \mathcal{A}^m$) et/ou de ses composantes ($\mathcal{X}_t^{m'} \subset \mathcal{X}_t^m$). On peut considérer une suite de systèmes emboîtés $\boldsymbol{\mathcal{S}} = \{\mathcal{S}^m\}_{m>0}$ (on note $\mathcal{S}^m \in \boldsymbol{\mathcal{S}}$ et $\mathcal{S}^m \subset \mathcal{S}^{m'}$ si $m < m'$). Dans ce cas, à t fixé, la famille $\{\mathcal{F}_t^m\}_{m>0}$ forme une filtration.

On appelle P^* probabilité sur $(\Omega, \mathcal{F})$ qui régit le système (c'est celle que l'on veut estimer) et on note $P^*_{\mathcal{F}^m}$ sa restriction à $\mathcal{F}^m$. L'idée est d'approcher et d'estimer $P^*_{\mathcal{F}^m}$ par une probabilité $P_{\mathcal{F}}^{\mathcal{S}^m}$ donnée par des lois naturelles liant les composantes de $\mathcal{S}^m$.

Revenons à un système $\mathcal{S}^1$ et à l'éventuel lien causal entre deux composantes j et k de $\mathcal{S}^1$. Si on fait l'hypothèse de l'existence d'un système $\mathcal{S}^m$ "parfait pour $\mathcal{S}^1$" c'est-à-dire que

$P^{\mathcal{S}^m}_{\mathcal{F}^1} = P^*_{\mathcal{F}^1}$, alors nous pouvons tenter de définir un lien causal entre j et k par la définition suivante.

Définition 1.3.7 (Influence causale) *Une composante j a une influence causale sur la composante k dans $\mathcal{S}^1$ s'il existe un système $\mathcal{S}^M$ parfait pour $\mathcal{S}^1$, tel que $X_j \rightarrow\rightarrow_{\mathcal{S}^M} X_k$ sous P^*.*

Evidemment, cette définition reste très abstraite car la vraie vie n'est pas parfaite et elle ne nous dit même pas si on est proche d'un système parfait !

1.3.5 Lien avec l'indépendance conditionnelle

Il peut être tentant, comme c'est le cas dans pour les modèles non dynamiques ou les processus discrets comme dans Dawid (60) ou Eichler *et. al* (71) de vouloir exprimer une condition d'indépendance locale en terme d'indépendance conditionnelle. Dans (71), la non causalité forte de Granger s'exprime pour un système discret et avec nos notations de la manière suivante :

$$X_{ks} \perp\!\!\!\perp_{\mathcal{F}_{-jt}} \mathcal{X}_{jt}, t = 0, 1, \ldots;\ s = t+1, t+2, \ldots, t+h,$$

Il est tentant de généraliser cette notion aux processus continus en remplaçant la condition 1.3.5 par la suivante

$$\mathcal{X}_{k\tau} \perp\!\!\!\perp_{\mathcal{F}_{-jt}} \mathcal{X}_{jt-}, 0 \leq s < t \leq \tau.$$

Mais, cette condition n'a pas de sens car souvent les événements de $\mathcal{X}_{kt}$ connaissant $\mathcal{X}_{kt-}$ sont de probabilité un ou zero (par exemple un saut en t lorque le compensateur est absolument continu) et la condition, toujours vérifiée n'a pas de sens. Dans (8), nous montrons que la condition (1.1) ci-dessous est équivalente à SCLI pour une classe de processus de diffusion avec sauts.

$$X_{k\tau} \perp\!\!\!\perp_{\mathcal{F}_{-jt-}} \mathcal{X}_{jt-}, 0 \leq t \leq \tau. \tag{1.1}$$

Nous montrons aussi dans le Lemme 4 de (7) que, sous certaines conditions, deux groupes de composantes qui ne s'influencent pas mutuellement sont indépendants conditionnellement à l'information au temps 0. Ce résultat d'indépendance conditionnelle nous permet de démontrer que sous certaines conditions, l'influence d'un processus non influencé sur une composante d'un système simple est une influence causale suivant la définition 1.3.7. En ce qui concerne un processus non influencé, on pense évidemment à l'attribution randomisée d'un traitement dans les essais cliniques.

1.4 Conclusion

Les travaux présentés dans ce chapitre me sont chers car ils représentent la reprise de mes activités de recherche, une vision nouvelle des mathématiques et de leurs utilisations et de nouvelles collaborations. Je ne compte pas présenter des perspectives ou prolongements directs puisque je n'ai pas travaillé sur ces thématiques ces quatre dernières années. Mais il est

clair que la formalisation de la tribu d'observation (équation (2.19)) dans l'étude des processus BAR avec données manquantes présentée au chapitre suivant a été grandement facilitée par notre travail préalable sur cette problématique. Le travail sur les Processus Markoviens Déterministes par Morceaux (voir Chapitre 3), qui sont une extension des processus de sauts en est aussi une suite naturelle. Ma contribution au projet ANR Fautocoes notamment par la définition et l'encadrement de la thèse de Romain Azaïs (section 3.4) portant sur l'inférence des caractéristiques d'un PDMP et notamment sur le taux de sauts est aussi dans cette lignée. La confrontation très brève avec les critères de choix de modèles pourra d'ailleurs nous être utile sur ce sujet pour le choix de paramètres de lissage ou de partitions de l'espace, nous y reviendrons.

Les travaux sur l'indépendance locale m'ont amusée et je pourrais être tentée de m'y remettre notamment pour réfléchir comme le fait (118) sur les changements de probabilités nécessaires pour mesurer efficacement l'influence d'un facteur sur un événement d'intérêt quand l'attribution de ce facteur n'a pas été (ou ne peut pas) être randomisée. Enfin les études sur la causalité présentées ici restent très théoriques ; mais dans des modèles paramétriques au moins, il pourrait être intéressant de quantifier des influences comme le fait Chambaz (47) pour les modèles structuraux.

Chapitre 2

Processus de bifurcation

Ce chapitre présente les travaux réalisés au sein de l'Institut Mathématique de Bordeaux (IMB) qui ont débuté sous l'impulsion de Bernard Bercu et en collaboration avec Benoîte de Saporta (IMB) et se sont prolongés en collaboration avec Benoîte de Saporta (IMB) et Laurence Marsalle de l'Université de Lille. Ils correspondent aux articles publiés (6), (9), (11) et aux articles soumis (15) et (14). Après avoir rappelé le contexte à la section 2.1 et la structure des arbres binaires à la section 2.2, nous présentons à la section 2.3 (resp. 2.4) les différents modèles pour la généalogie (resp. de processus autorégressifs de bifurcation) étudiés. Nous donnons en 2.5 et 2.6 les types de résultats obtenus et les méthodes pour les démontrer. Enfin nous présentons en 2.8 une étude de simulation et des résultats sur des données de division cellulaire.

2.1 Contexte

Dans ce chapitre nous appelons processus de bifurcation des processus indexés par un arbre binaire et ainsi adaptés à l'étude de données de division cellulaire. Nous considérons deux types de processus : d'une part les processus indexés par un arbre binaire et à valeurs dans $\{0,1\}^{\mathbb{N}}$ modélisant la présence ou l'absence d'une cellule dans la généalogie et que nous appellerons dans la suite processus de généalogie et d'autre part les processus autorégressifs de bifurcation (acronyme BAR en anglais pour Bifurcating AutoRegressive processes). Les processus de généalogie sont intimement liés, nous le verrons, à un processus de Galton-Watson dont les propriétés, que nous utiliserons, sont connues. Les processus BAR sont une adaptation des processus autorégressifs (AR) pour les données structurées par un arbre binaire. Ils ont été introduits par Cowan and Staudte (55) pour les données de division cellulaire quand chaque individu d'une génération donne naissance à deux individus dans la génération suivante. Les processus BAR modélisent une caractéristique quantitative liée à chaque cellule et observée sur plusieurs générations descendant d'une cellule initiale. Ils permettent de prendre en compte simultanément les effets de la généalogie (par la régression sur la caractéristique de la mère) et ceux de l'environnement dans l'évolution de la caractéristique étudiée (ajout du bruit).

La définition proposée dans (55), d'un processus BAR est la suivante. La cellule initiale est nommée 1, et les deux descendants d'une cellule n, sont eux $2n$ et $2n+1$. Soit X_n la

caractéristique quantitative de la cellule n. Alors, le processus BAR symétrique d'ordre 1 est défini récursivement pour tout $n \geq 1$, par

$$\begin{cases} X_{2n} &= a + bX_n + \epsilon_{2n}, \\ X_{2n+1} &= a + bX_n + \epsilon_{2n+1}. \end{cases}$$

La suite $(\epsilon_{2n}, \epsilon_{2n+1})$ est une suite de bruits représentant les effets de l'environnement, les paramètres a, b sont des nombres réels inconnus vérifiant $|b| < 1$. Dans (55), la suite des bruits $(\epsilon_{2n}, \epsilon_{2n+1})$ était indépendante et équidistribuée de loi normale ; une corrélation étant toutefois possible entre ϵ_{2n} et ϵ_{2n+1} ; cette corrélation entre X_{2n} et X_{2n+1}, due au même environnement, se rajoute à celle induite par la généalogie. Par la suite plusieurs auteurs ont étudié ce modèle en en proposant des extensions. Huggins and Basawa (87) et Basawa and Zhou (39; 136) font des hypothèses plus générales sur le bruit. Dans (87), Huggins and Basawa étudient un BAR d'ordre supérieur, dans ce cas, l'effet de la généalogie vient non seulement de la mère mais aussi de la grand-mère et des ascendants d'ordre supérieur. En ce qui concerne l'inférence sur les paramètres (a, b) et les résultats asymptotiques qui l'accompagnent, Huggins and Basawa dans (87) ont proposé un estimateur du maximum de vraisemblance pour l'observation de plusieurs petits arbres indépendants. L'estimateur de maximum de vraisemblance pour un arbre unique quand le nombre de cellules croît vers l'infini a été étudié par Huggins dans (86) pour le modèle BAR d'origine, par Huggins et Basawa (88) pour des BAR gaussiens d'ordre supérieur et par Zhou and Basawa (136) pour une BAR d'ordre 1 avec bruit exponentiel. Zhou et Basawa, quant à eux, ont étudié dans (135) l'estimateur des moindres carrés. Dans toutes ces publications, le processus BAR est supposé stationnaire ; la série chronologique admet alors une représentation par une fonction holomorphe.

Nos travaux, qui ne supposent pas le régime stationaire, font suite à ceux de Guyon (81; 82) qui introduit les processus BAR asymétriques. Avant d'en détailler les aspects mathématiques, il nous paraît important de préciser que l'étude de l'asymétrie des BAR a été motivée par une question biologique relative au vieillissement des organismes unicellulaires de type E. Coli. En effet, la mesure du vieillissement des organismes unicellulaires se fait en quantifiant la dissymétrie dans le mécanisme de reproduction entre les cellules ayant hérité du vieux pôle de leur mère et celles ayant hérité du nouveau pôle (cf paragraphe 2.8 et Figure **??**). D'après (122), cette dissymétrie peut apparaître dans la reproduction elle-même : les cellules dites "jeunes" ayant en moyenne un nombre de descendants plus élevé que les cellules dites "âgées" ; elle peut aussi se manifester dans une dissymétrie des caractéristiques quantitatives de la cellule, par exemple le taux de croissance ou la masse. L'étude de données de division d'E. coli présentées dans (122) est donc à l'origine des travaux de Guyon (81; 82) et a indirectement ou directement motivé les nôtres : indirectement lorsque notre but a été de généraliser les hypothèses mathématiques de (81) (à un BAR d'ordre p avec des hypothèses plus souples sur le bruit) dans (6) ou d'étudier les BAR à coefficients aléatoires dans l'article soumis récemment (14). Plus directement lorsque nous avons proposé un estimateur tenant compte des cellules manquantes dans (9) ou après les études de simulations et l'applications du modèle à un arbre, nous avons proposé une approche "multi-arbres" dans (15). L'article de Delmas et Marsalle (64) est aussi dans la ligne directe de Guyon (81). Nous nous proposons de mettre ces travaux en perspective dans ce chapitre.

2.2 Quelques notations sur les arbres binaires

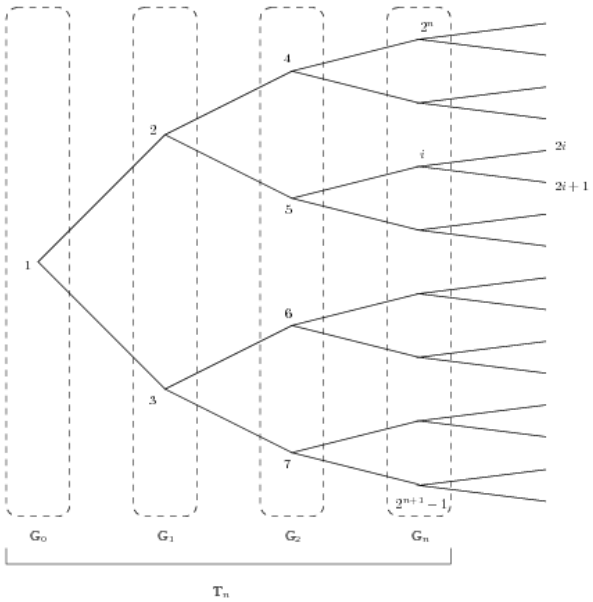


FIGURE 2.1 – Arbre associé à un processus auto-regressif de bifurcation.

Nous donnons ici quelques notations sur les arbres binaires : ceux-ci modélisant une généalogie quand chaque individu d'une génération donne naissance à deux individus dans la génération suivante. La Figure 2.1 en donne une illustration. Chaque noeud de l'arbre représente un individu ou une cellule et le noeud 1 représente l'ancêtre. Pour tout $n \geq 1$ la n^e génération $\mathbb{G}_n$ est définie par

$$\mathbb{G}_n = \{2^n, 2^n + 1, \ldots, 2^{n+1} - 1\}.$$

La génération initiale $\mathbb{G}_0 = \{1\}$, contient l'ancêtre original, et ses descendants de la première génération constituent $\mathbb{G}_1 = \{2, 3\}$. L'individu n est dans la génération $\mathbb{G}_{r_n}$ avec $r_n = [\log_2(n)]$ où $[x]$ désigne la partie entière de x. Les filles de la cellule n ont pour étiquette $2n$ et $2n + 1$ et inversement la mère de la cellule n est $[n/2]$ et ses ancêtres sont $[n/2], [n/2^2], \ldots, [n/2^{r_n}]$. Le sous arbre de tous les individus de l'ancêtre jusqu'à la génération n est noté $\mathbb{T}_n = \bigcup_{l=0}^{n} \mathbb{G}_l$. On a alors 2^n individus dans la génération $\mathbb{G}_n$ et $|\mathbb{T}_n| = 2^{n+1} - 1$ dans le sous-arbre $\mathbb{T}_n$.

2.3 Modèle pour la généalogie

Cette section est dédiée à la modélisation de la généalogie des cellules. D'une part, il convient de distinguer deux types de cellules, les cellules paires et les cellules impaires. D'autre part, comme nous l'avons dit plus haut, une cellule peut mourir et la branche de l'arbre naissant de cette cellule est coupée. Comme il est signalé dans (122), les taux de mort peuvent mesurer eux aussi le vieillissement et être différents suivant le type de la mère et suivant le type de la fille. Pour mesurer cette asymétrie, nous pouvons estimer les paramètres de reproduction

de ce processus et tester l'asymétrie éventuelle. Il convient de définir un processus $(\delta_k)_{k\in\mathbb{T}}$ à valeurs dans $\{0,1\}^{\mathbb{N}}$ qui modélise la présence ou l'absence d'une cellule et tel que si une cellule est absente, ses descendantes le sont aussi. La Figure 2.2 donne l'exemple d'une réalisation de $(\delta_k)_{k\leq 32}$ pour $n=4$ générations.

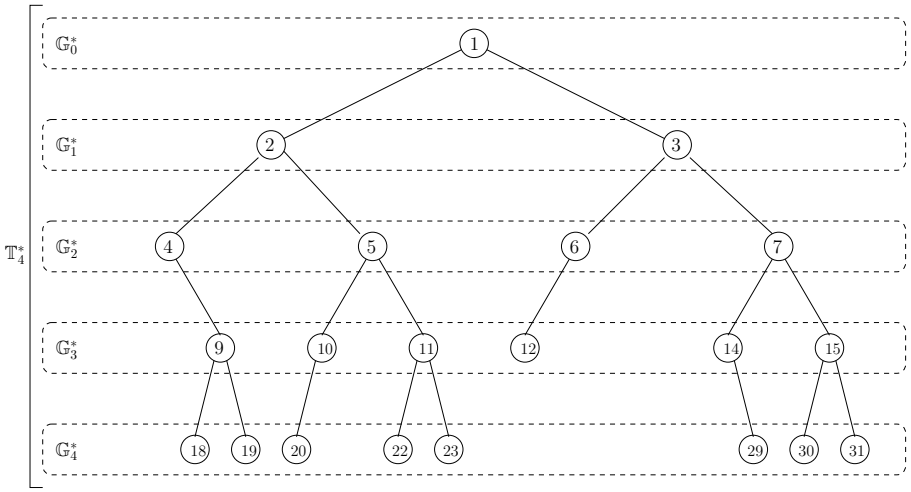


FIGURE 2.2 – Arbre associé aux données observées de l'arbre de la Figure 2.1.

2.3.1 Définition

Pour définir $(\delta_k)_{k\in\mathbb{T}}$, on pose $\delta_1=1$ et la suite est définie de manière récursive pour tout $k\geq 1$ par

$$\delta_{2k}=\delta_k\zeta_k^0 \quad \text{et} \quad \delta_{2k+1}=\delta_k\zeta_k^1, \tag{2.1}$$

avec $(\boldsymbol{\zeta}_k=(\zeta_k^0,\zeta_k^1))$ une suite indépendante de vecteurs aléatoires dans $\{0,1\}^2$. La variable ζ_k^i donne le nombre de descendants de type i de la cellule k. On distingue les lois des deux suites i.i.d. et indépendantes entre elles $(\boldsymbol{\zeta}_k, k\in 2\mathbb{N}^*)$ et $(\boldsymbol{\zeta}_k, k\in 2\mathbb{N}+1)$. Leur loi est donnée par les probabilités $p^{(i)}(j_0,j_1)$ pour $(i,j_0,j_1)\in\{0,1\}^3$; $p^{(i)}(j_0,j_1)$ étant la probabilité qu'un individu de type i donne naissance à j_0 cellule de type 0 et j_1 cellules de type 1. Des estimateurs de ses probabilités de descendance seront donnés à la section 2.4.3

Dans le cas de l'arbre complet, le nombre d'individus par génération est déterministe et égal à 2^n pour la génération n. Maintenant, ce cardinal est un nombre aléatoire $|\mathbb{G}_n^*|$ où $\mathbb{G}_n^*$ est l'ensemble des individus observés dans la génération n, défini par :

$$\mathbb{G}_n^*=\{k\in\mathbb{G}_n : \delta_k=1\} \quad \text{et de même} \quad \mathbb{T}_n^*=\{k\in\mathbb{T}_n : \delta_k=1\}.$$

Parmi ces individus, nous distinguons les individus de chacun des deux types en posant :

$$Z_n^0=|\mathbb{G}_n^*\cap 2\mathbb{N}| \quad \text{et} \quad Z_n^1=|\mathbb{G}_n^*\cap(2\mathbb{N}+1)|, \tag{2.2}$$

Z_n^0 (resp. Z_n^1) est le nombre d'individu de type 0 (resp 1) de la génération n et bien sûr $|\mathbb{G}_n^*|=Z_n^0+Z_n^1$ et $|\mathbb{T}_n^*|=\sum_{\ell=0}^n(Z_\ell^0+Z_\ell^1)$.

2.3.2 Processus de Galton-Watson associé et propriétés

Le processus $(\boldsymbol{Z}_n, n \geq 0)$, issu de notre processus de généalogie $(\delta_k)_{k\in\mathbb{T}}$ et défini pour $n \geq 1$ par $\boldsymbol{Z}_n = (Z_n^0, Z_n^1)$ est un processus de Galton-Watson bi-type de loi de reproduction très spécifique (chaque individu a au plus une cellule de chaque type) qui garantit des moments à tout ordre. Ses propriétés, asymptotiques notamment, sont décrites dans Harris (84). Nous donnons ici celles qui sont utiles dans la suite notamment les conditions de non-extinction du processus. Pour cela, on définit la matrice de descendance $\boldsymbol{P}$ par

$$\boldsymbol{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

avec $p_{i0} = p^{(i)}(1,0) + p^{(i)}(1,1)$ et $p_{i1} = p^{(i)}(0,1) + p^{(i)}(1,1)$, pour $i \in \{0,1\}$. La probabilité $p_{ij} = \mathbb{E}[\zeta_{2+i}^j]$ est le nombre attendu de descendants de type j d'un individu de type i. Lorsque tous les termes de la matrice $\boldsymbol{P}$ sont positifs, $\boldsymbol{P}$ admet une valeur propre dominante strictement positive et simple, (voir Theorem 5.1 de (84)), nous la noterons π. Le paramètre π détermine le comportement asymptotique du processus, il joue le rôle du nombre attendu d'enfants dans le cas du Galton Watson standard et est relié à l'extinction du processus. Cette extinction est de probabilité inférieure à 1 lorsque le Galton-Watson est dit sur-critique i.e. lorsque $\pi > 1$. Si on définit l'extinction du Galton-Watson par $\mathcal{E} = \cup_{n\geq 1}\{\boldsymbol{Z}_n = (0,0)\}$, il est clair que cette extinction est aussi celle de $(\delta_k)_{k\in\mathbb{T}}$: $\mathcal{E} = \bigcup_{n\geq 1}\{|\mathbb{G}_n^*| = 0\}$. Aussi les résultats asymptotiques des estimateurs seront établis sur l'ensemble de non-extinction $\overline{\mathcal{E}}$, complémentaire de $\mathcal{E}$. Nous faisons l'hypothèse suivante pour garantir une probabilité non-nulle à $\overline{\mathcal{E}}$:

(HO) Les termes de la matrice $\boldsymbol{P}$ sont positifs : pour tout $(i,j) \in \{0,1\}^2$, $p_{ij} > 0$, et sa valeur propre dominante vérifie : $\pi > 1$.

Sous l'hypothèse (**HO**), on a $\mathbb{P}(\mathcal{E}) < 1$ et π^n est un grand $\mathcal{O}$ déterministe de $\boldsymbol{Z}_n$, $|\mathbb{G}_n^*|$ et $|\mathbb{T}_n^*|$. Plus précisément, il existe une variable aléatoire positive W telle que

$$\lim_{n\to+\infty} \frac{\boldsymbol{Z}_n}{\pi^n} = \lim_{n\to+\infty} \frac{\pi - 1}{\pi^{n+1} - 1} \sum_{\ell=0}^{n} \boldsymbol{Z}_\ell = W\boldsymbol{z} \quad \text{p.s..} \tag{2.3}$$

où $\boldsymbol{z} = (z^0, z^1)$ est le vecteur propre à droite pour la valeur propre π de $\boldsymbol{P}$ vérifiant $z^0 + z^1 = 1$. On a alors que $\{W = 0\} = \mathcal{E}$ p.s., ou encore que l'événement $\{W > 0\}$ est l'ensemble de non-extinction $\overline{\mathcal{E}}$ de $(\boldsymbol{Z}_n)$ à un ensemble négligeable près. Les propriétés (2.2) et (2.3) entrainent alors

$$\lim_{n\to+\infty} \frac{|\mathbb{G}_n^*|}{\pi^n} = \lim_{n\to+\infty} \frac{\pi - 1}{\pi^{n+1} - 1} |\mathbb{T}_n^*| = W \quad \text{p.s.} \tag{2.4}$$

Le lemme suivant, utilisé de nombreuses fois dans nos travaux est une conséquence directe des propriétés du processus de généalogie.

Lemme 2.3.1 *Sous l'hypothèse* (**HO**)*, on a*

$$\lim_{n\to+\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \frac{\pi^n}{|\mathbb{T}_n^*|} = \frac{\pi - 1}{\pi} \frac{1}{W} \mathbb{1}_{\overline{\mathcal{E}}} \quad p.s.$$

2.4 Modèles de BAR

2.4.1 BAR asymétrique d'ordre p

Modèle

L'étude du modèle présenté dans cette section et correspondant à l'article (6) est à l'initiative de Bernard Bercu (IMB) et en collaboration avec Benoîte de Saporta (IMB). Le modèle est un BAR asymétrique d'ordre p défini comme suit. Soit p un entier non nul. Le processus BAR(p) asymétrique est défini pour tout $k \geq 2^{p-1}$, par

$$\left\{ \begin{array}{lcl} X_{2k} & = & a_0 + \sum_{\ell=1}^{p} a_k X_{[\frac{k}{2^{\ell-1}}]} + \epsilon_{2k}, \\ X_{2k+1} & = & b_0 + \sum_{\ell=1}^{p} b_k X_{[\frac{k}{2^{\ell-1}}]} + \epsilon_{2k+1}. \end{array} \right. \tag{2.5}$$

Les états initiaux $\{X_k,\ 1 \leq k \leq 2^{p-1}-1\}$ sont les ancêtres du processus et $(\epsilon_{2k}, \epsilon_{2k+1})$ est le processus de bruit. Les paramètres $(a_0, a_1, \ldots a_p)$ and $(b_0, b_1, \ldots, b_p)$ sont des nombres réels inconnus que nous cherchons à estimer. Nous supposons que les matrices companion $p \times p$ $\boldsymbol{A}$ and $\boldsymbol{B}$ définies par

$$\boldsymbol{A} = \begin{pmatrix} a_1 & a_2 & \cdots & a_p \\ 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} b_1 & b_2 & \cdots & a_p \\ 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

vérifient la propriété de contraction suivante

$$\beta = \max\{\|\boldsymbol{A}\|, \|\boldsymbol{B}\|\} < 1. \tag{2.6}$$

Ce processus est une généralisation directe du BAR(p) symétrique étudié par Huggins, Basawa et Zhou (87; 135). Dans le cas particulier où $p = 1$, il correspond au modèle étudié par Guyon dans (81).

Estimateur

Notons $\boldsymbol{\theta}$ le vecteur des paramètres du modèle et $\widehat{\boldsymbol{\theta}}_n$ son estimateur, ils sont définis par

$$\boldsymbol{\theta} = \begin{pmatrix} a_0 \\ \vdots \\ a_p \\ b_0 \\ \vdots \\ b_p \end{pmatrix} \qquad \text{and} \qquad \widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{a}_{0,n} \\ \vdots \\ \widehat{a}_{p,n} \\ \widehat{b}_{0,n} \\ \vdots \\ \widehat{b}_{p,n} \end{pmatrix}.$$

Nous estimons $\boldsymbol{\theta}$ à partir des donnés X_k des individus k jusqu'à la $n$ième génération, c'est-à-dire l'observation complète du sous arbre $\mathbb{T}_n$. Pour cela nous utilisons l'estimateur des moindres carrés $\widehat{\boldsymbol{\theta}}_n$ qui minimise

$$\Delta_n(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k \in \mathbb{T}_{n-1,p-1}} (X_{2k} - a_0 - \sum_{\ell=1}^{p} a_k X_{[\frac{k}{2^{\ell-1}}]})^2 + (X_{2k+1} - b_0 - \sum_{\ell=1}^{p} a_k X_{[\frac{k}{2^{\ell-1}}]})^2.$$

Pour $n \geq p$, cet estimateur est donné par

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\Sigma}_{n-1}^{-1} \sum_{k \in \mathbb{T}_{n-1,p-1}} \begin{pmatrix} X_{2k} \\ X_{2k}\mathbb{X}_k \\ X_{2k+1} \\ X_{2k+1}\mathbb{X}_k \end{pmatrix} \tag{2.7}$$

où $\mathbb{X}_n = (X_n, X_{[\frac{n}{2}]}, \ldots, X_{[\frac{n}{2^{p-1}}]})^t$ et la matrice $(p+1) \times (p+1)$ $\boldsymbol{\Sigma}_n$ définie par $\boldsymbol{\Sigma}_n = \mathbf{I}_2 \otimes \boldsymbol{S}_n$ avec $\mathbf{I}_2$ matrice identité et $\otimes$ qui est symbolise le produite de Kronecker et

$$\boldsymbol{S}_n = \sum_{k \in \mathbb{T}_{n,p-1}} \begin{pmatrix} 1 & \mathbb{X}_k^t \\ \mathbb{X}_k & \mathbb{X}_k\mathbb{X}_k^t \end{pmatrix}.$$

Dans le cas $p = 1$, $\boldsymbol{\theta} = (a_0, a_1, b_0, b_1)^t$ et l'estimateur s'écrit

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\Sigma}_{n-1}^{-1} \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} X_{2k} \\ X_k X_{2k} \\ X_{2k+1} \\ X_k X_{2k+1} \end{pmatrix}, \tag{2.8}$$

où, pour tout $n \geq 0$, $\boldsymbol{\Sigma}_n = \begin{pmatrix} \boldsymbol{S}_n & 0 \\ 0 & \boldsymbol{S}_n \end{pmatrix}$ et $\boldsymbol{S}_n$ est donnée simplement par $\boldsymbol{S}_n = \sum_{k \in \mathbb{T}_n} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}$. Nous ne donnons pas ici, les estimateurs des paramètres σ^2 et ρ que le lecteur peut consulter dans (6) et qui seront donnés plus bas dans un cas d'observation plus général lorsque $p = 1$.

Hypothèses sur la suite des bruits

Nous ne donnons pas dans cette synthèse l'énoncé rigoureux des hypothèses qui peut être consulté dans (6). L'hypothèse essentielle est que le bruit est une différence de martingale qui vérifie des conditions de moments et que les cellules d'une même génération qui ne sont pas soeurs sont conditionnellement indépendantes sachant les générations précédentes. Plus précisément, si $\mathbb{F} = (\mathcal{F}_n)$ est la filtration naturelle associée au processus BAR(p) avec $\mathcal{F}_n$ la σ-algèbre engendrée par les individus jusqu'à la n^e génération, le degré exigé des moments variant suivant les résultats, nous posons pour $q \leq 4$,

(HN.1-q) Pour tout $n \geq 0$ et tout $k \in \mathbb{G}_{n+1}$, ϵ_k est dans L^q et

$$\sup_{n \geq 0} \sup_{k \in \mathbb{G}_{n+1}} \mathbb{E}[\epsilon_k^q | \mathcal{F}_n] < \infty \qquad \text{p.s.}$$

De plus, il existe $\sigma^2 \in (0, +\infty)$, $|\rho'| \in [0, 1)$ tels que :
– $\forall n \geq 0$ and $k \in \mathbb{G}_{n+1}$,

$$\mathbb{E}[\epsilon_k | \mathcal{F}_n] = 0, \qquad \mathbb{E}[\epsilon_k^2 | \mathcal{F}_n] = \sigma^2, \text{ p.s.}$$

– $\forall n \geq 0 \quad \forall k \neq l \in \mathbb{G}_{n+1}$ avec $[k/2] = [l/2]$,

$$\mathbb{E}[\epsilon_k \epsilon_l | \mathcal{F}_n] = \rho = \rho' \sigma^2 \text{p.s.}$$

(HN.2) Pour tout $n \geq 0$ les vecteurs aléatoires $\{(\epsilon_{2k}, \epsilon_{2k+1}), k \in \mathbb{G}_n\}$ sont conditionnellement indépendants sachant $\mathcal{F}_n$.

Discussion

Ce modèle généralise celui de Guyon (81) dans plusieurs directions. En effet Guyon étudie un BAR d'ordre 1 lorsque le bruit est indépendant et identiquement distribué et possède des moments à tout ordre. Même si dans le cas $p = 1$, l'estimateur que nous proposons est le même que celui de Guyon, nous allégeons ces hypothèses d'indépendance et de moments. Nous verrons plus loin que nous utilisons des outils différents pour démontrer les résultats de convergence, basés sur des résultats de convergence de martingales comme ceux de Duflo (69) et que cette méthode nous permet de donner une vitesse et une loi forte quadratique en plus de la convergence p.s. et du TCL établis par Guyon.

Les modèles présentés dans la suite de cette section sont tous issus de travaux en collaboration avec Benoîte de Saporta (IMB) et Laurence Marsalle (Université de Lille).

2.4.2 BAR(1) avec données manquantes

Pour poursuivre le travail initié en (6) sur le modèle de la section 2.4.1, il nous a paru intéressant de proposer une méthode d'estimation qui puissent tenir compte des nombreuses données manquantes dues soit à la mort des cellules, soit à un défaut de mesure (voir section 2.8 pour un exemple de recueil des données). Le modèle de BAR sous-jacent est donc un BAR(1) défini par (2.5) pour $p = 1$ que nous redonnons ici pour modifier les notations des coefficients

$$\left\{ \begin{array}{lcccccl} X_{2k} & = & a & + & b\, X_k & + & \epsilon_{2k} \\ X_{2k+1} & = & c & + & d\, X_k & + & \epsilon_{2k+1}. \end{array} \right.$$

sous l'hypothèse analogue à (2.6) : $\beta = \max\{|b|, |d|\} < 1$.

Modèle

Les cellules pouvant mourir ou ne pas être observées, les X_k ne sont donc pas tous observés voire même pas tous définis en cas de mort. On va donc allier le processus de généalogie δ_k défini en 2.2 et le BAR. Rappelons que si $\delta_{2k} = 1$ ou $\delta_{2k+1} = 1$ alors $\delta_k = 1$. On pose alors

$$\left\{ \begin{array}{lcccccll} X_{2k} & = & a & + & b\, X_k & + & \epsilon_{2k} & \text{si} \quad \delta_{2k} = 1, \\ X_{2k+1} & = & c & + & d\, X_k & + & \epsilon_{2k+1} & \text{si} \quad \delta_{2k+1} = 1, \end{array} \right. \tag{2.9}$$

On s'appuiera donc sur les X_k observés pour faire de l'inférence sur les paramètres (a, b, c, d). Une approche naïve serait de faire apparaître un indicateur de présence δ_k devant chaque X_k dans la formule (2.8) de $\widehat{\boldsymbol{\theta}}_n$ et dans celle de la matrice $\boldsymbol{S}_n$, mais nous allons voir que cette approche donne un estimateur différent de l'estimateur "des moindres carrés observés" défini ci-dessous.

Estimateur

Nous proposons l'estimateur des moindres carrés de $\boldsymbol{\theta} = (a, b, c, d)^t$ à partir des X_k observés dans le sous-arbre $\mathbb{T}_n^*$. Cet estimateur est celui qui minimise

$$\Delta_n(\boldsymbol{\theta}) = \sum_{k\in\mathbb{T}_{n-1}} \delta_{2k}(X_{2k} - a - bX_k)^2 + \delta_{2k+1}(X_{2k+1} - c - dX_k)^2.$$

Par conséquent, pour $n \geq 1$, il est donné par

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{a}_n \\ \widehat{b}_n \\ \widehat{c}_n \\ \widehat{d}_n \end{pmatrix} = \boldsymbol{\Sigma}_{n-1}^{-1} \sum_{k\in\mathbb{T}_{n-1}} \begin{pmatrix} \delta_{2k}X_{2k} \\ \delta_{2k}X_kX_{2k} \\ \delta_{2k+1}X_{2k+1} \\ \delta_{2k+1}X_kX_{2k+1} \end{pmatrix}, \tag{2.10}$$

où, pour tout $n \geq 0$,

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} \boldsymbol{S}_n^0 & 0 \\ 0 & \boldsymbol{S}_n^1 \end{pmatrix}, \quad \text{et} \quad \boldsymbol{S}_n^i = \sum_{k\in\mathbb{T}_n} \delta_{2k+i} \begin{pmatrix} 1 & X_k \\ X_k & X_k^2 \end{pmatrix}.$$

Cet estimateur est à rapprocher de celui du modèle complètement observé donné en (2.8). Lorsque le modèle est complètement observé, tous les δ_k valent 1 et les estimateurs définis en (2.8) et (2.10) sont les mêmes. Cependant il est différent d'un estimateur "naïf" qui consiste à rajouter δ_k devant X_k ; en effet l'expression de la matrice $\boldsymbol{\Sigma}_n$ fait apparaître l'indicateur de la présence d'une fille de type pair (δ_{2k}) pour $\boldsymbol{S}_n^0$ et de type impair (δ_{2k+1})pour $\boldsymbol{S}_n^1$. Nous donnons maintenant les estimateurs de la variance du bruit σ^2 et de la corrélation ρ. Nous donnons pour cela leurs estimateurs empiriques donnés par

$$\widehat{\sigma}_n^2 = \frac{1}{|\mathbb{T}_n^*|} \sum_{k\in\mathbb{T}_{n-1}} (\widehat{\epsilon}_{2k}^2 + \widehat{\epsilon}_{2k+1}^2), \quad \widehat{\rho}_n = \frac{1}{|\mathbb{T}_{n-1}^{*01}|} \sum_{k\in\mathbb{T}_{n-1}} \widehat{\epsilon}_{2k}\widehat{\epsilon}_{2k+1}, \tag{2.11}$$

où pour tout k dans $k \in \mathbb{G}_n$,

$$\begin{cases} \widehat{\epsilon}_{2k} &= \delta_{2k}(X_{2k} - \widehat{a}_n - \widehat{b}_nX_k), \\ \widehat{\epsilon}_{2k+1} &= \delta_{2k+1}(X_{2k+1} - \widehat{c}_n - \widehat{d}_nX_k), \end{cases}$$

et

$$\mathbb{T}_n^{*01} = \{k \in \mathbb{T}_n : \delta_{2k}\delta_{2k+1} = 1\},$$

l'ensemble $\mathbb{T}_{n-1}^{*01}$ est l'ensemble des cellules de $\mathbb{T}_{n-1}$ qui ont exactement deux filles. De ses formules, on peut déduire aisément les estimateurs de σ^2 et ρ dans le cas complètement observé. Il est à noter que dans le cas général, les dénominateurs $|\mathbb{T}_n^*|$ et $|\mathbb{T}_{n-1}^{*01}|$ sont des termes aléatoires dépendants du processus d'observation alors qu'ils sont déterministes dans le cas complètement observé.

Hypothèses

Les hypothèses qui sont faîtes sur ce modèle sont **(HN.1-8)** et **(HN.2)** sur la suite des bruits du BAR auxquelles s'ajoute **(HO)** pour le processus d'observation. Enfin nous rajoutons l'hypothèse d'indépendance entre le BAR et le processus d'observation :

(HI) Les suites (δ_k) et $(\boldsymbol{\zeta}_k)$ sont indépendantes des suites (X_k) et (ϵ_k).

Discussion

Ce modèle généralise bien sûr (2.5) lorsque $p = 1$. C'est aussi une extension de Delmas-Marsalle (64) qui modélisent les données manquantes par un processus de Galton-Watson ; mais celui-ci n'autorise pas des lois de reproduction différentes suivant le type de la mère. La méthode nous permet de donner une vitesse de convergence et la loi forte quadratique.

2.4.3 Observation de plusieurs arbres

Les études de simulations sur le modèles à données manquantes publiées dans (11) et résumées au paragraphe 2.8, ainsi que l'application du modèle sur chacune des généalogies d' E.Coli de l'étude (122) ont mis en évidence un manque de puissance des tests d'asymétrie. Or les donnée issues d'expériences de division cellulaire sont généralement des mesures réalisées sur plusieurs descendances dans des conditions identiques mais indépendantes. C'est pourquoi nous avons jugé opportun de proposer un estimateur des paramètres du BAR et du Galton-Watson adapté à l'observation d'un nombre fixe d'arbres de descendance, indépendants et identiquement distribués. Ce travail, est l'objet d'un article soumis (15).

Modèle

Nous voulons ici faire de l'inférence sur les paramètres d'un BAR(1) à partir de l'observation de m généalogies i.i.d. $(\delta_{(j,k)})_{1\leq j\leq m}$ vérifiant les hypothèses de la section 2.2 et des processus BAR(1) associés à ces m généalogies et eux aussi i.i.d. Pour $1 \leq j \leq m$, la première cellule de l'arbre j est notée $(j, 1)$ et pour $k \geq 1$, les deux filles de la celule (j, k) sont notées $(j, 2k)$ et $(j, 2k+1)$. La présence de la cellule k dans la généalogie j est donnée par $\delta_{(j,k)}$ et la caractéristique quantitative de cette cellule est notée $X_{(j,k)}$. Suivant la définition (2.9), nous avons donc pour $1 \leq j \leq m$ et $k \geq 1$,

$$\left\{ \begin{array}{lcl} X_{(j,2k)} & = & a + bX_{(j,k)} + \epsilon_{(j,2k)}, \quad \text{si} \quad \delta_{(j,2k)} = 1 \\ X_{(j,2k+1)} & = & c + dX_{(j,k)} + \epsilon_{(j,2k+1)}, \quad \text{si} \quad \delta_{(j,2k+1)} = 1 \end{array} \right. \tag{2.12}$$

Non-extinction

Dans le but de donner les propriétés asymptotiques des estimateurs lorsque le nombre de cellules dans l'ensemble des m arbres tend vers l'infini (ce qui signifie qu'au moins l'un des arbres a un nombre infini de cellules), on définit comme dans la section 2.2 l'ensemble des caractéristiques de chaque arbre. Pour $1 \leq j \leq m$, le nombre de cellules présentes dans la génération n (resp. le sous arbre des n premières générations) de l'arbre j est donné par

$$\mathbb{G}^*_{j,n} = \{k \in \mathbb{G}_n : \delta_{(j,k)} = 1\}, \quad \mathbb{T}^*_{j,n} = \{k \in \mathbb{T}_n : \delta_{(j,k)} = 1\}.$$

Le nombre total de cellules observées dans l'ensemble des m arbres à la génération n et jusqu'à la génération n respectivement est donc :

$$|\mathbb{G}^*_n| = \sum_{j=1}^{m} |\mathbb{G}^*_{j,n}| \quad \text{et} \quad |\mathbb{T}^*_n| = \sum_{j=1}^{m} |\mathbb{T}^*_{j,n}|.$$

L'extinction de la généalogie j s'écrit alors $\mathcal{E}_j = \bigcup_{n\geq 1}\{|\mathbb{G}^*_{j,n}| = 0\}$. Si son complémentaire i.e., l'ensemble de non-extinction de l'arbre j est noté $\overline{\mathcal{E}}_j$, la non-extinction dans l'ensemble des m arbres correspond à l'union des $\overline{\mathcal{E}}_j$, soit

$$\overline{\mathcal{E}} = \bigcup_{j=1}^{m} \overline{\mathcal{E}}_j = \{\lim_{n\to\infty} |\mathbb{T}^*_n| = \infty\}.$$

Sur l'événement $\overline{\mathcal{E}}$ certains arbres peuvent s'éteindre, à des générations différentes de surcroît, pourvu qu'un au moins un d'entre eux soit infini.

Hypothèses

Chaque généalogie doit vérifier **(HO)**, chaque BAR les hypothèses **(HN.1-8)** et **(HN.2)** et pour chaque j, l'indépendance **(HI)** doit être vérifiées. A ces hypothèses sur chaque arbre, s'ajoute l'indépendances entre les m réalisations :

(HI.B) Les suites $(\varepsilon_{(1,k)})_{k\geq 2}, (\varepsilon_{(2,k)})_{k\geq 2}, \ldots, (\varepsilon_{(m,k)})_{k\geq 2}$ sont indépendantes. Les suites aléatoires $(X_{(j,1)})_{1\leq j\leq m}$ sont indépendantes et indépendantes des suites des bruits.

(HI.G) Les suites $(\delta_{(1,k)})_{k\geq 2}, (\delta_{(2,k)})_{k\geq 2}, \ldots, (\delta_{(m,k)})_{k\geq 2}$ sont indépendantes.

(HI.BG) Pour tout $1 \leq j \leq m$, la suite $(\delta_{(j,k)})_{k\geq 1}$ est indépendante des suites $(X_{(j,k)})_{k\geq 1}$ et $(\varepsilon_{(j,k)})_{k\geq 2}$.

Estimateur de la loi de reproduction

Même si la loi de reproduction commune aux processus de généalogie $(\delta_{j,k})$ est liée à celle du Galton-Watson associé $(\boldsymbol{Z}_{j,n})$, il serait réducteur d'utiliser les estimateurs donnés dans la littérature pour les Galton-Watson multi-types étudiés par exemple dans (36), (80) ou (100). En effet, notre contexte d'observation du Galton-Watson est spécifique puisque l'on connaît exactement la présence ou l'absence de chaque cellule de l'arbre binaire à travers l'observation des $(\delta_{(j,k)})$ ce qui est plus précis que l'observation du nombre de cellules de chaque type dans chaque génération donné par les $(\boldsymbol{Z}_{j,n})$ habituellement utilisés dans la littérature. On donne donc les estimateurs empiriques des probabilité de reproduction définies à la section 2.2 à l'aide de l'observation des cellules jusqu'à la génération n. Pour i, l_0, l_1 dans $\{0, 1\}$

$$\widehat{p}^{(i)}_n(l_0, l_1) = \frac{\sum_{j=1}^{m}\sum_{k\in\mathbb{T}_{n-2}} \delta_{(j,2k+i)}\phi_{l_0}(\delta_{(j,2(2k+i))})\phi_{l_1}(\delta_{(j,2(2k+i)+1)})}{\sum_{j=1}^{m}\sum_{k\in\mathbb{T}_{n-2}} \delta_{(j,2k+i)}}, \tag{2.13}$$

avec $\phi_0(x) = 1-x$, $\phi_1(x) = x$. Par exemple pour estimer $p^0(0,1)$ la probabilité qu'une cellule de type pair ait 0 fille de type pair et 1 de type impair, on fait le quotient de la somme des $\delta_{(j,2k)}(1-\delta_{(j,4k)})\delta_{(j,4k+1)}$, indicateurs des filles paires qui ont une unique fille de type impair et de la somme des $\delta_{(j,2k)}$ indicateurs des filles paires. A partir de cet estimateur, on peut aisément donner un estimateur du rayon spectral π par exemple ou des nombres espérés de descendants de chaque type suivant celui de la mère. Ces estimateurs sont détaillés dans (11).

Estimateur des paramètres du BAR

Cette fois-ci, $\widehat{\boldsymbol{\theta}}_n$ est l'estimateur des moindres carrés observés sur les m arbres. Il minimise

$$\Delta_n(\boldsymbol{\theta}) = \sum_{j=1}^{m} \sum_{k \in \mathbb{T}_{n-1}} \delta_{(j,2k)} (X_{(j,2k)} - a - bX_{(j,k)})^2 + \delta_{(j,2k+1)} (X_{(j,2k+1)} - c - dX_{(j,k)})^2,$$

et il est donné par

$$\widehat{\boldsymbol{\theta}}_n = \boldsymbol{\Sigma}_{n-1}^{-1} \sum_{j=1}^{m} \sum_{k \in \mathbb{T}_{n-1}} \begin{pmatrix} \delta_{(j,2k)} X_{(j,2k)} \\ \delta_{(j,2k)} X_{(j,k)} X_{(j,2k)} \\ \delta_{(j,2k+1)} X_{(j,2k+1)} \\ \delta_{(j,2k+1)} X_{(j,k)} X_{(j,2k+1)} \end{pmatrix} \tag{2.14}$$

avec, pour $i \in \{0, 1\}$,

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} \boldsymbol{S}_n^0 & 0 \\ 0 & \boldsymbol{S}_n^1 \end{pmatrix}, \quad \boldsymbol{S}_n^i = \sum_{j=1}^{m} \boldsymbol{S}_{j,n}^i, \quad \boldsymbol{S}_{j,n}^i = \sum_{k \in \mathbb{T}_n} \delta_{(j,2k+i)} \begin{pmatrix} 1 & X_{(j,k)} \\ X_{(j,k)} & X_{(j,k)}^2 \end{pmatrix}.$$

Cet estimateur est à rapprocher de l'estimateur obtenu pour un arbre unique en (2.10) ou des sommes sur les arbres apparaissent dans la matrice normalisatrice Σ_n et au numérateur. Nous ne donnons pas ici les estimateurs de σ^2 et ρ qui sont facilement généralisables à partir des formules (2.11).

Discussion

Le modèle et les estimateurs proposés dans cette section permettent de proposer un estimateur des lois de reproduction et des paramètres du BAR à partir de plusieurs arbres observés en tenant compte des données manquantes et du nombre différents de générations de chaque arbre. C'est exactement le schéma des données de (122) que nous discutons à la section 2.8.

2.4.4 BAR à coefficients aléatoires

Nous présentons dans cette section un modèle de BAR(1) à coefficients aléatoires qui généralise le modèle BAR(1) à données manquantes (2.9). En effet, dans la définition du BAR(1) (2.9), les coefficients d'autorégression b et d sont déterministes. Pour amener de la souplesse dans le modèle BAR standard, on peut rendre ces coefficients aléatoires en leur ajoutant un bruit de moyenne nulle. Ceci permet de rendre compte d'un aléa supplémentaire dû à l'environnement comme l'irrégularité de la concentration de nutriments et/ou un lieu de vie différent pour la cellule mère et les cellules filles par exemple.

A notre connaissance, seuls deux papiers traitent à ce jour des BAR à coefficients aléatoires : le premier de Bui and Huggins (44) date de 1999 alors que le second de Blandin (41) est lui très récent. Dans Bui et Huggins, non seulement le BAR est symétrique mais les termes déterministes a et c sont supposés nuls. De plus, les bruits sont gaussiens et sans corrélation entre le terme constant et le terme de régression. Blandin dans (41), se place dans la lignée des BAR (81; 6; 42) en étudiant un modèle de BAR asymétrique à coefficients aléatoires

complètement observé et propose un estimateur des moindres carrés pondérés. Le processus d'observation (δ_k) que nous proposons est cependant un peu moins général que celui proposé au paragraphe 2.3 (voir plus bas **(H.Obis)**. En effet nous supposons pour des raisons que nous expliquerons plus loin et comme dans (64) que les taux de reproduction des cellules ne dépendent pas du type de la mère.

Modèle

$$\left\{ \begin{array}{rcl} X_{2k} & = & (b + \eta_{2k})X_k + (a + \varepsilon_{2k}), \quad \text{si} \quad \delta_{2k} = 1 \\ X_{2k+1} & = & (d + \eta_{2k+1})X_k + (c + \varepsilon_{2k+1}) \quad \text{si} \quad \delta_{2k+1} = 1. \end{array} \right. \tag{2.15}$$

Hypothèses

(HN.1-4γ) Les variables aléatoires ε_2, η_2, ε_3, η_3 er X_1 ont des moments jusqu'à l'ordre 4γ, pour $\gamma \geq 1$, de plus

$$\begin{array}{c} \mathbb{E}[\varepsilon_2] = \mathbb{E}[\varepsilon_3] = 0, \;\; \mathbb{E}[\varepsilon_2^2] = \mathbb{E}[\varepsilon_3^2] = \sigma_\varepsilon^2 > 0 \;\; \text{et} \;\; \mathbb{E}[\varepsilon_2\varepsilon_3] = \rho_\varepsilon, \\ \mathbb{E}[\eta_2] = \mathbb{E}[\eta_3] = 0, \;\; \mathbb{E}[\eta_2^2] = \mathbb{E}[\eta_3^2] = \sigma_\eta^2 > 0 \;\; \text{et} \;\; \mathbb{E}[\eta_2\eta_3] = \rho_\eta, \\ \mathbb{E}[\varepsilon_{2+i}\eta_{2+j}] = \rho_{ij}, \;\; \text{pour } (i,j) \in \{0,1\}, \;\; \text{et } \rho = \frac{1}{2}(\rho_{01} + \rho_{10}). \end{array}$$

(HN.2bis) Pour $k \geq 1$, les vecteurs $(\varepsilon_{2k}, \eta_{2k}, \varepsilon_{2k+1}, \eta_{2k+1})_{k\geq 1}$ sont indépendants, équidistribués et de plus indépendants de X_1.

(HN.3) Il existe $1 \leq \kappa \leq \gamma$ tel que

$$\frac{p_0 + p_{01}}{m}\mathbb{E}[(b+\eta_2)^{4\kappa}] + \frac{p_1 + p_{01}}{m}\mathbb{E}[(d+\eta_3)^{4\kappa}] < 1.$$

(H.Obis) Pour tout $(i, j_0, j_1) \in \{0,1\}^3$, $p^0(j_0, j_1) = p^1(j_0, j_1) = p(j_0, j_1)$ et la moyenne de la loi de reproduction $\pi = p(0,1) + p(1,0) + 2p(1,1)$ est supérieure à 1 : $\pi > 1$.

(H.Ibis) La suite des $(\boldsymbol{\xi}_k)_{k\in\mathbb{T}}$ est indépendante de la suite des bruits $(\varepsilon_{2k}, \eta_{2k}, \varepsilon_{2k+1}, \eta_{2k+1})_{k\in\mathbb{T}}$ et de X_1.

Ces hypothèses sont bien sûr une adaptation au cas du BAR à coefficients aléatoires des hypothèses des sections précédentes. **(HN.1-4γ)** est une hypothèses de moments sur le bruit et **(HN.1-2bis)** une hypothèse d'indépendance. **(HN.3)**, dont le rôle est d'assurer la non-explosion de l'autorégression, généralise $\max\{|b|, |d|\} < 1$ du cas déterministe. Si on se ramène au cas simple ($\eta_2 = \eta_3 = 0$) et complètement observé, cette équation se réduit à $(b^{4\kappa} + d^{4\kappa})/2 < 1$. D'autre part, nous supposons comme dans (64) que les taux de reproduction des cellules ne dépendent pas du type de la mère, c'est la première partie de **(H.Obis)** ; ainsi le nombre espéré π de filles pour une cellule devient le critère de non-extinction et les équivalents du nombre de cellules présentes à la génération n ou dans le sous-arbre $\mathbb{T}_n$ donnés par l'équation (2.4) restent valable sous cette hypothèse.

Estimateur

L'estimateur est celui des moindre carrés qui minimise

$$\Delta_n(\boldsymbol{\theta}) = \sum_{k\in\mathbb{T}_{n-1}} \delta_{2k}(X_{2k} - a - bX_k)^2 + \delta_{2k+1}(X_{2k+1} - c - dX_k)^2.$$

Malgré la présence du bruit supplémentaire (η_{2k}, η_{2k+1}) l'expression de $\widehat{\theta}_n$ est la même que lorsque ce bruit est nul et est donc donnée par (2.10).

En revanche, les résidus de la régression mélangent les deux bruits. Ces derniers et leurs estimateurs connaissant les n premières génération sont définis, pour $k \in \mathbb{G}_n$, par

$$\left\{\begin{array}{lcl} \epsilon_{2k} &=& \delta_{2k}(\varepsilon_{2k} + \eta_{2k}X_k), \\ \epsilon_{2k+1} &=& \delta_{2k+1}(\varepsilon_{2k+1} + \eta_{2k+1}X_k), \end{array}\right. \quad \left\{\begin{array}{lcl} \widehat{\epsilon}_{2k} &=& \delta_{2k}(X_{2k} - \widehat{a}_n - \widehat{b}_nX_k), \\ \widehat{\epsilon}_{2k+1} &=& \delta_{2k}(X_{2k+1} - \widehat{c}_n - \widehat{d}_nX_k). \end{array}\right. \tag{2.16}$$

Pour estimer $\boldsymbol{\sigma} = (\sigma_\varepsilon^2, \rho_{00}, \rho_{11}, \sigma_\eta^2)^t$, termes de variance et de covariance des bruits $(\varepsilon_2, \eta_2, \varepsilon_3, \eta_3)$, on procède comme dans (107) en utilisant une méthode des moindres carrés entre le bruit $(\epsilon_{2k}, \epsilon_{2k+1})$ sachant les X_k observés (voir plus loin en (2.19) la définition des tribus $\mathcal{F}_\ell^{\mathcal{O}}$) et le résidu observé correspondant $(\widehat{\epsilon}_{2k}, \widehat{\epsilon}_{2k+1})$ en minimisant

$$\Delta'_n(\boldsymbol{\sigma}) = \frac{1}{2}\sum_{\ell=1}^{n-1}\sum_{k\in\mathbb{G}_\ell} (\widehat{\epsilon}_{2k}^2 - \mathbb{E}[\epsilon_{2k}^2|\mathcal{F}_\ell^{\mathcal{O}}])^2 + (\widehat{\epsilon}_{2k+1}^2 - \mathbb{E}[\epsilon_{2k+1}^2|\mathcal{F}_\ell^{\mathcal{O}}])^2.$$

Les hypothèses du modèle assurent

$$\widehat{\boldsymbol{\sigma}}_n \;=\; \boldsymbol{U}_{n-1}^{-1} \sum_{k\in\mathbb{T}_{n-1}} \left(\widehat{\epsilon}_{2k}^2 + \widehat{\epsilon}_{2k+1}^2, 2X_k\widehat{\epsilon}_{2k}^2, 2X_k\widehat{\epsilon}_{2k+1}^2, X_k^2(\widehat{\epsilon}_{2k}^2 + \widehat{\epsilon}_{2k+1}^2)\right)^t, \tag{2.17}$$

où

$$\boldsymbol{U}_n = \sum_{k\in\mathbb{T}_n} \begin{pmatrix} \delta_{2k} + \delta_{2k+1} & 2\delta_{2k}X_k & 2\delta_{2k+1}X_k & (\delta_{2k} + \delta_{2k+1})X_k^2 \\ 2\delta_{2k}X_k & 4\delta_{2k}X_k^2 & 0 & 2\delta_{2k}X_k^3 \\ 2\delta_{2k+1}X_k & 0 & 4\delta_{2k+1}X_k^2 & 2\delta_{2k+1}X_k^3 \\ (\delta_{2k} + \delta_{2k+1})X_k^2 & 2\delta_{2k}X_k^3 & 2\delta_{2k+1}X_k^3 & (\delta_{2k} + \delta_{2k+1})X_k^4 \end{pmatrix}.$$

De même l'estimateur du vecteur des paramètres $\boldsymbol{\rho} = (\rho_\varepsilon, \rho, \rho_\eta)^t$ est l'estimateur des moindres carrés qui minimise

$$\Delta''_n(\boldsymbol{\rho}) = \frac{1}{2}\sum_{\ell=1}^{n-1}\sum_{k\in\mathbb{G}_\ell} (\widehat{\epsilon}_{2k}\widehat{\epsilon}_{2k+1} - \mathbb{E}[\epsilon_{2k}\epsilon_{2k+1}|\mathcal{F}_\ell^{\mathcal{O}}])^2,$$

il est donné par

$$\widehat{\boldsymbol{\rho}}_n = \boldsymbol{V}_{n-1}^{-1} \sum_{k\in\mathbb{T}_{n-1}} \left(\widehat{\epsilon}_{2k}\widehat{\epsilon}_{2k+1}, 2X_k\widehat{\epsilon}_{2k}\widehat{\epsilon}_{2k+1}, X_k^2\widehat{\epsilon}_{2k}\widehat{\epsilon}_{2k+1}\right)^t, \tag{2.18}$$

où

$$\boldsymbol{V}_n = \sum_{k\in\mathbb{T}_n} \delta_{2k}\delta_{2k+1} \begin{pmatrix} 1 & 2X_k & X_k^2 \\ 2X_k & 4X_k^2 & 2X_k^3 \\ X_k^2 & 2X_k^3 & X_k^4 \end{pmatrix}.$$

Discussion

Il est remarquable que les estimateurs des paramètres $(a, b, c, d)^t$ du BAR s'expriment de la même manière que dans le modèle (2.9). Cependant les X_k intervenant dans les résidus de la régression, nous aurons besoin d'hypothèse de moments supplémentaires pour les convergence. Lorsque $\sigma_\eta^2 = 0$, les estimateurs des moindres carrés de σ_ε^2 et ρ_ε coïncident avec les empiriques donnés dans les sections précédentes ou dans (9). Notons enfin que ρ_{01} et ρ_{10} ne sont pas identifiables et que nous estimons $\rho = (\rho_{01} + \rho_{10})/2$.

2.5 La convergence presque sûre

Nous donnons dans cette section et la suivante, les résultats asymptotiques obtenus pour les estimateurs et discutons les différentes méthodes de démonstration que nous avons utilisées suivant les modèles. Les résultats donnés ici ont été obtenus pour tous les modèles cités plus haut mais les hypothèses notamment de moments peuvent différer d'un modèle à l'autre. Pour les différents modèles présentés dans la section 2.4, nous avons le résultat de consistance donné par le théorème suivant :

Théorème 2.5.1 *Lorsque n tend vers l'infini, l'estimateur $\widehat{\boldsymbol{\theta}}_n$ (resp. $\widehat{\boldsymbol{\sigma}}_n$, $\widehat{\boldsymbol{\rho}}_n$) converge presque sûrement vers $\boldsymbol{\theta}$ (resp. $\boldsymbol{\sigma}$, $\boldsymbol{\rho}$) sur l'ensemble de non-extinction $\overline{\mathcal{E}}$. Plus précisément :*

$$\begin{aligned} \lim_{n\to\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\widehat{\boldsymbol{\theta}}_n &= \boldsymbol{\theta}\mathbb{1}_{\overline{\mathcal{E}}} \quad p.s. \qquad et \\ \lim_{n\to\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\widehat{\boldsymbol{\sigma}}_n &= \boldsymbol{\sigma}\mathbb{1}_{\overline{\mathcal{E}}}\ p.s., \qquad \lim_{n\to\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\widehat{\boldsymbol{\rho}}_n = \boldsymbol{\rho}\mathbb{1}_{\overline{\mathcal{E}}}\ p.s. \end{aligned}$$

Cet énoncé est bien sûr valable dans le cas complètement observé, les ensembles $(|\mathbb{G}_n^*| > 0)$ et $\mathbb{1}_{\overline{\mathcal{E}}}$ étant des ensembles de probabilité 1 dans ce cas, on retrouve bien la convergence p.s. énoncée dans (6). Pour ce qui est de la démonstration de ces propriétés, dans tous les cas, elles reposent sur des résultats de lois des grands nombres. Si on regarde attentivement par exemple l'estimateurs proposé pour $\widehat{\boldsymbol{\theta}}_n$ en (2.10), si on montre séparément les convergences $\boldsymbol{\Sigma}_{n-1}/|\mathbb{T}_{n-1}^*|$ et $\boldsymbol{\Sigma}_{n-1}\widehat{\boldsymbol{\theta}}_n/|\mathbb{T}_{n-1}^*|$, on se rend compte que l'on doit établir les lois des grands nombres suivantes : $\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\frac{1}{|\mathbb{T}_n^*|}\sum_{k\in\mathbb{T}_n}\delta_{2k+i}X_k^q$ pour $i \in \{0,1\}$ et $q \in \{0,1,2\}$ et $\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}\frac{1}{|\mathbb{T}_n^*|}\sum_{k\in\mathbb{T}_n}\delta_{2k+i}X_k\varepsilon_{2k+i}$.

Notons que dans (6; 9), la convergence p.s. est montrée stricto sensu via la propriété de vitesse (2.24) que nous verrons dans la section suivante. Mais pour démontrer cette vitesse, nous devons établir des quantités de lois des grands nombres. Nous distinguons ci-dessous deux méthodes pour démontrer ces lois des grands nombres : la méthode utilisant les propriétés de convergence de martingales que nous utilisons dans (6; 9; 11; 15) et la méthode chaînes de Markov bifurquantes utilisée dans (14) à la suite de (81) et (64).

2.5.1 Méthode martingale

Il est à noter que les méthodes martingales pour l'inférence dans les BAR ont été utilisées pour la première fois dans notre article (6). La démonstration des lois des grands nombres impliquant les X_k repose sur le lemme 5.2. de (9) qui donne la limite $\lim_{n\to\infty}\sum_{\ell=0}^n \boldsymbol{A}_{n-\ell}\boldsymbol{B}_\ell = \boldsymbol{AB}$ lorsque la série des matrices $\boldsymbol{A}_\ell$ converge vers $\boldsymbol{A}$ et la suite des vecteurs $\boldsymbol{B}_\ell$ converge

vers $\boldsymbol{B}$. On montre par exemple dans les Lemme 6.3. et 6.4 de (9) que les vecteurs $\boldsymbol{H}_n(q) = (H_n^0(q), H_n^1(q))^t$ lorsque $H_n^i(q) = \sum_{k\in\mathbb{T}_n} \delta_{2k+i} X_k^q$, s'écrivent à peu de choses près comme une somme du type $\sum_{\ell=0}^n \boldsymbol{A}_{n-\ell}(q)\boldsymbol{B}_\ell(q)$. Les propriétés de contraction du BAR assurent la convergence de la séries des matrices. La convergence du vecteur $\boldsymbol{B}_\ell(q)$ repose, elle, sur de nouvelles lois des grands nombres. C'est pour démontrer ces lois des grands nombres que nous utilisons, à de multiples reprises, la loi des grands nombres pour les martingales réelles de carré intégrable donnée par exemple dans le Théorème 1.3.15 de (69). Ce résultat dit que là où la limite de son processus croissant prévisible converge, la martingale converge vers une variable aléatoire finie p.s. tandis que le quotient de la martingale par son crochet tend vers 0 ailleurs. Nous renvoyons le lecteur à Duflo (69) pour une formulation plus précise. Le nombre de lois des grands nombres à établir pour obtenir la convergence de nos estimateurs est impressionnant, les lemmes 5.3, 5.4, 5.5, 5.6 et 5.7 du seul papier (9) en sont des exemples.

Quant aux difficultés apportées par l'introduction des données manquantes par rapport au modèle BAR de départ, nous en donnons une idée dans ce paragraphe. Elle multiplie les filtrations à considérer pour montrer les différentes lois des grands nombres. A la filtration $\mathbb{F} = (\mathcal{F}_n)$ déjà définie plus haut (hypothèses) s'ajoutent deux filtrations associées au processus de généalogie : $\mathcal{Z}_n = \sigma\{\boldsymbol{\zeta}_k, k \in \mathbb{T}_n\}$, $\mathcal{O}_n = \sigma\{\delta_k, k \in \mathbb{T}_n\}$ et bien sûr la filtration des X_k observés : $\mathbb{F}^{\mathcal{O}} = (\mathcal{F}_n^{\mathcal{O}})$ définie par

$$\mathcal{F}_n^{\mathcal{O}} = \mathcal{O} \vee \sigma\{\delta_k X_k, k \in \mathbb{T}_n\} = \mathcal{O} \vee \sigma\{X_k, k \in \mathbb{T}_n^*\}, \tag{2.19}$$

où $\mathcal{O} = \sigma\{\delta_k, k \in \mathbb{T}\}$. La possibilité d'extinction complique aussi l'exercice. Les processus croissants des martingales sont de l'ordre de $|\mathbb{T}_n^*|$ qui tend vers l'infini sur l'ensemble de non-extinction uniquement. C'est pourquoi, grâce au Lemme 5.1. de (9) utilisé avec $a_n = \pi^n$ nous démontrons d'abord les résultats des lois des grands nombres avec π^n remplaçant $|\mathbb{T}_n^*|$. Le lemme 2.3.1 permet de revenir facilement à $|\mathbb{T}_n^*|$.

En ce qui concerne l'approche multi-arbre, nous avons vu que l'écriture de l' estimateur se déduisait aisément de celle d'un estimateur "monoarbre". Mais ce nouvel estimateur n'est pas une fonction des m estimateurs mono-arbres, il faut donc en redémontrer la consistance. Le lemme suivant, démontré dans (15) et qui dit que si une loi des grands nombres est valable sur l'ensemble de non-extinction d'un arbre donné, alors elle l'est aussi sur celui des m arbres, permet de simplifier cette démonstration.

Lemme 2.5.1 *Soit $(x_{(1,n)}), \ldots, (x_{(m,n)})$ m suites de nombres réels tels que pour $1 \le j \le m$, on a la limite presque sûre*

$$\lim_{n\to\infty} \frac{\mathbb{1}_{\{|\mathbb{G}_{j,n}^*|>0\}}}{|\mathbb{T}_{j,n}^*|} \sum_{k\in\mathbb{T}_n} x_{(j,k)} = \ell \mathbb{1}_{\overline{\mathcal{E}}_j}, \tag{2.20}$$

alors sous les hypothèses **(HO)** *et* **(HI.G)**, *on a*

$$\lim_{n\to\infty} \frac{\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}}{|\mathbb{T}_n^*|} \sum_{j=1}^m \sum_{k\in\mathbb{T}_n} x_{(j,k)} = \ell \mathbb{1}_{\overline{\mathcal{E}}} \quad p.s.$$

2.5.2 Méthode chaînes de Markov bifurquantes

Introduites par Guyon dans (81), les chaînes de Markov bifurcantes (BMC) sont des chaînes de Markov indexées par un arbre binaire ; ainsi un modèle de BAR peut être considéré comme un BMC particulier. Guyon a établi des lois des grands nombres et un théorème central limite associés aux BMC qui lui permettent de démontrer des résultats de consistance et un TCL pour l'estimateur des paramètres du BAR complètement observé sous des hypothèses discutées plus haut. Dans (64), Delmas & Marsalle généralisent le modèle de Guyon à un arbre de Galton-Watson bi-type pouvant éventuellement s'éteindre. Leur résultat généralise les résultats de Guyon mais aussi d'autres résultats de lois des grands nombres pour des processus de Markov indexés par des arbres aléatoires mais sans extinction possible. Depuis, Bansaye & al (38) ont étudié des extensions au cas continu. Dans notre cas, c'est le processus (X_n^*) égal à X_n lorsque celui-ci est observé et à un point cimetière sinon, défini par

$$X_n^* = X_n \mathbb{1}_{\{\delta_n=1\}} + \partial \mathbb{1}_{\{\delta_n=0\}}, \tag{2.21}$$

qui est une BMC sur l'arbre de Galton-Watson. En effet, il vérifie la propriété

$$\mathbb{E}\left[\prod_{k\in\mathbb{G}_n} f_k(X_{2k}^*, X_{2k+1}^*) \;\middle|\; \sigma(X_j^*, j\in\mathbb{T}_n)\right] = \prod_{k\in\mathbb{G}_n} Pf_k(X_k^*).$$

avec le noyau de transition P défini sur $\mathbb{T}^*$ pour f fonction mesurable et positive de $\overline{\mathbb{R}}^3$ dans $\mathbb{R}$ par

$$\begin{aligned} Pf(x) &= p(1,1)\mathbb{E}\left[f(x,(b+\eta_2)x+a+\varepsilon_2,(d+\eta_3)x+c+\varepsilon_3)\right] \\ &\quad +p(1,0)\mathbb{E}\left[f(x,(b+\eta_2)x+a+\varepsilon_2,\partial)\right] \\ &\quad +p(0,1)\mathbb{E}\left[f(x,\partial,(d+\eta_3)x+c+\varepsilon_3)\right]+(1-p_{01}-p_0-p_1)f(x,\partial,\partial). \end{aligned} \tag{2.22}$$

A partir de cette chaîne de Markov bifurquante, on fabrique la chaîne de Markov réelle induite (Y_n) qui correspond à la suite des valeurs de X_k que l'on obtiendrait si on descendait au hasard (un hasard qui respecte les proportions données par les $p(j_0,j_1)$!) le long d'un arbre de Galton Watson conditionnellement à la non-extiction et que l'on note les valeurs des X_k de chaque noeud visité. Pour définir un tel processus, on procède en deux étapes :

- Si on pose $a_2 = a$, $b_2 = b$, $a_3 = c$ and $b_3 = d$, la suite $(A_n, B_n)_{n\geq 1}$ est une suite i.i.d. de même loi que $(a_{2+\zeta}+\varepsilon_{2+\zeta}, b_{2+\zeta}+\eta_{2+\zeta})$, où ζ est une Bernoulli de paramètre $(p_{01}+p_1)/m$ indépendante des $(\varepsilon_2,\eta_2,\varepsilon_3,\eta_3)$.
- On pose $Y_0 = X_1^* = X_1$ et on définit Y_{n+1} de manière récursive par

$$Y_{n+1} = A_{n+1} + B_{n+1}Y_n. \tag{2.23}$$

D'après (43; 62), la chaîne auto-régressive (Y_n) admet une loi invariante μ et sous les hypothèses de moments **(HN.4γ)** et de contraction du BAR **(HN.3)**, on établit aussi l'ergodicité géométrique pour les fonctions polynomiales de degré inférieur à 2κ, on a alors l'existence d'une constante positive c telle que si ν est la loi de X_1, pour $n\in\mathbb{N}$, on a $\left|\mathbb{E}_\nu[f(Y_n)] - \langle\mu, f\rangle\right| \leq c\|B_1\|_{4\kappa}^n$.

Grâce à cette propriété, on adapte la démonstration du Théorème 3.1 de (64) pour démontrer le résultat suivant :

Théorème 2.5.2 *Sous les hypothèses du BAR à coefficients aléatoires, pour toute fonction f dans F_q défini par ,*

$$F_q = vect\{x^\alpha y^\beta \mathbb{1}_{\mathbb{R}}(y),\ x^\alpha z^\tau \mathbb{1}_{\mathbb{R}}(z),\ x^\alpha y^\beta z^\tau \mathbb{1}_{\mathbb{R}^2}(y,z), \quad (\alpha,\beta,\tau) \in \mathbb{N}^3,\ 0 \leq \alpha+\beta+\tau \leq q\},$$

on a les lois des grands nombres

$$\lim_{n\to\infty} \frac{1}{\pi^n} \sum_{k\in\mathbb{G}_n^*} f(X_k^*, X_{2k}^*, X_{2k+1}^*) = \langle \mu, Pf\rangle W \quad p.s.$$

L'ensemble de fonctions F_κ est plus général que l'ensemble F du théorème 3.1. de (64) et ne nécessite pas des moments à tout ordre de nos bruits $(\varepsilon_2, \eta_2, \varepsilon_3, \eta_3)$. Cependant nous ne montrons le résultat non pas pour tous les BMC mais pour la classe de BMC déduite du BAR qui nous intéresse. Le théorème est très puissant et permet d'obtenir aisément des lois des grands nombres pour des sommes de polynômes impliquant les X_k, X_{2k}, X_{2k+1} et d'exprimer les limites en fonction des moments de la loi invariante μ. En voici un exemple :

$$\lim_{n\to\infty} \frac{\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}}}{|\mathbb{T}_{n-1}^*|} \sum_{k\in\mathbb{T}_{n-1}^*} \delta_{2k}\delta_{2k+1} X_k^q X_{2k+i} = p_{01}(a_{2+i}\mathbb{E}[Y_\infty^q] + b_{2+i}\mathbb{E}[Y_\infty^{q+1}])\mathbb{1}_{\overline{\mathcal{E}}}$$

La variable aléatoire Y_∞ a pour loi μ et ses moments sont calculés récursivement grâce à la relation d'autorégression (2.23) et aux lois de A_1 et B_1.

2.6 Vitesse et loi forte quadratique

Le résultat de convergence p.s. peut être affiné en donnant une vitesse et une loi forte quadratique. Nous donnons ici la version du résultat dans le cadre du modèle (2.9).

Théorème 2.6.1 *Sous les hypothèses* **(HN.1-8)**, **(HN.2)**, **(HO)** *et* **(HI)**, *on a la vitesse de convergence presque sûre, sur l'ensemble de non-extinction $\overline{\mathcal{E}}$ donnée par*

$$\mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|^2 = \mathcal{O}\left(\frac{\log|\mathbb{T}_{n-1}^*|}{|\mathbb{T}_{n-1}^*|}\right)\mathbb{1}_{\overline{\mathcal{E}}} \qquad p.s. \tag{2.24}$$

de plus, on a la loi forte quadratique

$$\lim_{n\to\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \frac{1}{n}\sum_{\ell=1}^{n} |\mathbb{T}_{\ell-1}^*| (\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta})^t \boldsymbol{\Sigma} (\widehat{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}) = 4\frac{\pi-1}{\pi}\sigma^2 \mathbb{1}_{\overline{\mathcal{E}}} \qquad p.s. \tag{2.25}$$

où $\boldsymbol{\Sigma}$ est une matrice 4×4 finie, limite sur l'ensemble de non-extinction de $\frac{\boldsymbol{\Sigma}_n}{|\mathbb{T}_n^*|}$.

Remarquons tout d'abord que $\frac{\log|\mathbb{T}_{n-1}^*|}{|\mathbb{T}_{n-1}^*|}$ est de l'ordre de n/π^n avec $\pi = 2$ dans le cas complètement observé. La vitesse est toujours la même dans le cas des coefficients aléatoires (on la démontre sous les hypothèses de la section 2.4.4 si $\kappa \geq 4$). De la même manière, si on définit

$$\sigma_n^2 = \frac{1}{|\mathbb{T}_n^*|}\sum_{k\in\mathbb{T}_{n-1}^*} (\delta_{2k}\varepsilon_{2k}^2 + \delta_{2k+1}\varepsilon_{2k+1}^2), \quad \rho_n = \frac{1}{|\mathbb{T}_{n-1}^{*01}|}\sum_{k\in\mathbb{T}_{n-1}^*} \delta_{2k}\varepsilon_{2k}\delta_{2k+1}\varepsilon_{2k+1}.$$

on a un résultat de vitesse de convergence de $\widehat{\sigma}_n^2$ (resp $\widehat{\rho}_n$) vers σ_n^2 (resp ρ_n). Ce résultat est donné dans le Théorème 3.3 de (9) dont on déduit aisément la version complètement observée, on démontre aussi un résultat analogue pour les termes de covariance des coefficients aléatoires dans les Théorèmes 3.3 et 3.4 de (14).

Quant à la loi forte quadratique, on l'obtient aussi pour tous les modèles avec des matrices normalisatrices un peu différentes dans le cas coefficients aléatoires à cause des X_k présents dans le bruit $(\epsilon_{2k}, \epsilon_{2k+1})$.

Ces vitesses reposent sur des résultats de convergence de martingales. En effet, on peut faire apparaître une terme martingale dans $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}$ et démontrer des résultats de convergence pour cette martingale. Le processus $(\boldsymbol{M}_n)$ défini par les équations suivantes est une $(\mathcal{F}_n^{\mathcal{O}})$-martingale sous les hypothèses du modèle.

$$\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} = \boldsymbol{\Sigma}_{n-1}^{-1} \sum_{k \in \mathbb{T}_{n-1}} (\epsilon_{2k}, X_k\epsilon_{2k}, \epsilon_{2k+1}, X_k\epsilon_{2k+1})^t = \boldsymbol{\Sigma}_{n-1}^{-1} \boldsymbol{M}_n, \tag{2.26}$$

avec

$$\boldsymbol{M}_n = \sum_{k \in \mathbb{T}_{n-1}} (\epsilon_{2k}, X_k\epsilon_{2k}, \epsilon_{2k+1}, X_k\epsilon_{2k+1})^t . \tag{2.27}$$

Une première remarque sur $\boldsymbol{M}_n$ est que le nombre de terme de $\boldsymbol{M}_{n+1} - \boldsymbol{M}_n$ croît exponentiellement puisqu'il vaut exactement 2^n dans le cas observé et il est de l'ordre de π^n sur l'ensemble de non-extinction sinon. Cette particularité de la martingale $\boldsymbol{M}_n$ ne nous a pas permis d'appliquer les résultats de vitesse de convergence ou de loi forte quadratique pour les martingales vectorielles standard que l'on trouve par exemple dans le théorème 2.II.8 de Duflo (69). Nous avons dû l'adapter et le démontrer pour plusieurs martingales dans chacun de nos travaux pour finalement le systématiser dans le résultat suivant correspondant au théorème 5.1 de (14).

Théorème 2.6.2 *Soit $(\boldsymbol{M}_n)$ une $\mathbb{F}^{\mathcal{O}}$-martingale à valeurs dans $\mathbb{R}^d$ sur l'arbre de Galton-Watson $\mathbb{T}^*$: $\boldsymbol{M}_n = \sum_{\ell=1}^n \sum_{k \in \mathbb{G}_\ell^*} \boldsymbol{W}_k$, avec $\boldsymbol{W}_k = (w_k^1, w_k^2, \ldots, w_k^p)^t$. Sous les quatre hypothèses suivantes :*

(A.1) *$(\boldsymbol{M}_n)$ est de carré intégrable.*

On note $< \boldsymbol{M} >_n = \sum_{\ell=0}^{n-1} \boldsymbol{\Gamma}_\ell$ son processus croissant avec

$$\boldsymbol{\Gamma}_n = \mathbb{E}[\Delta \boldsymbol{M}_{n+1} \Delta \boldsymbol{M}_{n+1}^t \mid \mathcal{F}_n^{\mathcal{O}}].$$

(A.2) *$|\mathbb{T}_{n-1}^*|^{-1} < \boldsymbol{M} >_n$ converge p.s. vers une matrice semi-définie positive $\boldsymbol{\Gamma}$ sur $\overline{\mathcal{E}}$.*

(A.3) *La $p \times p$ $\mathbb{F}^{\mathcal{O}}$-martingale $(\boldsymbol{K}_n)$ à valeur matricielle définie par*

$$\boldsymbol{K}_n = \sum_{\ell=1}^n |\mathbb{T}_\ell^*|^{-1} (\Delta \boldsymbol{M}_{\ell+1} \Delta \boldsymbol{M}_{\ell+1}^t - \mathbb{E}[\Delta \boldsymbol{M}_{\ell+1} \Delta \boldsymbol{M}_{\ell+1}^t \mid \mathcal{F}_\ell^{\mathcal{O}}])$$

est de carré intégrable et les processus croissants de chacun de ses termes sont des $\mathcal{O}(n)$ p.s. sur $\overline{\mathcal{E}}$.

La suite des matrices $p \times p$ symétriques $(\boldsymbol{\xi}_n)$ adaptées à $\mathbb{F}^{\mathcal{O}}$, vérifie

(A.4) *$|\mathbb{T}_n^*|^{-1} \boldsymbol{\xi}_n$ admet une limite définie positive $\boldsymbol{\xi}$ p.s. sur $\overline{\mathcal{E}}$.*

On a alors que $\boldsymbol{M}_n^t \boldsymbol{\xi}_{n-1}^{-1} \boldsymbol{M}_n = \mathcal{O}(n)$ *et* $\|\boldsymbol{M}_n\|^2 = \mathcal{O}(n\pi^n)$ *p.s. sur* $\overline{\mathcal{E}}$.

Si de plus les termes de $(\boldsymbol{M}_n)$ *satisfont*
(A.5) $\sup_n \mathbb{E}[(\pi^{-n/2} \sum_{k \in \mathbb{G}_n^*} w_k^i)^4 \mid \mathcal{F}_{n-1}^{\mathcal{O}}] < \infty$ *p.s.,*
alors pour tout $\delta > 1/2$, *on a* $\|\boldsymbol{M}_n\|^2 = o(n^\delta \pi^n)$ *p.s. et*

$$\lim_{n\to\infty} \mathbb{1}_{\{|\mathbb{G}_n^*|>0\}} \frac{1}{n} \sum_{\ell=1}^{n} \boldsymbol{M}_\ell^t \boldsymbol{\xi}_{\ell-1}^{-1} \boldsymbol{M}_\ell = tr(\boldsymbol{\Gamma}\boldsymbol{\xi}^{-1}) \mathbb{1}_{\overline{\mathcal{E}}} \quad p.s.$$

Ce théorème est appliqué plusieurs fois pour montrer les vitesses de convergence : à $\boldsymbol{M}_n$ et $\boldsymbol{\xi}_n$ qui est une sorte de version "standardisée" du processus croissant de $\boldsymbol{M}_n$ pour démontrer le théorème 2.6.1 ; à deux reprises pour démontrer la vitesse de convergence de $\widehat{\sigma}_n^2$ vers σ_n^2 et deux fois encore pour la convergence de $\widehat{\rho}_n$ vers ρ_n. C'est la martingale $(\boldsymbol{M}_n)$ ou successivement le vecteur constitué de ses deux premières ou deux dernières composantes qui est utilisée comme martingale principale dans ces convergences. Les matrices normalisatrices étant plus compliquées, notamment dans le cadre des coefficients aléatoires (voir par exemple les lemmes 5.8 et 5.9 de (14) pour voir des application du théorème 2.6.2 à ces résultats).

2.7 Normalité asymptotique

Nous obtenons aussi des résultats de normalité asymptotique pour nos estimateurs. Ceux-ci sont utiles pour en déduire des intervalles de confiance asymptotiques et des tests de Wald détaillés dans (15). Ces résultats sont établis sous la probabilité conditionnelle à la non-extinction $\mathbb{P}_{\overline{\mathcal{E}}}$ défini par $\mathbb{P}_{\overline{\mathcal{E}}}(A) = \mathbb{P}(A \cap \overline{\mathcal{E}})/\mathbb{P}(\overline{\mathcal{E}})$.

Nous donnons tout d'abord le résultat pour les paramètres du processus de généalogie qui sont aussi ceux du Galton-Watson associé. Notons

$$\boldsymbol{p}^{(i)} = (p^{(i)}(0,0), p^{(i)}(0,1), p^{(i)}(1,0), p^{(i)}(1,1))^t$$

le vecteur des 4 probabilités de reproduction pour les mères de type i, $\boldsymbol{p} = ((\boldsymbol{p}^{(0)})^t, (\boldsymbol{p}^{(1)})^t)^t$ le vecteur des 8 probabilité de reproduction et $\widehat{\boldsymbol{p}}_n$ son estimateur empirique donné par (2.13)

Théorème 2.7.1 *Sous les hypothèses* **(HO)** *et* **(HI-G)**, *on la convergence*

$$\sqrt{|\mathbb{T}_{n-1}^*|}(\widehat{\boldsymbol{p}}_n - \boldsymbol{p}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tilde{\boldsymbol{V}}) \quad \text{sur } (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}}),$$

avec $\tilde{\boldsymbol{V}}$ *défini au Théorème 4 de (15).*

Nous donnons l'énoncé du résultat pour les coefficients du BAR dans la version "coefficients aléatoires".

Théorème 2.7.2 *Si* $\kappa \geq 8$, *on a le théorème central limite*

$$|\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}) \quad \text{sur } (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}}) \tag{2.28}$$

où $\boldsymbol{\Sigma}$ *(resp.* $\boldsymbol{\Gamma}$ *) est la limite de* $\boldsymbol{\Sigma}_n/|\mathbb{T}_{n-1}^*|$ *(reps.* $<\boldsymbol{M}>_n/|\mathbb{T}_{n-1}^*|$*) sur l'ensemble de non-extinction.*

Si de plus $\kappa \geq 16$,

$$|\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\boldsymbol{\sigma}}_n - \boldsymbol{\sigma}) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \boldsymbol{U}^{-1}\boldsymbol{\Gamma}^{\boldsymbol{\sigma}}\boldsymbol{U}^{-1}\Big) \quad sur \ \ (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}}), \tag{2.29}$$

et

$$|\mathbb{T}_{n-1}^*|^{1/2}(\widehat{\boldsymbol{\rho}}_n - \boldsymbol{\rho}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{V}^{-1}\boldsymbol{\Gamma}^{\boldsymbol{\rho}}\boldsymbol{V}^{-1}) \quad sur \ \ (\overline{\mathcal{E}}, \mathbb{P}_{\overline{\mathcal{E}}}), \tag{2.30}$$

où les matrices $\boldsymbol{U}$ *et* $\boldsymbol{V}$*, les limites de* $\boldsymbol{U}_n/|\mathbb{T}_{n-1}^*|$ *et* $\boldsymbol{V}_n/|\mathbb{T}_{n-1}^*|$*,* $\boldsymbol{\Gamma}^{\boldsymbol{\sigma}}$ *et* $\boldsymbol{\Gamma}^{\boldsymbol{\rho}}$ *sont définis respectivement dans la Proposition 4.14 et les équations (6.1) et (6.2) de (14).*

Ce théorème va permettre de donner des intervalles de confiance asymptotiques ou des tests de Wald pour l'asymétrie du processus de généalogie ou le processus BAR que l'on peut trouver dans (11) et (15). Pour démontrer cette normalité asymptotique, nous avons utilisé les résultats sur les suites de différences de martingales donnés dans les Propositions 7.8 et 7.9 d'Hamilton (83) dans le cadre complètement observé. Dans le cadre partiellement observé, le terme de normalisation $|\mathbb{T}_{n-1}^*|^{1/2}$ étant aléatoire, nous avons dû utiliser le théorème de la limite centrale appliqué à des suites de martingales triangulaires donné par Duflo (69) et rappelé dans le Théorème 14 de (15). Que ce soit le résultat d'Hamilton ou celui de Duflo, on ne peut pas les appliquer aux différences de la martingales $(\boldsymbol{M}_n)$ définie par 2.27 à cause du nombre exponentiel de termes dans chaque génération. Il faut donc changer de filtration et considérer la filtration induite par les soeurs définie par

$$\mathcal{G}_p^{\mathcal{O}} = \mathcal{O} \vee \sigma\{\delta_1 X_1, \ (\delta_{2k}X_{2k}, \delta_{2k+1}X_{2k+1}), \ 1 \leq k \leq p\}$$

et pour la suite des différences de martingales $\boldsymbol{D}_k = (\epsilon_{2k}, X_k\epsilon_{2k}, \epsilon_{2k+1}, X_k\epsilon_{2k+1})^t$. Il faut noter aussi que le TCL est conditionnel à la non-extinction du processus de généalogie.

2.8 Etude de simulation et application au problème de vieillissement d'E. Coli

Nous l'avons dit dans l'introduction de ce chapitre, les données de Stewart & al (122) ont motivé les travaux avec données manquantes et multi-arbres (9), (11) et (15).

Nous rappelons la problématique biologique et la structure des données récolées.

2.8.1 Problématique biologique

La question est celle du vieillissement des organismes unicellulaire. Un organisme unicellulaire se divise en son milieu pour donner naissance à deux cellules. Au cours de la division, deux nouvelles membranes ou pôles sont créés, chacune d'elles constituant l'une des extrémités de chacune des filles. Une cellule a donc un pôle "jeune" et un pôle plus "âgé". Lors de la division suivante, on peut donc distinguer les deux nouvelles filles : l'une a hérité du pôle jeune de sa mère et l'autre du pôle plus âgé. Et ainsi de suite. Ce phénomène est schématisé à la figure 2.3 issue de Stewart & al (122). Il est possible de suivre et de marquer ainsi toute une généalogie de cellules, de les indexer par un arbre binaire. Nous choisissons l'étiquette $2n$ (resp. $2n+1$) pour la fille "jeune" (resp. "âgée") de la cellule n. Par abus de notation, nous parlerons de type pair et impair pour ces cellules. A noter que les cellules dont le label est une

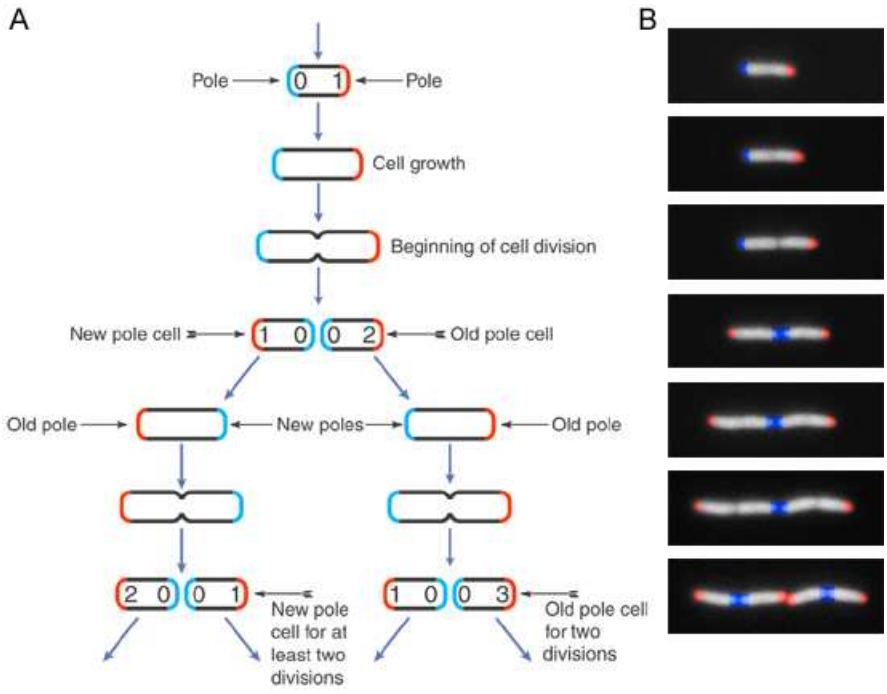


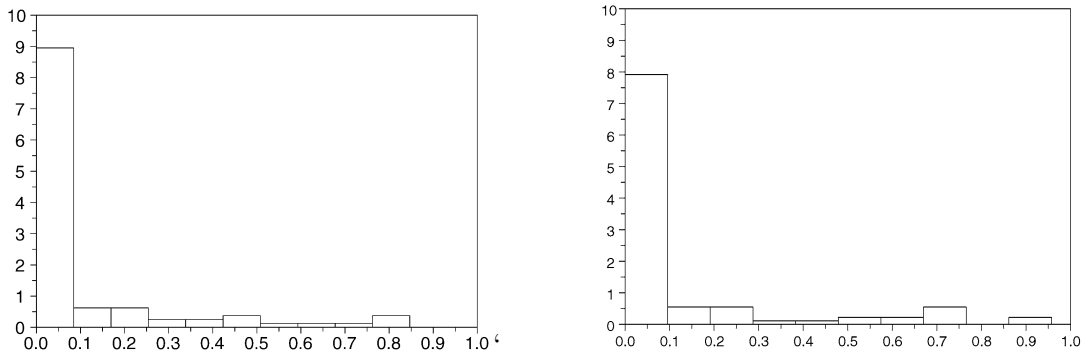
FIGURE 2.3 – Mécanisme de division, de Stewart, Madden, Paul and Taddei, *PLoS Biol.*, 2005.

puissance de deux (2^k) sont d'une lignée qui a toujours hérité du pôle jeune de sa mère alors que les cellules $2^k - 1$ ont toujours hérité du pôle "âgé", la généalogie des autres cellules est donnée par leur écriture en base deux. Si l'organisme vieillit, alors ce vieillissement se traduit par une dissymétrie entre les deux cellules filles que l'on mesure dans la loi de reproduction elle-même (les cellules paires ont une probabilité plus forte d'avoir deux filles, ...) ou par des caractéristiques physiologiques différentes (taux de croissance, biomasse, longueur, etc...).

Pour mesurer ce vieillissement, Stewart & al ont filmé 94 descendances d'Escherichia Coli et mesuré différents paramètres de chacune des cellules dont leur taux de croissance. Ces données rassemblent des informations sur 22394 cellules (11189 paires and 11205 impaires). Le nombre de divisions dans les généalogies va de quatre à neuf. La plupart des données manquantes sont dues à un défaut de mesure (cellules cachées par d'autres ou en dehors du champ de la caméra). On comptabilise 16 morts (non division) seulement dans ces 94 généalogies.

2.8.2 Application aux données et simulation

Ces données ont d'abord été traitées par (122) qui ont étudié le taux de croissance en réalisant un arbre moyen (à chaque noeud de l'arbre, est associé la moyenne des taux de croissance des cellules présentes à cette place de la généalogie) et un test de séries appariées a été utilisé pour comparer les taux de croissance et la biomasse de deux soeurs. Ils montrent ainsi un taux de croissance, une biomasse et une probabilité de reproduction plus faible des cellules de type impair. Cette analyse présente deux défauts majeurs à savoir que l'hypothèse

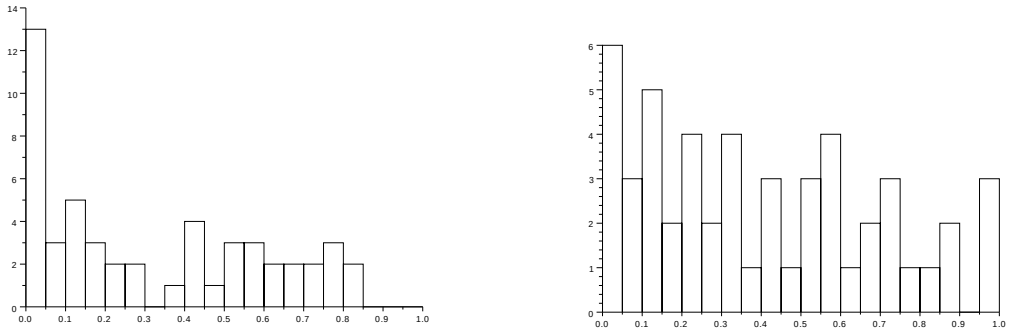


(a) $\boldsymbol{H_0} : (a, b) = (c, d)$ (b) $\boldsymbol{H_0} : a/(1-b) = c/(1-d)$

FIGURE 2.4 – Répartition des 94 p-valeurs des tests de symétrie du BAR de Guyon (81) sur les généalogies de Stewart & al (122).

d'indépendance entre chaque paire de soeurs n'est pas vérifiée (lien dû à la descendance) et que le nombre de cellules utilisées pour chaque moyenne n'est pas le même. Guyon (81) dont le modèle correspond au modèle donné par l'équation (2.5) lorsque $p = 1$ a réalisé les tests de Wald issus du TCL sur chacun des 94 arbres de généalogies. Le modèle ne gérant pas les données manquantes, nous ne savons pas comment ont été calculés les estimateurs avec les données réelles. Les résultats sont donnés dans les histogrammes de la Figure 2.4 et montrent que plus de 80 % des 94 p-valeurs sont inférieurs à 0.1 dans les deux cas ce qui correspond au rejet de l'hypothèse d'égalité des deux vecteurs (a, b) et (c, d) et des "points fixes" du BAR $a/(1-b)$ et $c/(1-d)$ donc de la symétrie de la division. C'est pourquoi nous avons été surprises par les résultats des tests de Wald issus du modèle (2.9) effectués sur les 51 généalogies présentant huit générations ou plus. Ces résultats sont présentés dans la Figure 2.5. Ce manque de puissance a été confirmé par des études de simulations présentées dans (11) et dont nous donnons un extrait dans le tableau 2.1 où des études de simulations sont présentées sous les deux hypothèses d'égalité et de non égalité des points fixes pour des descendances allant de 7 à 9 générations. On peut noter que le test est peu puissant pour des arbres de moins de 10 générations. Les tests ont aussi été réalisés sur les paramètres du processus de généalogie, ceux-ci ne montrent aucune différence significative entre les taux de reproduction des cellules paires et impaires, ceci quelle que soit la généalogie étudiée (cf Figure 1 de (11)). Ces résultats ne contredisent pas la possibilité d'un vieillissement d'E. Coli mesurable par une dissymétrie du mécanisme de reproduction ; la grande majorité des données manquantes étant causée par des défauts de mesure.

L'estimation globale sur les 94 généalogies des coefficients du BAR et du processus de reproduction grâce au modèle (2.12) est donnée dans les tableaux 1 à 3 de (15). Quant aux tests de Wald associés, l'égalité des couples $(a, b) = (c, d)$ est rejetée (p-value $= 10^{-5}$). Les deux points fixes du BAR ($a/(1-b)$ estimé 0.03773 et $c/(1-d)$ estimé à 0.03734) sont aussi significativement différents ($p = 2 \cdot 10^{-3}$) confirmant ainsi l'hypothèse d'une asymétrie dans la division d'E. coli. Pour ce qui est de la généalogie le nombre moyen de filles des cellules paires estimé à 1.2048, n'est pas significativement différent de celui des cellules impaires estimé à 1.2032 (p-valeur $= 0.9$).



(a) $\boldsymbol{H_0} : (a, b) = (c, d)$ (b) $\boldsymbol{H_0} : a/(1-b) = c/(1-d)$

FIGURE 2.5 – Répartition des 51 p-valeurs des tests de symétrie du BAR issus du modèle (2.9) sur les généalogies de 8 générations au moins de Stewart & al (122).

Gen	Sous $\boldsymbol{H_0} : a/1-b = c/(1-d)$			Sous $\boldsymbol{H_1}$		
	$p < 0.05$	$p < 0.01$	$p < 0.001$	$p < 0.05$	$p < 0.01$	$p < 0.001$
7	2.2	0.7	0	23.1	07.4	01.4
8	3.3	0.5	0.1	41.3	20.5	06.1
9	3.8	0.5	0	64.6	41.6	18.6
10	4.7	0.8	0	82.9	68.1	46.3
11	5.5	0.7	0.1	94.5	88.5	74.5

TABLE 2.1 – Proportions des p-valeurs inférieures aux seuils 0.05, 0.01 et 0.001 des tests d'asymétrie des points fixes du BAR (1000 replicas) $a = b = 0.5$ (sous (**H1**), $c = 0.5; d = 0.4$)

Enfin, en ce qui concerne le modèle avec coefficients aléatoires que nous appliquons à la généalogie la plus grande parmi les 94 de Stewart & al (122), les résultats des estimations sont les mêmes que dans (9) pour le modèle (2.9) mais les intervalles de confiance sont plus larges en présence des coefficients aléatoires (voir les tableaux 1 de chacun des papiers). Dans ce cadre il est bien sûr intéressant de calculer les variances σ_ε^2 et σ_η^2 et leur intervalle de confiance (qui nécessite l'estimation des moments d'ordre 4 du vecteur aléatoire $(\varepsilon_2, \eta_2, \varepsilon_3, \eta_3)$). Les tests de non-nullité de ces coefficients $H_0 : \sigma_\varepsilon^2 = 0$ (resp. $H_0 : \sigma_\eta^2 = 0$) ont pour p-valeur $p = 0.0799$ (resp. $p = 0.0671$).

2.9 Conclusion et perspectives

Même si notre étude des processus BAR en présence de données manquantes dans l'optique de mesurer l'asymétrie dans la division cellulaire est assez complète, il y a bien sûr encore beaucoup d'extensions possibles que nous aborderons ou non dans les mois ou années à venir. Le premier point est de lever l'hypothèse que les taux de reproduction des cellules ne dépendent pas du type de la mère dans le modèle à coefficients aléatoires présenté à la section 2.4.4 sans lequel nous n'avons plus de chaîne de Markov bifurcante associée. On nous a aussi suggéré de regarder un modèle multi arbres "à fragilité", c'est-à-dire avec bruit sur les coefficients du BAR propre à l'arbre : on est ici dans un cadre de coefficients aléatoires mais par arbre. On peut aussi étudier conjointement plusieurs paramètres liés aux cellules (masse, taux de croissance, ...) donnant ainsi un processus BAR multivarié. Un sujet sûrement ambitieux mais intéressant pour les biologistes, serait de chercher le lien entre le BAR et la mort des cellules (on suppose que la non-observation d'un paramètre est dûe à la mort de la cellule correspondante et non aux conditions de mesure). Cela nécessite de construire un modèle conjoint entre le BAR et le processus de généalogie et de lever bien sûr l'hypothèse d'indépendance **(H-I)**. Enfin, pour terminer cette liste non-exhaustive, on peut aussi, en se dégageant de la division cellulaire type E. Coli, étudier des modèles autorégressifs sur des arbres plus généraux que les arbres binaires.

Chapitre 3

Travail autour des PDMP

Ce chapitre décrit les travaux liés aux thématiques de l'équipe INRIA CQFD notamment les processus markoviens déterministes par morceaux que nous appellerons PDMP pour Piecewise Deterministic Markovian Processes en anglais. Ces processus sont au coeur des problématiques de l'équipe INRIA CQFD que ce soit par l'étude de problèmes théoriques, la recherche de résolution numérique des problèmes de contrôle ou leur utilisation pour modéliser des phénomènes de fonctionnement et de dégradation, notamment en milieu industriel. Pour développer ces outils notamment numériques, l'équipe utilise souvent la quantification optimale des variables aléatoires. Sous l'impulsion de François Dufour, j'ai travaillé avec ces thématiques notamment dans le cadre du projet ANR Fautocoes "Fault Tolerant Control for Embedded Systems" (n^o ANR-09-SEGI-004) du projet ARPEGE dont il est le responsable et de contrats industriels avec EADS Astrium et Thales Optronique.

Après avoir rappelé la définition des PDMP et le principe de la quantification optimale dans les deux premières sections, je présenterai l'utilisation des PDMP pour construire un modèle de propagation de fissures (section 3.3). La section 3.4 est dévolue à une étude plus théorique sur l'inférence de l'intensité de saut des PDMP. La section 3.5 revient à une étude plus appliquée de la modélisation d'un HUMS dans un cadre très spécifique. Enfin, la section 3.6 présente un travail plus statistique combinant quantification optimale et méthode SIR dans un modèle de régression semi-paramétrique.

3.1 Définition d'un PDMP

Dans les années 1980, les processus markoviens déterministes par morceaux ont été introduits en théorie des probabilités par M.H.A. Davis comme une classe générale de processus stochastiques non diffusifs (58; 59). Ils servent notamment à décrire des systèmes physiques dont la dynamique peut être perturbée par des événements ponctuels et aléatoires. Ces systèmes sont décrits par deux variables : à la variable euclidienne usuelle représentative de l'état, on adjoint une variable discrète, appelée mode ou régime, à valeurs dans un ensemble dénombrable. Dans ce contexte, la variable d'état est représentative des grandeurs physiques du système. Il peut s'agir, par exemple, de la position et de l'orientation d'un satellite dans l'espace ou encore de la pression dans un actionneur hydraulique. Le mode traduit quant à lui l'apparition des pannes ou des dysfonctionnements et caractérise ainsi le fonctionnement

du système, du régime nominal jusqu'à la panne complète. D'un point de vue mathématique, la trajectoire d'un tel processus peut se définir de façon très intuitive et par itération. Entre deux sauts, l'état du système suit une trajectoire régie par un flot indexé par le mode qui reste constant. Il existe deux types de sauts pouvant induire des ruptures sur le vecteur d'état ou sur le mode :

1) Le premier type de saut est de nature *déterministe*. Ce type de saut intervient lorsque la variable physique tente de franchir la frontière d'un ouvert.
2) Le second type de saut est *purement stochastique*. Il suit une loi *poissonienne* dont l'intensité est une fonction non linéaire du mode et de l'état.

Dans le contexte de la sûreté de fonctionnement, les processus markoviens déterministes par morceaux jouissent de propriétés particulièrement intéressantes.

1) Tout d'abord, les processus markoviens déterministes par morceaux possèdent une structure paramétrique très riche.
2) Ensuite, on peut souligner le caractère dynamique des composantes du vecteur d'état. Elles obéissent entre deux sauts à une équation différentielle et elles sont libres d'évoluer dans le temps, par opposition aux modèles constants par morceaux très souvent rencontrés dans la littérature. Cette caractéristique rend ces modèles particulièrement attractifs car très proches de la physique. Elle conduit le plus souvent à des modèles d'état régis par des équations différentielles.

Les PDMP dépendent de trois caractéristiques : le flot Φ, l'intensité de saut λ, qui régit la loi des temps entre les sauts et le noyau de transition Q qui donne la loi de la position juste après un saut. Un choix judicieux de l'espace d'état et des trois caractéristiques locales Φ, λ et Q permet de choisir des modèles traitant un large nombre de problèmes notamment en fiabilité et sûreté de fonctionnement, (voir (59) et (50)).

Nous donnons maintenant plus précisément la définition d'une classe de PDMP que nous utilisons dans ce chapitre.

On appelle $\mathcal{K}$ l'ensemble dénombrable des régimes possibles. Dans les applications pratiques, $\mathcal{K}$ sera assimilé à un ensemble fini. La composante physique du processus évolue dans un sous-ensemble ouvert E d'un espace métrique séparable $(\mathcal{E}, d)$. La dynamique du processus est donnée par les trois caractéristiques $(\lambda_k, Q_k, \Phi_k)_{k\in\mathcal{K}}$ qui dépendent du régime k.

– Quelque soit $k \in \mathcal{K}$, $\Phi_k : E \times \mathbf{R}_+ \to \overline{E}$ est un flot déterministe qui satisfait

$$\forall \xi \in E,\ \forall s, t \geq 0,\ \Phi_k(\xi, t+s) = \Phi_k(\Phi_k(\xi, t), s).$$

Pour $\xi \in E$, $t^\star(x_k)(\xi)$ est le temps d'atteinte de la frontière de E partant de ξ le long du flot Φ_k

$$t_k^\star(\xi) = \inf\{t > 0\ :\ \Phi_k(\xi, t) \in \partial E\},$$

avec la convention $\inf \emptyset = +\infty$.

– $\lambda_k : \overline{E} \to \mathbb{R}_+$ est le taux de saut dans le régime k. C'est une fonction mesurable qui vérifie

$$\forall \xi \in E,\ \exists \varepsilon > 0,\ \int_0^\varepsilon \lambda_k(\Phi_k(\xi, s))\mathrm{d}s < +\infty.$$

– Le noyau de Markov $Q_k : E \times \mathcal{E} \to [0,1]$, représente la transition du processus aux instants de sauts. Il vérifie

$$\forall \xi \in \overline{E},\ Q_k(\xi, \mathcal{K} \times \overline{E} \setminus \{(k,\xi)\} = 1 \quad \text{et} \quad Q_k(\xi, \mathcal{K} \times E) = 1.$$

Davis a démontré l'existence d'un espace de probabilité filtré $(\Omega, \mathcal{A}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}_{X_0})$ sur lequel le processus $(X_t)_{t\geq}0$ avec $X_t = (m_t, x_t)$ est bien défini (voir (59)). Ce processus est défini de manière itérative. La loi initiale de X_0 est données par ν_0. Partant de $(k_0, x_0) \in \mathcal{K} \times E$ sa trajectoire est définie comme suit. Le premier instant de saut T_1 est une variable aléatoire positive dont la loi est donnée par

$$\forall t \geq 0,\ \mathbf{P}_{\nu_0}(T_1 > t | X_0 = (k_0, x_0)) = \exp\left(-\int_0^t \lambda_{k_0}(\Phi_{k_0}(x_0, s))\mathrm{d}s\right) \mathbb{1}_{\{0\leq t<t^\star_{k_0}(x)\}}. \tag{3.1}$$

Cette définition de la fonction de survie permet de voir les deux types de saut possibles : le saut aléatoire de fonction de survie $\exp\left(-\int_0^t \lambda_{k_0}(\Phi_{k_0}(x_0,s))\mathrm{d}s\right)$ et le saut déterministe en $t^\star_{k_0}(x_0)$ si le saut aléatoire n'a pas eu lieu précédemment.
Ainsi, sur $[0, T_1[$, la trajectoire est donnée par

$$\begin{cases} m_t = k_0, \\ x_t = \Phi_{k_0}(x_0, t). \end{cases}$$

A l'instant T_1, le processus subit un saut. Il est initialisé en $Z_1 = (k_1, x_1)$ qui a pour loi $Q_{k_0}(\Phi_{k_0}(x_0, T_1), \cdot)$. Partant de X_{T_1} la loi du temps $S_2 = T_2 - T_1$ est donnée par 3.1 avec les nouveaux paramètres (k_1, x_1) et sur l'intervalle de temps $[T_1, T_2[$, le processus (X_t) suit la trajectoire

$$\begin{cases} m_t = k_1, \\ x_t = \Phi_{k_1}(x_1, t - T_1), \end{cases}$$

Et ainsi de suite de manière récursive.

Au PDMP est associée une chaîne de Markov $(Z_n, S_n)_{n\geq 0}$ définie à partir de $(X_t)_{t\geq 0}$ par $Z_n = X_{T_n}$, qui sont les positions juste après les sauts et les temps inter-sauts $S_n = T_n - T_{n-1}$ avec la convention $S_0 = 0$. Cette représentation des PDMP nous fait penser aux processus ponctuels marqués qui ont aussi une double représention mais les PDMP se différencient des processus marqués par la possibilité des sauts déterministes lorsque la variable physique atteint la frontière de l'ouvert.

Nous verrons que cette chaîne qui caractérise la trajectoire a beaucoup d'intérêt dans les problèmes qui nous concernent. Nous utiliserons plus loin la propriété que l'on peut trouver dans (59) (Chapitre 1, section 24) qui établit que cette chaîne est générée par un système dynamique. En effet, il existe deux fonctions mesurables φ et ψ et deux suites aléatoires indépendantes $(\varepsilon_n)_{n\geq 0}$ et $(\delta_n)_{n\geq 0}$, telles que pour tout $n \geq 1$,

$$\begin{cases} S_n &= \varphi(Z_{n-1}, \delta_{n-1}), \\ Z_n &= \psi(Z_{n-1}, S_n, \varepsilon_{n-1}). \end{cases} \tag{3.2}$$

D'autre part, la définition des PDMP n'est pas donnée ici dans sa généralité. On pourra consulter (59) pour une extension au cas l'ouvert où "vit" le processus n'est pas le même suivant le mode k.

3.2 Principe de la quantification optimale

Le mot “quantification" (“quantization” en anglais) est utilisé depuis les années 50 en ingénierie du signal ou de l'information. Quantifier signifiait discrétiser un signal continu avec un nombre fini de modalités ou “quantifieurs". Il est alors utile de choisir correctement la position des quantifieurs pour avoir une transmission efficace du signal. En mathématique, le problème de la quantification optimale est de trouver la meilleure approximation de la loi d'une variable aléatoire continue par une variable aléatoire discrète ayant un nombre fixé N de quantifieurs. D'abord utilisée dans le cas à une dimension, la méthode a été développée dans le cas multidimensionnel (voir par exemple Zador (134) ou Pagès (108)) puis largement utilisée en probabilité numérique. Ainsi elle est souvent utilisée en finance dans les problèmes d'arrêt optimal, de contrôle ou de filtrage (voir par exemple les articles de Pagés (110) et (109) et de Bally *et al.* (37)). Plus récemment, de Saporta *et al.* dans (63) utilisent la quantification dans le but de développer une méthode numérique pour l'arrêt optimal des PDMP (nous utiliserons et donnerons le principe de cette méthode pour la maintenance conditionnelle via un HUMS à la section 3.5).

La quantification optimale est bien adaptée à l'approximation d'espérance conditionnelle facile à calculer lorsque les variables en jeu sont discrètes. Nous l'utiliserons pour une approximation de loi conditionnelle à la section 3.6. Le paragraphe suivant donne le principe mathématique et quelques résultats de la quantification.

Soit X un vecteur aléatoire de l'espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ sur $\mathbb{R}^d$. On suppose que X est dans $\mathbf{L}^p$ pour $p \geq 1$. Le but de la quantification est d'approcher la loi de X par une loi discrète à N modalités. Soit γ_N une N-grille de $\mathbb{R}^d$ (ie un ensemble fini de cardinal N). On définit et note $\mathrm{Proj}_{\gamma_N}(x)$ comme le point de γ_N le plus proche de x pour la norme euclidienne. L'erreur de quantification de la variable X par la grille γ_N est définie par

$$Q_N^p(\mathbb{P}_X)(\gamma_N) = \|X - \mathrm{Proj}_{\gamma_N}(X)\|_p^p,$$

où $\|.\|_p$ désigne la norme dans $\mathbf{L}^p$.

L'existence (mais pas l'unicité) d'une N-grille optimale qui minimise $Q_N^p(\mathbb{P}_X)(\cdot)$ a été démontrée sous l'hypothèse que la loi de X ne charge pas les hyperplans. Pour tout vecteur aléatoire X dans $\mathbf{L}^p$ vérifiant cette hypothèse, on note $\hat{X}^N$ la projection de X sur une N-grille optimale. $\hat{X}^N$ est alors une variable aléatoire discrète avec un nombre fini de modalités qui vérifie la propriété de stationnarité conditionnelle $\mathbb{E}[X|\hat{X}^N] = \hat{X}^N$.

On peut utiliser un algorithme du gradient stochastique pour déterminer une grille optimale (voir les détails dans (10) ou surtout dans Pagès et Printems (110)).

Des résultats asymptotiques sur l'erreur de quantification ont été obtenus par Zador (134), puis par Pierce (115) jusqu'à une généralisation donnée dans le Corollary II.6.7 de Luschgy and Graph (2000) (78) que nous rappelons dans le résultat suivant

Théorème 3.2.1 *Si $\|X\|_{p+\delta}$ est fini pour $\delta > 0$, alors il existe des nombres réels $D_1, D_2, \tilde{N}$ tel que pour tout $N \geq \tilde{N}$,on ait*

$$\|X - \hat{X}^N\|_p^p \leq \frac{1}{N^{p/d}} \Big(D_1 \|X\|_{p+\delta}^{p+\delta} + D_2 \Big). \tag{3.3}$$

Nous utilisons ce résultat pour démontrer les propriétés de vitesses de convergence énoncées dans la section 3.6.

3.3 Modèles de propagation de fissures

3.3.1 Contexte

Les travaux présentés dans cette section se font dans le cadre de l'ANR FAUTOCOES "Fault Tolerant Control for Embedded Systems" (n^o ANR-09-SEGI-004) du projet ARPEGE.

Dans un premier temps, EADS Astrium, partenaire industriel de cette ANR, nous a sollicité sur des questions de propagation de fissures dans de l'aluminium que nous traitons en collaboration avec Marie Touzet et Monique Puiggali de l'Institut de Mécanique et d'Ingénierie de Bordeaux. L'objectif est de quantifier la probabilité qu'un matériel présentant un défaut de taille donnée atteigne le régime de rupture après un cycle de vie dont les contraintes sont connues.

La nature aléatoire de la fatigue est reconnue depuis plusieurs décennies. En effet, on observe un dispersion des grandeurs mesurées, même dans des conditions expérimentales identiques. La principale cause en est l'inhomogénéité des matériaux à l'échelle microscopique. Ainsi, des modèles stochastiques ont été développés pour décrire l'évolution de la fatigue et pour calculer les paramètres de la fiabilité.

La propagation des fissures se fait en trois phases :
- une phase d'initialisation des fissures, qui se passe au niveau microscopique,
- une phase de propagation,
- une phase de rupture.

Nous avons d'abord réalisé une revue de la littérature sur les modèles stochastiques de propagation de fissures (25).

Pour un joli article sur l'initialisation des fissures dans des plastiques, on pourra consulter Mézin & Vallois (105). Mais ici nous nous intéressons essentiellement à la phase de propagation. Une loi empirique très connue en mécanique modélise l'aspect macroscopique de cette phase en proposant une équation différentielle donnant la longueur de la fissure a_t en fonction du nombre de cycles de contrainte (que nous notons t par abus de langage) appliqués à la structure : c'est la loi de Paris-Erdogan (111) donnée par l'équation suivante :

$$\frac{da_t}{dt} = C(\Delta K_t)^m,$$

où ΔK_t est l'amplitude du facteur d'intensité de la contrainte à l'instant t pour un cycle d'amplitude $\Delta\sigma$ et vaut $\Delta K_t = \Delta\sigma F(\frac{a_t}{\omega})\sqrt{\pi a_t}$. La fonction F peut-être constante égale à 1 ou suivant la longueur de l'éprouvette ω, elle est un facteur de correction. Ainsi la longueur de la fissure satisfait l'équation différentielle

$$y' = C(\Delta\sigma\sqrt{\pi})^m F(\frac{y}{\omega})^{-\frac{m}{2}} y^{\frac{m}{2}} \tag{3.4}$$

où la condition initiale est du type $y(0) = a_0$. Les seuls paramètres inconnus de cette équation différentielle sont m et C qui sont des paramètres inhérents au matériau. C'est par le biais de ces paramètres que de nombreux auteurs ont "stochastisé" la loi de Paris pour rendre compte de la dispersion des phénomènes même si ceux-ci sont réalisés et reproduits dans des conditions strictement identiques comme dans par exemple les essais de propagation sur

aluminium de Virkler (128). De nombreuses publications proposent des modèles où C et/ou m sont considérés comme des variables aléatoires ou des processus comme par exemple dans (121; 133; 73; 131; 132). Les travaux de Chiquet et Limnios ont attiré notre attention : dans (49) et (50), ils modélisent la longueur de la fissure par un PDMP a_t qui vérifie l'équation :

$$\frac{da_t}{dt} = Ca_t^m \times X_t$$

avec C constante dépendant de la structure, m constante et X_t un processus Markovien de sauts. Différentes techniques statistiques sont proposées par les auteurs pour estimer les paramètres a et m et ceux de la dynamique de X_t.

3.3.2 Modèle

Il était important pour nous de proposer des modèles aléatoires de propagation de fissures reposant sur les lois connues de mécanique et dont les caractéristiques sont interprétables par les mécaniciens. Un premier travail sur le sujet a démarré avec le stage de 1ère année de master de Romain Azaïs, sous ma direction, qui a conduit à partir de données d'essais de propagation de fissures dans de l'aluminium due à Virkler (128), à la construction d'un modèle de PDMP à un seul saut reposant sur la loi empirique de propagation de Paris. Dans ce cas, C et m sont des variables aléatoires qui changent de valeur après un temps exponentiel. La longueur de la fissure est alors un PDMP dont le flot est donné par (3.4). L'ensemble des régimes possibles $\mathcal{K}$ est l'ensemble des couples de valeurs possibles pour (C, m). Chaque fissure a alors deux régimes de propagation successifs celui qui précède et celui qui suit le saut.

Les paramètres du modèle ont été estimés à l'aide des données de Virkler. A ce sujet, il faut noter que l'estimation n'est pas directe : chacune des fissures de Virkler doit être approchée par une courbe théorique dirigée par l'équation différentielle (3.4) à cinq paramètres : les valeurs de m et C avant et après le saut et le temps de saut. Pour rechercher ces paramètres, un algorithme de recuit simulé a été utilisé. A partir du modèle obtenu, nous proposons une méthode d'actualisation inspirée de Perrin (113) qui permet d'obtenir un faisceau de prédiction pour une fissure donnée à partir de l'observation du début de sa propagation. La méthode donne de bons résultats et a été présentée dans des rencontres nationales de Maîtrise des Risque (17) ou de mathématiques appliquées (23) au cours d'une session que j'ai organisée sur les modèles stochastiques de propagation de fissures (22).

Ce travail a repris récemment en collaboration avec A. Ben Abdessalem, post-doctorant en mécanique du projet ANR. Afin d'avoir une interprétation plus mécanique du changement de régime nous proposons un modèle de PDMP reposant sur la loi de propagation de Paris pour la première phase et de la loi de propagation de Forman, plus adaptée à la fin de la phase de propagation annonçant la rupture. L'estimation de l'instant de transition entre les deux régimes est intéressante pour les mécaniciens car elle donne des renseignements sur la transition vers la rupture et est plus interprétable que les résultats de la modélisation "Paris-Paris" qui donnait une bonne adéquation entre courbes théoriques et observées mais apportait peu de sens. Les perspectives sont d'utiliser les résultats de la transition vers la rupture pour proposer des critères d'arrêt. Ce travail a été présenté dans une conférence internationale de fatigue (20) et fait l'objet d'un article en cours de rédaction dans une revue de mécanique.

3.4 Estimation de l'intensité de saut d'un PDMP

Le second volet de cette thématique s'effectue dans le cadre de la thèse de Romain Azaïs, financée par le projet ANR Fautocoes, débutée en septembre 2010 et coencadrée avec François Dufour. L'objectif principal de cette thèse est de proposer des méthodes d'estimation des caractéristiques d'un PDMP par l'observation de plusieurs réalisations i.i.d. ou l'observation d'un processus en temps long. On suppose que le processus est parfaitement observé : les instants de sauts ou transition sont connus ainsi que les flots. (Ce n'était pas le cas dans l'exemple de fissuration décrit au paragraphe précèdent où il a fallu, une fois le type de flot (Paris ou Forman) et le nombre de sauts choisis, estimer les paramètres de ces flots et l'instant de saut à partir des données). C'est à partir de ces grandeurs estimées que nous avons "calé" les modèles.

Le flot étant par essence déterministe, il n'y a pas d'inférence à faire sur son éventuelle loi de probabilité par exemple. En revanche, l'estimation des intensités de saut et de la transition est à faire. Nous avons commencé par le problème de l'intensité de saut d'un PDMP observé en temps long sous des hypothèses d'ergodicité du processus. Nos travaux sur ce sujet ont donné lieu à un article accepté pour les Annales de l'IHP (12) et un article soumis (13). Dans la suite , nous donnons d'abord le contexte (section 3.4.1) puis présentons le principe de la méthode d'abord dans un cas simple (ensemble des valeurs possibles de Z_i dénombrable et noyau Q indépendant du temps passé depuis le dernier saut) (section 3.4.2) puis nous relâchons successivement l'hypothèse dénombrable (section 3.4.3) et l'hypothèse d'indépendance du temps intersauts à la section 3.4.4). D'autre part, il est possible de gérer le mode comme une composante de la variable physique, c'est pourquoi nous supposons dans toute la suite que $\mathcal{K}$ n'a qu'un élément, aussi les caractéristiques λ, Φ, Q ne dépendent pas de k.

3.4.1 Contexte

Notre but est de faire de l'inférence sur l'intensité de saut λ d'un PDMP comme défini en (3.1) dans le cadre de l'observation d'une réalisation d'un PDMP en temps long. Si on pose $\overline{\lambda}(\xi, t) = \lambda\big(\Phi(\xi, t)\big)$, on voit que cette intensité dépend d'une variable spatiale et du temps. Le cadre est ici spécifique à cause de la présence des temps de sauts déterministes $t^\star(\xi)$ qui jouent un rôle de censure déterministe et de la structure générale de l'espace E. A notre connaissance, l'estimation du taux saut d'un PDMP n'a pas été étudiée dans un cadre nonparamétrique même si on peut l'aborder dans la cadre paramétrique à l'aide des écritures de vraisemblance disponible dans (90) pour des PDMP sans saut déterministe ou dans (91) pour des processsus ponctuels très généraux. Cependant, nous sommes proches d'un problème de survie en présence de covariables ou de statistique des processus ponctuels. La littérature sur la statistique des processus ponctuels et son utilisation dans l'analyse de survie est très abondante, et nous invitons le lecteur à consulter les livres (35; 31) pour un très riche panorama sur le sujet ainsi que le joli article (30) pour un point de vue historique sur l'utilisation des martingales en analyse de survie. Dans le contexte qui nous intéresse, l'estimateur de Nelson-Aalen qui fait l'hypothèse d'une intensité de saut qui s'écrit sous forme multiplicative et qui s'applique notamment à l'estimation de l'intensité de saut d'un processus

ponctuel marqué à espace d'état fini est particulièrement intéressant. Cet estimateur a été introduit successivement par Nelson pour un usage graphique à la fin des années 60 (106), puis par Aalen quelques années plus tard (28; 27) et a connu de multiples développements et extensions. A noter l'article récent de Comte *et al.* (52) qui propose un estimateur pour les processus ponctuels en présence de covariables sous une hypothèse de décomposition multiplicative de l'intensité. La méthode de Nelson-Aalen permet d'estimer l'intensité cumulée $\Lambda(t) = \int_0^t \lambda(s)ds$ et le lissage de cet estimateur introduit par Ramlau-Hansen dans les années 80 (117) permet d'estimer directement λ par des méthodes à noyau.

En ce qui concerne les méthodes d'estimation nonparamétrique ou semiparamétrique lorsque la marque spatiale est dans un espace continu, on peut citer (35; 30; 101; 89) et les références qui les accompagnent. En particulier, McKeague et Utikal dans (103), ont estimé le taux de saut avec une covariable à valeur dans $[0, 1]$. Pour cela, ils lissent un estimateur de type Nelson-Aalen à la fois en temps et en espace. Li et Doss dans (99), font une approximation localement linéaire de la variable spatiale et prouvent eux aussi un résultat de consistance. Utikal dans (126; 127) estime le taux saut de deux classes spécifiques de processus ponctuels sous l'hypothèse qu'un processus dérivé soit une martingale. D'autres auteurs (40; 123; 57) ont aussi proposé une estimation non-paramétrique dans ce cadre à partir d'observations i.i.d. d'un même processus.

Les hypothèses, les résultats et les démonstrations de (12; 13) sont assez techniques. Dans la suite nous nous attachons à donner le cheminement et les grandes idées qui nous ont guidés mais qui resteront parfois empiriques ici. Dans la section suivante nous montrons comment adapter l'estimateur de Nelson-Aalen à une observation en temps long d'un même processus dans un cas simple.

3.4.2 Cas fini et transition non-dépendante du temps

Dans cette section nous montrons que l'utilisation de l'estimateur de Nelson-Aalen dans le cas de l'observation en temps long d'un processus est possible dans un cas simple. Nous voulons estimer $\overline{\lambda}(\xi, t) = \lambda\big(\Phi(\xi, t)\big)$. $\overline{\lambda}(\xi, t)$ s'interprète comme le taux de saut de $S_{n=1}$ sachant $Z_n = \xi$ pour tout n.

Les hypothèses que nous faisons dans cette section et la suivante portent sur la chaîne de Markov induite. Nous supposons que la transition de Z_{n-1} à Z_n ne dépend pas du temps S_{n-1} et l'équation (3.2) devient :

$$\left\{ \begin{array}{rcl} S_n & = & \varphi(Z_{n-1}, \delta_{n-1}), \\ Z_n & = & \psi(Z_{n-1}, \varepsilon_{n-1}). \end{array} \right. \tag{3.5}$$

Mêmes si nous ne nous sommes pas penchés sur la démonstration, ces conditions sont réalisées par exemple lorsque le noyau Q ne dépend pas du temps intersauts c'est-à-dire

$$\forall \xi \in E, \quad \forall t \geq 0, \quad \forall A \in \mathcal{E}, \quad Q\big(\Phi(\xi, t), A\big) = \overline{Q}(\xi, A). \tag{3.6}$$

D'après ces équations, on voit de manière heuristique que "sachant Z_{n-1}, S_n est indépendant du reste de l'histoire du processus (X_t)". (Voir les formulations précises des indépendances conditionnelles dans la Proposition 2.1 de (12)) et pour tout $i \in \mathbb{N}$, on peut écrire simplement l'intensité du processus ponctuel simple $N^{i+1}(t) = \mathbb{1}_{\{S_{i+1} \leq t\}}$ dans sa filtration naturelle augmentée de Z_i. C'est ce que donne le résultat suivant.

Lemme 3.4.1 *Pour tout $i \in \mathbb{N}$, le processus*

$$\forall 0 \leq t < t^\star(Z_i),\ M^{i+1}(t) = N^{i+1}(t) - \int_0^t \bar{\lambda}(Z_i, u)\mathbb{1}_{\{S_{i+1} \geq u\}}\mathrm{d}u, \tag{3.7}$$

est une martingale dans la filtration $(\sigma(Z_i) \vee \mathcal{F}_s^{i+1})_{0 \leq s < t^\star(Z_i)}$ avec $(\mathcal{F}_t^{i+1})_{t \geq 0}$ la filtration naturelle de N^{i+1}.

D'autre part nous supposons aussi que le noyau Q ne charge qu'un nombre fini de points que nous notons $\{x_1, \ldots, x_M\}$.

On veut alors estimer chacune des intensités cumulées $\Lambda(x_k, t) = \int_0^t \bar{\lambda}(x_k, s)ds$ pour $1 \leq k \leq M$. Et pour cela, nous allons utiliser la collection de processus $\mathbb{1}_{\{S_{i+1} \leq t\}}\mathbb{1}_{\{Z_i = x_k\}}$. Sachant $Z_i = x_k$, ces processus forment une collection de variables aléatoires de même loi d'où l'idée de les utiliser comme des réalisations d'une même loi comme dans l'estimateur de Nelson-Aalen. Même s'ils ne sont pas indépendants (les Z_i sont liés entre eux), cela n'empêche pas de trouver une filtration dans laquelle le processus ponctuels $N_n(x_k, t) = \sum_{i=0}^{n-1} \mathbb{1}_{\{S_{i+1} \leq t\}}\mathbb{1}_{\{Z_i = x_k\}}$ a une intensité multiplicative donnée dans le théorème suivant.

Théorème 3.4.1 *Soit $n \geq 1$. Posons $Y_n(x_k, t) = \sum_{i=0}^{n-1} \mathbb{1}_{\{S_{i+1} \geq t\}}\mathbb{1}_{\{Z_i = x_k\}}$, alors le processus $M_n(x_k, \cdot)$ défini par*

$$\forall 0 \leq t < t^\star(x_k),\ M_n(x_k, t) = N_n(x_k, t) - \int_0^t \bar{\lambda}(x_k, s)Y_n(x_k, s)\mathrm{d}s \tag{3.8}$$

est une $(\widetilde{\mathcal{F}}_t^n)_{0 \leq t < t^\star(x_k)}$ martingale à temps continu sous $\mathbb{P}_{\nu_0}$, avec

$$\forall 0 \leq t < t^\star(x_k),\ \widetilde{\mathcal{F}}_t^n = \mathcal{G}_n \vee \bigvee_{i=0}^{n-1} \mathcal{F}_t^{i+1} \qquad et \qquad \mathcal{G}_n = \sigma(Z_0, \ldots, Z_{n-1}).$$

Remarquons que le temps t qui intervient dans la définition de $N_n(x_k, t)$ n'a pas de sens par rapport à l'horloge de temps du processus (X_t), les réalisations de S_i ne sont pas simultanées mais successives et dépendantes. Le cadre utilisé ici est donc différent de l'estimateur de Aalen-Johansen introduit dans (33) et aussi présenté dans (35, Section IV.4) et dont nous avons parlé au chapitre I de ce document 1.2.2.

La décomposition multiplicative de l'intensité (i.e. produit du processus prévisible $Y_n(x_k, s)$ et d'un processus déterministe $\bar{\lambda}(x_k, s)$ dont on veut estimer l'intégrale $\Lambda(x_k, t)$), donnée par le théorème 3.4.1, nous permet d'estimer $\Lambda(x_k, t)$ par l'estimateur standard de Nelson-Aalen $\widehat{\Lambda}_n(x_k, t)$ donné par

$$\widehat{\Lambda}_n(x_k, t) = \sum_{i=0}^{n-1} \frac{1}{Y_n(x_k, S_{i+1})}\mathbb{1}_{\{S_{i+1} \leq t\}}\mathbb{1}_{\{Z_i = x_k\}} = \int_0^t \frac{\mathbb{1}_{\{Y_n(x_k, s) \neq 0\}}}{Y_n(x_k, s)} dN_n(x_k, s),$$

avec la convention usuelle $0/0 = 0$. Sous des hypothèses sur $Y_n(x_k, t)$ que l'on peut trouver dans (27; 35) et qui sont impliquées par l'ergodicité de Z_n par exemple, cet estimateur a de bonnes propriétés de convergence. L'écriture de cet estimateur sous sa forme intégrale montre son lien avec la martingale $M_n(x_k, t)$, puisque sous de bonnes hypothèses, on montre par le théorème de Lenglart que le processus $\widetilde{M}_n(x_k, t) = \int_0^t \frac{\mathbb{1}_{\{Y_n(x_k, s) \neq 0\}}}{Y_n(x_k, s)} dM_n(x_k, s)$ est une martingale qui converge vers 0 avec n à t fixé. Ceci implique la consistance de notre estimateur puisque $\tilde{M}_n(x_k, t) = \widehat{\Lambda}_n(x_k, t) - \int_0^t \lambda(x_k, s)\mathbb{1}_{\{Y_n(x_k, s) \neq 0\}}ds \approx \widehat{\Lambda}_n(x_k, t) - \Lambda(x_k, t)$.

3.4.3 Cas continu et transition non-dépendante du temps

Dans cette section, nous supposons toujours la propriété (3.15) vérifiée ce qui fait que (3.5) et le Lemme 3.4.1 le sont aussi. En revanche, nous supposons que le noyau Q est diffus i.e. qu'il ne charge pas les singletons de E. La méthode employée à la section précédente est alors obsolète car $\mathbb{1}_{\{Z_i=x\}}$ est nulle presque sûrement pour tout x de E. Sous des hypothèses de régularité de la fonction $\bar{\lambda}$, il parait cependant naturel de considérer un petit voisinage A de x et d'utiliser le processus ponctuel

$$N_n(A,t) = \sum_{i=0}^{n-1} \mathbb{1}_{\{Z_i\in A\}}\mathbb{1}_{\{S_{i+1}\leq t\}},$$

pour estimer $\Lambda(x,t)$ de la même manière que l'on a utilisé $N_n(x_k,t)$ pour estimer $\Lambda(x_k,t)$ à la section précédente.

Les propriétés d'indépendance conditionnelle déduite de (3.5), nous permettent de démontrer un résultat analogue au Théorème 3.4.1 et l'on montre que si $t^\star(A) = \inf_{z\in A} t^\star(z)$,

$$\forall 0 \leq t < t^\star(A),\ M_n(A,t) = N_n(A,t) - \int_0^t \sum_{i=0}^{n-1} \mathbb{1}_{\{Z_i\in A\}}\mathbb{1}_{\{S_{i+1}\geq u\}}\lambda(Z_i,u)\mathrm{d}u$$

est une martingale dans la filtration $\widetilde{\mathcal{F}}_t^n = \mathcal{G}_n \vee \bigvee_{i=0}^{n-1}\mathcal{F}_t^{i+1}$. Mais l'intensité de $N_n(A,t)$ n'est plus multiplicative car les $\lambda(Z_i,s)$ ne peuvent se mettre en facteur d'un terme qui serait

$$Y_n(A,u) = \sum_{i=0}^{n-1} \mathbb{1}_{\{Z_i\in A\}}\mathbb{1}_{\{S_{i+1}\geq u\}},$$

car ils ne sont pas tous égaux sachant $Z_i \in A$. De surcroit, les Z_i n'ont pas tous la même loi. En revanche, une hypothèse d'ergodicité sur la chaîne (Z_n) va entrainer des propriétés asymptotiques sur la chaîne (Z_n, S_{n+1}) et va permettre de traiter l'asymptotique de ce compensateur. C'est pourquoi nous posons l'hypothèse suivante sur (Z_n) :

H-Z Il existe une mesure de probabilité ν telle que, quelle que soit la loi initiale $\nu_0 = \delta_{\{x\}}$, la loi de probabilité ν_n de Z_n vérifie

$$\lim_{n\to+\infty} \|\nu_n - \nu\|_{TV} = 0,$$

où $\|.\|_{TV}$ désigne la norme en variation totale.

Grâce à cette hypothèse, on démontre à l'aide de résultats que l'on trouve notamment dans les livres de Meyn et Tweedie ou Hernandez-Lerma et Lasserre (104; 85), des propriétés d'ergodicité des chaînes (Z_n) et (Z_n, S_{n+1}) que l'on peut consulter à la proposition 4.5 de (12).

Sous des hypothèses supplémentaires de régularité du taux de saut $\bar{\lambda}$, de la densité conditionnelle f associée et du temps d'atteinte à la frontière $t^\star$ (pour les détails voir l'hypothèse 4.6. de (12)), on obtient les convergences suivantes

$$\lim_{n\to+\infty} \frac{Y_n(A,t)}{n} = \int_A G(z,t)\nu(\mathrm{d}z) \quad \mathbb{P}_x\text{-p.s.} \tag{3.9}$$
$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=0}^{n-1} \lambda(Z_i,t)\mathbf{1}_{\{Z_i\in A\}}\mathbf{1}_{\{S_{i+1}\geq t\}} = \int_A f(z,t)\nu(\mathrm{d}z) \quad \mathbb{P}_x\text{-p.s.}$$

avec $G(\xi,t) = \exp\left(-\int_0^t \overline{\lambda}(\xi,s)\mathrm{d}s\right)$ est la fonction de survie associée à la densité conditionnelle de S_{n+1} sachant Z_n $f(\xi,t) = \overline{\lambda}(\xi,t)G(\xi,t)$.

Ces convergences entraînent un résultat asymptotique pour le compensateur de la martingale $(M_n(A,t))$:

Proposition 3.4.2 *Soit*

$$\mathcal{B}_\nu^+ = \{A \in \mathcal{B}(E) \text{ : tel que } A \text{ est relativement compact, } \nu(A) > 0 \text{ et } \overline{A} \cap \partial E = \emptyset\}.$$

Pour $A \in \mathcal{B}_\nu^+$, $0 \leq t < t^\star(A)$ et $x \in E$, quand n tend vers l'infini, on a la convergence

$$\frac{\mathbb{1}_{\{Y_n(A,t)\neq 0\}}}{Y_n(A,t)} \sum_{i=0}^{n-1} \lambda(Z_i,t)\mathbf{1}_{\{Z_i\in A\}}\mathbf{1}_{\{S_{i+1}\geq t\}} \longrightarrow l(A,t) = \frac{\int_A f(z,t)\nu(\mathrm{d}z)}{\int_A G(z,t)\nu(\mathrm{d}z)} \mathbb{P}_x\text{-}p.s. \tag{3.10}$$

La fonction $l(A,t)$ peut s'interprèter comme le taux de saut de S_{n+1} sachant $Z_n \in A$ si Z_n est sous la loi stationnaire ν. Hormis le fait qu'il faille gérer un terme "de reste" $a_n(s)$ défini par

$$a_n(s) = \int_0^s \mathbb{1}_{\{Y_n(A,u)\neq 0\}} Y_n(A,u) \sum_{i=0}^{n-1} \left[\lambda(Z_i,u) - l(A,u)\right]\mathbf{1}_{\{Z_i\in A\}}\mathbf{1}_{\{S_{i+1}\geq u\}}\mathrm{d}u, \tag{3.11}$$

on montre de la même manière que dans la section précédente que l'estimateur de type Nelson-Aalen défini pour $0 \leq t < t^\star(A)$ par

$$\widehat{L}_n(A,t) = \sum_{i=0}^{n-1} \frac{1}{Y_n(A,S_{i+1})} \mathbb{1}_{\{S_{i+1}\leq t\}}\mathbb{1}_{\{Z_i\in A\}} = \int_0^t \frac{\mathbb{1}_{\{Y_n(A,s)\neq 0\}}}{Y_n(A,s)} \mathrm{d}N_n(A,s), \tag{3.12}$$

converge vers $L(A,t) = \int_0^t l(A,s)ds$. Ce résultat est énoncé dans le théorème suivant.

Théorème 3.4.2 *Pour $A \in \mathcal{B}_\nu^+$, $0 < t < t^\star(A)$ et $x \in E$, alors,*

$$\sup_{0\leq s\leq t} \left|\widehat{L}_n(A,s) - L(A,s)\right| \xrightarrow{\mathbf{P}_x} 0 \text{ quand } n \to +\infty.$$

Nous l'avons dit, ce résultat nécessite des hypothèses de régularité sur $\bar{\lambda}$, f et $t^\star$ mais, en revanche les hypothèses habituelles faites sur $Y_n(A,t)$, comme celle de (35)[Theorem IV.1.1] par exemple, sont directement impliquées par l'ergodicité de la chaîne Z_n,. En effet, lorsque A est de mesure non nulle pour la mesure invariante ν, la chaîne Z_n visite une infinité de fois l'ensemble A et $Y_n(A,t)$ tend vers l'infini comme démontré avec l'équation (3.9).

Sous des hypothèses proches de A3 et A4 dans (99), on montre, comme les auteurs de (99), un résultat de normalité asymptotique :

Théorème 3.4.3 *Soit $A \in \mathcal{B}_\nu^+$, $0 < t < t^\star(A)$ et $x \in E$. Supposons que la vitesse de convergence en probabilité dans* (3.10) *soit de l'ordre de $n^{-1/2}$, i.e.*

$$\sqrt{n}\left(\frac{\mathbb{1}_{\{Y_n(A,s)\neq 0\}}}{Y_n(A,s)}\sum_{i=0}^{n-1}\lambda(Z_i,t)\mathbf{1}_{\{Z_i\in A\}}\mathbf{1}_{\{S_{i+1}\geq t\}} - l(A,t)\right) \xrightarrow{\mathbf{P}_x} 0, \tag{3.13}$$

quand n tend vers l'infini. Alors, on a le théorème central limite,

$$\sqrt{n}\left(\widehat{L}_n(A,t) - L(A,t)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \int_0^t \frac{l(A,s)}{\int_A G(z,s)\nu(\mathrm{d}z)}\mathrm{d}s\right).$$

Sous l'hypothèse $\bar{\lambda}$ est uniformément Lipschitz par rapport à la variable spatiale, on a aisément que $\bar{\lambda}$ (resp Λ) et l (resp L) sont proches dans le sens suivant :

$$\begin{aligned} \forall z \in A \qquad \left|\bar{\lambda}(z,s) - l(A,s)\right| &\leq [\bar{\lambda}]_{Lip}\ \mathrm{diam}\, A, \\ \left|\Lambda(z,s) - L(A,s)\right| &\leq s\,[\bar{\lambda}]_{Lip}\ \mathrm{diam}\, A. \end{aligned}$$

Grâce à ce résultat, on montre que pour tout sous-ensemble compact de E, en considérant une partition suffisamment fine, on peut estimer uniformément $\Lambda(\xi,s)$ par $\widehat{L}_n(A_j,s)$, si $\xi \in A_j$.

Théorème 3.4.4 *Soit C sous-ensemble compact de E et $\xi \in E$. Pour tout $\varepsilon, \eta > 0$, il existe un entier N et une partition finie $P = (A_k)$ de C, tel que pour tout $n \geq N$, et tout $0 < t < \min_k t^\star(A_k)$,*

$$\mathbb{P}_\xi\Big(\sup_{x\in C}\sup_{0\leq s\leq t}\left|\sum_{k=1}^{|P|}\widehat{L}_n(A_k,s)\mathbb{1}_{\{\widehat{\nu}_n(A_k)>\frac{1}{\sqrt{n}}\}}\mathbb{1}_{\{x\in A_k\}} \;-\; \Lambda(x,s)\mathbb{1}_{\{x\in C'\}}\right| > \eta\Big) < \varepsilon,$$

avec $\widehat{\nu}_n(A) = \frac{1}{n}\sum_{i=0}^{n-1}\mathbf{1}_{\{Z_i\in A\}}$, et C' l'union de A_k de mesure invariante non-nulle : $C' = \bigcup_{\nu(A_k)>0} A_k$.

Bien sûr on ne peut estimer $\Lambda(x,s)$ que si la chaîne visite infiniment souvent un petit voisinage du point x, ce qui explique l'indicatrice $\mathbb{1}_{\{x\in C'\}}$. Dans (12) on montre que $\mathbb{1}_{\{\widehat{\nu}_n(A_k)>\frac{1}{\sqrt{n}}\}}$ tend vers $\mathbb{1}_{\{\nu(A_k)>0\}}$.

Afin d'approcher le taux de saut $\bar{\lambda}$, on va lisser le processus $(\widehat{L}_n(A,t))$ à l'aide d'un noyau K à support sur $[-1,1]$. Pour cela, pour $b > 0$ et $0 < t < t^\star(A)$, on définit l'estimateur de $l(A,.)$ sur $[0,t]$ par

$$\forall 0 \leq u \leq t,\ \widehat{l}_{n,b,t}(A,u) = \frac{1}{b}\int_0^t K\left(\frac{u-s}{b}\right)\mathrm{d}\widehat{L}_n(A,s). \tag{3.14}$$

On obtient les deux résultats suivants :

Proposition 3.4.3 *Soit $A \in \mathcal{B}_\nu^+$ et $0 < r_1 < r_2 < t < t^\star(A)$.Il existe une suite $(\beta_n)_{n\geq 0}$qui converge presque sûrement vers 0 tel que*

$$\sup_{r_1\leq s\leq r_2} \left|\widehat{l}_{n,\beta_n,t}(A,s) - l(A,s)\right| \xrightarrow{\mathbf{P}_x} 0 \text{ quand } n \to +\infty,$$

pour tout $x \in E$.

Théorème 3.4.5 *Soit C sous-ensemble compact de E et $\xi \in E$. Pour tout $\varepsilon, \eta > 0$, il existe un entier N et une partition finie $P = (A_k)$ de C, tel que pour tout $n \geq N$, et tout $0 < t < \min_k t^\star(A_k)$, il existe pour tout k une suite $(\beta_n(A_k))_{n\geq 0}$ (dépendant de t) qui converge presque sûrement vers 0, tel que pour tout $n \geq N$, et tout $0 < r_1 < r_2 < t$,*

$$\mathbb{P}_\xi\Big(\sup_{x\in C}\sup_{r_1\leq s\leq r_2}\left|\sum_{k=1}^{|P|}\widehat{l}_{n,\beta_n(A_k),t}(A_k,s)\mathbb{1}_{\{\widehat{\nu}_n(A_k)>\frac{1}{\sqrt{n}}\}}\mathbb{1}_{\{x\in A_k\}} \;-\; \bar{\lambda}(x,s)\mathbb{1}_{\{x\in C'\}}\right| > \eta\Big) < \varepsilon.$$

La section 5 de (12) présente des résultats de simulation suivant un exemple inspiré de (76) où le PDMP peut modéliser le mouvement d'une bactérie dans un espace clos à deux dimensions.

Les résultats de cette section et de la précédente font l'objet de l'article (12) accepté pour publication. Première pierre d'un travail plus général, les PDMP qui y sont traités (un seul flot Φ et Q non dépendant du temps), sont en fait des processus marqués non-homogènes de renouvellement, d'où le titre de l'article.

3.4.4 Q dépendant du temps

Cette section généralise la précédente au cas où le noyau Q dépend du temps écoulé depuis le dernier saut du PDMP, c'est-à-dire

$$\forall \xi \in E, \quad \forall t \geq 0, \quad \forall A \in \mathcal{E}, \quad Q\big(\Phi(\xi,t),A\big) = \overline{Q}(\xi,t,A). \tag{3.15}$$

Dans ce cas le système dynamique qui régit la chaîne (Z_n, S_{n+1}) est donné par (3.2) et et la propriété d'indépendance conditionnelle que l'on peut résumer de manière heuristique "Sachant Z_n, S_{n+1} est indépendant de l'histoire du PDMP" utilisée à la section précédente et qui permet d'étudier directement le processus d'intérêt à savoir le taux de saut $\lambda(z,t)$, ne tient plus. Désormais, c'est conditionnellement à (Z_n, Z_{n+1}) que S_{n+1} est indépendant de l'histoire du PDMP (on peut lire la formulation précise des indépendances conditionnelles dans la proposition 4.7. de (13)). C'est donc à travers l'étude de la loi de S_{n+1} sachant (Z_n, Z_{n+1}) que nous allons essayer d'estimer la loi de S_{n+1} sachant Z_n.

Introduisons d'abord quelques hypothèses et notations sur le noyau Q.

H-Q Le noyau de transition $\overline{Q}$ est diffus et il existe une mesure μ sur $(E,\mathcal{E})$ telle que pour tout ensemble mesurable B d'intérieur non-vide $\mu(B) > 0$ et telle que pour tout B, on ait

$$\forall \xi \in E,\ \forall s \geq 0,\ \overline{Q}(\xi,s,B) = \int_B \widetilde{Q}(\xi,s,y)\mu(\mathrm{d}y),$$

de plus $\widetilde{Q}(x,\cdot,y)$ une fonction continue en x, y vérifiant

$$\exists m > 0 \text{ tel que } \forall x \in E,\ \forall s \geq 0,\ \forall y \in E,\ \widetilde{Q}(x,s,y) \geq m.$$

Un premier résultat à établir est l'existence de l'intensité de saut de S_{n+1} sachant (Z_n, Z_{n+1}) qui s'exprime en fonction de f et G respectivement densité et fonction de survie associé à $\widetilde{\lambda}$

Proposition 3.4.4 *Pour tout entier n la loi conditionnelle de S_{n+1} sachant Z_n, Z_{n+1} s'exprime pour $t \geq 0$, par*

$$\mathbb{P}_{\nu_0}(S_{n+1} > t | Z_n, Z_{n+1}) = \exp\left(- \int_0^{t \wedge t^\star(Z_n)} \widetilde{\lambda}(Z_n, Z_{n+1}, s) \mathrm{d}s \right) \mathbb{1}_{\{0 \leq t < t^\star(Z_n)\}},$$

où le taux de saut $\widetilde{\lambda}$ est défini pour $x, y \in E$, $0 \leq t \leq t^\star(x)$, par

$$\widetilde{\lambda}(x, y, t) = \frac{f\widetilde{Q}(x, t, y)}{\int_t^{t^\star(x)} f\widetilde{Q}(x, s, y)\mathrm{d}s + G\widetilde{Q}(x, t^\star(x), y)}. \tag{3.16}$$

On peut en déduire l'intensité du processus de comptage

$$N_n(A, B_k, t) = \sum_{i=0}^{n-1} \mathbb{1}_{\{S_{i+1} \leq t\}} \mathbb{1}_{\{Z_i \in A\}} \mathbb{1}_{\{Z_{i+1} \in B_k\}}$$

en démontrant que

$$M_n(A, B_k, t) = N_n(A, B_k, t) - \sum_{i=0}^{n-1} \int_0^t \mathbb{1}_{\{Z_i \in A\}} \mathbb{1}_{\{Z_{i+1} \in B_k\}} \mathbb{1}_{\{S_{i+1} \geq u\}} \widetilde{\lambda}(Z_i, Z_{i+1}, u) \mathrm{d}u,$$

est une $(\mathcal{G}_n \vee \bigvee_{i=0}^{n-1} \mathcal{F}_t^{i+1})_{0 \leq t < t^\star(A)}$-martingale.

Une fois ce résultat établi, le tapis se déroule comme dans la section précédente. Grâce à l'hypothèse **(H-Q)**, on montre que la chaîne de Markov (Z_n, Z_{n+1}) (resp. (Z_n, Z_{n+1}, S_{n+1})) a des propriétés d'ergodicité et on note $\tilde{\nu}$ (resp. η) sa loi invariante. Les équations (15), (18), (22), (28), (29) et la remarque 3.6. de (13) donne l'expression de ces mesures en fonction de la loi invariante ν de la chaîne (Z_n). Et on démontre avec des techniques identiques à la section précédente que le processus (encore un estimateur de type Nelson-Aalen !)

$$\forall 0 \leq t < t^\star(A),\ \widehat{\widetilde{L}}_n(A, B_k, t) = \int_0^t Y_n(A, B_k, s)^+ \mathrm{d}N_n(A, B_k, s)$$

converge vers $\widetilde{L}(A, B_k, t) = \int_0^t \widetilde{l}(A, B, s)\mathrm{d}s$ où $\widetilde{l}(A, B, s)$ est une approximation du taux de saut $\tilde{\lambda}(x, y, t)$. $\widetilde{l}(A, B, s)$ s'interprète comme le taux de saut sous le régime stationnaire sachant x dans A et y dans B. De plus, en lissant $\widehat{\widetilde{L}}_n(A, B_k, t)$, on obtient $\widehat{\widetilde{l}}_{n,b,t}(A, B_k, t)$ un estimateur de $\widetilde{l}(A, B, .)$ (Proposition 2.5. de (13)).

Le taux de saut conditionnel à (Z_n, Z_{n+1}) étant estimé, il faut revenir à loi conditionnelle de S_{n+1} sachant Z_n, on utilise la relation 3.16 qui lie $\tilde{\lambda}$ que l'on sait estimer

aux caractéristiques f et G de cette loi conditionnelle via $\tilde{Q}$. Si on pose $\tilde{H}(x,y,t) = \int_t^{t^\star(x)} f\widetilde{Q}(x,s,y)\mathrm{d}s + G\widetilde{Q}(x,t^\star(x),y)$, on obtient la relation

$$H(x,y,t)\widetilde{\lambda}(x,y,t) = f(x,t)\widetilde{Q}(x,t,y),$$

et si on se souvient des propriétés de noyau **(H-Q)**, l'intégrale de $\widetilde{Q}$ contre la mesure μ vaut 1 et la relation devient $f(x,t) = \int_E H(x,y,t)\widetilde{\lambda}(x,y,t)\mu(dy)$. On a déjà un estimateur de $\widetilde{\lambda}$ et s'il existe une famille d'ensembles de diamètre fini $\{B_1,\dots,B_p\}$ jouant en un certain sens le rôle de partition de E (voir la formulation exacte dans Assumptions 2.1. de (13)), alors on peut montrer le résultat

Proposition 3.4.5 *Pour $\xi \in A$. Pour tout $0 \le t < t^\star(A)$,*

$$\left| f(\xi,t) - \sum_{k=1}^{p} \widetilde{l}(A,B_k,t)\widetilde{H}(A,B_k,t) \right| \le Cst_1 \operatorname{diam} A + Cst_2 \max_{1\le k\le p} \operatorname{diam}(A\times B_k),$$

avec $\widetilde{H}(A,B_k,t) = \frac{1}{\nu(A)}\int_{A\times B_k} H(x,y,t)\mu(\mathrm{d}y)\nu(\mathrm{d}x)$.

En fait H a une interprétation sous le régime stationnaire ν, en effet, pour $0 \le t < t^\star(A)$, on a

$$\int_{A\times B_k} H(x,y,t)\nu(\mathrm{d}x)\mu(\mathrm{d}y) = \mathbb{P}_\nu(S_1 > t, Z_1 \in B_k, Z_0 \in A),$$

ce qui entraîne immédiatement que $\widetilde{H}(A,B_k,t) = \mathbb{P}_\nu(S_1 > t, Z_1 \in B_k | Z_0 \in A)$. On peut alors l'estimer par sa version empirique

$$\widehat{p}_n(A,B_k,t) = \frac{\sum_{i=0}^{n-1} \mathbf{1}_{\{S_{i+1}>t\}}\mathbf{1}_{\{Z_{i+1}\in B_k\}}\mathbf{1}_{\{Z_i\in A\}}}{\sum_{i=0}^{n-1}\mathbf{1}_{\{Z_i\in A\}}}.$$

Nous avons maintenant tous les "ingrédients" pour fabriquer un estimateur de $f(\xi,t)$ et donner la convergence associée, ce qui est fait dans le théorème qui suit.

Théorème 3.4.6 *Soit C sous-ensemble compact de E et $\xi \in E$. Pour tout $\varepsilon,\eta > 0$, il existe un entier N et une partition finie $P = (A_l)$ de C, tel que pour tout $n \ge N$, et tout $0 < t < \min_k t^\star(A_l)$, il existe un couple (l,k) une suite $(\beta_n(A_l,B_k))_{n\ge 0}$ (dépendant de t) qui converge presque sûrement vers 0, alors l'estimateur de f défini par*

$$\widehat{f}_n(A,s) = \sum_{k=1}^{p} \widehat{\widetilde{l}}_{n,\beta_n(A,B_k),t}(A,B_k,s)\,\widehat{p}_n(A,B_k,s),$$

vérifie pour tout $n \ge N$, et tout $0 < r_1 < r_2 < t$,

$$\mathbb{P}_\xi\left(\sup_{x\in\mathcal{K}}\sup_{r_1\le s\le r_2}\left|\sum_l \widehat{f}_n(A_l,t)\mathbf{1}_{\{x\in A_l\}} - f(x,s)\right| > \eta\right) < \epsilon.$$

Enfin, je voudrais noter que Romain Azaïs a développé la mise à disposition de cette méthode par un package R nommé "EstSimPDMP" disponible sur le site du CRAN. D'autre part, ce problème d'estimation a été utilisé dans une application inattendue par Romain Azaïs et des collègues de l'UMR CNRS 5805 EPOC "Environnements et Paléoenvironnements Océaniques et Continentaux" pour déterminer des profils de fréquences d'ouverture et fermeture d'huitres suivant le régime de marée. Ces différents profils sont liés à la qualité de l'eau dans laquelle vivent les huitres!!

3.5 Modélisation d'un HUMS

Cette étude a débuté par le stage de master 2 à Thales Optronique de Camille Baysse sous ma direction. Il se prolonge dans le cadre d'une thèse CIFRE dans la même entreprise, débutée en novembre 2011 et co-encadrée avec Jérôme Saracco. Ce travail ne constitue pas une recherche théorique à proprement dit, mais une construction méthodologique pour répondre de manière réaliste aux besoins et aux questions de Thales.

Dans le cadre de l'optimisation de la fiabilité, Thales intégre des HUMS (pour Health Unit Monitoring System) dans ses différents équipements. Les HUMS enregistrent la mesure de différentes variables physiques caractérisant le matériel ainsi que les conditions environnementales et d'utilisation. L'objectif est d'utiliser toutes ces données pour fabriquer des indicateurs qui donnent une mesure de « l'état de santé » de l'équipement et alertent sur la nécessité d'une maintenance.

L'ensemble du travail a porté sur l'analyse du fonctionnement d'un équipement optronique. Cet appareil est équipé d'une "machine à froid" qui abaisse sa température interne jusqu'à une température très basse indispensable au fonctionnement. Le temps de mise en froid de l'appareil est évidemment un indicateur de l'état de la machine à froid. Un temps de mise en froid plus long pouvant indiquer une perte de puissance ou une éventuelle fuite. Nous avons donc utilisé ce temps comme indicateur de l'état de fonctionnement de la machine. Un premier objectif est, à partir des observations du temps de mise à froid, de déterminer le plus tôt possible un passage vers l'état dégradé afin de faire revenir l'équipement pour maintenance avant la panne. Un second, lié au premier, est de proposer un critère d'arrêt optimal pour proposer une date de maintenance dans un modèle dynamique, tenant compte non seulement de la panne de la machine à froid mais aussi des autres causes de panne (roulement à billes, électronique). Les deux paragraphes suivants exposent ces deux problèmes.

3.5.1 Modèle de Markov caché pour la détection d'un état dégradé

Pour ce faire, nous avons utilisé un modèle de chaînes de Markov cachées dans un brownien. Même si le cas général est exposé dans (26) et (18), nous ne présentons dans la suite que le modèle à deux états que nous utilisons pour modéliser le problème concret qui nous occupe. On note m_t la chaîne de Markov à temps continu modélisant l'état de la machine à froid : état de bon fonctionnement ou état dégradé. m_t n'est pas observée directement mais à travers les temps Y_t de mise à froid lors de la mise en marche de l'appareil. Lorsque l'appareil est en bon état de fonctionnement, ces temps, pour une même caméra, sont sensiblement égaux et les fluctuations peuvent être assimilées à un bruit centré. En revanche, si l'appareil est dans un état dégradé, ce temps augmente au fil des mises en marche jusqu'à la panne de l'appareil. Nous suivons le modèle proposé par Elliott dans (72), qui, pour de judicieuses raisons de calcul, assimile les différentes valeurs possibles de la chaîne $\boldsymbol{X}_t$ à chacun des vecteurs d'une base orthonormale $(e_1, \ldots, e_d)$ de $\mathbb{R}^d$ ($d = 2$ ici). On a alors un processus $\boldsymbol{X}_t$ à valeurs dans (e_1, e_2) et $(m_t = 1) \Leftrightarrow (\boldsymbol{X}_t = e_1)$. $\boldsymbol{X}_t$ est associé à une Q-matrice $\boldsymbol{A}$ que nous supposons constante au cours du temps. Nous posons

$$Y_t = Y_0 + \int_0^t < c, \boldsymbol{X}_s > ds + W_t, \tag{3.17}$$

où W_t est un processus brownien standard, $\boldsymbol{c}$ un vecteur de $\mathbb{R}^2$ et $< , >$ désigne le produit scalaire. En choisissant les composantes de $\boldsymbol{c}$: $c_1 = 0$ et $c_2 > 0$, le temps de mise en froid Y_t répond alors aux critères énoncés plus haut : il oscille autour d'une valeur constante lorsque $\boldsymbol{X}_t = e_1$ correspond à l'état de bon fonctionnement de m_t et il est égal à un processus croissant bruité lorsque $\boldsymbol{X}_t = e_2$. Les équations du filtrage établies par Elliott (72) permettent de donner la valeur de la probabilité conditionnelle $P(\boldsymbol{X}_t = e_1|\mathcal{Y}_t)$ égale tout simplement à l'espérance conditionnelle de la première composante de $\boldsymbol{X}_t$ sachant les observations de Y_s jusqu'à l'instant t.

Des études de simulation ont d'abord été menées pour tester la méthode et aussi la sensibilité du modèle à des erreurs de valeurs des coefficients de la matrice $\boldsymbol{A}$ et du vecteur $\boldsymbol{c}$. Le modèle a ensuite été "câlé" sur les données réelles constituées de 28 histoires de caméra dont 5 étaient tombées en panne. Nous supposons les 28 histoires indépendantes avec des paramètres $\boldsymbol{A}$ et $\boldsymbol{c}$ communs mais des valeurs de Y_0 pouvant varier suivant les appareils. Compte-tenu du faible nombre de données, nous avons augmenté la valeur du taux de passage vers l'état dégradé obtenu par une estimation standard de survie avec censure. Et nous avons arbitrairement donné une valeur très faible mais strictement positive au taux de transition de l'état dégradé vers l'état "sain" afin de donner de la souplesse au modèle.

La confrontation aux données réelles a été plus difficile que les simulations, les signaux des temps de mise en froid étant très fortement bruités. Pour pallier cette difficulté, nous avons d'abord régressé les temps de mise en froid sur les températures initiales de l'appareil : une température plus élevée au départ justifiant un temps de mise en froid plus long et ramené l'ensemble des temps de mise à froid à une même température initiale. Malgré cela, on observe encore un signal très bruité comme dans la figure 3.1 qui donne le signal après régression d'une des 28 caméras. A noter que les pics vers le bas sont sûrement dûs à une mise en marche juste après un arrêt, la température étant alors restée basse. Nous avons alors lissé le signal par une moyenne mobile sur 20 démarrages. La figure 3.2 donne le lissage correspondant au signal de la figure 3.1. C'est ce signal lissé que nous utilisons comme Y_t dans le modèle de Markov caché (3.17) pour déterminer l'état du système. L'inspection de l'histoire globale du signal de la figure 3.2 semble montrer une augmentation de Y_t entre les démarrages 100 et 160 suivie d'une baisse et d'un palier avant une nouvelle augmentation jusqu'à la panne à partir de 260 (lissage sur 20). Bien sûr, nous avons reçu le signal initial dans sa globalité mais l'objectif du HUMS est bien de donner, au fur et à mesure des démarrages, un indicateur sur l'état de santé de l'appareil. Le filtre proposé par Elliot répond à cette question puisqu'il est calculé de manière récursive et donne à chaque instant, la probabilité d'être dans l'état dégradé.

Nous donnons dans le figure 3.3 la valeur de cette probabilité correspondant au signal de la Figure 3.2. Il reste à déterminer un critère d'arrêt qui consiste à donner un seuil et le nombre de fois où la probabilité dépasse ce seuil pour déterminer l'arrêt. Ce critère doit à la fois minimiser les faux positifs et les faux négatifs. Avec les critères choisis, 3 des 5 caméras sont détectées avant la panne et aucune des autres n'est arrêtée. A noter que les deux pannes non détectées n'ont pas été précédée d'une période de croissance du temps de mise en froid.

Cette partie fait l'objet d'une communication à un congrès industriel international (18) et est en cours de rédaction pour publication.

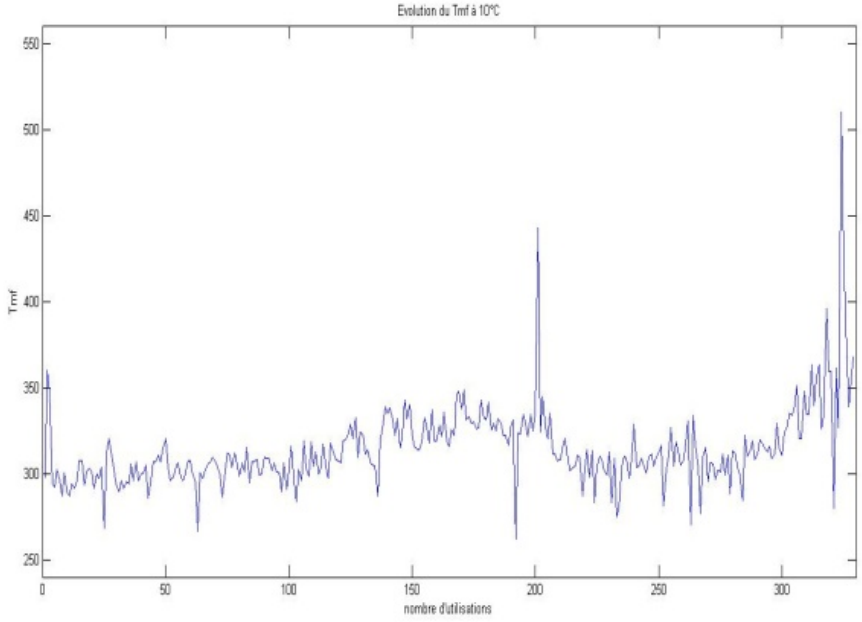


FIGURE 3.1 – Signal d'une caméra après regression à 10^o.

3.5.2 Arrêt optimal pour une maintenance conditionnelle

Ce travail, en cours, prolonge le précèdent dans le cadre de l'optimisation de la maintenance avec la prise en compte des autres causes de panne possible de l'appareil. Il s'agit de proposer à l'utilisateur une politique de maintenance optimisant un critère tenant compte du temps d'utilisation de l'équipement mais pénalisé par la survenue d'une panne. Cette partie, à laquelle collabore B. de Saporta, utilise les compétences de l'équipe INRIA CQFD en arrêt optimal avec les techniques numériques associées notamment la quantification optimale de variables aléatoires (63). Pour ce faire, nous utilisons un modèle simple de Processus Markoviens Déterministes par Morceaux (PDMP) auquel nous appliquons un méthode d'arrêt optimal. L'application se fait toujours sur l'équipement comportant la machine à froid du paragraphe précèdent

Cette fois-ci, nous supposons que la transition entre les états est observée. Le système a trois états possibles : soit il est en état de bon fonctionnement (état 1), soit sa machine à froid est dégradée (état 2), soit il est en panne. La transition de l'état 1 à l'état de panne 3 peut-être due à deux causes possibles que nous appellerons **RB** et **E**. Une seule cause possible pour la transition de l'état 1 à 2 que nous supposons observée. En plus des causes **RB** et **E**, la panne de la machine à froid vient augmenter le taux de transition de l'état 2 vers l'état 3. Les taux de transition sont donnés par avis d'expert. Etant donné que certains d'entre eux sont des transitions de Weibull, pour avoir un processus markovien, nous considérons le processus hybride $X_t = (m_t, t)$ où m_t est le processus discret égal à l'état du système et t le temps écoulé. Ce processus est clairement un PDMP (il ne le serait pas sans la composante t). Ce PDMP a trois régimes $\mathcal{K} = \{1, 2, 3\}$ dont le troisième est dégénéré. Nous voulons arrêter le

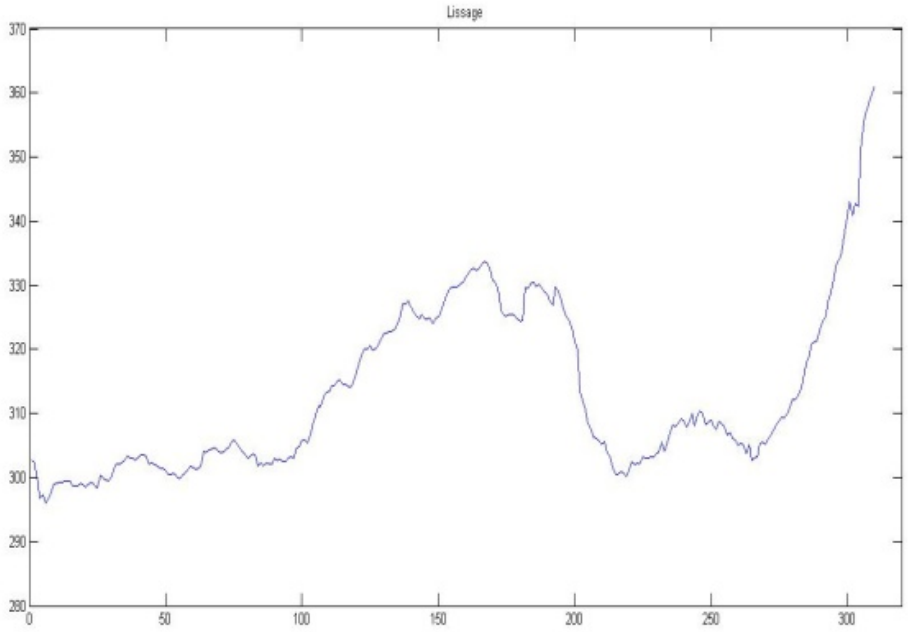


FIGURE 3.2 – Signal de la figure 3.1 après lissage .

système pour maintenance en un temps d'arrêt τ, adapté à la filtration du PDMP qui optimise une fonction de performance du type $\mathbb{E}[g(X_\tau)]$ en optimisant une fonction de performance. On est donc ramené à un problème d'arrêt optimal d'un PDMP. De nombreux auteurs ont travaillé sur l'arrêt optimal des PDMP dont (53; 54; 59; 79), mais nous allons utiliser les résultats de de Saporta *et al.* dans (63) qui donnent une méthode numérique d'arrêt optimal d'un PDMP basé sur la quantification de la chaîne de Markov sous-jacente. Dans notre cas cette chaîne est tout simplement la chaîne (X_{T_n}, T_n) pour $0 \leq n \leq 2$ car compte-tenu des transitions possibles et du fait que la panne est un état absorbant, notre processus a au plus deux transitions. L'idée de cette approximation est de remplacer l'opérateur continu de la programmation dynamique par un opérateur qui est d'une part discrétisé en temps sur une grille déterministe et d'autre part pour lequel la chaîne (X_{T_n}, T_n) est remplacé par sa version quantifiée. Des études poussées de simulation sont en cours, avec différentes fonctions de performance et une éventuelle différentiation des états de panne suivant la cause. Ce travail en cours fait l'objet d'une communication acceptée dans un congrès industriel national (19) .

3.6 Quantification optimale et méthode SIR

Issu d'une idée proposée par François Dufour, ce travail a fait l'objet du stage de master 2e année de Romain Azaïs sous ma direction et en collaboration avec Jérôme Saracco et qui est publié dans (10). Il s'agit d'associer la quantification optimale à la méthode de régression non paramétrique SIR (pour Sliced Inverse Regression). En régression, l'objectif principal est

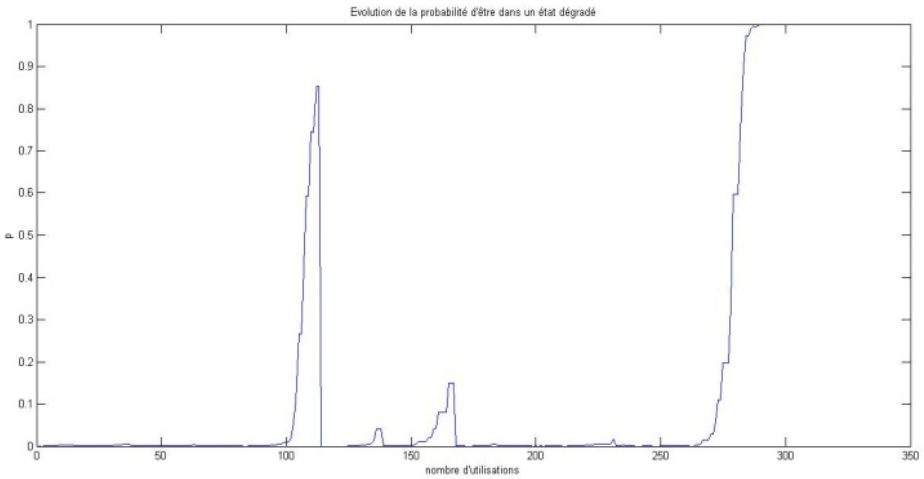


FIGURE 3.3 – Probabilité d'être dans l'état dégradé correspondant au signal lissé de la figure 3.2.

d'approcher de façon parcimonieuse la loi conditionnelle d'une variable d'intérêt Y sachant une variable explicative $\boldsymbol{X}$ de dimension d. Dans beaucoup de problèmes la dimension de la variable explicative $\boldsymbol{X}$ est grande et l'analyse statistique n'est pas facile. Un moyen usuel de pallier cette difficulté est de réduire la dimension de la variable explicative via une projection linéaire sans pour autant imposer de forme complètement paramétrique au modèle. Dans le modèle que nous considérons dans cette partie, $\boldsymbol{X}$ intervient seulement via sa projection $\boldsymbol{\beta'X}$ de la manière suivante :

$$Y = f(\boldsymbol{\beta'X}, \epsilon) \tag{3.18}$$

Ici, la fonction f et le paramètre $\boldsymbol{\beta}$ de dimension d, sont inconnus, la variable aléatoire ϵ est un bruit indépendant de $\boldsymbol{X}$. Une autre manière de définir le modèle (3.18) est de dire que Y est indépendant de $\boldsymbol{X}$ sachant $\boldsymbol{\beta'X}$. La variable $\boldsymbol{\beta'X}$ est alors une statistique exhaustive pour le problème (3.18). Dans ce cas, seul le sous-espace engendré par β est identifiable, cet espace est appelé espace EDR pour Effective Dimension Reduction.

La méthode SIR introduite par Duan and Li en 1991 (68) est une des méthodes pour identifier (ou plutôt estimer) cet espace. Le principe est de partager l'espace en plusieurs parties (tranches !). La matrice de covariance de $\boldsymbol{X}$ sachant la tranche de Y, notée $\hat{\Gamma}$ est alors importante puisque c'est par son intermédiaire que l'on peut estimer β. En effet sous des hypothèses simples sur la distribution de X et le modèle (3.18) que l'on peut consulter en $(\mathcal{A}_5)$ et $(\mathcal{A}_6)$ données ci-dessous, le vecteur propre principal de $(\mathbf{V}\mathrm{ar}(\boldsymbol{X}))^{-1}\hat{\boldsymbol{\Gamma}}$ engendre l'espace EDR. Ces deux matrices sont alors estimée par leur version empirique à partir des données.

Remarquons tout d'abord que le "tranchage" de la variable Y est une quantification (non-optimale) de cette variable. D'autre part, le problème de régression est un problème de loi conditionnelle de Y sachant $\boldsymbol{X}$ ou plutôt $\boldsymbol{\beta'X}$ ici. Nous avons vu à la section 3.2 que la quantification optimale était souvent utilisée pour approcher des espérance conditionnelles,

elle est donc particulièrement adaptée au problème qui nous intéresse ici.

3.6.1 Estimation de l'EDR

En ce qui concerne l'estimation de l'espace EDR, l'idée est de remplacer $\boldsymbol{X}$ par sa version quantifiée $\hat{X}_N$ dans la définition de la matrice $\boldsymbol{\Gamma}$ pour obtenir $\boldsymbol{\Gamma_N}$. Nous avons montré que sous des hypothèses standard sur la loi de $\boldsymbol{X}$ utilisées en SIR ($\mathcal{A}_{5,6}$), et les hypothèses standard sur $\boldsymbol{X}$ pour assurer sa quantification optimale ($\mathcal{A}_{7,8}$)le vecteur propre principal de la matrice $(\mathbf{V}\mathrm{ar}(\boldsymbol{X}))^{-1}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{N}}$ se rapproche de la direction du vecteur $\boldsymbol{\beta}$ quand N tend vers l'infini. Plus précisément, en gardant les numérotations des hypothèses de (10), on a :

Théorème 3.6.1 *Soit* $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{N}} = \mathbf{V}\mathrm{ar}(\mathbb{E}[\hat{\boldsymbol{X}}^{\boldsymbol{N}}|\hat{Y}])$, *la matrice de covariance de* $\mathbb{E}[\hat{\boldsymbol{X}}^{\boldsymbol{N}}|\hat{Y}]$ *où* $\hat{Y} := Proj_\gamma(Y)$ *est la projection de* Y *sur une grille non nécessairement optimale* γ *de* $\mathbb{R}$. *Posons*

$(\mathcal{A}_5)$ $\exists \hat{y} \in \gamma,\ \mathbb{E}[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\beta}|\hat{Y}=\hat{y}] \neq 0,$
$(\mathcal{A}_6)$ $\boldsymbol{X}$ *a une loi elliptique symétrique,*
$(\mathcal{A}_7)$ $\exists\ p \geq 1$ *s.t.* $\boldsymbol{X} \in \mathbf{L}^p \cap \mathbf{L}^q$ *with* $\frac{1}{p}+\frac{1}{q}=1,$
$(\mathcal{A}_8)$ *La loi* $\boldsymbol{X}$ *ne charge pas les hyperplans,*

Sous $(\mathcal{A}_{5\to 8})$, *quelle que soit la suite* $(\tilde{\boldsymbol{\beta}}_{\boldsymbol{N}})$ *tel que* $\tilde{\boldsymbol{\beta}}_{\boldsymbol{N}}$ *est un vecteur prinicipal de* $(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{N}})$, *nous avons*

$$\cos^2(\tilde{\boldsymbol{\beta}}_{\boldsymbol{N}}, \boldsymbol{\beta}) \to 1 \quad as\ N \to \infty.$$

avec $\cos^2(\tilde{\boldsymbol{\beta}}_{\boldsymbol{N}}, \boldsymbol{\beta}) = \dfrac{(\tilde{\boldsymbol{\beta}}'_{\boldsymbol{N}}\boldsymbol{\beta})^2}{(\tilde{\boldsymbol{\beta}}'_{\boldsymbol{N}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{N}}) \times (\boldsymbol{\beta}'\boldsymbol{\beta})}.$

En utilisant le résultat du théorème 3.2.1 et sous l'hypothèse supplémentaire que $\boldsymbol{X}$ est dans $\mathbf{L}^{p+\delta}$ ($\delta > 0$), on a un résultat un peu plus précis ; dans ce cas on peut montrer qu'il existe une suite $\boldsymbol{\beta_N}$ de vecteurs principaux de $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{N}}$ tel que la norme euclidienne de $(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{N}} - \boldsymbol{\beta})$ est un $\mathcal{O}\left(\frac{1}{N^{1/d}}\right)$.

3.6.2 Loi conditionnelle de Y sachant $\boldsymbol{X}$

La meilleure prévision de la variable d'intérêt Y connaissant la covariable $\boldsymbol{X}$ est donnée par la loi conditionnelle de Y sachant $\boldsymbol{X}$. Dans cette section, nous allons donner une approximation de cette loi conditionnelle. En préalable, remarquons que sans hypothèse sur $\boldsymbol{\beta}$ et/ou f , le paramètre $\boldsymbol{\beta}$ du modèle (3.18) n'est pas identifiable. En revanche, il le devient si on impose un signe à la première composante non nulle de $\boldsymbol{\beta}$ et une norme à $\boldsymbol{\beta}$ (notons que grâce à $(\mathcal{A}_5)$, β est différent du vecteur nul). Nous faisons désormais cette hypothèse sur $\boldsymbol{\beta}$ et les vecteurs de la suite $\boldsymbol{(\beta_N)}$ définie à la fin du paragraphe précèdent, vérifient la même hypothèse et sont alors aisément déductibles de n'importe quelle suite $(\tilde{\boldsymbol{\beta}}_{\boldsymbol{N}})$ définie par le théorème 3.6.1.

Pour approcher la loi conditionnelle de Y sachant $\boldsymbol{X}$, nous procédons en deux étapes. Tout d'abord, on estime $\boldsymbol{\beta}$ à l'aide d'une suite de vecteurs $(\boldsymbol{\beta_N})$. Dans un deuxième temps, on quantifie Y et $\boldsymbol{\beta}'_{\boldsymbol{N}}\boldsymbol{X}$ pour obtenir $\hat{Y}^m$ et $\widehat{\boldsymbol{\beta}'_{\boldsymbol{N}}\boldsymbol{X}}^m$ et l'on définit $\hat{\boldsymbol{P}}$ la matrice de transition de $\widehat{\boldsymbol{\beta}'_{\boldsymbol{N}}\boldsymbol{X}}^m$ vers $\hat{Y}^m$ par

$$\forall \hat{u} \in \gamma_m,\ \forall \hat{y} \in \delta_m,\ \hat{\boldsymbol{P}}(\hat{u}, \hat{y}) = \mathbb{P}(\hat{Y}^N = \hat{y} | \widehat{\boldsymbol{\beta}_N' \boldsymbol{X}}^m = \hat{u}),$$

où γ_m et δ_m sont les m-grilles optimales de quantification de $\boldsymbol{\beta}_N' \boldsymbol{X}$ and Y.

$\hat{P}$ dépend des deux paramètres N et m : N est le nombre lié à la quantification de X pour la détermination de β_N et le nombre m de la quantification de $\boldsymbol{\beta}_N' \boldsymbol{X}$ et Y.

La variable discrète $\hat{Y}^c$ telle que $(\widehat{\boldsymbol{\beta}_N' \boldsymbol{X}}^m, \hat{Y}^c)$ est une chaîne de Markov arrêtée de matrice de transition $\hat{P}$ est un prédicteur de Y sachant $\boldsymbol{\beta}_N' \boldsymbol{X}$. Le théorème suivant donne une vitesse de convergence de la loi de $\hat{Y}^c$ vers celle de Y sachant $\boldsymbol{X}$.

Théorème 3.6.2 *Sous $(\mathcal{A}_{5,6})$, si f est Lipschitz et si X et Y vérifient $(\mathcal{A}_{7,8})$ et sont dans $\mathbf{L}^{p+\delta}$, pour toute fonction Lipschitzienne ϕ, il existe A_1, A_2, A_3 dans $\mathbb{R}^3$, une suite (g_N) convergeant vers une limite strictement positive, et deux entiers $\overline{m}$ et $\overline{N}$ tel que quels que soient $m \geq \overline{m}$ et $N \geq \overline{N}$, on ait*

$$\left\| \mathbb{E}[\phi(Y) | \boldsymbol{\beta}' \boldsymbol{X}] - \mathbb{E}\left[\phi(\hat{Y}^m) | \widehat{\boldsymbol{\beta}_N' \boldsymbol{X}}^m \right] \right\|_1 \leq \frac{A_1}{N^{1/d}} + \frac{A_2}{m} g_N + \frac{A_3}{m}.$$

La preuve de ces convergences et de leur vitesse est donnée dans (10). On y présente aussi une étude de simulation. Cette étude montre que pour de grands échantillons (n= 1000) la performance de l'estimation de β est comparable avec celle de la méthode SIR ; elle est même meilleure lorsque la dépendance entre les deux variables est presque symétrique (cas $Y = (\boldsymbol{\beta}' \boldsymbol{X})^2 \exp\left(\frac{\boldsymbol{\beta}' \boldsymbol{X}}{\theta}\right) + \epsilon$, pour θ grand par exemple). D'autre part, l'aspect prévision avec ce type de méthode est original dans les problèmes de régression.

3.7 Conclusion et perspectives

Certaines des perspectives des thématiques présentées dans ce chapitre sont celles du projet ANR Fautocoes, qui prévoit la modélisation de la propagation d'une fissure soumise à un chargement aléatoire correspondant à un "cycle de vie type" de la structure étudiée. Ce modèle à construire avec les ingénieurs d'EADS sera utilisé par nos collègues de l'ANR Benoîte de Saporta et François Dufour, spécialistes des calculs d'espérance de temps de sortie de PDMP pour calculer des probabilités de rupture. Ce projet prévoit aussi l'estimation du noyau de transition Q du PDMP, question sur laquelle s'est déjà penchée notre étudiant Romain Azaïs (presque en autonomie !). Nous l'avons dit plus haut, l'estimation du taux de saut d'un PDMP dans un espace continu pose la question du choix de la partition (voir Théorème 3.4.6 par exemple) ou du paramètre de lissage : des critères de choix de modèles dans ce cadre non i.i.d. sont peut-être à développer. La thèse de Camille Baysse qui a démarré sur la modélisation d'un HUMS dans un cadre simple de machine à froid et qui sera rapidement mis en pratique par Thales pour la détection du mode dégradé, pose aussi des questions méthodologiques plus difficiles. D'une part, si on envisage la maintenance conditionnelle, le vrai problème à considérer n'est pas de l'arrêt optimal comme nous l'avons traité ici mais un problème de contrôle impulsionnel permettant de considérer l'équipement après réparation et d'optimiser une fonction performance sur un plus long terme (à noter qu'une étude théorique

sur le contrôle impulsionnel est l'objet d'un sujet de post-doc du projet ANR). D'autre part, la thèse prévoit la modélisation de HUMS dans un cadre beaucoup plus compliqué que celui de la machine à froid avec des maintenances conditionnelles qui pourraient dépendre de conditions d'utilisation (fréquences, choc, d'accélération, ...) mais aussi de conditions environnementales (température, hygrométrie,....). L'effet de ces covariables dépendant du temps sur les taux de panne est à étudier en préalable et pose de sérieux problèmes méthodologiques, les études de simulation utilisant les packages R adaptés aux modèles de survie à dépendance additive ou multiplicative présentés dans (101) montrent qu'en pratique la méthode d'estimation des paramètres ne fonctionnent pas lorsque les covariables dépendent du temps. Une des solutions seraient peut-être de discrétiser l'espace en comptant par exemple des dépassements de seuil critique pour ces covariables et revenir à des changements de régime qui nous sont plus familiers.

Chapitre 4

Soutien méthodologique aux chercheurs d'autres disciplines

Les travaux présentés dans ce chapitre correspondent à du soutien méthodologique en modélisation aux chercheurs d'autres disciplines. Ce travail n'est pas a priori de la recherche fondamentale en probabilités ou statistique mais il est une aide à la publication des collègues utilisant des données et il motive nos travaux.

4.1 Modélisation de la dynamique de l'ESCA

4.1.1 Contexte

L'esca de la vigne est une maladie du bois associée au développement de nécroses dans le bois des ceps qui impacte l'efficience des flux de sève, influençant à son tour la qualité du raisin. Cette maladie présente des symptômes foliaires de deux formes possibles : soit une forme lente qui ne se reconduit pas forcément d'une année sur l'autre, soit une forme apoplectique qui conduit presque systématiquement à la mort du cep. L'interdiction en 2001 du seul traitement chimique "curatif" en raison de sa toxicité sur l'homme a incité les phytopathologistes à se pencher sur l'épidémiologie de cette maladie afin de comprendre les facteurs associés à sa progression. La difficulté de reproduire les symptômes foliaires en conditions contrôlées et le développement des nécroses dans le bois durant une période longue constituent des freins à l'étude épidémiologique et au développement de modèle de compréhension de type mécaniste. Aussi, les études épidémiologiques entreprises à l'UMR SAVE sont basées sur l'analyse de la dynamique spatio-temporelle de la maladie, in natura, à l'échelle de la parcelle en intégrant les facteurs environnementaux et agronomiques. Pour ce faire l'UMR SAVE dispose de données de suivi de symptômes et de mort, relevés une fois par an au cep pour une vingtaine de parcelles de 2000 pieds environ depuis 2004.

C'est l'étude de ces données qui constitue la collaboration avec Lucia Guérin (INRA de Bordeaux UMR de santé végétale-Bordeaux Sup Agro) qui a débuté dans le cadre de tutorat de stages d'étudiants de notre master "Modélisation Ingénierie Mathématique Economique et Statistique", au département de Santé Végétale de l'lNRA Bordeaux notamment le stage en 2010 de Mamadou Cisse qui est actuellement en thèse à l'INRA de Rennes. Plus récemment, Amaury Labenne ancien étudiant de notre master, durant un CDD de 10 mois qui se termine

ce mois-ci a travaillé sur le sujet et dès le mois d'octobre 2012, Shuxiang Li débutera une thèse, financée par la région Aquitaine sur ce sujet.

4.1.2 Etude temporelle

Cette première partie porte sur l'étude de la dynamique temporelle de la maladie. Compte tenu du nombre d'observations réduit en temps (8 transitions observées), les techniques standard de traitement des séries temporelles n'est pas possible. Nous avons utilisé un modèle logistique à la parcelle afin d'estimer les risques de mort d'un cep en fonction de l'histoire de ses symptômes d'ESCA au cours des années précédentes. L'événement d'intérêt est la mort du cep pour une année fixée. L'étude met en évidence un risque accru de mort pour les ceps ayant exprimé des symptômes l'année précédent la mort. Le risque ne semble pas s'accumuler si des symptômes ont lieu plusieurs années de suite sauf pour certaines parcelles ce qui nous invite donc a étudier l'effet parcelle. Cette étude a donné lieu à une communication dans une conférence internationale (21) et à un article en cours de rédaction et qui sera prochainement soumis dans une revue de biologie végétale.

Pour explorer et comprendre les "effets parcelles", nous avons réalisé un modèle de Markov discret à 4 états et calculé les transitions empiriques correspondantes par année et parcelle. Une étude purement exploratoire n'a pas mis en évidence un "effet année" (ce qui pour nous est un résultat, car nous espérions capter un effet du climat au travers d'un "effet année"). En revanche, l'"effet passerelle" est réel et une enquête auprès des responsables de chaque parcelle a permis de mettre en évidence un lien entre les techniques de taille de la vigne et la probabilité pour un cep atteint de symptômes foliaires une année donnée d'être à nouveau symptomatique l'année suivante. Ce résultat est remarquable et pourrait conduire rapidement à des recommandations de conduites de parcelles auprès des vignerons. Ce travail est en collaboration avec Marie Chavent (IMB et équipe CQFD) et Amaury Labenne.

4.1.3 Etude spatiale et spatio-temporelle

Les techniques spatiales standard pour étudier l'agrégation spatiale des symptômes et ou de mort ont donné une nouvelle fois des résultats très variables suivant les parcelles et difficilement interprétables. En revanche pour répondre à la question "la structure spatiale des symptômes et/ou de la mort des ceps dans une parcelle donnée est-elle la même sur deux années consécutives ?", nous n'avons pas trouvé de tests adaptés dans la littérature. Aussi nous avons développé un test dit de MonteCarlo basé sur la vraisemblance. Le principe en est le suivant : à partie des données d'une parcelle A, on estime par des méthodes à noyau les probabilités théoriques d'être atteint pour chaque cep (package spatstat de R). On simule ensuite un échantillon de n réalisations indépendantes suivant ces probabilités, on écrit la vraisemblance correspondante pour la parcelle B et on regarde si elle appartient à l'intervalle à 95% des réalisations de cette vraisemblance dans l'échantillon. Par construction, ce test n'est pas symétrique dans le sens ou même si on veut tester l'hypothèse H_0 : "les parcelles A et B sont issues de la réalisation d'une même loi spatiale" la mise en oeuvre n'est pas symétrique et une parcelle A de loi uniforme ne sera pas rejetée par ce test pour des raisons spatiales mais pour des raisons de différence d'intensité en moyenne entre les deux parcelles. Une étude poussée de simulation a permis de mettre en évidence les propriétés et les limites

de la mise en pratique de ce test et a été présentée aux 1éres Journées R à Bordeaux en juillet dernier (24). Elle fait l'objet d'une publication méthodologique en cours de rédaction.

4.2 Normes et traitements en pneumologie

Au sein de l'Université de Bordeaux Segalen, j'ai développé une collaboration avec H. Guénard, PU-PH au CHU de Bordeaux et au laboratoire de physiologie ainsi qu'avec A. Chambellan du CHU de Nantes pour le deuxième volet des travaux présentés dans cette section.

Un premier travail qui a fait l'objet d'un article publié dans une revue internationale de pneumologie (5) a permis de définir des normes sur les paramètres de transfert des gaz dans le poumon. En effet, de nouveaux examens physiologiques mesurant le transfert de gaz dans le poumon sont désormais disponibles pour mesurer l'état de santé d'un patient. En revanche, il n'existait pas de normes permettant de décider au vu de chacun des résultats de ces examens si la mesure correspondait à un état pathologique ou non. Pour établir ces normes nous avons, pour chacune des variables, construit un modèle linéaire multiple avec deux pentes sur le paramètre âge. Les caractéristiques d'un patient étant données, il est alors facile de réaliser un intervalle de prédiction pour la valeur du paramètre physiologique qui donne alors une norme dépendante de covariables pour ce paramètre.

Il est connu que l'effet de la prise de bronchodilatateur n'est pas le même suivant les patients. Il est couramment admis que 40 % d'entre eux ne répondent pas au traitement. Cependant la réponse au traitement se mesure couramment sur son effet sur un voire deux paramètres physiologiques. Certains des patients pour lesquels le traitement n'a pas d'effet sur ces paramètres se sentent malgré tout globalement en meilleure forme. Comme de plus en plus de paramètres peuvent être mesurés en routine, il m'a été demandé en 2009 d'étudier l'effet de la prise de broncho-dilatateur et notamment de proposer un critère permettant de dire si un patient est répondant ou non à partir de la valeur de cinq paramètres mesurés avant et après une période de traitement. D'abord étonnée de cette question que je pensais réservée au médecin, je me suis prise au jeu en proposant tout simplement une classification avec ACP préalable des patients suivant leur réponse aux cinq paramètres espérant que le regroupement en deux classes homogènes fassent apparaître un groupe de patients avec amélioration des paramètres et un groupe sans grand changement. Le miracle s'est produit ! Comme le groupe des répondants se scindait naturellement en deux classes, il est apparu des effets différenciés du traitement : un groupe pour lequel l'amélioration se mesurait sur tous les paramètres dont celui couramment mesuré et un autre qui n'agissait que sur certains d'entre eux. L'ACP ayant été préalable à la classification, il a alors été facile de construire une règle à partir d'un indice linéaire en fonction de deux de ces variables. Les pneumologues impliqués ont été très intéressés notamment par la prise en compte d'une nouvelle variable dans le critère. Il a été décidé de lancer une nouvelle étude, dite de validation, sur un nouveau groupe de patients. Une fois les nouvelles données à disposition en 2011, se posait la question de valider la méthode sans toute façon pouvoir se baser sur un "gold-standard". Nous avons d'une part vérifié que l'ACP conduisait aux mêmes corrélations entre paramètres puis utilisé les indices

établis sur la cohorte d'origine pour constituer trois groupes. Ces groupes présentaient les mêmes caractéristiques que ceux de la cohorte initiale sur les cinq paramètres.

L'article correspondant a été soumis très récemment dans une revue internationale de pneumologie (16).

4.3 Conclusion et perspectives

Nous l'avons dit, ces collaborations ont d'abord démarré dans un cadre informel (autour de la machine à café pour les collègues pneumologues) ou pas forcément dans un objectif de production de recherche (encadrement de stages professionnels pour l'INRA) pour aboutir à un travail intéressant pour les différents partenaires. Finalement les problématiques de l'INRA sont des problématiques de mesure de risque (climatique, lié à la conduite de la passerelle, ...) pour l'élaboration éventuelle de recommandations et cette problématique ne diffère pas de celle des HUMS. Même si une première approche exploratoire n'a pas donné de profils caractéristiques d'année, il nous faudra tout de même creuser cette question par une modélisation. Les propriétés mathématiques du test de vraisemblance seront aussi à étudier. La collaboration avec l'INRA a débouché sur une thèse à la rentrée 2012 dont le projet a été défini par Lucia Guérin (INRA) et moi-même et a obtenu un financement de la région Aquitaine. Bien sûr cette thèse aura des aspects spatiaux que nous ne pourrons pas gérer sans une expertise extérieure (notamment pour tester l'éventuel lien suggéré par les phytopathologistes entre la résistivité du sol dont nous allons recevoir les mesures sur plusieurs passerelles).

Conclusion

Ce paragraphe sera bref, des conclusions et des perspectives ont été données à la fin de chaque chapitre. Je rappellerai juste que ces recherches se sont effectuées au sein de l'équipe INSERM de biostatistique E 0338 de 2003 à 2005. Après un bref passage dans l'équipe d'accueil Statistique et Applications dirigée par M. Nikouline, j'ai intégré l'Institut Mathématique de Bordeaux UMR 5251 en 2006 à sa création et j'en suis toujours membre. Je suis aussi membre de l'équipe INRIA CQFD depuis sa création en 2007. Une partie de mes recherches s'inscrit dans le projet ANR Fautocoes et dans des collaborations industrielles avec Astrium et Thales Optronique. Enfin je co-encadre les thèses de Romain Azaïs (financement ANR, début 09/2010) et de Camille Baysse (CIFRE Thales Optronique, début 09/2011) dont les travaux ont donné lieu à la rédaction de trois articles (un publié, un accepté, un soumis) et de nombreuses communications dans des conférences nationales et internationales. Une troisième thèse inscrite dans une école doctorale d'environnement démarre en cette rentrée.

Enfin, nos travaux sur les modèles de division cellulaire ont intéressé M. Rojo, directeur de recherche à l'Institut de Biochimie Génétique Cellulaire UMR 5095 de Bordeaux et spécialiste des mitochondries. Suite à nos échanges, nous avons le projet d'examiner les possibilités de modélisation stochastique de la division et la fusion des mitochondries. D'après M. Rojo, les régimes de fragmentation des mitochondries à l'échelle d'une cellule reflètent l'état de santé du sujet tout entier. J'attends donc avec impatience l'envoi de ce rapport pour aborder ce nouveau sujet avec mes collègues Benoîte de Saporta et Laurence Marsalle

Bibliographie

Publications personnelles citées

Articles

[1] Gégout-Petit, A. Approximate filter for the conditional law of a partially observed process in nonlinear filtering. *SIAM J. Control Optim. 36*, 4 (1998), 1423–1447 (electronic).

[2] Gégout-Petit, A., and Pardoux, E. Équations différentielles stochastiques rétrogrades réfléchies dans un convexe. *Stochastics Stochastics Rep. 57*, 1-2 (1996), 111–128.

[3] Commenges, D., and Gégout-Petit, A. Likelihood for generally coarsened observations from multistate or counting process models. *Scand. J. of Statist 34* (2007), 33–52.

[4] Commenges, D., Gégout-Petit, A., Joly, P., and Liquet, B. Choice between semi-parametric estimators of markov and non-markov multi-states model from generally coarsened observations. *Scand. J. of Statist 34* (2007), 432–450.

[5] Aguilaniu, B., Maitre, J., Gégout-Petit, A., Glénet, S., and Guénard, H. Co and no lung transfer, membrane conductance and capillary lung volume in a wide healthy european population : effects of age. *European Respiratory Journal 31* (2008), 1091–1097.

[6] Bercu, B., de Saporta, B., and Gégout-Petit, A. Asymptotic analysis for bifurcating autoregressive processes via a martingale approach. *Electron. J. Probab. 14* (2009), no. 87, 2492–2526.

[7] Commenges, D., and Gégout-Petit, A. A general dynamical statistical model with causal interpretation. *J. R. Stat. Soc. Ser. B Stat. Methodol. 71*, 3 (2009), 719–736.

[8] Gégout-Petit, A., and Commenges, D. A general definition of influence between stochastic processes. *Lifetime Data Anal. 16*, 1 (2010), 33–44.

[9] de Saporta, B., Gégout-Petit, A., and Marsalle, L. Parameters estimation for asymmetric bifurcating autoregressive processes with missing data. *Electron. J. Statist. 5* (2011), 1313–1353.

[10] Azaïs, R., Gégout-Petit, A., and Saracco, J. Optimal quantization applied to sliced inverse regression. *J. Statist. Plann. Inference 142*, 2 (2012), 481–492.

[11] de Saporta, B., Gégout-Petit, A., and Marsalle, L. Asymmetry tests for bifurcating autoregressive processes with missing data. *Statist. Probab. Lett. 82* (2012), 1439–1444.

[12] Azaïs, R., Dufour, F., and Gégout-Petit, A. Nonparametric estimation of the jump rate for non-homogeneous marked renewal processes. à paraître dans les Annales de l'IHP, arXiv :1202.2212.

Articles soumis

[13] Azaïs, R., Dufour, F., and Gégout-Petit, A. Nonparametric estimation of the jump rate for piecewise-deterministic markov processes. Submitted, arXiv :1202.2211.

[14] de Saporta, B., Gégout-Petit, A., and Marsalle, L. Random coefficient bifurcating autoregressive processes. submitted, arXiv :1205.3658.

[15] de Saporta, B., Gégout-Petit, A., and Marsalle, L. Statistical study of asymmetry in cell lineage data. submitted, arXiv :1205.4840.

[16] Chambellan, A., Gégout-Petit, A., Ben Saad, H., Perez, T., Benichou, M., Glérant, J.-C., and Guénard, H. Air trapping is the most sensitive target of the bronchodilator to salbutamol in copd : the trap (or trap'air) index.

Conférences avec actes

[17] Azais, R., Elegbede, C., Gégout-Petit, A., and Touzet, M. Estimation, simulation et prévision d'un modè de propagation de fissures par des processus markoviens déterministes par morceaux. In *Actes du congrès lambd-mu 17 17e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement 5-7 octobre 2010 La Rochelle* (France, 2010), pp. H5–C3.

[18] Baysse, C., Bihannic, D., Gégout-Petit, A., Prenat, M., and Saracco, J. Detection of a degraded operating mode of optronic equipment using Hidden Markov Model. In *PSAM 11 / ESREL 2012* (Helsinki, Finland, June 2012).

[19] Baysse, C., Bihannic, D., Gégout-Petit, A., Prenat, M., and Saracco, J. Optimisation de la maintenance d'un équipement optronique. In *Actes du congrès lambda-mu 18, 18e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement* (Tours, Oct. 2012).

[20] Ben Abdessalem, A., Azaïs, R., Gégout-Petit, A., Puiggali, M., and Touzet, M. Modelling of fatigue crack propagation using Piecewise Deterministic Markov Processes . In *ENGOTP2012 - 3rd International Conference on Engineering Optimization* (Rio de Janeiro, Brazil, July 2012).

[21] Guérin-Dubrana, L., Labrousse, J.-C., Bastien, S., Rey, P., and Gégout-Petit, A. Statistical analysis of the grapevines mortality associated Esca or Eutypa dieback foliar expression. In *8th International Workshop on Grapevine Trunk Diseases* (Valencia, Espagne, June 2012).

Autres conférences

[22] Gégout-Petit, A. Modèles probabilistes pour l'initiation et la propagation de fissures. Type : Conference digest, Aug 2010.

[23] Azais, R., Gégout-Petit, A., and Touzet, M. Modèlisation de propagation de fissure par un processus markovien déterministe par morceaux. Type : Conference digest, Aug 2010.

[24] Labenne, A., bonneu, F., Chavent, M., Gégout-Petit, A., and Guérin-Dubrana, L. Test de la vraisemblance entre deux motifs de points. In *1ères Journées R* (Bordeaux, France, July 2012).

Rapports techniques

[25] de Saporta, B., Dufour, F., Gégout-Petit, A., and Touzet, M. Modèles stochastiques pour la propagation de fissures. Tech. rep., Equipe Projet Contrôle de Qualité et Fiabilité Dynamique (CQFD), INRIA Bordeaux, EADS Astrium, 2007.

[26] Baysse, C., Gégout-Petit, A., and Saracco, J. Modèle de markov caché pour la détection d'un mode de fonctionnement dégradé d'un équipement optronique. Tech. rep., Equipe Projet Contrôle de Qualité et Fiabilité Dynamique (CQFD), INRIA Bordeaux, Thales Optronique, 2011.

Références dans le texte

[27] Aalen, O. Nonparametric inference for a family of counting processes. *Ann. Statist. 6*, 4 (1978), 701–726.

[28] Aalen, O. O. Statistical inference for a family of counting processes ProQuest LLC, Ann Arbor, MI, 1975. Thesis (Ph.D.)–University of California, Berkeley.

[29] Aalen, O. O. Dynamic modelling and causality. *Scand. Actuar. J.*, 3-4 (1987), 177–190.

[30] Aalen, O. O., Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. History of applications of martingales in survival analysis. *J. Électron. Hist. Probab. Stat. 5*, 1 (2009), 28.

[31] Aalen, O. O., Borgan, Ø., and Gjessing, H. K. *Survival and event history analysis.* Statistics for Biology and Health. Springer, New York, 2008. A process point of view.

[32] Aalen, O. O., and Frigessi, A. What can statistics contribute to a causal understanding ? *Scand. J. Statist. 34*, 1 (2007), 155–168.

[33] Aalen, O. O., and Johansen, S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Statist. 5*, 3 (1978), 141–150.

[34] Andersen, P. K. Multi-state models in survival analysis :a study oh nephropathy and mortality in diabetes. *Statis. Med. 7* (1988), 661–670.

[35] Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. *Statistical models based on counting processes.* Springer Series in Statistics. Springer-Verlag, New York, 1993.

[36] Asmussen, S., and Keiding, N. Martingale central limit theorems and asymptotic estimation theory for multitype branching processes. *Advances in Appl. Probability 10*, 1 (1978), 109–129.

[37] Bally, V., Pagès, G., and Printems, J. A quantization tree method for pricing and hedging multidimensional American options. *Math. Finance 15*, 1 (2005), 119–168.

[38] Bansaye, V., Delmas, J.-F., Marsalle, L., and Tran, V. Limit theorems for markov processes indexed by continuous time galton-watson trees. *Ann. Appl. Probab. 21* (2008), 2263–2314.

[39] Basawa, I. V., and Zhou, J. Non-Gaussian bifurcating models and quasi-likelihood estimation. *J. Appl. Probab. 41A* (2004), 55–64.

[40] Beran, J. Nonparametric regression with randomly censored survival data, 1981. Technical report, Dept. Statist. Univ. California, Berkeley.

[41] Blandin, V. Asymptotic results for bifurcating random coefficient autoregressive processes. ArXiv 1204.2926, 2012.

[42] Blandin, V. Limit theorems for bifurcating integer-valued autoregressive processes. ArXiv 1202.0470, 2012.

[43] Brandt, A. The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Adv. in Appl. Probab. 18*, 1 (1986), 211–220.

[44] Bui, Q. M., and Huggins, R. M. Inference for the random coefficients bifurcating autoregressive model for cell lineage studies. *J. Statist. Plann. Inference 81*, 2 (1999), 253–262.

[45] Bunge, M. *Causality : the place of the causal principle in modern science.* Dover Publications, New York, 1979.

[46] Cadarso-Suárez, C., Meira-Machado, L., Kneib, T., and Gude, F. Flexible hazard ratio curves for continuous predictors in multi-state models : an application to breast cancer data. *Stat. Model. 10*, 3 (2010), 291–314.

[47] Chambaz, A., Neuvial, P., and van der Laan M.J. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electron. J. Statist. 12* (2012), 1059–1099.

[48] Chen, B., Yi, G. Y., and Cook, R. J. Progressive multi-state models for informatively incomplete longitudinal data. *J. Statist. Plann. Inference 141*, 1 (2011), 80–93.

[49] Chiquet, J., and Limnios, N. Estimating stochastic dynamical systems driven by a continuous-time jump Markov process. *Methodol. Comput. Appl. Probab. 8*, 4 (2006), 431–447.

[50] Chiquet, J., Limnios, N., and Eid, M. Piecewise deterministic Markov processes applied to fatigue crack growth modelling. *J. Statist. Plann. Inference 139*, 5 (2009), 1657–1667.

[51] Daniel Commenges and Anne Gégout-Petit. Likelihood inference for incompletely observed stochastic processes : ignorability conditions. arXiv :math/0507151v1, 2005.

[52] Comte, F., Gaïffas, S., and Guilloux, A. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. Inst. H. Poincaré Probab. Statist. 47*, 4 (2011), 1171–1196.

[53] Costa, O. L. V., and Davis, M. H. A. Approximations for optimal stopping of a piecewise-deterministic process. *Math. Control Signals Systems 1*, 2 (1988), 123–146.

[54] Costa, O. L. V., Raymundo, C. A. B., and Dufour, F. Optimal stopping with continuous control of piecewise deterministic Markov processes. *Stochastics Stochastics Rep. 70*, 1-2 (2000), 41–73.

[55] Cowan, R., and Staudte, R. G. The bifurcating autoregressive model in cell lineage studies. *Biometrics 42* (1986), 769–783.

[56] Cox, D. R., and Wermuth, N. *Multivariate dependencies*, vol. 67 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996. Models, analysis and interpretation.

[57] Dabrowska, D. M. Nonparametric regression with censored survival time data. *Scand. J. Statist. 14*, 3 (1987), 181–197.

[58] Davis, M. H. A. Piecewise-deterministic Markov processes : a general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B 46*, 3 (1984), 353–388. With discussion.

[59] Davis, M. H. A. *Markov models and optimization*, vol. 49 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1993.

[60] Dawid, A. P. Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B 41*, 1 (1979), 1–31.

[61] Dawid, A. P. Causal inference using influence diagrams : the problem of partial compliance. In *Highly structured stochastic systems*, vol. 27 of *Oxford Statist. Sci. Ser.* Oxford Univ. Press, Oxford, 2003, pp. 45–81. With part A by Elja Arjas and part B by James M. Robins.

[62] de Saporta, B. Tail of the stationary solution of the stochastic equation $Y_{n+1} = a_n Y_n + b_n$ with Markovian coefficients. *Stochastic Process. Appl. 115*, 12 (2005), 1954–1978.

[63] de Saporta, B., Dufour, F., and Gonzalez, K. Numerical method for optimal stopping of piecewise deterministic markov process. *Ann. Appl. Probab. 20*, 5 (2010), 1607–1637.

[64] Delmas, J.-F., and Marsalle, L. Detection of cellular aging in a Galton-Watson process. *Stoch. Process. and Appl. 120* (2010), 2495–2519.

[65] Didelez, V. *Graphical models for event history analysis based on local independence*. Logos Verlag Berlin, Berlin, 2000. Dissertation, Universität Dortmund, Dortmund, 2000.

[66] Didelez, V. Graphical models for composable finite Markov processes. *Scand. J. Statist. 34*, 1 (2007), 169–185.

[67] Didelez, V. Graphical models for marked point processes based on local independence. *J. R. Stat. Soc. Ser. B Stat. Methodol. 70*, 1 (2008), 245–264.

[68] Duan, N., and Li, K.-C. Slicing regression : a link-free regression method. *Ann. Appl. Stat. 19*, 2 (1991), 505–530.

[69] Duflo, M. *Random iterative models*, vol. 34 of *Applications of Mathematics*. Springer-Verlag, Berlin, 1997.

[70] Eerola, M. *Probabilistic causality in longitudinal studies*, vol. 92 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994.

[71] Eichler, M., and Didelez, V. On Granger causality and the effect of interventions in time series. *Lifetime Data Anal. 16*, 1 (2010), 3–32.

[72] Elliott, R. J., Aggoun, L., and Moore, J. B. *Hidden Markov models*, vol. 29 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1995. Estimation and control.

[73] Farhangdoost, K., and Provan, J. W. A stochastic systems approach to fatigue reliability – an application to ti-6al-4v. *Engineering Fracture Mechanics 53*, 5 (1996), 687–706.

[74] Fleming, T. R. Nonparametric estimation for nonhomogeneous Markov processes in the problem of competing risks. *Ann. Statist. 6*, 5 (1978), 1057–1070.

[75] Florens, J.-P., and Fougere, D. Noncausality in continuous time. *Econometrica 64*, 5 (1996), 1195–1212.

[76] Fontbona, J., Guérin, H., and Malrieu, F. Quantitative estimates for the long time behavior of a PDMP describing the movement of bacteria. *Preprint* (2010).

[77] Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37* (1969), 424–438.

[78] Graph, S., and Luschgy, H. *Foundations of quantization for random vectors*, vol. 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

[79] Gugerli, U. S. Optimal stopping of a piecewise-deterministic Markov process. *Stochastics 19*, 4 (1986), 221–236.

[80] Guttorp, P. *Statistical inference for branching processes*. Wiley Series in Probability and Mathematical Statistics : Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.

[81] Guyon, J. Limit theorems for bifurcating Markov chains. Application to the detection of cellular aging. *Ann. Appl. Probab. 17*, 5-6 (2007), 1538–1569.

[82] Guyon, J., Bize, A., Paul, G., Stewart, E., Delmas, J.-F., and Taddéi, F. Statistical study of cellular aging. In *CEMRACS 2004—mathematics and applications to biology and medicine*, vol. 14 of *ESAIM Proc.* EDP Sci., Les Ulis, 2005, pp. 100–114 (electronic).

[83] Hamilton, J. D. *Time series analysis*. Princeton University Press, Princeton, NJ, 1994.

[84] Harris, T. E. *The theory of branching processes*. Die Grundlehren der Mathematischen Wissenschaften, Bd. 119. Springer-Verlag, Berlin, 1963.

[85] Hernández-Lerma, O., and Lasserre, J. B. *Markov chains and invariant probabilities*, vol. 211 of *Progress in Mathematics*. Birkhäuser Verlag, Basel, 2003.

[86] Huggins, R. M. Robust inference for variance components models for single trees of cell lineage data. *Ann. Statist. 24*, 3 (1996), 1145–1160.

[87] Huggins, R. M., and Basawa, I. V. Extensions of the bifurcating autoregressive model for cell lineage studies. *J. Appl. Probab. 36*, 4 (1999), 1225–1233.

[88] Huggins, R. M., and Basawa, I. V. Inference for the extended bifurcating autoregressive model for cell lineage studies. *Aust. N. Z. J. Stat. 42*, 4 (2000), 423–432.

[89] Jacobsen, M. *Statistical analysis of counting processes*, vol. 12 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1982.

[90] Jacobsen, M. *Point Process Theory and Applications : Marked Point and Piecewise Deterministic Processes*. Probability and its Applications. Birkhäuser, Boston-Basel-Berlin, 2006.

[91] Jacod, J. Multivariate point processes : predictable projection, Radon-Nikodým derivatives, representation of martingales. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 31* (1974/75), 235–253.

[92] Jacod, J., and Mémin, J. Caractéristiques locales et conditions de continuité absolue pour les semimartingales. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 35*, 1 (1976), 1–37.

[93] Jacod, J., and Shiryaev, A. N. *Limit theorems for stochastic processes*, second ed., vol. 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2003.

[94] Joly, P., Durand, C., Helmer, C., and Commenges, D. Estimating life expectancy of demented and institutionalized subjects from interval-censored observations of a multi-state model. *Stat. Model. 9*, 4 (2009), 345–360.

[95] Kallenberg, O. *Foundations of modern probability*, second ed. Probability and its Applications (New York). Springer-Verlag, New York, 2002.

[96] Keiding, N. Age-specific incidence and prevalence : a statistical perspective. *J. Roy. Statist. Soc. Ser. A 154*, 3 (1991), 371–412. With discussion.

[97] Kneib, T., and Hennerfeind, A. Bayesian semiparametric multi-state models. *Stat. Model. 8*, 2 (2008), 169–198.

[98] Lagakos, S. W., Sommer, C. J., and Zelen, M. Semi-Markov models for partially censored data. *Biometrika 65*, 2 (1978), 311–317.

[99] Li, G., and Doss, H. An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist. 23*, 3 (1995), 787–823.

[100] Maaouia, F., and Touati, A. Identification of multitype branching processes. *Ann. Statist. 33*, 6 (2005), 2655–2694.

[101] Martinussen, T., and Scheike, T. H. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York, 2006.

[102] McCrea, R. S. Sufficient statistic likelihood construction for age- and time-dependent multi-state joint recapture and recovery data. *Statist. Probab. Lett. 82*, 2 (2012), 357–359.

[103] McKeague, I. W., and Utikal, K. J. Inference for a nonlinear counting process regression model. *Ann. Statist. 18*, 3 (1990), 1172–1187.

[104] Meyn, S., and Tweedie, R. L. *Markov chains and stochastic stability*, second ed. Cambridge University Press, Cambridge, 2009.

[105] Mézin, U., and Vallois, P. Statistical analysis of unidirectional multicracking of coatings by a two-dimensional poisson point process. *Math. Mech. Solids.*

[106] Nelson, W. Hasard plotting for incomplete failure data. 27–52.

[107] Nicholls, D. F., and Quinn, B. G. *Random coefficient autoregressive models : an introduction*, vol. 11 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1982. Lecture Notes in Physics, 151.

[108] Pagès, G. A space quantization method for numerical integration. *J. Comput. Appl. Math. 89*, 1 (1998), 1–38.

[109] Pagès, G., Pham, H., and Printems, J. An optimal Markovian quantization algorithm for multi-dimensional stochastic control problems. *Stoch. Dyn. 4*, 4 (2004), 501–545.

[110] Pagès, G., Pham, H., and Printems, J. Optimal quantization methods and applications to numerical problems in finance. In *Handbook of computational and numerical methods in finance*. Birkhäuser Boston, Boston, MA, 2004, pp. 253–297.

[111] Paris, P., and Erdogan, F. A critical analysis of crack propagation laws. *Journal of basic engineering 85* (1963), 528–534.

[112] Pearl, J. *Causality*, second ed. Cambridge University Press, Cambridge, 2009. Models, reasoning, and inference.

[113] Perrin, F. *Prise en compte des données expérimentales dans les modèles probabilistes pour la prévision de la durée de vie des structures*. PhD thesis, Université Blaise Pascal Clermont-Ferrand II, 2008.

[114] Picard, J. Efficiency of the extended Kalman filter for nonlinear systems with small noise. *SIAM J. Appl. Math. 51*, 3 (1991), 843–885.

[115] Pierce, J. Asymptotic quantizing error for unbounded random variables. *IEEE Trans. Inform. Theory 16* (1970), 101–112.

[116] Pötter, U., and Blossfeld, H.-P. Causal inference from series of events. *European Sociological Review 17* (2001), 21–32.

[117] Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *Ann. Statist. 11*, 2 (1983), 453–466.

[118] Røysland, K. A martingale approach to continuous-time marginal structural models. *Bernoulli 17*, 3 (2011), 895–915.

[119] Salmon, W. *Scientific axplanation and the causal structure of the world*. Princeton University Press, Princeton, 1984.

[120] Schweder, T. Composable Markov processes. *J. Appl. Probability 7* (1970), 400–410.

[121] Shen, W., Soboyejo, A. B. O., and Soboyejo, W. O. Probabilistic modeling of fatigue crack growth in ti–6al–4v. *International Journal of Fatigue 23*, 10 (2001), 917–925.

[122] Stewart, E., Madden, R., Paul, G., and Taddei, F. Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biol. 3*, 2 (2005), e45.

[123] Stute, W. Conditional empirical processes. *Ann. Statist. 14*, 2 (1986), 638–647.

[124] Sweeting, M. J., Farewell, V. T., and De Angelis, D. Multi-state Markov models for disease progression in the presence of informative examination times : an application to hepatitis C. *Stat. Med. 29*, 11 (2010), 1161–1174.

[125] Titman, A. C. Computation of the asymptotic null distribution of goodness-of-fit tests for multi-state models. *Lifetime Data Anal. 15*, 4 (2009), 519–533.

[126] Utikal, K. J. Nonparametric inference for a doubly stochastic Poisson process. *Stochastic Process. Appl. 45*, 2 (1993), 331–349.

[127] Utikal, K. J. Nonparametric inference for Markovian interval processes. *Stochastic Process. Appl. 67*, 1 (1997), 1–23.

[128] Virkler, D., Hillberry, B., and Goel, P. The statistical nature of fatigue crack propagation. *J. Engng Mater Tech , Trans. ASME 101* (1979), 148–153.

[129] Wright, S. Correlation and causation. *J. Agric. Res. 20* (1921), 557–585.

[130] Wright, S. The method of paths coefficients. *Ann. Math. Statist. 5* (1934), 161–215.

[131] Wu, W. F., and Ni, C. C. A study of stochastic fatigue crack growth modeling through experimental data. *Probabilistic Engineering Mechanics 18*, 2 (2003), 107–118.

[132] Wu, W. F., and Ni, C. C. Probabilistic models of fatigue crack propagation and their experimental verification. *Probabilistic Engineering Mechanics 19*, 3 (2004), 247–257.

[133] Yang, J. N., and Manning, S. D. A simple second order approximation for stochastic crack growth analysis. *Engineering Fracture Mechanics 53*, 5 (1996), 677–686.

[134] Zador, P. *Development and evaluation of procedures for quantizing multivariate distributions*. PhD thesis, Stanford University, Stanford, CA 94305, 1963.

[135] Zhou, J., and Basawa, I. V. Least-squares estimation for bifurcating autoregressive processes. *Statist. Probab. Lett. 74*, 1 (2005), 77–88.

[136] Zhou, J., and Basawa, I. V. Maximum likelihood estimation for a first-order bifurcating autoregressive process with exponential errors. *J. Time Ser. Anal. 26*, 6 (2005), 825–842.