Estimation de la fonction d'intensité d'un processus ponctuel par complexité minimale

Jocelyn Nembé

THÈSE

présentée par

Jocelyn NEMBÉ

pour obtenir le grade de Docteur
de l'Université de Joseph Fourier de Grenoble
(Arrêté ministériel du 23 juillet 1984)

Spécialité : mathématiques appliquées

Estimation de la Fonction d'Intensité d'un Processus Ponctuel par Complexité Minimale.

Soutenue le 29 Octobre 1996

Président :	Alain	LE BRETON
Rapporteurs :	Denis	BOSQ
	Monique	BERTRAND
Examinateurs :	Anestis	ANTONIADIS
	Claudine	ROBERT
Directeur :	Gérard	GRÉGOIRE

Thèse préparée au sein du **L**aboratoire de **M**odélisation et **C**alcul

THÈSE

présentée par

Jocelyn NEMBÉ

pour obtenir le grade de Docteur
de l'Université de Joseph Fourier de Grenoble
(Arrêté ministériel du 23 juillet 1984)

Spécialité : mathématiques appliquées

Estimation de la Fonction d'Intensité d'un Processus Ponctuel par Complexité Minimale.

Soutenue le 29 Octobre 1996

Président :	Alain	LE BRETON
Rapporteurs :	Denis	BOSQ
	Monique	BERTRAND
Examinateurs :	Anestis	ANTONIADIS
	Claudine	ROBERT
Directeur :	Gérard	GRÉGOIRE

Thèse préparée au sein du **L**aboratoire de **M**odélisation et **C**alcul

Cette thèse a été préparée au sein du Laboratoire de Modélisation et Calcul à l'Institut de Mathématiques Appliquées de Grenoble (IMAG). Je tiens à remercier tous les membres de l'équipe "Statistique et Modélisation Stochastique", les uns pour leur disponibilité, et les autres pour l'amitié qu'ils m'ont témoigné pendant plusieurs années.

Mes remerciements vont en particulier à Monsieur Gérard GRÉGOIRE, Professeur à l'Université Lumière (Lyon II), qui a assuré la direction de ce travail, pour sa disponibilité, sa patience et la compréhension dont il a toujours fait preuve. Par ses remarques et son orientation, j'ai énormément appris durant ces quatres années passées en sa compagnie. Je tiens également à le remercier pour toute l'aide matérielle qu'il m'a apportée pendant toute la durée de ce travail.

Je remercie Monsieur Alain LE BRETON, Professeur à l'Université Joseph-Fourier (Grenoble), qui m'a fait l'honneur de présider le jury de cette thèse.

Que Monsieur Denis BOSQ, Professeur à l'Université Paris VI, et Madame Monique BERTRAND, Professeur à l'Université de Rennes, qui ont accepté d'être rapporteurs de ce travail, trouvent ici l'expression de ma reconnaissance pour le temps qu'ils y ont consacré, et pour les remarques judicieuses qu'ils m'ont faites.

Mes vifs remerciements à Monsieur Anestis ANTONIADIS, Professeur à l'Université Pierre-Mendès FRANCE pour m'avoir fait l'honneur de participer à ce jury, et d'avoir pris le temps d'examiner cette thèse.

Madame Claudine ROBERT, Professeur à l'Université Joseph-Fourier a également accepté d'examiner ce travail. Je la remercie pour le temps qu'elle y a consacré, ainsi que pour ses encouragements.

Cette thèse ne ce serait pas achevée sans l'aide et la présence de mes proches. Je remercie toute ma famille, et spécialement mon Père pour son appui affectif et matériel. La communauté gabonaise de Grenoble, a été une seconde famille pour moi. Je remercie tous mes frères pour leur présence constante, et enfin ma fiancée, qui m'a tant soutenu et supporté.

Liste des figures

Liste des exemples

Table des matières

NOTATIONS

N : processus ponctuel.

Y : processus prévisible associé.

$(\Gamma_n)_n$: suite croissante d'ensemble de fonctions candidates.

$m(n)$: dimension de Γ_n dans les cas paramétrique et semi-paramétrique.

$(\phi_1, \phi_2, \ldots, \phi_{m(n)})$: base de Γ_n dans les cas paramétrique et semi-paramétrique.

δ_n : précision à laquelle sont codés les coefficients de la décomposition des fonctions de Γ_n sur la base $\phi_1, \ldots, \phi_{m(n)}$ dans les cas paramétrique et semi-paramétrique.

Γ :limite de la suite $(\Gamma_n)_n$.

P_α : loi de probabilité du processus ayant généré les données.

α : fonction d'intensité associée à la loi du processus ayant généré les données.

β : fonction d'intensité candidate.

$\mathcal{L}_\beta$: vraisemblance basée sur la fonction β.

$H_\alpha(\alpha)$: entropie du processus.

$d(\alpha \parallel \beta)$: distance de Kullback entre les deux fonctions d'intensité. Redondance du code basé sur la fonction d'intensité β.

$\overline{\Gamma}$: fermeture de Γ au sens de la distance de l'entropie.

$C_n(\cdot)$: complexité d'une fonction de Γ_n.

$R_n(\alpha)$: indice de résolvabilité.

$d_H(\alpha, \beta)$: distance de Hellinger entre les deux donctions d'intensité.

$D_H(\mathcal{L}_\alpha, \mathcal{L}_\beta)$: distance de Hellinger entre les deux vraisemblances.

Chapitre 1

Introduction.

Une réalisation de processus ponctuel sur un intervalle $[0, T]$ (qu'on pourra prendre pour le temps), peut être considérée comme une famille de points, dont le nombre total et les coordonnées sur l'axe de temps, sont aléatoires. Une vaste classe de processus ponctuels comprenant les processus de Poisson et les processus associés aux modèles de durées de vie avec censure, peut être représentée à l'aide du modèle à intensité multiplicative (modèle de Aalen). Dans ce modèle, l'intensité stochastique est de la forme $\lambda(s) = \alpha(s)Y(s)$ pour tout $s \in [0, T]$, où α est une fonction déterministe strictement positive, et Y un processus prévisible. La recherche d'un estimateur lisse de la fonction d'intensité s'est développée dans deux directions principales. La première consiste à lisser des estimateurs discontinus de l'intégrale indéfinie de α. La seconde consiste à rechercher directement l'estimateur dans un espace de fonctions lisses. Dans cette deuxième direction on trouve en particulier des méthodes de projection orthogonale, et des méthodes basées sur la fonction de vraisemblance. Cette dernière n'admettant pas de maximum dans un cadre général, l'existence d'un estimateur du maximum de vraisemblance ne peut être assurée que par l'imposition de contraintes, relâchées au fur et à mesure que la taille de l'échantillon croît, ou par l'introduction d'une pénalité dépendant de la lissitude des fonctions candidates. La méthode étudiée ici peut formellement être considérée comme la minimisation de la somme de deux termes, où l'un est l'opposé de la log-vraisemblance, et l'autre un terme de pénalité, représentant la complexité de Kolmogorov de la fonction d'intensité candidate. Le premier terme mesure l'adéquation de la fonction candidate aux données. C'est aussi la longueur d'un code de Shannon basé sur la loi de probabilité définie par la fonction d'intensité candidate.

Le second mesure la difficulté liée à la description de l'algorithme calculant la fonction candidate, et peut être considéré comme la longueur d'un code pour cette fonction. La somme des deux termes représente ainsi la longueur d'un code défini d'une part par un préfixe codant une fonction d'intensité, d'autre part par un code pour les réalisations d'un processus ponctuel (tronquées à une certaine précision) défini à partir de la fonction codée dans le préfixe. Cette méthode, proche de la méthode de description minimale développée à l'origine par Rissanen, a été étudiée dans le cadre de l'estimation de densité par un certain nombre d'auteurs et a permis d'obtenir des estimateurs fortement consistants. Un lien direct entre l'erreur d'estimation statistique et la redondance du code ainsi construit peut être établi (voir [9]). Le principe de la méthode de complexité minimale est de trouver un compromis entre la longueur de description des fonctions candidates (i.e. la longueur des algorithmes associés à ces fonctions) et leur adéquation aux données. Indépendamment de cet aspect inférentiel, les codes ainsi construits présentent des propriétés remarquables du strict point de vue du codage. Les résultats d'optimalité sont présentés à travers un indice de résolvabilité. Cette quantité que nous définissons dans le cadre des processus ponctuels, a été utilisée par [9] dans le contexte de l'estimation de densités de probabilité. L'indice de résolvabilité représente la redondance moyenne minimale du code dans la famille des codes candidats.

Le chapitre 2 présente les résultats et les définitions de bases qui serviront dans la suite. Sont ainsi exposées, afin que le document se suffise à lui-même dans une large mesure, des notions élémentaires de calculabilité ainsi que la complexité des fonctions réelles. Les éléments de la théorie des processus ponctuels nécessaires à l'utilisation du modèle de Aalen sont aussi fournis dans ce chapitre, de même que quelques éléments de la théorie du codage.

Le chapitre 3 traite du problème du codage universel de réalisations de processus ponctuels dans le modèle de Aalen. Nous obtenons des majorations de l'indice de résolvabilité pour différents types de familles candidates. Selon le type de famille, le code associé à une fonction sera un code sur des entiers décrivant soit les coefficients (tronqués à une certaine précision) du développement de la fonction sur une base, soit l'indice de la fonction au sein d'une famille finie de fonctions. Si la fonction n'admet pas un développement fini sur la base considérée, ou si on ne connaît pas a priori un ensemble fini susceptible de la contenir, elle ne peut alors être codée. Dans ce cas, nous montrons qu'il existe néanmoins pour tout échantillon de taille n, un code en deux parties, dont la longueur ne diffère du code de Shannon basé sur la fonction inconnue que d'un terme de

l'ordre de $\log n$ pour un échantillon de n réalisations. Ce résultat étend la propriété d'optimalité en moyenne des codes universels, et suggère l'utilisation du code basé sur la fonction d'intensité estimée dans le cas où la fonction associée à la loi du processus est inconnue, et dans celui où elle ne peut être codée. Afin d'illustrer cette approche, nous proposons une application au problème de la compression d'images numériques par un modèle de chaîne de Markov avec censure.

L'inférence statistique est abordée au chapitre 4. Ce chapitre est consacré à l'étude des conditions de consistance de l'estimateur de complexité minimale et à la détermination de vitesses de convergence. La recherche de la fonction inconnue α dans un espace de fonctions positives permet d'obtenir des estimateurs presque-sûrement consistants de la fonction d'intensité en distance de Hellinger. Le premier résultat de consistance montre que si l'estimateur appartient à l'une des familles de fonctions candidates, alors presque-sûrement il existe une taille d'échantillon à partir de laquelle l'estimateur est identiquement égal à cette fonction. Le second résultat montre que, dans le cas le plus général, si la fonction d'intensité inconnue peut être approchée au sens de l'entropie par des fonctions candidates, alors l'estimateur est presque-sûrement consistant au sens de la distance de Hellinger. Des hypothèses légèrement plus fortes sur la famille des fonctions candidates (continuité), suffisent à établir la consistance presque-sûre de l'estimateur de complexité minimale au sens de l'entropie, et par conséquent au sens de la norme L_1. Cette étude est complétée par la détermination des vitesses de convergence, effectuée à travers l'indice de résolvabilité. Et nous montrons en particulier que l'erreur d'estimation est presque-sûrement majorée au sens de la distance de Hellinger par l'indice de résolvabilité, ce qui établit à nouveau un lien entre l'optimalité de ce principe de codage et la précision de l'estimation de la loi du processus. Des versions plus fortes de ces derniers résultats sont données dans le cas de modèles spécifiques tels que les processus de Poisson non-homogènes et les modèles de durées de vie avec censure. Pour ces deux cas particuliers, la vitesse de convergence en distance de Kullback-Leibler est presque-sûrement majorée par une fonction de l'indice de résolvabilité. Des exemples de vitesses de convergence sont données pour des familles de fonctions candidates particulières (polynomiales, trigonométriques et splines). Au chapitre 5 des résultats de normalité asymptotique sont établis, et des tests exponentiels sont construits. Sous des conditions relativement peu contraignantes, deux résultats de normalité asymptotique sont montrés dans ce chapitre. Le premier, est de type séquentiel; le second, montre sous des hypothèses plus fortes, la normalité asymptotique de l'estimateur obtenu sur tout l'intervalle de temps, et autorise la construction

d'intervalles de confiance. Enfin le chapitre 6 présente une étude numérique complète et des résultats de simulation dans le cadre des modèles de durées avec censure.

Le travail effectué ici peut être généralisé ou affiné dans les directions suivantes :

- La plupart des résultats présentés ici dans le cas d'observations i.i.d., devraient pouvoir se généraliser au cas de processus Markoviens. Dans ce cadre en effet, on dispose d'inégalités de grandes déviations équivalentes à celles qui servent ici à établir les résultats de convergence.

- L'étude des vitesses de convergence montre qu'elles sont majorées en distance de Hellinger par l'indice de résolvabilité. La recherche de l'estimateur dans plusieurs familles de bases de fonctions possédant des propriétés d'approximation différentes, peut peut-être améliorer l'indice de résolvabilité, et ainsi, les vitesses de convergence.

- La détermination de résultats de normalité asymptotique pour ce type d'estimateur (en dehors du cadre de l'histogramme) quand la fonction estimée n'appartient pas à l'une des familles candidates, demeure un problème ouvert. Une étude peut être menée dans ce sens.

Chapitre 2

Définitions et résultats préliminaires.

2.1 Calculabilité et complexité

Quels problèmes admettent une solution algorithmique ? Un algorithme est une succession d'instructions, qui à partir d'un certain nombre d'éléments de base, ou d'entrée, produit le résultat désiré. Pour donner un exemple très simpliste : à partir d'un certain nombre d'ingrédients, farine, sucre, oeufs (qui constituent l'entrée), et en suivant une recette (suite d'instructions), on obtient un gateau (résultat désiré ou sortie). L'ensemble {entrée, suite d'instructions, sortie}, donne une définition intuitive de la notion d'algorithme. La question fondamentale qui a été à l'origine de la théorie de la calculabilité, est la suivante :

" Quels problèmes peuvent être résolus par un algorithme ? "

En 1931, le théorème d'incomplétude de Gödel a permis de montrer qu'il existe dans tout système formel défini par des règles d'opérations logiques et des axiomes, des propositions vraies, qui ne peuvent pas être démontrées à l'intérieur du système. Le problème posé était de concevoir un algorithme qui accepte une proposition en entrée, et qui après une suite d'opérations logiques, indique en sortie que la proposition est vraie ou fausse. A la suite de Gödel, de nombreux autres auteurs, dont Church, Kleene, Post, et Turing, ont découvert quantité de problèmes qui ne possédaient pas de solution algorithmique. Chacun de ces auteurs a tenté de donner une définition plus précise du mot algorithme. Gödel le définit comme étant un ensemble de règles pour exprimer une fonction mathématique à partir de fonctions de base, dites simples. Church utilise un formalisme appelé " calcul lambda ", tandis que Turing utilise une machine

hypothétique appelée "machine de Turing". Toutes ces définitions se sont révélées équivalentes, et permettent chacune de caractériser l'ensemble des problèmes qui admettent une solution à partir d'une suite d'instructions propres au formalisme considéré. Les machines de Turing constituent néanmoins le modèle le plus général utilisé. Nous retiendrons donc que les problèmes qui admettent une solution algorithmique, sont ceux qui peuvent être formalisés par une machine de Turing.

Les machines de Turing. Une machine de Turing est un modèle de calculateur, ou d'automate, possédant une entrée et des mémoires représentées par plusieurs rubans, une sortie constituée d'un ruban, et un dispositif interne de fonctionnement qui en détermine les actions. Les rubans sont composés de cases, sur lesquelles une tête de lecture-écriture permet de lire et d'écrire les symboles d'un alphabet donné (généralement 0,1 et un séparateur noté B). Son fonctionnement est séquentiel, et caractérisé par un ensemble d'états. Une partie de ces états provoque l'arrêt de la machine, ce sont des états terminaux. Certains états terminaux sont dits accepteurs, ils correspondent à un déroulement "normal" [1] des opérations, et d'autres sont dits non-accepteurs. La machine peut également ne pas s'arrêter[2]. Quand la machine s'arrête, le résultat est la chaîne de caractères inscrite sur le ruban de sortie si l'état final est accepteur, il est indéfini sinon. La suite des actions est spécifiée par une fonction de transition d'états, qui associe à un état donné et à la valeur lue sur le ruban, l'action à exécuter et l'état suivant de la machine. Pour une définition plus précise, voir définition A.1 en annexe.

Fonctions partielles et fonctions récursives. Considérons des machines dont l'alphabet est composé des symboles 0, 1 et du séparateur B. Les entrées et sorties de la machine de Turing sont alors des chaînes binaires. On admettra d'autre part, qu'il est possible d'associer une chaîne binaire unique à chaque machine de Turing, qui en décrit toutes les caractéristiques. Une telle chaîne sera aussi appelée un programme. Dans la suite on identifiera donc chaque machine de Turing à une chaîne binaire. Chaque machine de Turing à p rubans d'entrée calcule une fonction d'un domaine $I \subseteq \mathrm{N}^p$ vers N : la fonction qui associe à chaque entrée le contenu du ruban de sortie de la machine de Turing quand elle s'arrête dans un état accepteur, et qui

[1] Un déroulement est considéré normal selon ce que l'utilisateur attend du fonctionnement de la machine, c'est une notion relative.

[2] C'est l'équivalent d'un programme qui entre en boucle infinie

n'est pas définie sinon. L'ensemble de toutes les fonctions qui peuvent être représentées par une machine de Turing, représente une classe de fonctions appelées fonctions partielles récursives, ou fonctions partielles. Si la fonction s'arrête dans un état accepteur pour toute les entrées, elle est dite récursive ou récursive totale.

Définition 2.1. (Fonction partielle récursive)

Chaque machine de Turing à p rubans d'entrée définit une fonction partielle de p-uplets d'entiers à valeurs dans l'ensemble des entiers strictement positifs. Une telle fonction est dite partielle récursive ou calculable. Si cette fonction est définie pour toute entrée et que la machine de Turing s'arrête pour toutes les entrées, la fonction est dite récursive totale ou récursive.

Complexité d'une chaîne binaire. Les fonctions partielles permettent de définir la complexité d'une chaîne binaire au sens de Kolmogorov et au sens de Chaitin. Pour ces définitions, nous admettrons qu'il est possible d'énumérer tous les programmes de machine de Turing, et qu'il existe des fonctions partielles A et U qui acceptent en entrée un programme p de machine de Turing, et qui calculent la fonction représentée par le programme p. Les chaînes binaires représentant les programmes peuvent appartenir à un ensemble prefix-free (voir définition 2.2 ci-dessous). C'est le cas pour la complexité de Chaitin. A une chaîne binaire x, on associe alors l'ensemble de tous les programmes de machines de Turing produisant x comme résultat d'une exécution, et la longueur du plus petit de ces programmes au sein de cet ensemble définit la complexité[3] de x.

Définition 2.2. *Un ensemble de chaînes est dit prefix-free si aucune chaîne n'est le préfixe d'une autre. Si $x_1x_2\ldots x_n$ est une chaîne, alors pour $1 \leq k < n$, $x_1\ldots x_k$ n'en est pas une.*

Définition 2.3. (Complexité).

Dans les définitions suivantes p appartient à un ensemble de chaînes binaires, chacune représentant le codage d'une machine de Turing. On note $\{0,1\}^*$ l'ensemble de toutes les chaînes binaires.

[3] Chaque programme binaire est identifié à une chaîne de caractères et la longueur d'un programme est le nombre d'éléments de la chaîne.

1 *La complexité de Kolmogorov d'une chaîne binaire* $x(n) \in \{0,1\}^n$ *par rapport à une fonction partielle récursive* $A : \{0,1\}^* \times \mathrm{N} \to \{0,1\}^*$ *est définie par*

$$K_A(x(n)|n) = \min_{A(p,n)=x(n)} l(p)$$

où $l(\cdot)$ *désigne la longueur de la chaîne, et* N *est l'ensemble des entiers naturels.*

2 *Soit* $U : \{0,1\}^* \to \{0,1\}^*$ *une fonction partielle récursive à domaine prefix-free. On définit la complexité de Chaitin d'une chaîne binaire* x *par rapport à* U *par*

$$C_U(x) = \min_{U(p)=x} l(p)$$

La complexité d'une fonction numérique f, de $I \subseteq \mathrm{N}^p$ vers N peut être spécifiée de plusieurs manières à partir de la définition 2.3. Une première approche consiste à identifier chaque fonction entière à son graphe, c'est à dire à l'ensemble des couples $(x, f(x))$, quand x varie dans I.

Complexité d'une fonction définie par son graphe. Si f est partielle récursive, ou récursive, on pourra caractériser l'ensemble des couples (x, y) tels que $y = f(x)$. Il existe donc une équivalence entre le calcul de f et le calcul de la fonction indicatrice de l'ensemble :

$$\mathcal{G}(f) = \{(x, y) \in \mathrm{N}^p \times \mathrm{N} : y = f(x)\}.$$

Inversement, pour tout ensemble A de $\mathrm{N}^p \times \mathrm{N}$ on a les définitions suivantes.

Définition 2.4. (Ensemble récursif ou récursivement énumérable)
A est récursivement énumérable si A est vide ou si A est l'ensemble des valeurs d'une fonction récursive. De manière équivalente, il existe une machine de Turing M, telle que pour tout élément de A, M s'arrête dans un état accepteur, et pour tout élément de $\bar{A}$, *M ne s'arrête pas ou s'arrête dans un état non accepteur. A est un ensemble récursif si* $\mathbf{I}_A$ *est une fonction récursive. De manière équivalente A est récursif si A est accepté par une machine de Turing qui s'arrête toujours.*

Exemple 2.1 : Ensembles récursifs.

Les ensembles suivants sont récursifs : l'ensemble des nombres premiers, l'ensemble des multiples d'un nombre donné. Il existe un algorithme définissant une fonction qui vaut 1 quand un nombre est dans

l'ensemble et 0 sinon. Les ensembles suivants sont récursivement énumérables : tout ensemble récursif, l'ensemble des x tels qu'il existe y, z et $n > 2$, vérifiant $x^n + y^n = z^n$. L'ensemble des entiers i tel qu'il existe un bloc de i zéros consécutifs dans le développement décimal de π.

Remarque : Dans le cas des ensembles récursivement énumérables (mais non récursifs), on peut trouver une fonction qui vaut 1 quand l'élément appartient à l'ensemble, mais qui est indéfinie sinon. A un instant donné on ne sait pas si l'élément analysé appartient à l'ensemble; l'exécution pouvant ne jamais se terminer, il peut être impossible de conclure sur un temps d'exécution fini.

◇

Ces définitions permettent de déterminer la complexité d'une fonction entière à partir de son graphe. Les programmes considérés dans toute la suite appartiennent à un ensemble prefix-free de programmes. Pour simplifier, on considère le cas d'une fonctions β de $I \subseteq \mathbf{N}$ vers $\mathbf{N}$. Si la fonction β est calculable sur le domaine $I \subseteq \mathbf{N}$, alors son graphe $\mathcal{G} = \{(x, \beta(x)), x \in I\}$ est un ensemble récursivement énumérable. En effet si cette fonction est calculable, alors il existe une machine de Turing M qui, à toute entrée $x \in I$, associe en sortie, l'entier $\beta(x)$. Pour tout couple (x, y), tel que $(x, y) \in I \times \mathbf{N}$, il est alors possible de déterminer si $y = \beta(x)$. En énumérant récursivement l'ensemble $I \times \mathbf{N}$, et en faisant fonctionner la machine de Turing M sur chaque entrée, on obtient ainsi une énumération récursive du graphe de β. La complexité de la fonction β sur I est alors égale à la complexité de son graphe. Si β est une fonction de $\mathbb{Q}$ dans $\mathbb{Q}$, on sait qu'il existe une bijection σ entre $\mathbf{N}$ et $\mathbb{Q}$. A tout couple $(x, \beta(x)) \in \mathbb{Q} \times \mathbb{Q}$ peut être associé un couple d'entiers $(\sigma(x), \sigma(\beta(x))$. La complexité de la fonction β sur une partie A de l'ensemble des nombres rationnels sera (comme précédemment) égale à la complexité de la fonction indicatrice du graphe défini par :

$$\mathcal{G} = \{(\sigma(x), \sigma(\beta(x)), x \in A\}.$$

Considérons maintenant le cas d'une fonction β d'un sous-ensemble dénombrable D de $\mathbf{R}$ dans un sous-ensemble dénombrable R de $\mathbf{R}$. Soit $\gamma(x, a) : D \times \mathbf{N} \to R$ une fonction rationnelle telle que :

$$\forall x \in D, \forall a \in \mathbf{N}, |\gamma(x, a) - \beta(x)| \leq 2^{-a} \tag{2.1}$$

La complexité de la fonction réelle β est alors égale à celle de la fonction rationnelle γ^* de complexité minimale parmi toutes celles qui vérifient (2.1) [4].

Dans le cas le plus général, il n'existe pas de moyen de déterminer une expression, même approchée de ces complexités, mais si le but recherché est la transmission de données à partir d'un code variable (resp. l'inférence statistique), il suffit de se restreindre à des ensembles particuliers, possédant de bonnes propriétés d'approximation par rapport à un ensemble suffisamment grand, susceptible de contenir le code recherché (resp. la fonction inconnue), et de s'intéresser au sein de ces ensembles, non pas à la complexité intrinsèque, mais à une complexité relative aux ensembles considérés[5]. Supposons pour cela que la fonction f puisse être représentée par une chaîne binaire $x(f)$, de telle manière que deux fonctions différentes n'admettent pas la même représentation. Si les fonctions considérées ont un domaine de définition I et sont de la forme :

$$\beta(s) = \sum_{j=0}^{m} \beta_j \phi_j(s), \quad s \in I,$$

et qu'il existe une fonction calculable ϕ telle que

$$\phi_j(s) = \phi(j, s), \quad j = 0, \ldots, m, \quad s \in I,$$

à chaque fonction peut être associée une chaîne $x(\beta) = x(\phi) + x(\beta_0, \ldots, \beta_m)$, où $x(\phi)$ représente le programme décrivant ϕ, et $x(\beta_0, \ldots, \beta_m)$ est la description des coefficients de β sur le ruban d'une machine de Turing. Pour que cette décomposition soit unique, il faut et il suffit que le procédé de représentation des coefficients sur le ruban le soit également. La complexité intrinsèque de chaque fonction sera ainsi déterminée par un terme $x(\phi)$, commun à toutes les fonctions de la famille, représentant le préfixe d'un programme décrivant la famille pour une machine de Turing, et d'un second terme $x(\beta_0, \ldots, \beta_m)$, décrivant les coefficients (tronqués à une certaine précision par exemple dans le cas de coefficients réels) de la décomposition de la fonction dans la base. Le processus qui permet de représenter de manière optimale les coefficients par des chaînes de

[4] On peut également définir la complexité d'une mesure de probabilité : la complexité d'une mesure de probabilité P sur un espace mesurable Ω est définie comme étant la complexité de la fonction réelle $\log 1/P$ restreinte à l'espace dénombrable $\pi = \bigcup_n \pi_n$ où π_n est une suite de partitions dénombrables qui génèrent Ω. La loi de probabilité sera dite calculable s'il existe un programme sur un ordinateur universel avec domaine prefix-free qui énumère récursivement les ensembles de la forme $\{x, a, g(x, a)\}$, où $g(x, a)$ est la fonction rationnelle la plus simple qui approxime P, pour tout $x \in \pi_n$ et pour toute précision a, c'est-à-dire $|\log 1/P(x) - g(x, a)| < 2^{-a}$.

[5] Ces notions seront détaillées dans la section 3.2

symboles sur le ruban d'une machine de Turing est appelé un codage ou une représentation des entiers. La complexité relative de toute fonction de cette famille se réduit alors simplement à la longueur d'un code pour un vecteur d'entiers. Ce point de vue sera explicité dans les sections 2.3.1 et 3.2.

Historique La notion de complexité de Kolmogorov est due chronologiquement à R. J. Solmonoff, à A. N. Kolmogorov et G. J. Chaitin. En effet, en 1960 R. J. Solomonoff publia un rapport de recherche sur une théorie générale de l'inférence inductive, présentant les idées de base de la complexité algorithmique comme moyen de résoudre certains problèmes résultant de l'application de la règle de Bayes en statistique. L'intérêt de Solomonoff était d'utiliser la complexité comme un auxiliaire servant à obtenir une probabilité a priori à des fins d'inférence statistique; les travaux de Kolmogorov par contre, se sont surtout concentrés sur la notion de complexité. Ce dernier propose la complexité descriptive d'un objet comme définition de l'information intrinsèque qu'il contient et comme une mesure de son caractère aléatoire. En 1965 il démontre le théorème d'invariance pour la complexité ainsi définie. En 1955 et 1956, Kolmogorov introduit la notion d'ϵ-entropie d'un ensemble de fonctions dans un espace métrique, qui l'amena plus tard au treizième problème de Hilbert. C'est dans la période des années 1950 que Kolmogorov a considéré une approche de la complexité à travers les fonctions récursives. En 1966 G. J. Chaitin propose le produit nombre d'états par nombre de symboles utilisés par une machine de Turing comme mesure de la complexité d'un algorithme, et plus tard (en 1969) démontre également un théorème d'invariance. Chacune de ces trois approches a contribué à élargir le champ d'application de la complexité algorithmique, mais ce sont incontestablement les travaux de Kolmogorov qui ont donné une telle portée à la notion de complexité descriptive. La notion de machine de Turing donne un modèle de calcul qui permet de caractériser l'ensemble des fonctions calculables. Toutes ces notions générales d'informatique théorique, ont été implémentées dans différents langages de programmation. Parmi les langages évolués, le Pascal peut servir de support à l'exposition de la théorie de la calculabilité, voir [1] par exemple. Cette théorie a eu de nombreuses applications, entre autres la simulation de machines déterministes et indéterministes, l'implémentation de la théorie des fonctions récursives, la dérécursification et les théorèmes du point fixe. La définition de la complexité algorithmique et la démonstration du théorème d'universalité sont dus à [45]. Des notions similaires apparaissent dans [71] et [20, 19].

La notion de machine de Turing fut introduite par [74]. Indépendamment, [57, 58] introduisit un modèle semblable d'automate. De nombreux autres modèles de machine ont été proposés, mais toutes les machines décrites se sont révélées imitables par une machine de Turing, voir par exemple [65], [52], [23], [66], [37], [21] (Théorème 2.2). Chaitin [22] (Théorème 1), établit l'existence d'une machine de Turing universelle dont le domaine est prefix-free, et l'universalité de la complexité ainsi définie. Les éléments de base de la théorie de la calculabilité peuvent être trouvés par exemple dans [66] et plus récemment [56].

2.2 Processus ponctuels

2.2.1 Processus ponctuels et martingales.

Un processus ponctuel est intuitivement une variable aléatoire dépendant du temps, et n'évoluant que par des sauts d'amplitude unité, le nombre de sauts étant presque-sûrement fini sur [0,T]. Nous pouvons en donner la définition formelle suivante : un processus ponctuel est un processus stochastique $(N(t))_{t\geq 0}$ adapté à une filtration $(\mathcal{F}_t)_{t\geq 0}$ avec $N(0) = 0$, $N(t) < \infty$ presque-sûrement, et dont les trajectoires sont presque-sûrement continues à droite, constantes par morceaux, de sauts +1. Soit un espace probabilisé $(\Omega, \mathcal{F}, P)$ sur lequel on définit une filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfaisant les "conditions habituelles", et un processus ponctuel $N = (N(t))_{0\leq t\leq T}$, adapté à $(\mathcal{F}_t)$.

Définition 2.5. (Processus vectoriel)
Un processus vectoriel k-dimensionnel $(N_1, \ldots, N_k)$ est appelé processus ponctuel multivarié si :

1. *Chaque N_j, $j = 1, \ldots, k$, est un processus ponctuel.*

2. *La probabilité que deux composantes sautent en même temps est nulle.*

Théorème 2.6. *Soit $N_1, \ldots, N_k$ un processus ponctuel multivarié. Il existe k processus croissants prévisibles $A_1, \ldots, A_k$ tels que $A_j(0) = 0$ p.s, $A_j(t) < \infty$ p.s. pour tout $t > 0$, et $M_j = N_j - A_j$ est une martingale locale, $j = 1, \ldots, k$. Si chaque A_j est un processus continu p.s. :*
1) $< M_j >= A_j$ est l'unique processus croissant prévisible, c'est-à-dire tel que $A_j(0) = 0$ p.s., $A_j(t) < \infty$ p.s., pour tout $t > 0$, et tel que $M_j^2 - A_j$ soit une martingale locale, $j = 1, \ldots, k$.

2) Si $i \neq j$, $< M_i, M_j > (t) = 0$ *p.s., i.e.* $M_i M_j$ *est une martingale locale.*

3) Si l'on suppose de plus :

H1 : $E(N_j(T)) < \infty$, $j = 1, \ldots, k$, *alors* M_j *est une martingale de carré intégrable, et* $M_j^2 - A_j$ *est une martingale pour* $j = 1, 2, \ldots, k$.

Une condition suffisante pour que le processus croissant A associé au processus ponctuel N dont les instants de sauts sont notés $\tau_1, \tau_2, \ldots$ soit à trajectoires p.s. continues est que l'hypothèse suivante soit vérifiée :

H2 : Les fonctions de répartition $P(\tau_{k+1} - \tau_k \leq t | \mathcal{F}_{\tau_k})$ sont des fonctions absolument continues de t.

2.2.2 Le modèle multiplicatif de Aalen .

Soit N un processus ponctuel sur $[0, T]$ de compensateur A par rapport à $(\mathcal{F}_t)$ et P. Sous des conditions satisfaites dans les situations les plus fréquentes, les trajectoires de A sont des fonctions absolument continues (voir propositions 3.1 et 3.2 de [13]). Il existe donc un processus adapté non-négatif λ à trajectoires continues à gauche, et possédant des limites à droite en tous points, tel que $A(t) = \int_0^t \lambda(u)du$. Le processus λ est appelé processus d'intensité (ou intensité stochastique) du processus ponctuel N par rapport à $((\mathcal{F}_t), P)$. Le lemme suivant explicite le lien entre λ et l'évolution de N conditionnellement au passé :

Lemme 2.7. (Aalen 1975).

Supposons que le processus d'intensité λ *soit majoré par une variable aléatoire intégrable, alors les relations suivantes sont vérifiées :*

a) $\lim_{h \to 0} h^{-1} E\left(N(t+h) - N(t) \mid \mathcal{F}_t\right) = \lambda(t+)$.

b) $\lim_{h \to 0} h^{-1} P\left(N(t+h) - N(t) = 1 \mid \mathcal{F}_t\right) = \lambda(t+)$.

Le modèle multiplicatif de Aalen est caractérisé par la donnée d'un couple (N, Y) sur $(\Omega, (\mathcal{F}_t), P_\alpha)$, où Y est un processus non-négatif, prévisible, continu à gauche, et N un processus ponctuel d'intensité stochastique :

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

En formulation abrégée, on dit que N est un processus de Aalen admettant α comme fonction d'intensité.

Exemple 2.2 : Modèle multiplicatif.

Soit un processus de naissance et de mort dont la taille de la population à l'instant t est $X(t)$. Soient $\lambda(t)$ et $\nu(t)$ les probabilités de naissance et de mort, et (N_λ, N_ν) le processus bivarié comptant le nombre de naissances et de morts dans la population. Soit $(\mathcal{F}_t)$ la tribu engendrée par $\{X(s), s \leq t\}$. Le processus (N_λ, N_ν) est un processus ponctuel d'intensité $(\lambda(t)X(t), \nu(t)X(t))$ par rapport à $(\mathcal{F}_t)$.

En situation d'échantillonage, on suppose qu'il existe une suite de couples (N_i, Y_i) sur $(\Omega, (\mathcal{F}_{i,t}), P_\alpha)$ qui sont i.i.d sous P_α pour toute fonction α sur $[0,T]$ satisfaisant les contraintes données plus haut, et tels que N_i a pour intensité stochastique relativement à $(\mathcal{F}_{i,t})$ et P_α :

$$\lambda_i(t) = \alpha(t)Y_i(t), \quad t \in [0,T].$$

Un choix naturel pour Ω consiste à prendre l'espace canonique $\prod_{i=1}^{\infty} E_i$, où les E_i sont les copies d'un même espace E. Les processus (N_i, Y_i) sont alors définis comme les applications projection $\Omega \to E_i$. Soient :

$$\overline{N}_n = \sum_{i=1}^{n} N_i, \quad \overline{Y}_n = \sum_{i=1}^{n} Y_i, \text{ et } \overline{\lambda}_n(t) = \sum_{i=1}^{n} \lambda_i(t) = \alpha(t)\overline{Y}_n(t).$$

Le processus $(\overline{N}_n, \overline{Y}_n)$ est un processus de Aalen avec pour intensité stochastique $\overline{\lambda}_n$ (par rapport à P_α et à la tribu engendrée par la réunion des $(\mathcal{F}_{i,t})$. Dans toute la suite P_0 désigne la probabilité P_α obtenue pour $\alpha \equiv 1$. Dans le cadre de la modélisation qui vient d'être décrite, en situation d'échantillon de taille n, P_α restreinte à l'information associée à cet échantillon est absolument continue par rapport à P_0. C'est ce que précise le résultat suivant (voir par exemple [39]).

Proposition 2.8. *Soit P_α la probabilité associée à la fonction d'intensité α, et P_0 la probabilité associée à $\alpha \equiv 1$:*

a) P_α est absolument continue par rapport à P_0 sur $\mathcal{F}_T^n = \bigvee_{i=1}^n \mathcal{F}_{i,T}$ la dérivée de Radon-Nikodym de P_α par rapport à P_0 s'écrit :

$$dP_\alpha/dP_0 = \exp\left(-\int_0^T (\alpha(s)-1)\overline{Y}_n(s)ds + \int_0^T \log(\alpha(s))d\overline{N}_n(s)\right), \tag{2.2}$$

b) Le couple $(\overline{N}_n, \overline{Y}_n)$ est une statistique exhaustive pour le paramètre α étant donné les observations $(N^n, Y^n) = ((N_1, Y_1), \ldots, (N_n, Y_n))$.

Dans toute la suite, on note $\mathcal{L}_\alpha(N^n, Y^n)$ la dérivée de Radon-Nikodym dP_α/dP_0. C'est la vraisemblance des observations (N^n, Y^n) pour la valeur α du paramètre fonctionnel.
Les notions introduites dans ce paragraphe peuvent être généralisées sans difficulté à une situation multivariée. On considère alors des couples (N, Y) où $N = (N^1, \ldots, N^k)$ est un processus ponctuel k-dimensionnel et $Y = (Y^1, \ldots, Y^k)$ un vecteur de k processus non-négatifs, continus à gauche et limités à droite. Chaque N^i est un processus de Aalen d'intensité $\lambda^i(t) = \alpha^i(t)Y^i(t)$. Nous montrons à la fin du chapitre 3 comment une telle modélisation multivariée pourrait être utilisée pour la description de l'évolution d'une image. Pour les notions générales relatives aux processus ponctuels introduites dans cette section, le lecteur intéressé pourra consulter par exemple [41].

2.3 Codage universel

L'estimateur de complexité minimale est obtenu par la minimisation d'un critère composé de deux termes. Le premier représente la complexité d'une fonction déterministe ou d'une mesure de probabilité, le second (celui auquel nous nous intéressons dans ce chapitre) décrit un code de Shannon pour les données, construit à partir de la fonction décrite dans la première partie. Ce chapitre présente les éléments de la théorie du codage qui justifient cette approche, ainsi que les principales définitions qui serviront dans la suite. La théorie mathématique de l'information a son origine dans les travaux de Shannon [67]. L'auteur modélise les systèmes de communication par des processus aléatoires. Il donne une théorie probabiliste de l'information, et utilise des résultats d'ergodicité afin d'obtenir des théorèmes de codage et des résultats d'optimalité dans les systèmes de communication idéaux. Des multiples aspects considérés par Shannon, nous retiendrons uniquement le codage d'une source à travers un canal non bruité. Une source est la donnée d'un espace de base, et d'un processus aléatoire à temps discret (X_n) à valeurs dans cet espace (dans le cadre des applications qui nous intéressent, l'espace sera $\mathbf{R}^k$, ou un espace euclidien k-dimensionnel). Par hypothèse, les symboles transmis à travers le canal ne peuvent être perdus. Dans l'étude des systèmes de compression des données, le canal est non bruité, au sens où le vecteur reçu $\hat{X}_n$ est identique au vecteur transmis X_n pour tout n. Une suite de longueurs de codes (et par abus de langage, de codes) est dite optimale si la longueur moyenne des mots de codes (relativement à la taille de l'échantillon) converge vers l'entropie. La

longueur de description excédant l'entropie est la redondance du code. Elle représente la longueur additionnelle due au fait que le codage n'est pas effectué avec la "vraie" distribution. Les codes universels sont des descriptions de données issues d'une source de loi inconnue, telles que les longueurs de codes soient asymptotiquement optimales pour une large classe de distributions de source candidates. Le codage universel de sources a été développé par Davisson [28]. Dans le cas de codes de longueurs variables, le meilleur code en moyenne est basé sur la distribution ayant généré les données : l'espérance de la longueur de ce code est égale à l'entropie, qui est une borne inférieure de l'espérance de la longueur de code pour les codes les plus largement utilisés dans le cadre qui nous intéresse, i.e les codes uniquement décodables dont nous donnons une définition plus loin (2.3.2). Davisson propose deux méthodes pour construire des codes universels. La première consiste à construire des codes de Shannon par rapport à des mélanges de distributions [30]). La seconde méthode est basée sur des descriptions en deux blocs. Le premier décrit une distribution estimée et le second, un code de Shannon basé sur la distribution estimée. Plutôt que coder les données avec une distribution P fixée par avance, déterminer l'estimateur $\hat{P}_n$ de la distribution permet de construire des codes adaptatifs. Dans [28] (théorème 7), l'auteur montre comment obtenir des codes universels pour une classe de sources stationnaires ergodiques à alphabet fini en utilisant un code à deux blocs basé sur des histogrammes conditionnels. Le codage universel est aussi traité par [25], [73], [32],[44], [50], [47], [29]. Les résultats du chapitre 3 montrent que l'opposé de la log-vraisemblance basé sur la loi de probabilité du processus ayant généré les données, représente la longueur d'un code universel pour les instants de sauts d'un processus ponctuel N, et la trajectoire d'un processus observé Y, tous discrétisés à une précision convenable.

2.3.1 Principe et définitions de base.

Soit $\mathcal{A}$ un ensemble fini qu'on appellera alphabet, et $\mathcal{A}^*$ l'ensemble des suites formées d'éléments de $\mathcal{A}$. Une suite quelconque d'éléments de $\mathcal{A}$ est appelée mot, et la longueur $l(s)$ d'un mot s, est le nombre de symboles de $\mathcal{A}$ contenus dans s. Soit X une variable aléatoire à valeurs dans $\mathcal{O}$, de loi P (discrète). Le couple $\mathcal{S} = (\mathcal{O}, P)$ est appelé une source. Une application injective f d'un ensemble dénombrable d'objets $\mathcal{O}$ vers $\mathcal{A}^*$ est appelée un code. Les nombres entiers $l(f(x)), x \in \mathcal{O}$ sont appelés longueurs de code des éléments de $\mathcal{O}$. Pour une source $\mathcal{S} = (\mathcal{O}, P)$

donnée, la longueur moyenne d'un code est définie par :

$$\overline{l}(f) = \sum_{x \in \mathcal{O}} P(x) \cdot l(f(x)).$$

Cette quantité représente l'espérance de la longueur de code associée au code f, pour la source $\mathcal{O}$. Le but du codage universel est de déterminer des fonctions (ou codes) f, produisant les codes les plus courts en moyenne. L'entropie d'une source $\mathcal{S} = (\mathcal{O}, P)$ est définie par :

$$H(P) = -\sum_{x \in \mathcal{O}} P(x) \log_b(P(x)),$$

où b est le cardinal de l'alphabet $\mathcal{A}$, et $\log_b$ le logarithme en base b.

Exemple 2.3 : Premier exemple de codage de source.

Soit $\mathcal{A} = \{0, 1\}$, l'ensemble $\mathcal{A}^*$ représente l'ensemble des chaînes binaires. L'ensemble $\mathcal{O}$ peut être pris égal à N par exemple. A toute variable aléatoire X, à valeurs dans N de loi de probabilité P, peut être associée une source $\mathcal{S} = (\mathrm{N}, P)$. Un procédé de codage de N, sera défini pour cette source, par une application injective de N vers $\{0, 1\}^*$, associant à tout entier naturel une chaîne binaire. Un exemple de procédé de codage, est pour tout $n \in \mathrm{N}$, la représentation binaire de n. Ce procédé associe à tout entier n, un mot de code de longueur $\lceil \log_2(n) \rceil + 1$.

Exemple 2.4 : Second exemple de codage de source.

Supposons que $\mathcal{O}$ soit un ensemble dénombrable de lois de probabilités associées à des processus ponctuels observés sur $[0, T]$, dont la fonction d'intensité α, appartient à une famille paramétrique de dimension m variable, et supposons que les coefficients de la décomposition de la fonction sur la base sont tronqués à une précision maximale δ. L'ensemble $\mathcal{A}$ peut être pris égal à $\{0, 1\}$, et $\mathcal{A}^*$, l'ensemble des représentations binaires d'entiers. Soit P une probabilité sur $\mathcal{O}$ on a donc une source $(\mathcal{O}, P)$. Un exemple de procédé de codage pour les éléments de $\mathcal{O}$, consiste à associer à toute loi de probabilité, de fonction d'intensité $\alpha(s) = \sum_{j=1}^{m} \alpha_j \phi_j(s)$, $s \in [0, T]$, le vecteur $\overline{\alpha} = (m, \delta, \overline{\alpha}_1, \ldots, \overline{\alpha}_m)$, où $\overline{\alpha}_j, j = 1, \ldots, m$ est la représentation binaire de l'entier $2^\delta \alpha_j$. Un mot de code pour $\overline{\alpha}$ peut être obtenu par la concaténation de mots de codes pour des entiers. Pour que le décodage soit possible, les différentes parties du code doivent pouvoir être distinguées sans avoir recours à un séparateur (du type virgule ou espace). Un tel code associerait donc à toute loi de probabilité P_α, un mot de code de longueur $\sum_{j=1}^{m} q(2^\delta \alpha_j) + q(m) + q(\delta)$, où q est la longueur d'un code sur les entiers, autorisant un décodage sans séparateur.

Un code est dit à longueurs variables si les mots de code peuvent avoir des longueurs différentes. Quand la distribution de probabilité P de la source n'est pas uniforme, le codage de longueur variable est préférable à celui de longueur fixe, car intuitivement, les meilleurs codes en moyenne seront ceux qui affectent les mots de code les plus longs aux symboles les moins probables. Néanmoins les codes de longueurs variables peuvent présenter un inconvénient, la longueur des mots de code transmis n'est pas connue a priori, et c'est un problème lorsqu'on effectue le décodage. L'ajout d'un séparateur indiquant au décodeur qu'un mot de code se termine, ne résoud pas le problème. En effet, l'introduction d'un séparateur après chaque mot sur un message de longueur n augmente la longueur totale d'au moins n positions binaires. La notion de décodabilité unique permet de résoudre ce problème. Un code est uniquement décodable, si un message ne peut pas être interprété de deux manières différentes. Deux messages codés identiques, ont nécessairement le même nombre de mots de code, et peuvent être identifiés mot à mot. Un code uniquement décodable peut néanmoins nécessiter la réception de tout le message avant de pouvoir être décodé, de tels codes ne sont pas instantanés. La classe des codes instantanés peut facilement être caractérisée par une propriété, celle d'être prefix-free. Si $c = x_1 \dots x_n$ est un mot de code, alors pour tout k, tel que $1 \leq k < n$, le mot $x_1 \dots x_k$ n'est pas un mot de code (cf. définition 2.2). Un code est instantané ssi il est prefix-free. On résume ces notions dans la définition suivante.

Définition 2.9. (Code uniquement décodable. Code instantané.)

a) *Un code C est uniquement décodable si toute suite d'éléments de l'alphabet correspond au plus à un message.*

b) *Un code est dit instantané si pour tout message reçu, chaque mot de code peut être décodé dès qu'il est reçu.*

Remarque : Tout code instantané est uniquement décodable, mais un code uniquement décodable n'est pas nécessairement instantané.

◇

Le théorème suivant de Kraft [46] donne une condition nécessaire et suffisante pour l'existence d'un code instantané. Un code b-aire, un code dont l'alphabet est constitué des chiffres compris entre 0 et b.

Théorème 2.10.

1) Si C est un code instantané b-aire de longueurs de code $l_1, \ldots, l_n$, alors ces longueurs doivent satisfaire l'inégalité de Kraft : $\sum_{i=1}^{n} b^{-l_i} \leq 1$.

2) Si les nombres $l_1 \ldots l_n$ vérifient l'inégalité de Kraft, alors il existe un code instantané b-aire dont les mots de code ont pour longueur $l_1, \ldots, l_n$.

Le théorème de Kraft énonce que si les longueurs $l_1, l_2, \ldots, l_n$ vérifient l'inégalité de Kraft alors il existe un code instantané avec ces longueurs de mots de code. C'est un théorème d'existence. Le théorème n'énonce pas que tout code dont les longueurs des mots de code vérifient cette propriété est instantané. L'inégalité de Kraft est aussi une condition nécessaire et suffisante pour qu'il existe un code uniquement décodable. La condition est suffisante car tout code instantané est uniquement décodable. [55] démontre que cette inégalité est aussi une condition nécessaire.

Théorème 2.11. *Si $C = \{c_1, c_2, \ldots, c_n\}$ est un code b-aire uniquement décodable, alors les longueurs des mots de code $l_1, \ldots, l_n$ doivent satisfaire l'inégalité de Kraft : $\sum_{i=1}^{n} b^{-l_i} \leq 1$.*

L'emploi conjugué des théorèmes de Kraft et de McMillan fournit le résultat suivant :

Théorème 2.12. *S'il existe un code uniquement décodable dont les longueurs des mots de code sont $l_1, \ldots, l_n$, alors il existe aussi un code instantané dont les mots de code ont les mêmes longueurs.*

En effet, s'il existe un code uniquement décodable dont les longueurs sont $l_1, \ldots, l_n$, alors par le théorème de McMillan, ces longueurs doivent satisfaire l'inégalité de Kraft. Si les longueurs vérifient l'inégalité de Kraft, alors par le 2) du théorème 2.10, il existe un code instantané b-aire dont les mots de code ont pour longueurs $l_1, \ldots, l_n$.

2.3.2 Codes optimaux en moyenne.

Le résultat fondamental du codage dans un canal non-bruité énoncé par Shannon, est que la longueur moyenne minimale du code dans une famille de codes uniquement décodables est minorée par l'entropie $H(P)$ de la source. De plus on a le résultat suivant :

Théorème 2.13. *Il existe un code instantané dont la longueur moyenne vérifie*

$$H(P) \leq L \leq H(P) + 1.$$

Le résultat précédent suggère alors de définir comme mesure de l'efficacité d'un code le rapport entre sa longueur moyenne et l'entropie de la source. Ce critère est à la base du principe d'universalité, qui peut être formulé de la manière suivante :

Définition 2.14. *Un code f d'un ensemble d'objets $\mathcal{O}$ vers un ensemble de descriptions $\mathcal{A}^*$ est universel, s'il existe une constante c, telle que*

$$\frac{\sum_{x \in \mathcal{O}} P(x) l(f(x))}{\max\{H(P), 1\}} \leq c,$$

pour toute distribution P telle que $0 < H(P) < \infty$.

Dans [68], l'auteur propose un code affectant à tout élément x de l'espace des objets un élément de $\mathcal{A}^*$ de longueur $\lceil -\log P(x) \rceil$, où P est la loi de probabilité de la source. Ce code vérifie clairement la propriété d'universalité, et l'inégalité de Kraft. Le code de Shannon peut être construit de la manière suivante. Soit une source à n éléments $\mathcal{S} = (\mathcal{O}, P)$, et un arrangement de ces mots correspondant à des probabilités décroissantes $p_1, \ldots, p_n$. Soit $P_r = \sum_{i=1}^{r-1} p_i$, pour $r = 1, \ldots, n$. Le code binaire de l'ensemble des n éléments vers $\{0,1\}^*$ est obtenu en codant r par un nombre binaire $E(r)$ obtenu en tronquant le développement binaire de P_r à la longueur $l(E(r))$ de manière à obtenir :

$$-\log p_r \leq l(E(r)) \leq 1 - \log p_r.$$

Le code ainsi obtenu, est le code de Shannon. Il a la propriété que les éléments hautement probables de la source sont codés avec des mots de codes courts, et les éléments les moins probables, avec les mots de codes les plus longs. De plus, on a :

$$2^{-l(E(r))} \leq p_r < 2^{-l(E(r))+1}.$$

Le code de l'élément d'indice r diffère des codes des éléments d'indices $r+1$ à travers au moins une position binaire, puisque pour tout i tel que $r+1 \leq i \leq n$, on a :

$$P_i \geq P_r + 2^{-l(E(r))},$$

et ainsi les chiffres du développement binaire de P_r et P_i ne peuvent être identiques sur l'ensemble des $l(E(r))$ premières positions. La fonction E est donc uniquement décodable. Aucune valeur de E n'étant le préfixe d'une autre valeur, l'ensemble des mots de codes est un ensemble prefix-free. Le message source peut être reconstitué séquentiellement. En posant $L = \sum_{r=1}^{n} p_r l(E(r))$ la longueur moyenne du code et $H(P) = -\sum_r p_r \log p_r$. Ce qui précède fournit les inégalités suivantes :

$$-\sum_r p_r \log p_r \leq \sum_r p_r l(E(r)) \leq \sum_r p_r (1 - \log p_r) = 1 - \sum_r p_r \log p_r,$$

ce qui montre que

$$H(P) < L < 1 + H(P).$$

Chapitre 3

Compression optimale des données.

On observe sur un intervalle de temps $[0, T]$, la réalisation de n processus ponctuels, de même loi, d'intensité stochastique $\lambda(s)$, $s \in [0, T]$. A partir d'une partition κ_m de l'intervalle de temps $[0, T]$, il est possible de construire une partition π_m de l'espace des réalisations du couple (N, Y), basée sur κ_m (voir section 3.1). Sur les atomes de cette partition, un code à deux parties peut être défini. La première est un code pour une fonction d'intensité, et la seconde est un code de Shannon pour les données observées, basé sur la fonction d'intensité correspondante. Sous des hypothèses convenables, toute loi de probabilité P_β associée à un modèle de Aalen admet une dérivée de Radon-Nikodym par rapport à une loi donnée P_0. Pour une précision suffisamment fine, l'opposé de la log-vraisemblance calculée sur les données discrétisées, coincide avec la longueur du code de Shannon à un terme près ne dépendant que de la mesure de base P_0. Cette approximation permet de considérer l'opposé de la log-vraisemblance comme la longueur à une constante près d'un code de Shannon pour les données discrétisées, et justifie le fait de définir la longueur du code en terme de vraisemblance, et non de probabilités sur les atomes de la partition (lemme 3.2). Les fonctions candidates sont supposées appartenir à un espace fonctionnel donné ou à une suite croissante d'espaces. L'application qui associe à toute fonction la longueur de son code est appelée fonction de complexité. Le choix des fonctions de complexité est soumis à des contraintes liées à la décodabilité unique du message transmis. Ces fonctions de complexité seront de deux types : algorithmiques, ou descriptionnelles. La complexité algorithmique mesure le contenu en information d'un objet individuel, et représente la longueur minimale d'un code pour cet objet. La complexité descriptionnelle par contre, ne cherche pas à décrire complètement

l'objet, mais cherche à l'identifier au sein d'un ensemble auquel il appartient. Les fonctions de complexité seront définies soit pour des fonctions admettant une décomposition sur une base, soit dans un espace de fonctions de cardinal fini. Dans le premier cas, le codage des coefficients de la décomposition de la fonction dans la base considérée suffira à la décrire complètement. Dans le second cas, (purement théorique) une complexité égale au log du cardinal de l'ensemble peut être affectée à chaque fonction, ce qui revient à en donner l'indice au sein de l'ensemble. Si la famille considérée fait elle-même partie d'un ensemble de familles, il suffira d'indicer chaque famille pour décrire complètement les fonctions de l'ensemble de toutes les familles.

Modèles de familles candidates.

Dans toute la suite la fonction d'intensité α associée à la loi P_α du processus appartient à l'espace L_2 des fonctions de carré intégrable sur $[0,T]$. Soit $\Gamma^{(n)}$ une famille disjointe de sous-ensembles dénombrables de L_2, et $\Gamma_n = \Gamma_{n-1} \cup \Gamma^{(n)}$ avec $\Gamma_0 = \Gamma^{(0)}$. La suite $(\Gamma_n)_n$ est une suite croissante d'ensembles, et on notera $\Gamma = \cup_{n\geq 0}\Gamma^{(n)}$ sa limite.

Cas paramétrique et semi-paramétrique. Dans tout ce paragraphe, $m(n)$ est considéré comme constant dans le cas paramétrique, et comme une suite entière croissante dans le cas semi-paramétrique. Soient $\phi_1, \ldots, \phi_{m(n)}$, une famille de fonctions orthonormées qui engendrent un espace $\mathcal{E}_m$ contenu dans $L_2([0,T])$; nous noterons $\widetilde{\Gamma}_n$ le sous-ensemble des fonctions de $\mathcal{E}_m$ positives sur $[0,T]$. L'ensemble Γ_n est alors par définition, l'ensemble obtenu en tronquant les coefficients des fonctions de $\widetilde{\Gamma}_n$, à une précision δ_n. Dans l'espace $\widetilde{\Gamma}_n$, on définit la projection de α par :

$$\tilde{\alpha}_n = \arg\min_{\beta\in\widetilde{\Gamma}_n} \| \alpha - \beta \|_2 .$$

Par continuité de la norme L_2, le minimum sur $\widetilde{\Gamma}_n$ de $\| \alpha - \beta \|_2$, est également le minimum sur le sous-ensemble fermé $\mathcal{B}_n = \{\beta \in \widetilde{\Gamma}_n \ :\| \alpha - \beta \|_2 \leq K\}$, pour une constante K convenablement choisie. Cet ensemble étant convexe et fermé, l'existence et l'unicité du minimum sont assurés. L'ensemble Γ_n est dénombrable. Tout voisinage de $\tilde{\alpha}_n$ au sens de la distance induite par la norme L_2 contient donc un ensemble fini de fonctions de Γ_n. L'application qui à β associe $\| \alpha - \beta \|_2$, admet dans cet ensemble un minimum qui n'est pas nécessairement unique et qui sera noté α_n^*.

Par définition :

$$\alpha_n^* = \arg\min_{\beta \in \Gamma_n} \| \alpha - \beta \|_2 .$$

Remarque : S'il existe n_0 tel que $\alpha \in \Gamma_{n_0}$, alors pour tout $n \geq n_0$ on aura également $\alpha_n^* \equiv \tilde{\alpha}_n \equiv \alpha$.

◇

Cas totalement non-paramétrique. Dans le cas totalement non-paramétrique, on considère que la fonction d'intensité est bornée inférieurement, et qu'elle appartient à $W_{2,+}^r = \{\beta : \beta(s) \geq a^{-1}, \| \beta^{(k)} \|_2 < \infty, k = 1, 2, \ldots, r\}$, avec $a > 0$. Dans ce cas, on utilise l'ε-entropie H_ε d'un ensemble de fonctions W, qui est le log du cardinal du plus petit réseau de fonctions $\tilde{\gamma}$ tel que pour toute fonction $\beta \in W$ il existe une fonction $\tilde{\gamma}$ satisfaisant $\| \beta - \tilde{\gamma} \|_\infty < \varepsilon$. Des résultats de [12] montrent que pour tout ε suffisamment petit, l'ε-entropie de la boule de Sobolev est bornée par $c(1/\varepsilon)^{1/r}$ où c dépend uniquement de a et de r. Une suite ε_n décroissante permet de définir une suite $\Gamma_n(\varepsilon_n)$ de réseaux de $W_{2,+}^r$. La complexité $C_n(\cdot)$ est par définition le logarithme du cardinal de $\Gamma_n(\varepsilon_n)$. Si $\alpha \in W_{2,+}^r$, par définition il existe au moins une fonction $\tilde{\gamma}_n \in \Gamma_n(\varepsilon_n)$ telle que $\| \alpha - \tilde{\gamma}_n \|_\infty < \varepsilon_n$.

Définition 3.1. *Dans le cas totalement non-paramétrique, on notera*

$$\tilde{\alpha}_n \equiv \alpha_n^*,$$

une fonction arbitraire de $\Gamma_n(\varepsilon_n)$ contenue dans la boule de centre α et de rayon ε_n.

On suppose, et ce sera le cas dans les situations étudiées, qu'en cas de non-unicité on dispose d'un procédé permettant de définir $\tilde{\alpha}_n$ de manière unique.
Dans la suite on pourra être conduits à supposer vérifiées certaines des conditions ci-dessous.

Conditions sur Γ_n et C_n.

G1 Il existe $0 < x < 1$, et $b > 0$ tel que les fonctions de complexité satisfont l'inégalité :

$$\sum_{\beta \in \Gamma_n} 2^{-xC_n(\beta)} \leq b. \tag{3.1}$$

G2 Les fonctions de complexité satisfont

$$C_n = o(n). \tag{3.2}$$

G3 Soit $\alpha_n^* = \arg\min_{\beta \in \Gamma_n} \| \alpha - \beta \|_2$, alors :

$$\lim_{n \to \infty} \| \alpha - \alpha_n^* \|_\infty = 0. \tag{3.3}$$

G4 Il existe une suite croissante a_n et deux constantes f_0 et f_1 telles que pour toute fonction $\beta \in \Gamma_n$, on ait l'inégalité

$$f_0 a_n^{-1} \leq \beta \leq f_1 a_n. \tag{3.4}$$

De même, certains énoncés feront éventuellement appel aux hypothèses suivantes.

Hypothèses sur le modèle.

H3 Le processus Y est borné P_0-presque-sûrement par une constante D_Y.

$$P_0(\sup_s Y(s) \leq D_Y) = 1. \tag{3.5}$$

H4 Il existe deux constantes θ_0 et θ_1 telles que :

$$0 < \theta_0 \leq E_\alpha(Y(s)) = \theta_\alpha(s) \leq \theta_1. \tag{3.6}$$

Remarque : Les conditions $G1$ à $G4$ portent uniquement sur les fonctions candidates et les fonctions de complexité. Il est aisé de déterminer en pratique des longueurs de codes et des ensembles de fonctions candidates qui les vérifient. Les hypothèses $H3$ et $H4$ sont vérifiées par exemple dans le cas de processus de Poisson avec $\theta_0 = \theta_1 = 1$, et $D_Y = 1$.

◇

3.1 Approximation de la longueur du code de Shannon.

Soit $\beta \in \Gamma_n$ une fonction d'intensité, P_β la probabilité associée et (N, Y) un processus sur $(\Omega, (\mathcal{F}_t))$ muni de la probabilité P_β. Si l'on définit une partition finie ou dénombrable $\pi = (A_i, i \in J)$ de l'espace des réalisations du processus (N, Y), on peut associer à (N, Y) un processus de loi discrète $(P_\beta(A_i), i \in J)$, et donc un code de Shannon.

En situation d'échantillonage, nous montrons sous certaines conditions qu'il est possible de construire une suite de partitions $(\pi_{m(n)})$ associées à des précisions $(\delta_{m(n)})$ de plus en plus fines des observations sur $[0, T]$ telles que la log-vraisemblance $\log \mathcal{L}_\beta(N^n, Y^n)$ se comporte asymptotiquement comme $\log P_\beta(A) - \log P_0(A)$, où A est l'élément de la partition $\pi_{m(n)}$ contenant l'observation (N^n, Y^n). Clairement ce résultat tend à justifier la démarche consistant à définir la longueur de description de l'observation (N^n, Y^n) sous la probabilité P_β par le logarithme de l'inverse de la vraisemblance du processus d'origine.

Dans la suite, nous nous restreignons à une classe de processus de la forme :

$$Y(s) = \sum_{j=0}^{\tilde{y}} e_j \mathbf{I}_{B_j}(s),$$

avec $B_0 = [0, t_1[$, $B_{\tilde{y}} = [t_{\tilde{y}}, T]$, et pour $0 < j < \tilde{y}$, $B_j = [t_j, t_{j+1}[$. $\tilde{y}$ est le nombre de changements d'états de Y. Nous supposons que Y prend ses valeurs dans un ensemble E_Y de cardinal fini. Les processus de Poisson sont dans cette classe avec $\tilde{y} = 0$, $Card(E_Y) = 1$, et $e_0 = 1$. Il en est de même pour les processus associés aux durées censurées, on a alors $\tilde{y} \in \{0, 1\}$. Nous notons $\tau_1, \ldots, \tau_{\tilde{n}}$ les dates de saut du processus N.

L'approximation $(\widetilde{N}, \widetilde{Y})$

Etant donné une précision $\delta_m = 2^{-m}$, nous notons $(\widetilde{N}, \widetilde{Y})$, le couple (N, Y) vu à la précision δ_m. De manière précise, soit la partition de $[0, T]$ en intervalles I_j de longueurs $1/2^m$ définis par :

$$I_j = [jT/2^m, (j+1)T/2^m[, \text{ pour } j = 0, \ldots, 2^m - 2,$$

et $I_{2^m-1} = [(2^m - 1)T/2^m, T]$. Nous notons σ_Y^m (resp. σ_N^m) l'application de $\overline{1, \tilde{y}_n} = \{1, \ldots, \tilde{y}_n\}$ (resp. de $\overline{1, \tilde{n}_n}$) dans $\overline{0, 2^m - 1}$ qui, à chaque date de changement d'état de Y (resp. de N) associe l'indice de l'intervalle I_j où il se produit.

Soient $\tilde{t}_i$ et $\tilde{\tau}_i$ les dates t_i et τ_i tronquées à la précision δ_m. $\widetilde{Y}$ et $\widetilde{N}$ sont alors définis par

$$\widetilde{Y}(s) = \begin{cases} Y(\tilde{t}_j) \text{ si } s \in I_j \text{ et } j \in \sigma_Y^m \\ Y(s) \text{ sinon }, \end{cases}$$

et

$$\tilde{N}(s) = \begin{cases} N(\tilde{t}_j) \text{ si } s \in I_j \text{ et } j \in \sigma_N^m \\ N(s) \text{ sinon.} \end{cases}$$

Notons que si $(N^n, Y^n) = ((N_1, Y_1), \ldots, (N_n, Y_n))$ est un échantillon, et $(\tilde{N}^n, \tilde{Y}^n) = ((\tilde{N}_1, \tilde{Y}_1), \ldots, (\tilde{N}_n, \tilde{Y}_n))$ l'échantillon observé à la précision δ_m, alors $\widetilde{\overline{N}_n} = \overline{\tilde{N}_n}$ et $\widetilde{\overline{Y}_n} = \overline{\tilde{Y}_n}$. Nous notons $\mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n)$ la fonction

$$\exp\left(-\int_0^T (\beta(s)-1)\overline{\tilde{Y}_n}(s)ds + \int_0^T \ln\beta(s) d\overline{\tilde{N}_n}(s)\right).$$

La partition π_m associée à δ_m.

La partition π_m est définie comme la famille des ensembles de la forme

$$A = A(\tilde{n}_n, \tilde{y}_n, e_0, \ldots, e_{\tilde{y}_n}, \sigma_Y^m, \sigma_N^m). \tag{3.7}$$

(N^n, Y^n) appartient à A si : $\overline{N}_n([0,T]) = \tilde{n}$, les points de sauts du processus cumulé $\overline{N}_n$, sont dans les intervalles $I_{\sigma_N^m(j)}$, pour $j = 1, \ldots, \tilde{n}_n$, $\overline{Y}_n$ parcourt les états $e_0, \ldots, e_{\tilde{y}_n}$, et les changements d'état de $\overline{Y}_n$ ont lieu dans les intervalles $I_{\sigma_Y^m}(j)$. Clairement les ensembles A sont mesurables, et constituent une partition de l'espace des réalisations de (N_n, Y_n). Le résultat que nous démontrons ci-dessous fera appel à la condition suivante :

G5 : Il existe C telle que pour tout n et tout $\beta \in \Gamma_n$ on ait :

$$\forall x, y \in [0,T], |\beta(x) - \beta(y)| \leq C|x-y|. \tag{3.8}$$

Utilisant $1 - 1/a \leq \ln(a) \leq a - 1$ avec $a = \beta(x)/\beta(y)$, on vérifie facilement que G4 et G5 impliquent

$$\forall x, y \in [0,T], |\ln\beta(x) - \ln\beta(y)| \leq Ca_n|x-y|. \tag{3.9}$$

Lemme 3.2. *Supposons les conditions G4 et G5 satisfaites. Soit $(\delta_{m(n)}) = 2^{-m(n)}$ une suite de précisions dyadiques, $(N^n, Y^n) = ((N_1, Y_1), \ldots, (N_n, Y_n))$ un échantillon du couple (N, Y) et $(\tilde{N}^n, \tilde{Y}^n)$ le vecteur des approximations de (N^n, Y^n) à la précision $\delta_{m(n)}$. On a P_β presque-sûrement*

$$\left|\log 1/\mathcal{L}_\beta(N^n, Y^n) - \log 1/\mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n)\right| \leq Ca_n\delta_{m(n)}(\tilde{y}_n + \tilde{n}_n),$$

où $\tilde{y}_n$ est le nombre de changements d'état de $\overline{Y}_n$ et $\tilde{n}_n = \overline{N}_n([0,T])$.

Preuve : Avec des notations évidentes, le processus cumulé $\overline{Y}_n$ peut s'écrire sous la forme

$$\overline{Y}_n(s) = \sum_{j=0}^{\tilde{y}_n} e_j^n \mathbf{I}_{B_j^n}(s),$$

et $\overline{\widetilde{Y}}_n$ sous la forme

$$\overline{\widetilde{Y}}_n(s) = \sum_{j=0}^{\tilde{y}_n} e_j^n \mathbf{I}_{\widetilde{B}_j^n}(s),$$

où $\widetilde{B}_j^n = [\tilde{t}_j, \tilde{t}_{j+1}[$ si $B_j^n = [t_j, t_{j+1}[$. On a alors

$$\frac{\mathcal{L}_\beta(N^n, Y^n)}{\mathcal{L}_\beta(\widetilde{N}^n, \widetilde{Y}^n)} = \exp\left(- \int_0^T (\beta(s) - 1)(\overline{Y}_n(s) - \overline{\widetilde{Y}}_n(s))ds + \int_0^T \ln \beta(s) d(\overline{N}_n(s) - \overline{\widetilde{N}}_n(s)) \right).$$

Pour le premier terme dans l'exponentielle, on peut écrire :

$$\begin{aligned}
\left| \int_0^T (\beta(s) - 1)(\overline{Y}_n(s) - \overline{\widetilde{Y}}_n(s))ds \right| &\leq \sum_{j=0}^{\tilde{y}_n} \int_0^T |1 - \beta(s)| \left| \mathbf{I}_{B_j^n}(s) - \mathbf{I}_{\widetilde{B}_j^n}(s) \right| ds \\
&\leq \sup_s |1 - \beta(s)| \sum_{j=0}^{\tilde{y}_n} \int_0^T \left| \mathbf{I}_{B_j^n}(s) - \mathbf{I}_{\widetilde{B}_j^n}(s) \right| ds \\
&\leq 2 \sup_s |1 - \beta(s)| \, \delta_m (1 + \tilde{y}_n) \\
&\leq C a_n \delta_m \tilde{y}_n \qquad (3.10)
\end{aligned}$$

Pour le second terme, utilisant la condition G5 on a :

$$\begin{aligned}
\left| \int_0^T \ln \beta(s) d(\overline{N}_n(s) - \overline{\widetilde{N}}_n(s)) \right| &\leq \sum_{j=1}^{\tilde{n}_n} |\ln \beta(\tilde{\tau}_j) - \ln \beta(\tau_j)| \\
&\leq C a_n \sum_{j=1}^{\tilde{n}_n} |\tilde{\tau}_j - \tau_j| \\
&\leq C a_n \delta_m \tilde{n}_n \qquad (3.11)
\end{aligned}$$

Par 3.10 et 3.11, on obtient :

$$\log \frac{\mathcal{L}_\beta(N^n, Y^n)}{\mathcal{L}_\beta(\widetilde{N}^n, \widetilde{Y}^n)} \leq C a_n \delta_{m(n)} (\tilde{y}_n + \tilde{n}_n),$$

ce qui termine la démonstration.

□

Le lemme ci-dessus est important pour le contrôle de l'erreur lors du codage d'observations mais est aussi déterminant dans la démonstration du résultat énoncé ci-dessous.

Proposition 3.3. *Supposons les conditions G4 et G5 satisfaites. Soit* $(\delta_{m(n)}) = 2^{-m(n)}$ *une suite de précisions dyadiques,* $(\pi_{m(n)})$ *la suite de partitions de* Ω *associées,* $(N^n, Y^n) = ((N_1, Y_1), \ldots, (N_n, Y_n))$ *un échantillon du couple* (N, Y) *et* A *l'élément de la partition* $\pi_{m(n)}$ *contenant* (N^n, Y^n). *On a* P_β *presque-sûrement*

$$|\log 1/P_\beta(A) - \log 1/P_0(A) - \log 1/\mathcal{L}_\beta(N^n, Y^n))| \leq C a_n \delta_{m(n)}(\tilde{y}_n + \tilde{n}_n), \tag{3.12}$$

où $\tilde{y}_n$ *est le nombre de changements d'état de* $\overline{Y}_n$ *et* $\tilde{n}_n = \overline{N}_n([0,T])$.
Si $a_n = O(n^q)$, $q > 0$, $\tilde{n}_n = O(n)$ *presque-sûrement, alors* $n^{1+q} = o(2^{m(n)})$ *assure la convergence vers 0 de l'erreur d'approximation ci-dessus.*

Preuve : Observons que

$$\begin{aligned} |\log 1/P_\beta(A) - \log 1/P_0(A) \quad - \quad & \log 1/\mathcal{L}_\beta(N^n, Y^n))| \leq \\ & \left|\log 1/P_\beta(A) - \log 1/P_0(A) - \log 1/\mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n))\right| \\ + \quad & \left|\log 1/\mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n)) - \log 1/\mathcal{L}_\beta(N^n, Y^n))\right|. \end{aligned} \tag{3.13}$$

Le lemme (3.2) nous a donné une majoration pour le 2ème terme de la partie droite de l'inégalité. Considérons le 1er terme.

$$\begin{aligned} P_\beta(A) &= \int_A \mathcal{L}_\beta(N^n, Y^n) dP_0 \\ &= \mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n) \int_A \exp\left(-\int_0^T (\beta(s) - 1)(\overline{Y}_n(s) - \overline{\tilde{Y}}_n(s)) ds\right) \\ &\cdot \exp\left(\int_0^T \ln \beta(s) d(\overline{N}_n(s) - \overline{\tilde{N}}_n(s))\right) dP_0, \end{aligned}$$

car $\mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n)$ est constante sur A. En reprenant le raisonnement utilisé dans la démonstration du lemme 3.2, on obtient que l'exponentielle figurant dans l'intégrale est majorée sur A par $C a_n \delta_{m(n)}(\tilde{y}_n + \tilde{n}_n)$. Il s'ensuit

$$P_\beta(A) \leq \mathcal{L}_\beta(\tilde{N}^n, \tilde{Y}^n) \exp\left(C a_n \delta_{m(n)}(\tilde{y}_n + \tilde{n}_n)\right) P_0(A).$$

Prenant le logarithme, utilisant l'inégalité 3.13 et faisant à nouveau appel au lemme 3.2 on obtient le résultat.

□

Remarque : Dans le cas des processus associés à des modèles de durées, card(E_Y)=2, $\tilde{y}_n \leq n$ et $\overline{N}_n[0,T] \leq n$ presque-sûrement. Si la famille Γ_n vérifie G4 avec $a_n = n^q$, la convergence presque-sûre vers 0 de l'erreur d'approximation (3.12) est assurée pour $n^{1+q} = o(2^{m(n)})$.

◇

Remarque : Il en est de même dans le cas des processus de Poisson. En effet si N est un processus de Poisson de fonction d'intensité élément de Γ_n, il découle de la loi du logarithme itéré que $\tilde{n}_n = \overline{N}_n([0,T])$ se comporte en $O(n)$ presque-sûrement.

◇

Remarque : Le terme $-\log \mathcal{L}_\beta(\widetilde{N}^n, \widetilde{Y}^n)$ est à une constante près (indépendante de β), la longueur d'un code de Shannon pour les réalisations des processus discrétisés $\widetilde{N}^n$ et $\widetilde{Y}^n$, tandis que l'opposé de la log-vraisemblance $-\log \mathcal{L}_\beta(N^n, Y^n)$, à partir de laquelle nous obtiendrons tous les résultats asymptotiques, ne peut pas en toute rigueur admettre une telle interprétation, car il n'existe pas de code de longueur finie pour les réalisations du couple (N^n, Y^n).

◇

Le théorème de Shannon (théorème du codage dans un canal non-bruité) montre que l'entropie est une borne inférieure de l'espérance de la longueur de tout code uniquement décodable. La longueur du code de Shannon a, par construction, une espérance égale à un bit près à l'entropie de la source, cette longueur est donc quasiment optimale dans la classe des codes uniquement décodables.

Proposition 3.4. *Soit $\alpha \in \Gamma_n$, $\pi_{m(n)}$ une partition de l'espace Ω et A_n l'élément aléatoire défini par (N,Y). On a :*

1. *La longueur moyenne de tout code uniquement décodable $l(\cdot)$ défini sur la partition $\pi_{m(n)}$ est supérieure à la longueur moyenne du code de Shannon basé sur α :*

$$E_\alpha \left(l(A_n) - \log 1/P_\alpha(A_n) \right) \geq 0.$$

2. *Soit u_n une suite telle que $\sum_n 2^{-u_n} < \infty$, ($u_n = \log(n) + 2\log\log(n)$ par exemple). P_α presque-sûrement, pour n suffisamment grand :*

$$l(A_n) \geq \log 1/P_\alpha(A_n) - u_n.$$

Preuve : Soit $Q(A_n) = 2^{-l(A_n)}$ pour tout $A_n \in \pi_{m(n)}$. Par hypothèse sur l, Q définit une sous-probabilité. En utilisant les propriétés du logarithme, on a :

$$\begin{aligned} E_\alpha\left[\log\left(\frac{P_\alpha(A_n)}{Q(A_n)}\right)\right] &= \sum_{A_n\in\pi_m^n} P_\alpha(A_n)\log\left(\frac{P_\alpha(A_n)}{Q(A_n)}\right) \\ &\geq \sum_{A_n\in\pi_m^n} P_\alpha(A_n)\left(1-\frac{Q(A_n)}{P_\alpha(A_n)}\right)\frac{1}{\ln 2} \\ &= \left(1-\sum_{A_n\in\pi_m^n} Q(A_n)\right)\frac{1}{\ln 2} \\ &\geq 0, \end{aligned}$$

d'où la conclusion. Pour montrer le second résultat, il suffit de remarquer, que l'on a, en utilisant l'inégalité de Markov :

$$\begin{aligned} P_\alpha\left(l(A_n) \leq -\log P_\alpha(A_n) - u_n\right) &= P_\alpha\left(Q(A_n)2^{-u_n} \geq P_\alpha(A_n)\right) \\ &\leq 2^{-u_n}. \end{aligned}$$

Par hypothèse cette borne est le terme général d'une série convergente, et le lemme de Borel-Cantelli permet de conclure.

□

Les résultats des propositions 3.4 et 3.3 permettent de conclure que $\log 1/\mathcal{L}_\alpha$ constitue à une constante près en moyenne, mais également P_α presque-sûrement à une suite (de l'ordre de $\log(n)$) près, une borne inférieure pour la longueur du code dans la famille des codes uniquement décodables définis sur $\pi_{m(n)}$. D'où l'optimalité de ce principe de codage.

3.2 Codage variable de réalisations de processus ponctuels

En général, la fonction d'intensité α de la loi ayant généré les données est inconnue. Intuitivement, un code basé sur un estimateur de cette fonction aura des longueurs sensiblement égales à celles du code basé sur α. Un tel procédé de codage nécessite également le codage de la fonction d'intensité estimée. A l'issue de la section 2.3.1, on a vu qu'un code pour une fonction admettant une décomposition suivant une base pouvait se ramener à un code pour ses coefficients. Le prochain paragraphe détaille la définition des longueurs de codes pour différents types de familles candidates.

Fonctions de complexité.

Dans la section 2.1, nous avons montré que le programme binaire $x(\beta)$ pour une fonction β admettant une décomposition sur une base (ϕ) est la concaténation de deux chaînes :

$$x(\beta) = x(\phi) + x(\beta_0, \ldots, \beta_m),$$

où $x(\phi)$ est un programme pour les fonctions de base, et $x(\beta_0, \ldots, \beta_m)$, un programme pour les coefficients de la fonction β[1]. Si les coefficients varient dans un intervalle borné, le codage binaire classique suffit à les représenter, mais dans le cas de coefficients variant librement dans N, ce codage ne satisfait plus l'inégalité de Kraft. Déterminer une représentation pour des entiers naturels qui satisfait l'inégalité de Kraft, équivaut à trouver un code uniquement décodable pour une source dénombrable. On peut résumer en disant que le premier terme $x(\phi)$ du programme calculant β correspond à la théorie exposée dans la section 2.1, et le second, $x(\beta_0, \ldots, \beta_m)$, correspond à la théorie de la section 2.3 pour une source dénombrable d'entiers. Certains auteurs ont étudié les performances des codes universels pour des sources dénombrables d'entiers. Dans [32], l'auteur construit des codes universels pour des entiers issus d'une source, sans connaissance a priori de la loi de probabilité de la source. Plus précisément, étant donné un processus à valeur dans l'ensemble des entiers naturels, on cherche à déterminer une suite de longueurs de codes $(L^*(1), L^*(2), \ldots)$, qui satisfait l'inégalité de Kraft, et qui est solution du problème minimax :

$$\min_{L} \sup_{P \in \mathcal{P}} \sum_{i \in \mathrm{N}} [P(i)L(i)]/H(P), \tag{3.14}$$

où $\mathcal{P}$ est un ensemble de probabilités.

Définition 3.5. *(Suite de longueurs optimale.) Soit $P = (P(1), P(2), \ldots)$ une probabilité définie sur N^*, qui vérifie :*

i) $P(i) < 1$ pour tout $i \in \mathrm{N}$, et il existe M tel que $P(i) \geq P(i+1)$ pour tout $i > M$.

ii) $H(P) = -\sum P(i) \log P(i) = \infty$,

[1]Un tel programme consiste en fait à recopier le vecteur des coefficients inscrit sur des rubans de la machine, (faisant office de mémoire) sur le ruban de sortie. Sa longueur est à une constante additive près, égale à la longueur de la représentation du vecteur des coefficients

et $L = (L(1), L(2), \ldots)$ *une suite de longueurs d'entiers qui vérifie l'inégalité de Kraft. Toute solution du problème*

$$\min_{L} \sup_{P} \lim_{N\to\infty} \left(\frac{\sum_{i=1}^{N} P(i)L(i)}{-\sum_{i=1}^{N} P(i)\log P(i)} \right), \tag{3.15}$$

est appelée une suite de longueurs optimales.

Exemple 3.1 : Longueur de code optimale.

Soit $\log^*(x) = \log(x) + \log\log(x) + \ldots$, où la somme ne comprend que des termes positifs. [49] montrent que $\sum_k 2^{-\log^*(k)} = c$, où $c \approx 2.865064$. Toute suite de longueurs telle que $L(k) \geq \log^*(k) + \log c$, satisfait donc l'inégalité de Kraft, de plus, de telles suites sont optimales au sens de la définition 3.5 (pour une preuve, voir par exemple [63]).

Remarque : La vitesse de croissance des suites optimales est bornée, en effet [63] théorème 2 montre que toute suite optimale L vérifie :

$$\log(k) \leq L(k) \leq \log(k) + r(k),$$

où $r(k)/\log(k) \to 0$, et $r(k) \to \infty$ quand $k \to \infty$.

◇

Dans la suite, ces suites de longueurs optimales, (ou codes optimaux) seront notées $C(\cdot)$, ou $C_n(\cdot)$ si le code est autorisé à varier avec la taille de l'échantillon.

Dans le cas totalement non-paramétrique, le plus petit ε-réseau associé à l'ensemble, fournit un ensemble fini de fonctions, qui peut donc être indexé. Un code de longueurs inférieures ou égales au logarithme du cardinal de l'ensemble permet ainsi d'en identifier toutes les fonctions. Cette dernière complexité est descriptionnelle, et a des liens avec l'approche métrique de l'estimation développée entre autres par Lecam, Birgé et Yatracos. Pour des familles de fonctions paramétriques, ou des suites de familles paramétriques dont la dimension croît avec la taille de l'échantillon, la complexité d'une fonction se décompose en la somme de plusieurs termes, dont l'un, (codant les coefficients, tronqués à une certaine précision, de la décomposition de la fonction sur une base) représente la longueur d'un code universel pour un vecteur d'entiers, et les autres, décrivent un ensemble de paramètres liés à l'espace considéré (type de famille, dimension de l'espace, précision, etc ...).

1. **Cas paramétrique.** α appartient à une partie d'un sous-espace isomorphe à $\mathbf{R}^m$. On peut définir pour tout $\beta \in \Gamma_n$:

$$C_n(\beta) = c_0 + C(\delta_n) - m \log \delta_n + C([\beta]), \tag{3.16}$$

où $C([\beta])$ est la longueur d'un code optimal pour la partie entière des coefficients de la décomposition de β sur la base (ϕ), c_0 est la longueur d'un code décrivant la base et $C(\delta_n)$ la longueur d'un code pour la précision. Dans cet exemple, le code utilisé pour la partie fractionnaire du vecteur des coefficients est autorisé à croître avec la taille de l'échantillon, puisqu'on autorise une précision dépendant de cette taille. Les coefficients des fonctions de Γ_n sont tronqués à une précision $\delta_n = 2^{-d_n}$ où d_n est une suite croissante qui représente le nombre de positions binaires dans le développement fractionnaire du coefficient.

2. **Cas semi-paramétrique.** Dans ce cas, nous supposons que la fonction α peut être approximée au sens de la distance L_∞ sur l'intervalle $[0, T]$ par des fonctions de la suite d'espaces $(\Gamma_n)_n$ (i.e. condition G3). Pour tout $\beta \in \Gamma_n$ on définit :

$$C_n(\beta) = c_0 + C(\delta_n) + C(m(n)) - m(n) \log \delta_n + C([\beta]), \tag{3.17}$$

où $C([\beta])$, c_0 et $C(\delta_n)$ sont définis de manière identique, $C(m(n))$ représentant la longueur d'un code universel pour la dimension.

3. **Cas totalement non-paramétrique.** α est recherchée dans une partie finie d'un espace de Sobolev. La complexité affectée à chaque fonction est de l'ordre du log du cardinal de l'ensemble. Cette complexité, purement descriptionnelle, revient à donner l'indice de la fonction au sein de l'ensemble. Elle a pour expression :

$$C_n(\beta) = c(1/\varepsilon_n)^{1/r}, \tag{3.18}$$

où c dépend des éléments a et de r intervenant dans la définition de $W^r_{2,+} = \{\beta : \beta(s) \geq a^{-1}, \| \beta^{(k)} \|_2 < \infty, k = 1, 2, \ldots, r\}$.

Quand aucune spécification ne sera nécessaire, $C_n(\cdot)$ désignera l'une quelconque de ces complexités.

3.2.1 Bornes supérieures de la redondance

L'équivalence à une constante additive près entre la vraisemblance et les probabilités calculées sur les atomes d'une partition, permet de déduire des résultats d'optimalité en moyenne pour les longueurs $(-\log \mathcal{L}_\beta)$, qui sont approximativement les longueurs d'un code de Shannon pour des réalisations de processus ponctuels. Après avoir spécifié les familles de fonctions candidates, nous rappelons les résultats d'optimalité du code de Shannon basé sur la loi P_α du processus ayant généré les données, et nous montrons, que même quand la fonction d'intensité associée à cette loi, ne peut pas être décrite par un code de longueur finie, il existe des codes obtenus à partir d'une suite de fonctions de Γ_n, dont la redondance converge vers zéro à une vitesse dépendant des propriétés d'approximation de la famille de fonctions candidates.

Définition 3.6. (Complexité.) *La complexité $\mathcal{C}(N^n, Y^n)$ des données (N^n, Y^n) relativement aux fonctions de complexité $C_n(\cdot)$ et aux ensembles Γ_n est définie par*

$$\mathcal{C}(N^n, Y^n) = \min_{\beta \in \Gamma_n} \left(C_n(\beta) + \log 1/\mathcal{L}_\beta(N^n, Y^n) \right) \tag{3.19}$$

où $\mathcal{L}_\beta$ est la vraisemblance obtenue pour la probabilité associée à la fonction β.

Remarque : Etant donné n observations du processus (N, Y), la complexité stochastique est une mesure de la qualité de la meilleure compression possible des données relativement aux fonctions C_n et à Γ_n. Cette complexité sera d'autant plus faible que la qualité de l'approximation de α par des fonctions de Γ_n sera bonne et que le taux de croissance des fonctions C_n sera faible.

◇

L'espérance de la différence entre la longueur du code élaboré à partir de la fonction ayant généré les données et un code basé sur une fonction d'intensité β quelconque (la redondance du code), sera également utilisée comme mesure de l'erreur d'estimation. Bien que n'étant pas une distance, cette quantité est proche de la distance L_2 et lui est localement équivalente. Elle est équivalente à la distance L_2 dans le cas de fonctions bornées (voir lemme 4.7), et elle domine la distance L_1.

Définition 3.7. *La distance $d(\alpha \parallel \beta)$ entre deux fonctions d'intensité α et β déduite de la distance de Kullback-Leibler des vraisemblances associées est :*

$$d(\alpha \parallel \beta) = E_\alpha \left[\log \left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta} \right) \right] \tag{3.20}$$

Un calcul direct permet de montrer que :

Proposition 3.8. *La distance de Kullback-Leibler $d(\alpha \parallel \beta)$ admet l'expression*

$$d(\alpha \parallel \beta) = \frac{1}{\ln 2} \int_0^T \left[\alpha(s) \log \frac{\alpha(s)}{\beta(s)} + \beta(s) - \alpha(s) \right] \theta_\alpha(s) ds,$$

où $\theta_\alpha(s) = E_\alpha(Y(s))$.

Preuve :

$$\log\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right) = \frac{1}{\ln 2} \int_0^T (\alpha(s) - \beta(s)) Y(s) ds + \frac{1}{\ln 2} \int_0^T \log\left(\frac{\alpha(s)}{\beta(s)}\right) dN(s).$$

En utilisant $dN(s) = \alpha(s)Y(s)ds + dM(s)$ et $E_\alpha\left(dN(s)\right) = \alpha(s)E_\alpha(Y(s))ds$, on obtient :

$$\begin{aligned} E_\alpha\left[\log\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right)\right] &= \frac{1}{\ln 2} \int_0^T (\beta(s) - \alpha(s)) E_\alpha\left(Y(s)\right) ds + \frac{1}{\ln 2} \int_0^T \log\left(\frac{\alpha(s)}{\beta(s)}\right) E_\alpha\left(dN(s)\right) \\ &= \frac{1}{\ln 2} \int_0^T (\beta(s) - \alpha(s)) \theta_\alpha(s) ds + \frac{1}{\ln 2} \int_0^T \log\left(\frac{\alpha(s)}{\beta(s)}\right) \alpha(s) \theta_\alpha(s) ds \\ &= \frac{1}{\ln 2} \int_0^T \left[\alpha(s) \log \frac{\alpha(s)}{\beta(s)} + \beta(s) - \alpha(s) \right] \theta_\alpha(s) ds \end{aligned}$$

□

L'ensemble des fonctions pouvant être approchées par des fonctions de Γ, au sens de l'entropie, est la fermeture au sens de l'information que nous noterons $\overline{\Gamma}$.

Définition 3.9. *La fermeture $\overline{\Gamma}$ de Γ est l'ensemble des fonctions α pour lesquelles*

$$\inf_{\beta \in \Gamma} d(\alpha \parallel \beta) = 0.$$

Définition 3.10. *Soit un échantillon de taille n d'un processus de Aalen (N, Y) sur $(\Omega, (\mathcal{F}_t))$ muni de la probabilité P_α. La redondance $R_n(\alpha \parallel \beta)$ du code basé sur la fonction β est définie par*

$$R_n(\alpha \parallel \beta) = \frac{1}{n} E_\alpha \left[\left(\log(1/\mathcal{L}_\beta(N^n, Y^n)) + C_n(\beta) \right) - \log\left(1/\mathcal{L}_\alpha(N^n, Y^n)\right) \right].$$

Il découle immédiatement de la définition ci-dessus que la redondance $R_n(\alpha \parallel \beta)$ vérifie

$$R_n(\alpha \parallel \beta) = d(\alpha \parallel \beta) + \frac{1}{n} C_n(\beta).$$

La redondance minimale parmi les codes basés sur des fonctions d'un ensemble Γ_n est une quantité qui jouera un rôle fondamental dans la suite : l'indice de résolvabilité.

Définition 3.11. *La redondance minimale entre le code basé sur α et les codes variables basés sur des fonctions de Γ_n, est appelée l'indice de résolvabilité. Elle a pour expression :*

$$R_n(\alpha) = \min_{\beta \in \Gamma_n} \left(d(\alpha \parallel \beta) + \frac{1}{n} C_n(\beta) \right).$$

Remarque : Si la condition G2 est vérifiée, une condition nécessaire et suffisante pour que l'indice de résolvabilité $R_n(\alpha)$ converge vers zéro est que α appartienne à la fermeture $\overline{\Gamma}$.

◇

La proposition suivante donne une condition suffisante pour que la complexité des fonctions α_n^* demeure bornée, dans le cas de suites de familles paramétriques.

Proposition 3.12. *Dans le cadre de suites de familles paramétriques, si la famille $(\phi_1, \ldots, \phi_{m(n)})$ est orthonormale, et $C_n(\cdot) = o(n)$, alors $C_n(\alpha_n^*) = o(n)$.*

Preuve : Par définition, $C_n(\beta) = c_0 + C([\beta]) - m(n) \log \delta_n$. Par hypothèse, pour $\beta \in \Gamma$ fixé, on a $C_n(\beta) = o(n)$. Cette relation implique que la partie constante de l'expression de la complexité, $c_0 - m(n) \log \delta_n$ se comporte également en $o(n)$. Pour montrer que $C_n(\alpha_n^*) = o(n)$, il suffit alors de montrer que $C([\alpha_n^*]) = o(n)$. Soit w une suite de longueurs optimale (cf. définition 3.5) telle que $C([\beta]) = \sum_{j=1}^{m(n)} w([\beta_j])$. On rappelle que w satisfait la contrainte $w(k) < \log k + r(k)$, où $r(k) = o(\log k)$. Soit $\tilde{\alpha}_n$, l'approximation de α dans $\tilde{\Gamma}_n$. On a :

$$\begin{aligned} C([\alpha_n^*]) &= \sum_{j=1}^{m(n)} w([\alpha_{nj}^*]) \\ &= \sum_{j=1}^{m(n)} w([\tilde{\alpha}_{n,j}]) \text{ car } [\tilde{\alpha}_{n,j}] = [\alpha_{nj}^*] \\ &\leq \sum_{j=1}^{m(n)} ([\tilde{\alpha}_{n,j}])^2 \end{aligned}$$

$$\begin{aligned} &\leq \sum_{j=1}^{m(n)} (\tilde{\alpha}_{n,j})^2 \\ &= \| \tilde{\alpha}_n \|_2^2 \\ &\leq \| \alpha - \tilde{\alpha}_n \|_2^2 + \| \alpha \|_2^2 \end{aligned}$$

Par définition de $\tilde{\alpha}_n$, $\| \alpha - \tilde{\alpha}_n \|_2$ converge vers 0. Pour n suffisamment grand, on a donc :

$$C([\alpha_n^*]) \leq 2 \| \alpha \|_2^2 .$$

Cette relation montre que :

$$C_n(\alpha_n^*) \leq c_0 + 2 \| \alpha \|_2^2 - m(n) \log \delta_n .$$

□

On rappelle que dans les cas paramétriques et semi-paramétriques, par définition,

$$\tilde{\alpha}_n = \arg \min_{\beta \in \widetilde{\Gamma}_n} \| \alpha - \beta \|_2 .$$

Dans le cas totalement non-paramétrique $\tilde{\alpha}_n = \alpha_n^*$ désigne une fonction arbitraire du réseau prise dans la boule de centre α et de rayon ε_n, on pose $\mu_n = \| \alpha - \tilde{\alpha}_n \|_2$.

Proposition 3.13. *Supposons que α et la famille $\widetilde{\Gamma}_n$ satisfont la condition G3, et que le modèle vérifie l'hypothèse H3. Soit δ_n la précision à laquelle les fonctions de Γ_n sont considérées dans les cas paramétriques et semi-paramétriques, et soit μ_n l'erreur d'approximation de α par $\tilde{\alpha}_n$. La redondance du code optimal parmi les codes basés sur les fonctions de Γ_n vérifie :*

a) *Dans le cas paramétrique :*

$$R_n(\alpha) = O\left(\delta_n^2\right) + O\left(-n^{-1} \log \delta_n\right). \tag{3.21}$$

b) *Dans le cas semi-paramétrique :*

$$R_n(\alpha) = O\left(\mu_n^2 + m(n)\delta_n^2\right) + O\left(-m(n)n^{-1} \log \delta_n\right). \tag{3.22}$$

c) *Dans le cas totalement non-paramétrique :*

$$R_n(\alpha) = O\left(n^{-1}{\varepsilon_n}^{-1/r}\right) + O(\mu_n^2), \tag{3.23}$$

où ε_n est le rayon des boules définissant Γ_n.

Preuve : La distance de Kullback entre la fonction α et sa meilleure approximation dans Γ_n, est voisine du carré de la distance L_2 entre les deux fonctions.

$$\begin{aligned} d(\alpha \parallel \alpha_n^*) &= \int_0^T \left[\alpha(s) \log \frac{\alpha(s)}{\alpha_n^*(s)} + \alpha_n^*(s) - \alpha(s)\right] \theta_\alpha(s)\, ds \\ &\leq \int_0^T \left[\alpha(s) \left(\frac{\alpha(s)}{\alpha_n^*(s)} - 1\right) + \alpha_n^*(s) - \alpha(s)\right] \theta_\alpha(s)\, ds \\ &= \int_0^T (\alpha(s) - \alpha_n^*(s))^2 \frac{1}{\alpha_n^*(s)} \theta_\alpha(s)\, ds \\ &\leq \sup_{0 \leq s \leq T} \left(\frac{\theta_\alpha(s)}{\alpha_n^*(s)}\right) \int_0^T (\alpha(s) - \alpha_n^*(s))^2\, ds, \end{aligned}$$

où la première inégalité est une conséquence de $\log(x) \leq x - 1$, appliquée à $x = \alpha(s)/\alpha_n^*(s)$. Par la condition G3, pour tout $\varepsilon < \alpha_0 = inf_s \alpha(s)$, il existe $n(\varepsilon)$ tel que $\alpha_n^* > \alpha_0 - \varepsilon$. Pour un tel ε et pour n suffisamment grand, on a $\sup_s \alpha_n^*(s)^{-1} \leq (\alpha_0 - \varepsilon)^{-1}$. Il existe donc par l'hypothèse 3.6, une constante K_1 telle que :

$$d(\alpha \parallel \alpha_n^*) \leq K_1 \int (\alpha(s) - \alpha_n^*(s))^2\, ds.$$

Par définition :

$$R_n(\alpha) = \min_{\beta \in \Gamma_n} \left(d(\alpha \parallel \beta) + \frac{1}{n} C_n(\beta)\right) \leq d(\alpha \parallel \alpha_n^*) + \frac{1}{n} C_n(\alpha_n^*).$$

En tenant compte de la relation entre la distance de Kullback et la norme L_2, on obtient :

$$R_n(\alpha) \leq K_1 \parallel \alpha - \alpha_n^* \parallel_2^2 + \frac{1}{n} C_n(\alpha_n^*) \leq K_2 \mu_n^2 + K_3 m(n) \delta_n^2 + \frac{1}{n} C_n(\alpha_n^*),$$

où K_2 et K_3 sont également des constantes. Dans le cas paramétrique, m est fixé, et $\mu_n = 0$, on obtient ainsi :

$$\begin{aligned} R_n(\alpha) &= O\left(\delta_n^2\right) + O\left(n^{-1} C_n(\alpha_n^*)\right) \\ &= O\left(\delta_n^2\right) + O\left(-n^{-1} \log \delta_n\right) \end{aligned}$$

Dans le cas semi-paramétrique on obtient :

$$\begin{aligned} R_n(\alpha) &= O\left(\mu_n^2 + m(n)\delta_n^2\right) + O\left(n^{-1} C_n(\alpha_n^*)\right) \\ &= O\left(\mu_n^2 + m(n)\delta_n^2\right) + O\left(-\frac{m(n)}{n} \log \delta_n\right) \end{aligned}$$

Dans le cas totalement non-paramétrique :

$$R_n(\alpha) \leq \frac{1}{n} H_{\varepsilon_n} + \mu_n^2 \leq \frac{1}{n} c(1/\varepsilon_n)^{1/r} + T\varepsilon_n^2$$

$$R_n(\alpha) = O\left(n^{-1}{\varepsilon_n}^{-1/r}\right) + O(\varepsilon_n^2) \tag{3.24}$$

□

Les bornes précédentes permettent de déterminer une précision pour les coefficients des fonctions de base dans les cas paramétrique et semi-paramétrique, ainsi qu'un rayon pour les boules du réseau dans le cas totalement non-paramétrique. Ces deux quantités ne sont pas optimales compte tenu du fait que la minimisation s'effectue sur des majorants de l'indice de résolvabilité.

Proposition 3.14. *Sous les conditions de la proposition précédente, les vitesses de convergence de l'indice de résolvabilité $R_n(\alpha)$ fournies dans les cas paramétrique et semi-paramétrique, sont optimisées pour une suite de précisions (δ_n^*) satisfaisant :*

$$\delta_n^* = O\left(n^{-\frac{1}{2}}\right).$$

Cette suite de précisions permet d'obtenir dans le cas paramétrique :

$$R_n(\alpha) = O\left(\frac{\log n}{n}\right), \tag{3.25}$$

et dans le cas des suites de familles paramétriques :

$$R_n(\alpha) = O\left(\mu_n^2 + m(n)\frac{\log n}{n}\right). \tag{3.26}$$

Dans le cas totalement non-paramétrique, le rayon optimal des boules de Sobolev est :

$$\varepsilon_n^* = O(n^{-\frac{r}{2r+1}}).$$

Ce rayon permet d'obtenir la borne optimale :

$$R_n(\alpha) = O\left(n^{\frac{-2r}{2r+1}}\right). \tag{3.27}$$

Preuve : Cas paramétrique.

Dans le cas paramétrique, m est fixé, et $\mu_n = 0$, on obtient ainsi :

$$R_n(\alpha) = O\left(\delta_n^2\right) + O\left(-\frac{m}{n}\log\delta_n\right).$$

En dérivant par rapport à δ_n dans le membre de droite, on obtient à l'optimum :

$$2\delta_n^* - \frac{m}{n\delta_n^*} = 0.$$

Soit

$$\delta_n^* = O\left(n^{-\frac{1}{2}}\right).$$

On obtient alors :

$$\begin{aligned} R_n(\alpha) &= O\left(n^{-1}\right) + O\left(\frac{1}{n}\log(n)\right) \\ &= O\left(\frac{\log n}{n}\right) \end{aligned}$$

Cas semi-paramétrique.

$$R_n(\alpha) = O\left(\mu_n^2 + m(n)\delta_n^2\right) - O\left(\frac{m(n)}{n}\log\delta_n\right).$$

δ_n^* est solution de :

$$m(n)\delta_n - \frac{m(n)}{n\delta_n} = 0.$$

On obtient également comme précision optimale :

$$\delta_n^* = O\left(n^{-\frac{1}{2}}\right).$$

La résolvabilité correspondante satisfait :

$$\begin{aligned} R_n(\alpha) &= O\left(\mu_n^2 + m(n)n^{-1}\right) + O\left(m(n)\frac{\log(n)}{n}\right) && (3.28) \\ &= O\left(\mu_n^2 + m(n)\frac{\log(n)}{n}\right) && (3.29) \end{aligned}$$

Cas totalement non-paramétrique.

$$R_n(\alpha) = O\left(T\varepsilon_n^2 + \frac{1}{n}c(1/\varepsilon_n)^{1/r}\right).$$

En dérivant par rapport à ε, l'optimum est solution de :

$$-\frac{1}{n}\varepsilon_n^{-\frac{1}{r}-1} + 2\varepsilon_n = 0.$$

Le minimum est obtenu pour :

$$\varepsilon_n^* = O\left(n^{-\frac{r}{2r+1}}\right).$$

La résolvabilité vérifie alors

$$\begin{aligned} R_n(\alpha) &= O\left(\varepsilon_n^2 + \frac{1}{n}c(1/\varepsilon_n)^{1/r}\right) \\ &= O\left(n^{-\frac{2r}{2r+1}}\right) + O\left(\frac{1}{n}n^{-\frac{1}{2r+1}}\right) \\ &= O\left(n^{\frac{-2r}{2r+1}}\right) \end{aligned}$$

□

3.2.2 Optimalité presque-sûre.

Le paragraphe précédent étudiait la moyenne de la longueur de code additionnelle, dans la description des observations, due à l'utilisation de fonctions de Γ_n à la place de la fonction α. On s'intéresse maintenant au comportement, trajectoire à trajectoire, de cette longueur additionnelle. Le lemme suivant est la base de la démonstration du résultat principal mais sera également utile au chapitre 4 dans l'étude des vitesses de convergence.

Lemme 3.15. *Si la condition G3 et l'hypothèse H3 sont vérifiées, alors l'espérance du rapport entre la vraisemblance basée sur la loi ayant généré les données, et la vraisemblance basée sur sa meilleure approximation dans Γ_n vérifie :*

$$E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right) \leq \exp\left(Knv_n^2\right),$$

où $v_n = \| \alpha - \alpha_n^* \|_2$ et K est une constante dépendant de α et D_Y.

Preuve : L'espérance du rapport admet l'écriture :

$$\begin{aligned} E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right) &= \int \exp\left(\int_0^T (\alpha_n^*(s) - 2\alpha(s) + 1)\overline{Y}_n(s)ds + \int_0^T \log\frac{\alpha^2}{\alpha_n^*}(s)d\overline{N}_n(s)\right) dP_0^n \\ &= \int \exp\left(\int_0^T (\alpha_n^*(s) - 2\alpha(s))\overline{Y}_n(s)ds\right) \\ &\cdot \exp\left(\int_0^T \frac{\alpha^2}{\alpha_n^*}(s)\overline{Y}_n(s)ds - \int_0^T \left(\frac{\alpha^2}{\alpha_n^*}(s) - 1\right)\overline{Y}_n(s)ds\right) \\ &\cdot \exp\left(\int_0^T \log\frac{\alpha^2}{\alpha_n^*}(s)d\overline{N}_n(s)\right) dP_0^n \end{aligned}$$

La fonction $\gamma_n \equiv \alpha^2/\alpha_n^*$ est strictement positive et bornée par hypothèse sur Γ_n et α. Soit $\mathcal{L}_{\gamma_n}$, la vraisemblance basée sur γ_n.

$$\begin{aligned} E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right) &= \int \mathcal{L}_{\gamma_n}(N^n, Y^n) \exp\left(\int_0^T \left(\alpha_n^*(s) - 2\alpha(s) + \frac{\alpha^2}{\alpha_n^*}(s)\right) \overline{Y}_n(s) ds\right) dP_0 \\ &\leq \int \mathcal{L}_{\gamma_n}(N^n, Y^n) \exp\left(\frac{1}{\inf_s \alpha_n^*} \int_0^T (\alpha(s) - \alpha_n^*(s))^2 \overline{Y}_n(s) ds\right) dP_0 \end{aligned}$$

Par la condition $G3$, pour tout ε on peut trouver n suffisamment grand tel que $\inf_s \alpha_n^*(s) > \inf_s \alpha(s) - \varepsilon = \alpha_0 - \varepsilon > 0$. En utilisant $\overline{Y}_n(s) \leq nD_Y$, et en posant $v_n = \| \alpha - \alpha_n^* \|_2$, et $K = \frac{D_Y}{\alpha_0 - \epsilon}$ on obtient :

$$E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right) \leq \exp\left(Knv_n^2\right) \int \mathcal{L}_{\gamma_n}(N^n, Y^n) dP_0 = \exp\left(Knv_n^2\right).$$

□

Définition 3.16. *On appelle redondance ponctuelle la quantité :*

$$\frac{1}{n}\left[\log 1/\mathcal{L}_\beta(N^n, Y^n) + C_n(\beta) - \log 1/\mathcal{L}_\alpha(N^n, Y^n)\right].$$

Cette quantité est la redondance de code observée pour une réalisation donnée, ramenée à la taille de l'échantillon. On peut également la considérer comme un estimateur empirique de la redondance. La proposition 3.18 permet de majorer cette quantité en fonction de la vitesse d'approximation de α dans la famille des fonctions candidates, et du taux de croissance des fonctions de complexité. Le lemme ci-dessous donne la vitesse de la redondance ponctuelle de la meilleure approximation de la fonction α dans Γ_n.

Lemme 3.17. *Soit α_n^* la meilleure approximation de α dans Γ_n, et $v_n = \| \alpha - \alpha_n^* \|_2$. Si la famille Γ_n vérifie la condition G3 et le processus Y l'hypothèse H3, alors P_α presque-sûrement, l'inégalité suivante est vérifiée pour n suffisamment grand :*

$$\frac{1}{n}\left(\log 1/\mathcal{L}_{\alpha_n^*} + C_n(\alpha_n^*) - \log 1/\mathcal{L}_\alpha\right) \leq \varepsilon_n,$$

avec $\varepsilon_n \approx \frac{C_n(\alpha_n^*)}{n} + (1+\mu)\frac{\log n}{n} + v_n^2$, et $\mu > 0$ une constante qui peut être choisie arbitrairement petite.

Preuve : Soit $0 < \varepsilon$ et A_n définis par $A_n = \{2^{-C_n(\alpha_n^*)}\mathcal{L}_{\alpha_n^*}(N^n, Y^n) \geq \mathcal{L}_\alpha(N^n, Y^n)2^{-n\varepsilon_n}\}$. D'après l'inégalité de Markov,

$$P_\alpha(\overline{A_n}) \leq E_\alpha(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}})2^{C_n(\alpha_n^*)-n\varepsilon_n}.$$

D'après le lemme 3.15, on a

$$E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right) \leq \exp\left(Knv_n^2\right)\int \mathcal{L}_{\gamma_n}(N^n, Y^n)dP_0 = \exp\left(Knv_n^2\right),$$

où $v_n = \| \alpha - \alpha_n^* \|_2$, d'où,

$$P_\alpha(\overline{A_n}) \leq 2^{C_n(\alpha_n^*)-n\varepsilon_n}2^{Knv_n^2} \tag{3.30}$$

Par le lemme de Borel-Cantelli, une condition suffisante pour que A_n soit vérifié presque-sûrement pour n suffisamment grand est que :

$$2^{C_n(\alpha_n^*)-n\varepsilon_n+Knv_n^2} \approx \frac{1}{n^{1+\mu}}, \tag{3.31}$$

pour $\mu > 0$, où $f(n) \approx g(n)$ signifie qu'il existe $0 < K_1 < K_2 < \infty$ tels que pour n suffisamment grand, $K_1 \leq f(n)/g(n) \leq K_2$.

□

Pour établir la proposition qui suit, il suffit de remarquer que la longueur du code minimal dans Γ_n est par définition inférieure à celle du code basé sur α_n^*. La borne obtenue dans le lemme 3.17 est donc également valable pour ce code.

Proposition 3.18. *La redondance ponctuelle du code de longueur minimale parmi les codes basés sur les fonctions de Γ_n satisfait P_α-presque-sûrement :*

$$\begin{aligned}\min_{\beta\in\Gamma_n}\frac{1}{n}[\log 1/\mathcal{L}_\beta(N^n, Y^n) \;+\; & C_n(\beta) - \log 1/\mathcal{L}_\alpha(N^n, Y^n)] \\ = \;& O\left(n^{-1}C_n(\alpha_n^*)\right) + O\left(v_n^2\right) + O\left(\frac{\log n}{n}\right).\end{aligned}$$

Ce résultat est une conséquence immédiate du lemme 3.17. La déviation de la redondance ponctuelle par rapport à la redondance $R_n(\alpha)$ se comporte donc en $O(n^{-1}\log n)$.

3.2.3 Chaîne de Markov avec censure pour la compression d'images numériques.

Dans cette section, nous montrons brièvement comment le principe de description minimale d'observations de processus ponctuels pourrait s'appliquer à la modélisation d'une image numérique observée sur un intervalle de temps $[0, T]$. On considère pour cette application, un modèle de chaîne de Markov avec censure.

Soit $E = \{1, \ldots, m\}$ l'espace fini des états d'une chaîne de Markov à temps continu. Soit $\alpha^{ij}(t)$ la probabilité de transition infinitésimale entre les états i et j. On suppose que toutes les fonctions α^{ij} sont des fonctions de L_2, strictement positives et continues à gauche. Soit X un processus de Markov à espace d'états E. Soit $N^{ij}(\cdot)$ le processus comptant le nombre de transitions directes entre les états i et j, et $Y^i(t) = 1$ si $X(t) = i$, 0 sinon. On suppose pour $i = 1, \ldots, m$, que Y^i a des trajectoires continues à gauche. Soit N le processus ponctuel multivarié formé par tous les processus N^{ij}, $1 \leq i, j \leq m$, et soit $\mathcal{N}_t$ la tribu engendrée par N.

$$N(\cdot) = \begin{pmatrix} N^{11}(\cdot) & N^{12}(\cdot)\ldots & N^{1j}(\cdot)\ldots & N^{1m}(\cdot) \\ \vdots & & \vdots & \vdots \\ N^{i1}(\cdot) & \ldots & N^{ij}(\cdot) & N^{im}(\cdot) \\ \vdots & & \vdots & \vdots \\ N^{m1}(\cdot) & N^{m2}(\cdot)\ldots & N^{mj}(\cdot)\ldots & N^{mm}(\cdot) \end{pmatrix} \quad \text{et} \quad Y(\cdot) = \begin{pmatrix} Y^1(\cdot) \\ \vdots \\ Y^i(\cdot) \\ \vdots \\ Y^m(\cdot) \end{pmatrix}$$

[2] (section 5D) montre qu'on est encore dans le cadre d'un modèle multiplicatif. Chaque composante $N^{ij}(\cdot)$ de N a un processus d'intensité $\lambda^{ij} = Y^i(t)\alpha^{ij}(t)$ par rapport à $\mathcal{N}_t$. Si on observe l'échantillon, $(N_1, \ldots, N_n)$ de processus ponctuels multivariés et $(Y_1, \ldots, Y_n)$ les processus prévisibles associés, ce modèle permet de décrire l'évolution au cours d'une période de temps finie, d'un ensemble de pixels (point-écran). On suppose pour cela que les états de E représentent les états possibles d'un pixel (intensité lumineuse, couleur, niveau de gris, etc ...), codés d'une manière adéquate. Le processus N^{ij}, compte pour un pixel donné, le nombre de transitions de l'état i vers l'état j. Les composantes du processus multivarié $Y(\cdot)$, indiquent l'état dans lequel se trouve le pixel. Les dates de changement d'état du pixel sont considérées comme des points de sauts du processus multivarié N. Si on considère un échantillon de pixels, on aura les vecteurs

de processus multivariés $N^{(n)} = (N_1, \ldots, N_n)$ et $Y^{(n)} = (Y_1, \ldots, Y_n)$, de même loi, représentant les processus associés aux points d'un écran. On suppose que les processus sont indépendants dans l'espace ((N_i, Y_i) et (N_j, Y_j) sont indépendants pour $i \neq j$). Soient

$$\overline{Y}^i = \sum_{k=1}^{n} \overline{Y}_k^i \text{ et } \overline{N}^{ij} = \sum_{k=1}^{n} \overline{N}_k^{ij}.$$

L'échantillon admet la vraisemblance :

$$\mathcal{L}_\beta(N^{(n)}, Y^{(n)}) = \exp\left(\sum_{i,j=1}^{m}\left(\int_0^T (1-\beta^{ij}(s))\overline{Y}^i(s)ds + \int_0^T \log\left(\beta^{ij}(s)\right) d\overline{N}^{ij}(s)\right)\right).$$

Sur l'intervalle de temps $[0,T]$, le mouvement est représenté par la suite d'images prises à un intervalle de temps régulier δ_m, qui définit ainsi une partition de l'intervalle de temps $[0,T]$, que nous pouvons comparer à la partition κ_m du modèle général (cf. lemme 3.2). Chaque pixel peut alors être codé par un atome $A = A(\tilde{n}, e_0, \ldots, e_{\tilde{n}}, m, \sigma_{\tilde{n}}^m)$. Ce vecteur de dimension aléatoire, représente un pixel qui a varié $\tilde{n}$ fois dans l'intervalle $[0,T]$ en prenant successivement les valeurs $e_0, \ldots, e_{\tilde{n}}$. L'entier m représente le nombre d'intervalles de la partition κ_m, et l'application $\sigma_{\tilde{n}}^m$ est l'application qui associe à chaque date de changement d'état l'intervalle de la partition dans lequel il s'est produit. Dans cet exemple, on se trouve dans des conditions moins générales que celles du lemme 3.2. Le code variable pour ces réalisations, basé sur une fonction candidate β, de complexité $Q_n(\beta)$, a pour longueur :

$$\sum_{i,j=1}^{m}\left(\int_0^T (\beta^{ij}(s)-1)\tilde{Y}^i(s)ds - \int_0^T \log\left(\beta^{ij}(s)\right) d\tilde{N}^{ij}(s)\right) + Q_n(\beta), \tag{3.32}$$

où $\tilde{N}$ et $\tilde{Y}$ sont les versions discrétisées des processus N et Y (cf. 3.2). La proposition 3.4 montre que si dans l'expression 3.32, on utilise la longueur du code de Shannon basé sur la fonction qui a généré les données, alors le code résultant est optimal dans la classe des codes uniquement décodables. Si de plus le minimum est pris dans une suite d'ensembles possédant de bonnes propriétés d'approximation au sens de la norme L_2, la proposition 3.18, fournit (au facteur m près), une borne supérieure de la longueur de code minimale. Si α est la fonction d'intensité associée à la loi du processus multivarié,

$$\alpha_n^* = Arg\min_{\beta\in\Gamma_n}\left(\sum_{i,j=1}^{m} \| \alpha^{ij} - \beta^{ij} \|_2^2\right)^{\frac{1}{2}},$$

et

$$\begin{aligned}\min_{\beta\in\Gamma_n}\frac{1}{n}[\log 1/\mathcal{L}_\beta(\tilde{N}^n,\tilde{Y}^n)+Q_n(\beta) &- \log 1/\mathcal{L}_\alpha(N^n,Y^n)]\\ &= O\left(n^{-1}Q_n(\alpha_n^*)\right)+O\left(mv_n^2\right)+O\left(\frac{\log n}{n}\right),\end{aligned}$$

avec $Q_n(\alpha_n^*)=\sum_{ij}C_n({\alpha_n^*}^{ij})$, et C_n étant une fonction de complexité du type de celles définies en 3.16, 3.17 et 3.18.

Dans le modèle précédent, seuls les instants de sauts $\tau_1,\ldots,\tau_{\tilde{n}}$, et les valeurs du processus Y (0 ou 1) en ces instants sont codées. Le codage d'une réalisation $A=(\tilde{n},e_0,\ldots,e_{\tilde{n}},m,\sigma_{\tilde{n}}^m)$, peut se faire à partir d'un code optimal q pour les entiers. Le codage de $\tilde{n}$ et m nécessite $q(\tilde{n})$ et $q(m)$ bits. Les valeurs de $e_0,\ldots,e_{\tilde{n}}$ nécessitent $\tilde{n}+1$ bits, car chaque valeur est codée par 0 ou 1. $\sigma_{\tilde{n}}^m$ représente une application d'un ensemble à $\tilde{n}$ éléments vers un ensemble à m éléments. Le nombre total de ces applications est $C_m^{\tilde{n}}$. Le codage de l'une de ces applications nécessite alors au plus $q(C_m^{\tilde{n}})$ bits. La longueur de description totale de la réalisation associée à un pixel est alors :

$$\tilde{n}+1+q(m)+q(C_m^{\tilde{n}}).$$

Chapitre 4

Estimation d'une fonction d'intensité.

4.1 Introduction.

Un certain nombre de méthodes ont été proposées pour traiter le problème de l'estimation de la fonction d'intensité d'un procesus ponctuel. Pour en citer quelques unes, [75] proposent des estimateurs analogues aux estimateurs à noyaux utilisés dans le cadre de l'estimation de densités de probabilité par Parzen et Rosenblatt. Breslow [15], Kalbfleisch et Prentice [40], proposent des estimateurs de type histogramme. Ces estimateurs ont été étudiés et généralisés par de nombreux auteurs, voir [61], [4], [11], [72]. Aalen [2, 3] introduit la théorie des martingales de carré intégrable dans l'estimation de la fonction d'intensité d'un processus ponctuel. L'auteur se place dans un cadre particulier : le modèle multiplicatif (dit de Aalen). Dans ce modèle, l'intensité stochastique du processus est de la forme $\lambda(s) = \alpha(s)Y(s)$ où $Y(\cdot)$ est un processus prévisible, et $\alpha(\cdot)$ est une fonction déterministe. L'idée de base de la méthode de Aalen consiste à faire l'analogie entre accroissement de martingale et bruit : l'écriture formelle caractérisant l'évolution du processus ponctuel N, i.e. $N(ds) = \alpha(s)Y(s)ds + dM(s)$ indique clairement que $N(ds)$ est une observation bruitée de $\alpha(s)Y(s)ds$. Aalen estime l'intensité cumulée $t \to \int_0^t \alpha(u)du$. Puisqu'aucune information ne peut être obtenue sur l'ensemble aléatoire où Y s'annule (l'intensité stochastique observée est nulle) on estime en fait :

$$B_\alpha(t) = \int_0^t \alpha(u)\mathbf{I}_{\{Y(u)>0\}}du.$$

Soit M la martingale définie par $M(t) = N(t) - \int_0^t \alpha(u)Y(u)du$. Sous des hypothèses convenables, la différence

$$\int_0^t \left(\mathbf{I}_{\{Y(u)>0\}}/Y(u)\right) dM(u) = \int_0^t \left(\mathbf{I}_{\{Y(u)>0\}}/Y(u)\right) dN(u) - B_\alpha(t),$$

intégrale d'un processus prévisible par rapport à une martingale (ou à une martingale locale), est aussi une martingale, d'où l'estimateur :

$$\hat{B_\alpha}(t) = \int_0^t \left(\mathbf{I}_{\{Y(u)>0\}}/Y(u)\right) dN(u).$$

De nombreux articles basés sur les mêmes techniques, proposent des estimateurs non-paramétriques de la fonction d'intensité, et non de l'intégrale indéfinie. [60] lisse l'estimateur de Aalen par une méthode de noyau, [53, 54] recherchent l'estimateur dans des espaces de fonctions orthogonales, [42] utilise la méthode des tamis de Grenander. L'estimateur est recherché dans un espace défini en termes de contraintes de régularité, et ces contraintes sont relâchées au fur et à mesure que la taille de l'échantillon croît de manière à assurer la convergence de l'estimateur. Antoniadis, et Grégoire [5] utilisent une méthode de maximum de vraisemblance pénalisé, la fonctionnelle de pénalité étant définie par le carré de la norme de la dérivée. Antoniadis, Grégoire et McKeague [6] proposent de lisser l'estimateur de Aalen par un noyau d'ondelettes. L'estimateur proposé ici, est basé sur la minimisation de la longueur totale de la complexité des données observées, ou de manière équivalente, d'un code universel (voir chapitre 2, section 2.3), c'est le principe de complexité minimale que nous avons développé dans les chapitres antérieurs.

Dans le cadre de l'estimation paramétrique de densités de probabilité, [62, 63] utilise le principe de description minimale pour déterminer la dimension des paramètres et leurs valeurs. Il montre dans le cas d'une famille paramétrique de dimension k, que les termes dominants de la longueur de description totale sont $(k/2)\log n + \log 1/P_{\theta^*}$, où n est la taille de l'échantillon et θ^* est l'estimateur du maximum de vraisemblance. Rissanen suggère la longueur de description minimale comme critère de sélection de modèle. Un critère équivalent fut proposé d'un point de vue bayésien par G. Schwarz dans le cadre des familles exponentielles. [8] montre la consistance d'estimateurs non paramétriques de mesures de probabilité, en se basant sur la convergence de tests exponentiels; [35] présentent des résultats de convergence d'un estimateur de la densité de probabilité, basé sur un histogramme à noeuds équidistants, dont la complexité dépend du nombre de classes. Les premiers résultats présentant des vitesses de convergence dans le

cas général furent obtenus dans [9]. On peut également citer [64] qui présente des vitesses de convergence pour l'histogramme et enfin [59] qui présentent une approche orientée théorie de la décision de la méthode d'estimation par complexité minimale.

L'estimateur que nous présentons dans le cadre de l'estimation de la fonction d'intensité d'un processus ponctuel est consistant et asymptotiquement normal. D'autre part, si la fonction inconnue appartient à l'une des familles candidates, l'estimateur la découvre presque-sûrement. Enfin la méthode autorise la recherche de la fonction dans une librairie pouvant contenir une quantité illimitée de bases de fonctions. Le problème d'inférence peut être formulé de la manière suivante : étant donné l'observation d'une réalisation de n couples aléatoires (N_i, Y_i), définis sur $(\Omega, (\mathcal{F}_t))$ muni de la probabilité P_α pour laquelle ces couples sont i.i.d., on désire estimer la fonction inconnue α par une fonction de $\Gamma_n \subset \Gamma$, où (Γ_n) est une suite de fonctions régulières.

Définition 4.1. *Soit* $(N^n, Y^n) = ((N_1, Y_1), \ldots, (N_n, Y_n))$ *un échantillon du couple* (N, Y).

L'estimateur de complexité minimale est défini par :

$$\begin{aligned} \hat{\alpha}_n &= Arg \min_{\beta \in \Gamma_n} C(N^n, Y^n) \\ &= Arg \min_{\beta \in \Gamma_n} \left(\log 1/\mathcal{L}_\beta(N^n, Y^n) + C_n(\beta) \right), \end{aligned} \tag{4.1}$$

En l'absence d'unicité parmi les minimiseurs de (4.1), l'estimateur est la fonction qui minimise $C_n(\cdot)$, et en cas d'ex-aequo, il suffit de choisir parmi les fonctions d'intensité, celle qui a le plus petit indice dans Γ_n (Γ_n étant dénombrable).

Proposition 4.2. *Il existe presque-sûrement au moins une fonction, et au plus un ensemble fini de fonctions maximisant* $2^{-C_n(\beta)}\mathcal{L}_\beta(N^n, Y^n)$. *Ainsi l'estimateur de complexité minimale* $\hat{\alpha}_n$ *existe presque-sûrement.*

Preuve : Soit

$$m(N^n, Y^n) = \sum_{\beta \in \Gamma_n} 2^{-C_n(\beta)}\mathcal{L}_\beta(N^n, Y^n) \tag{4.2}$$

En effet, supposons N^n fixé, et soit ν un réel quelconque strictement positif. Si la suite 4.2 converge, il y a dans Γ_n au plus un nombre fini de fonctions β vérifiant $2^{-C_n(\beta)}\mathcal{L}_\beta(N^n, Y^n) \geq \nu$. La fonction réalisant le maximum se trouve donc parmi celles-ci et existe nécessairement.

□

Remarque : Le problème de la maximisation de la fonction de vraisemblance n'admet pas de solution en l'absence de contraintes. Des estimateurs non-paramétriques du maximum de vraisemblance sont obtenus en imposant des contraintes sur l'espace des fonctions candidates. [42] utilise des tamis définis en bornant les fonctions candidates et leurs dérivées premières. [5] par contre, pénalisent la dérivée première, éliminant ainsi les fonctions trop irrégulières au sens de leurs variations. Aucune hypothèse n'est faite dans notre approche sur la lissitude des fonctions candidates. Les fonctions de complexité doivent uniquement satisfaire l'inégalité de Kraft

$$\sum_{\beta \in \Gamma_n} 2^{-C_n(\beta)} \leq 1,$$

et croître moins vite que la taille de l'échantillon : $C_n = o(n)$. La pénalisation implique donc que l'estimateur sera recherché parmi les fonctions qui ont la description la plus courte, et ces fonctions ne sont pas nécessairement les plus lisses.

◇

Nous rappelons dans la suite les conditions exigées dans le choix des fonctions de complexité, des espaces de fonctions candidates, et les hypothèses faites sur le modèle.

Conditions sur Γ_n et C_n.

G1 Il existe $0 < x < 1$, et $b > 0$ tel que les fonctions de complexité satisfont l'inégalité :

$$\sum_{\beta \in \Gamma_n} 2^{-xC_n(\beta)} \leq b.$$

G2 Les fonctions de complexité satisfont

$$C_n = o(n).$$

G3 Soit $\alpha_n^* = \arg\min_{\beta \in \Gamma_n} \| \alpha - \beta \|_2$, alors :

$$\lim_{n \to \infty} \| \alpha - \alpha_n^* \|_\infty = 0.$$

G4 Il existe une suite croissante a_n et deux constantes f_0 et f_1 telles que pour toute fonction $\beta \in \Gamma_n$, on ait l'inégalité

$$f_0 a_n^{-1} \leq \beta \leq f_1 a_n.$$

Rappel des hypothèses sur le modèle utilisées dans certaines preuves.

Hypothèses sur le modèle.

H3 Le processus Y est borné P_0-presque-sûrement par une constante D_Y.

$$P_0(\sup_s Y(s) \leq D_Y) = 1.$$

H4 Il existe deux constantes θ_0 et θ_1 telles que :

$$\theta_0 \leq E_\alpha(Y(s)) = \theta_\alpha(s) \leq \theta_1.$$

4.2 Résultats préliminaires

Dans cette section nous établissons des relations entre les distances obtenues entre les densités, et des distances entre les fonctions d'intensité correspondantes. Sur les densités, nous utiliserons la distance de Hellinger et la distance de Kullback-Leibler.

Définition 4.3. (Distance de Hellinger.) *La distance de Hellinger entre deux densités $\mathcal{L}_\beta$ et $\mathcal{L}_\gamma$ est définie par*

$$D_H^2(\mathcal{L}_\beta, \mathcal{L}_\gamma) = \frac{1}{2}\int_\Omega \left(\sqrt{\mathcal{L}_\beta} - \sqrt{\mathcal{L}_\gamma}\right)^2 dP_0.$$

La distance de Hellinger entre deux fonctions d'intensité β et γ est définie par :

$$d_H^2(\gamma, \beta) = \frac{1}{2}\int_0^T \left(\sqrt{\gamma(s)} - \sqrt{\beta(s)}\right)^2 ds.$$

Afin d'obtenir des résultats de convergence pour la distance de Hellinger entre les fonctions d'intensité, on établit les relations suivantes, qui lient les distances entre les densités et les distances entre les fonctions d'intensité.

Lemme 4.4. *Supposons l'hypothèse H3 vérifiée, et soit D_Y la constante qui y intervient (cf. 3.5). La distance de Hellinger entre deux densités $\mathcal{L}_\gamma$ et $\mathcal{L}_\beta$, est liée à la distance de Hellinger entre deux fonctions d'intensité γ et β par la relation :*

$$D_H^2(\mathcal{L}_\gamma, \mathcal{L}_\beta) \leq D_Y d_H^2(\gamma, \beta).$$

De plus, s'il existe une constante $D_0 > 0$ telle que $Y(s) \geq D_0$ P_0 presque-sûrement, alors on a également :

$$D_H^2(\mathcal{L}_\gamma, \mathcal{L}_\beta) \geq 1 - \exp\left(-D_0 d_H^2(\gamma, \beta)\right).$$

Dans le cas des processus de Poisson, cette dernière condition est vérifiée avec $D_0 = 1$.

Preuve :

$$D_H^2(\mathcal{L}_\gamma, \mathcal{L}_\beta) = 1 - \rho, \text{ où } \rho = \int_\Omega \sqrt{\mathcal{L}_\gamma \mathcal{L}_\beta} dP_0.$$

$$\begin{aligned}
\mathcal{L}_\gamma \mathcal{L}_\beta &= \exp\left(-\int_0^T (\gamma(s) + \beta(s) - 2) Y(s) ds + \int_0^T \ln(\gamma(s)\beta(s)) dN(s)\right) \\
&= \exp\left(-\int_0^T (\gamma(s) + \beta(s)) Y(s) ds + 2\int_0^T \sqrt{\gamma(s)\beta(s)} Y(s) ds\right) \\
&\cdot \exp\left(-2\int_0^T \left(1 - \sqrt{\gamma(s)\beta(s)}\right) Y(s) ds + 2\int_0^T \ln\left(\sqrt{\gamma(s)\beta(s)}\right) dN(s)\right).
\end{aligned}$$

En posant $\zeta \equiv \sqrt{\gamma\beta}$, en utilisant l'inégalité de Jensen on obtient :

$$\begin{aligned}
\rho = \int_\Omega \sqrt{\mathcal{L}_\gamma \mathcal{L}_\beta} dP_0 &= \int_\Omega \exp\left(-\frac{1}{2}\int_0^T \left(\sqrt{\gamma(s)} - \sqrt{\beta(s)}\right)^2 Y(s) ds\right) dP_\zeta \\
&= E_\zeta\left(\exp\left(-\frac{1}{2}\int_0^T \left(\sqrt{\gamma(s)} - \sqrt{\beta(s)}\right)^2 Y(s) ds\right)\right) \\
&\geq \exp\left(-\frac{1}{2}\int_0^T \left(\sqrt{\gamma(s)} - \sqrt{\beta(s)}\right)^2 \theta_\zeta(s) ds\right) \\
&\geq \exp\left(-D_Y d_H^2(\gamma, \beta)\right),
\end{aligned}$$

soit :

$$\begin{aligned}
D_H^2(\mathcal{L}_\gamma, \mathcal{L}_\beta) = 1 - \rho &\leq 1 - \exp\left(-D_Y d_H^2(\gamma, \beta)\right) \\
&\leq D_Y d_H^2(\gamma, \beta),
\end{aligned}$$

d'où le premier résultat.

Si P_0 presque-sûrement, $Y(s) \geq D_0$, on obtient :

$$\begin{aligned} D_H^2(\mathcal{L}_\gamma, \mathcal{L}_\beta) &= 1 - \int_\Omega \exp\left(-\frac{1}{2}\int_0^T \left(\sqrt{\gamma(s)} - \sqrt{\beta(s)}\right)^2 Y(s)ds\right) dP_\zeta \\ &\geq 1 - \exp\left(-D_0 d_H^2(\gamma, \beta)\right) \end{aligned}$$

□

Le résultat suivant servira à établir plusieurs inégalités :

Lemme 4.5.

$$\begin{aligned} P_\alpha(\mathcal{L}_\alpha(N^n, Y^n) \leq\ & c\mathcal{L}_\beta(N^n, Y^n) \text{ et } (N^n, Y^n) \in B) \leq \\ & cP_\beta(\mathcal{L}_\alpha(N^n, Y^n) \leq c\mathcal{L}_\beta(N^n, Y^n) \text{ et } (N^n, Y^n) \in B) \end{aligned} \tag{4.3}$$

Pour tout ensemble mesurable B et pour toute constante positive c. En particulier :

$$P_\alpha(\mathcal{L}_\alpha(N^n, Y^n) \leq c\mathcal{L}_\beta(N^n, Y^n)) \leq c. \tag{4.4}$$

Preuve : Soit $E_{n,\beta} = \{\mathcal{L}_\alpha(N^n, Y^n) \leq c\mathcal{L}_\beta(N^n, Y^n)\}$.

$$\begin{aligned} P_\alpha(\mathcal{L}_\alpha(N^n, Y^n) &\leq c\mathcal{L}_\beta(N^n, Y^n) \text{ et } (N^n, Y^n) \in B) \\ &= \int_{B \cap E_{n,\beta}} \mathcal{L}_\alpha(N^n, Y^n) dP_0^n \\ &\leq c \int_{B \cap E_{n,\beta}} \mathcal{L}_\beta(N^n, Y^n) dP_0^n \\ &= cP_\beta(\mathcal{L}_\alpha(N^n, Y^n) \leq c\mathcal{L}_\beta(N^n, Y^n) \text{ et } (N^n, Y^n) \in B) \end{aligned}$$

□

Remarque : Les densités $\mathcal{L}_\beta$ étant toujours strictement positives, il s'ensuit que toutes les mesures P_β sont équivalentes. Il suffit donc que Y soit minoré par D_0 P_0 presque-sûrement pour qu'elle le soit aussi P_β presque-sûrement.

◇

Proposition 4.6. *Pour tout $c > 0$, on a*

$$P_\alpha(\mathcal{L}_\alpha(N^n, Y^n) \leq c\mathcal{L}_\beta(N^n, Y^n)) \leq \sqrt{c} 2^{-nD_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta)} \tag{4.5}$$

Preuve : Il est clair que

$$P_\alpha(\mathcal{L}_\alpha(N^n,Y^n) \leq c\mathcal{L}_\beta(N^n,Y^n)) = P_\alpha\Big(\sqrt{\mathcal{L}_\alpha(N^n,Y^n)} \leq \sqrt{c}\sqrt{\mathcal{L}_\beta(N^n,Y^n)}\Big)$$

Soit $E_{n,\beta} = \{\mathcal{L}_\alpha(N^n,Y^n) \leq c\mathcal{L}_\beta(N^n,Y^n)\}$.

$$P_\alpha(E_{n,\beta}) = \int_{E_{n,\beta}} \mathcal{L}_\alpha(N^n,Y^n)dP_0 \leq \sqrt{c}\int_{E_{n,\beta}} \sqrt{\mathcal{L}_\alpha(N^n,Y^n)}\sqrt{\mathcal{L}_\beta(N^n,Y^n)}dP_0.$$

En posant $\rho = \int_\Omega \sqrt{\mathcal{L}_\alpha(N,Y)\mathcal{L}_\beta(N,Y)}dP_0$, on obtient $P_\alpha(E_{n,\beta}) \leq \sqrt{c}\rho^n$.

L'inégalité

$$D_H^2(\mathcal{L}_\alpha,\mathcal{L}_\beta) = \frac{1}{2}\int(\sqrt{\mathcal{L}_\alpha} - \sqrt{\mathcal{L}_\beta})^2 dP_0 = (1-\rho) \leq -\ln\rho$$

permet d'obtenir $\rho \leq 2^{-D_H^2(\mathcal{L}_\alpha,\mathcal{L}_\beta)}$, et finalement

$$P_\alpha(E_{n,\beta}) \leq 2^{-nD_H^2(\mathcal{L}_\alpha,\mathcal{L}_\beta)}. \tag{4.6}$$

□

Lemme 4.7. *La distance de Kullback-Leibler $d(\nu \parallel \beta)$ est localement équivalente à la norme L_2. Pour deux fonctions ν et β suffisamment proches au sens de la norme L_2, il existe deux constantes $k_1(\nu)$ et $k_2(\nu)$ telles que :*

$$k_1(\nu) \parallel \nu - \beta \parallel_2^2 \leq d(\nu \parallel \beta) \leq k_2(\nu) \parallel \nu - \beta \parallel_2^2 .$$

Preuve : Pour toute fonctionnelle J dérivable au second ordre de Taylor on a :

$$J(v) = J(u) + (\dot{J}(u), v-u) + \frac{1}{2}(\ddot{J}(u)(v-u), v-u) + o(\parallel u - v \parallel_2), \tag{4.7}$$

pour u et v tels que $\parallel u - v \parallel < 1$, et pour la norme associée au produit scalaire. Soit U un sous-ensemble de $L_2([0,T])$, et $\nu \in U$. On définit J sur U par

$$J(\beta) = d(\nu \parallel \beta) = \frac{1}{\ln 2}\int_0^T \left[\nu(s)\ln\frac{\nu(s)}{\beta(s)} + \beta(s) - \nu(s)\right]\theta_\nu(s)ds.$$

Il est clair que $\beta \longrightarrow d(\nu \parallel \beta)$ est deux fois différentiable au sens de Gateaux, et les dérivées premières et secondes dans la direction γ et ψ sont données respectivement par :

$$\begin{aligned} \dot{d}(\nu \parallel \beta)(\gamma) &= \frac{1}{\ln 2}\int_0^T \left(1 - \frac{\nu}{\beta}(s)\right)\gamma(s)\theta(s)ds \\ \ddot{d}(\nu \parallel \beta)(\gamma)(\psi) &= \frac{1}{\ln 2}\int_0^T \left(\frac{\nu(s)}{\beta^2(s)}\right)\psi(s)\gamma(s)\theta(s)ds. \end{aligned}$$

Ainsi en utilisant 4.7 avec $u = \nu$ et $v = \beta$, les deux premiers termes du membre de droite de 4.7 sont nuls. Le troisième terme a pour expression

$$\frac{1}{2\ln 2}\int_0^T \left(\frac{1}{\nu(s)}\right)(\nu(s) - \beta(s))^2\theta_\nu(s)ds,$$

et on peut trouver deux constantes

$$k_2(\nu) = 2\ln(2)\frac{\sup_s \theta_\nu(s)}{\inf_s \nu(s)} \text{ et } k_1(\nu) = 2\ln(2)\frac{\inf_s \theta_\nu(s)}{\sup_s \nu(s)},$$

telles que :

$$k_1(\nu) \parallel \nu - \beta \parallel_2^2 \leq d(\nu \parallel \beta) \leq k_2(\nu) \parallel \nu - \beta \parallel_2^2 .$$

□

Dans le cas où Γ_n est un ensemble de fonctions uniformément bornées, on obtient un résultat plus fort. C'est l'objet du lemme suivant. Il nécessite l'hypothèse H4 (cf. 3.6).

Lemme 4.8. *Soit α la fonction d'intensité associée à la loi du processus. Supposons que l'hypothèse H4 soit vérifiée et que Γ soit un ensemble de fonctions uniformément bornées par deux constantes $0 < f_0 < f_1 < \infty$, i.e. :*

$$\forall \beta \in \Gamma, \forall s \in [0,T], \quad f_0 < \beta(s) < f_1.$$

Alors pour tout β dans Γ, on peut trouver deux constantes K_1 et K_2 indépendantes de α et β telles que :

$$K_1 \cdot d(\alpha \parallel \beta) \leq \parallel \alpha - \beta \parallel_2^2 \leq K_2 \cdot d(\alpha \parallel \beta). \tag{4.8}$$

Preuve : On cherche premièrement une fonction ψ satisfaisant

$$\alpha \ln \frac{\alpha}{\beta} + \beta - \alpha \geq \psi(\alpha,\beta)(\alpha - \beta)^2.$$

En posant $u = \frac{\alpha}{\beta}$, on doit trouver $\psi(\alpha, \beta)$ majorée par la fonction :

$$\frac{1}{\beta}\frac{u\ln(u) + 1 - u}{(u-1)^2} = \frac{1}{\beta}g(u).$$

Il suffit de prendre $\psi(\alpha, \beta) = \frac{1}{\beta}\inf g(u)$. La fonction g est strictement décroissante, et atteint son minimum pour $\sup_{s\in[0,T]}\left(\frac{\alpha(s)}{\beta(s)}\right) < u^* = f_1/f_0$. On prend ainsi $\psi(\alpha, \beta) = \frac{1}{f_0}g(u^*)$ pour obtenir la relation

$$(\alpha(s) - \beta(s))^2 \leq \frac{f_0}{g(u^*)}\Big(\alpha(s)\ln\frac{\alpha(s)}{\beta(s)} + \beta(s) - \alpha(s)\Big). \tag{4.9}$$

On a également :

$$\begin{aligned} d(\alpha \parallel \beta) &= \frac{1}{\ln 2}\int\Big(\alpha(s)\ln\frac{\alpha(s)}{\beta(s)} + \beta(s) - \alpha(s)\Big)\theta(s)\,ds \\ &\geq \frac{\theta_0}{\ln 2}\int\Big(\alpha(s)\ln\frac{\alpha(s)}{\beta(s)} + \beta(s) - \alpha(s)\Big)\,ds. \end{aligned} \tag{4.10}$$

Finalement, de (4.9) et (4.10) on déduit :

$$\| \alpha - \beta \|_2^{\,2} \leq K_2 \cdot d(\alpha \parallel \beta) \text{ avec } K_2 = \frac{f_0 \ln 2}{\theta_0 \cdot g(f_1/f_0)}.$$

La seconde partie de la preuve découle des inégalités

$$\begin{aligned} d(\alpha \parallel \beta) &= \frac{1}{\ln 2}\int\Big(\alpha(s)\ln\frac{\alpha(s)}{\beta(s)} + \beta(s) - \alpha(s)\Big)\theta(s)\,ds \\ &\leq \frac{1}{\ln 2}\int\Big[\alpha(s)(\frac{\alpha(s)}{\beta(s)} - 1) + \beta(s) - \alpha(s)\Big]\theta(s)\,ds. \\ &= \frac{1}{\ln 2}\int\Big(\alpha(s) - \beta(s)\Big)^2\frac{1}{\beta(s)}\theta(s)\,ds \\ &\leq \frac{1}{\ln 2}\Big(\sup_{0\leq s\leq T}\frac{\theta(s)}{\beta(s)}\Big)\int\Big(\alpha(s) - \beta(s)\Big)^2\,ds, \end{aligned}$$

où la première inégalité est une conséquence de $\ln(x) \leq x - 1$ appliquée à $x = \alpha(s)/\beta(s)$. Finalement, en posant $K_1 = \frac{f_0}{\theta_1}\ln 2$, on a l'inégalité de gauche dans 4.8, ce qui termine la preuve.

□

Lemme 4.9. *Soit γ une fonction strictement positive, et supposons l'hypothèse H4 (cf. 3.6) satisfaite pour γ. Si (γ_n) est une suite de fonctions positives vérifiant $\lim_{n\to\infty} d(\gamma \parallel \gamma_n) = 0$, alors on a également $\lim_{n\to\infty} \| \gamma - \gamma_n \|_1 = 0$.*

Preuve : On a

$$d(\gamma \parallel \gamma_n) = \frac{1}{\ln 2} \int_0^T \left[\gamma(s) \ln \frac{\gamma(s)}{\gamma_n(s)} + \gamma_n(s) - \gamma(s) \right] \theta_\gamma(s) ds,$$

où $\theta_\gamma(s) = E_\gamma(Y(s))$. En posant $u_n = \gamma_n/\gamma$, la dernière expression peut encore s'écrire

$$\begin{aligned} d(\gamma \parallel \gamma_n) &= \frac{1}{\ln 2} \int_0^T [-\ln(u_n) + u_n - 1] \gamma(s) \theta_\gamma(s) ds \\ &= \frac{1}{\ln 2} \int_0^T v_n(s) \gamma(s) \theta_\gamma(s) ds, \end{aligned}$$

où v_n est une fonction positive. Puisque $d(\gamma \parallel \gamma_n)$ tend vers 0, il s'ensuit que v_n tend vers 0 dans $L_1(\gamma\theta)$, et donc aussi dans L_1, par hypothèse sur γ et θ_γ. D'autre part v_n se comporte comme $(u_n - 1)^2$ uniformément sur $[0, T]$. Par suite u_n tend vers 1 dans L_2, et le résultat en découle immédiatement.

□

4.3 Consistance de l'estimateur.

Si la fonction inconnue appartient à Γ, alors le prochain résultat montre que P_α presque-sûrement, elle est découverte pour n suffisamment grand. Néanmoins pour une taille d'échantillon donnée, on ne sait pas si elle a été découverte.

Théorème 4.10. *Supposons la condition G1 vérifiée, et qu'il existe $D_0 > 0$ tel que P_0 presque-sûrement, pour tout $s \in [0, T]$, $Y(s) \geq D_0$. Soit*

$$\alpha_n = Arg \min_{\substack{\beta \in \Gamma_n \\ \beta \neq \alpha}} d_H^2(\alpha, \beta).$$

Si α appartient à Γ et si il existe $\nu > 0$ tel que $d_H^2(\alpha, \alpha_n) \geq (1 + \nu) n^{-1} \log n$, alors

$$P_\alpha(\hat{\alpha}_n \equiv \alpha \text{ à partir d'un certain rang }) = 1.$$

Preuve : On veut montrer que

$$P_\alpha\left(\hat{\alpha}_n \not\equiv \alpha \text{ infiniment souvent }\right) = 0.$$

Cela revient à montrer que

$$P_\alpha\left(\liminf \bigcap_{\substack{\beta \in \Gamma_n \\ \beta \neq \alpha}} \{\mathcal{L}_\beta(N^n, Y^n) 2^{-C_n(\beta)} < \mathcal{L}_\alpha(N^n, Y^n) 2^{-C_n(\alpha)}\} \right) = 1.$$

Si on pose $A_n = \bigcap_{\substack{\beta\in\Gamma_n \\ \beta\neq\alpha}} \{\mathcal{L}_\beta(N^n,Y^n)2^{-C_n(\beta)} < \mathcal{L}_\alpha(N^n,Y^n)2^{-C_n(\alpha)}\}$, alors il suffira de montrer par le lemme de Borel-Cantelli, que : $\sum_n P_\alpha(\overline{A}_n) < \infty$. Pour cela, on utilise respectivement le la proposition 4.6 et le lemme 4.4.

$$\begin{aligned}
P_\alpha(\overline{A}_n) &= P_\alpha\left(\bigcup_{\substack{\beta\in\Gamma_n \\ \beta\neq\alpha}} \left(\mathcal{L}_\alpha 2^{-C_n(\alpha)} \leq \mathcal{L}_\beta 2^{-C_n(\beta)}\right)\right) \\
&\leq \sum_{\substack{\beta\in\Gamma_n \\ \beta\neq\alpha}} P_\alpha\left(\mathcal{L}_\alpha 2^{-C_n(\alpha)} \leq \mathcal{L}_\beta 2^{-C_n(\beta)}\right) \\
&\leq \sum_{\substack{\beta\in\Gamma_n \\ \beta\neq\alpha}} 2^{\frac{-C_n(\beta)+C_n(\alpha)}{2}} 2^{-nD_H^2(\mathcal{L}_\alpha,\mathcal{L}_\beta)} \\
&\leq \sum_{\substack{\beta\in\Gamma_n \\ \beta\neq\alpha}} 2^{\frac{-C([\beta])+C([\alpha])}{2}} 2^{-n(1-e^{-D_0 d_H^2(\alpha,\beta)})} \\
&\leq 2^{\frac{C([\alpha])}{2}} 2^{-n(1-e^{-D_0 d_H^2(\alpha,\alpha_n)})} \sum_{\substack{\beta\in\Gamma_n \\ \beta\neq\alpha}} 2^{-\frac{C([\beta])}{2}} \\
&\leq 2^{\frac{C([\alpha])}{2}} 2^{-n(1-e^{-D_0 d_H^2(\alpha,\alpha_n)})}.
\end{aligned}$$

Une condition suffisante pour que cette borne soit le terme général d'une série convergente est qu'on puisse trouver $\nu > 0$ arbitrairement petit tel que

$$2^{-n(1-e^{-D_0 d_H^2(\alpha,\alpha_n)})} \leq n^{-(1+\nu)}.$$

Cela est réalisé dès que

$$d_H^2(\alpha,\alpha_n) \geq (1+\nu)\frac{\log n}{n},$$

ce qui est vrai par hypothèse.

□

Remarque : La condition "il existe $D_0 > 0$ tel que P_0 presque-sûrement, pour tout $s \in [0,T]$, $Y(s) \geq D_0$" peut être supprimée, en imposant une contrainte du type $D_H^2(\mathcal{L}_\alpha,\mathcal{L}_{\alpha_n}) \geq (1+\nu)n^{-1}\log n$ sur les vraisemblances, au lieu de l'imposer sur les fonctions d'intensité. Néanmoins, cette contrainte est moins aisée à interpréter par rapport aux fonctions d'intensité. Elle est plus forte que la contrainte $d_H^2(\alpha,\alpha_n) \geq (1+\nu)n^{-1}\log n$ (cf. lemme 4.4), et donc renforce la contrainte sur la vitesse d'approximation de α par les fonctions des espaces Γ_n, tout en affaiblissant la contrainte sur le processus Y.

◇

Théorème 4.11. *Si les conditions G1, G3 sont vérifiées, et s'il existe D_0 tel que $Y(s) > D_0$ P_0 presque-sûrement, alors $\hat{\alpha}_n$, estimateur de complexité minimale de la fonction d'intensité, converge P_α p.s. au sens de la distance de Hellinger, i.e.*

$$d_H^2(\alpha, \hat{\alpha}_n) \to 0 \quad P_\alpha \text{ presque-sûrement.}$$

D'autre part on a aussi P_α presque-sûrement, $C_n(\hat{\alpha}_n) = o(n)$.

Preuve : Soit $D(\delta)$ un sous-ensemble de Γ_n,

$$A_{n,\beta} = \{2^{-C_n(\beta)}\mathcal{L}_\beta(N^n, Y^n) \geq \mathcal{L}_\alpha(N^n, Y^n)2^{-n\varepsilon}\} \text{ et } B_n(\delta) = \bigcap_{\beta \in D(\delta)} \overline{A}_{n,\beta}.$$

Par le lemme 3.17, il existe α^* tel que P_α p.s pour n suffisamment grand,

$$\frac{1}{n}\log\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha^*}}(N^n, Y^n) + \frac{1}{n}C_n(\alpha^*) < \varepsilon.$$

Cette inégalité peut encore être formulée par :

$$\{2^{-C_n(\alpha^*)}\mathcal{L}_{\alpha^*}(N^n, Y^n) \geq \mathcal{L}_\alpha(N^n, Y^n)2^{-n\varepsilon}\}. \tag{4.11}$$

D'autre part si l'évènement $B_n(\delta)$ est réalisé, alors toute fonction $\beta \in D(\delta)$ vérifie

$$\mathcal{L}_\alpha 2^{-n\varepsilon} \geq \mathcal{L}_\beta 2^{-C_n(\beta)}. \tag{4.12}$$

Par les relations (4.11) et (4.12), on peut conclure que si B_n est réalisé, alors $\hat{\alpha}_n \notin D(\delta)$. Dans la suite nous montrons que pour $D_1(\delta) = \{\beta \ : C_n(\beta) > n\delta\}$ et $D_2(\delta) = \{\beta \ : D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta) > \delta\}$, l'évènement $B_n(\delta)$ est réalisé pour n suffisamment grand. Plus précisément, on montre que $P_\alpha\left(\limsup \overline{B}_n(\delta) = 0\right)$, ce qui revient à montrer que $\sum_n P_\alpha(\overline{B}_n(\delta)) < \infty$, la conclusion découlant alors du lemme de Borel-Cantelli.

a)On montre que $\hat{\alpha}_n \notin D_1(\delta)$.

Par l'inégalité (4.4) on a :

$$P_\alpha(\overline{B}_n(\delta)) = P_\alpha\left(\bigcup_{\beta \in D_1(\delta)} A_{n,\beta}\right) \leq \sum_{\beta \in D_1(\delta)} 2^{n\varepsilon - C_n(\beta)}.$$

$$\begin{aligned} P_\alpha(\overline{B}_n(\delta)) &\leq \sum_{\beta \in D_1(\delta)} 2^{-xC_n(\beta)}2^{-(1-x)C_n(\beta)}2^{n\varepsilon} \\ &\leq \sum_{\beta \in D_1(\delta)} 2^{-xC_n(\beta)}2^{-n\delta(1-x)}2^{n\varepsilon} \\ &= \sum_{\beta \in D_1(\delta)} 2^{-xC_n(\beta)}2^{-n(\delta(1-x)-\varepsilon)} \end{aligned}$$

Par la condition G1, on obtient $\sum_{\beta\in D_1(\delta)} P_\alpha(A_{n,\beta}) \leq b2^{-n(\delta(1-x)-\varepsilon)}$. Il suffit de choisir $\varepsilon < \delta(1-x)$ pour s'assurer que cette série converge et que pour $\beta \in D_1(\delta)$ on a $P_\alpha(\liminf B_n) = 1$, et donc $\hat{\alpha}_n \notin D_1(\delta)$.

b)On montre que $\hat{\alpha}_n \notin D_2(\delta)$.

Par les l'inégalités (4.4) et (4.6), on a

$$P_\alpha(\overline{B}_n(\delta)) \leq \sum_{\beta\in D_2(\delta)} P_\alpha(A_{n,\beta}) \leq \sum_{\beta\in D_2(\delta)} \min\{2^{n\varepsilon-C_n(\beta)}, 2^{\frac{n\varepsilon-C_n(\beta)}{2}-n\delta}\}. \tag{4.13}$$

L'inégalité $\min\{y,z\} \leq y^u z^{1-u}$, valable pour $u<1$ appliquée au membre de droite de (4.13) en prenant $u = 2x-1$, $0<x<1$ fournit une majoration par

$$2^{(n\varepsilon-C_n(\beta))(2x-1)} \cdot 2^{(n\varepsilon-C_n(\beta)-2n\delta)(1-x)}. \tag{4.14}$$

L'exposant de (4.14) s'écrit :

$$(n\varepsilon - C_n(\beta))(2x-1) + (n\varepsilon - C_n(\beta) - 2n\delta)(1-x)$$

$$= -C_n(\beta)((2x-1)+(1-x)) + n\varepsilon((2x-1)+(1-x)) - 2n\delta(1-x)$$

$$= -xC_n(\beta) - n(2\delta(1-x) - \varepsilon x).$$

Choisissant x tel que $\sum_{\beta\in\Gamma_n} 2^{-x.C_n(\beta)} \leq b_2$, ce qui est possible d'après G1, nous obtenons

$$\sum_{\beta\in D_2(\delta)} P(\{\mathcal{L}_\alpha(N^n,Y^n)2^{-n\varepsilon} \leq \mathcal{L}_\beta(N^n,Y^n)2^{-C_n(\beta)}\}) \leq b_2 2^{-n((\delta(1-x))-\varepsilon x)}.$$

Il suffit maintenant de choisir ε tel que $0<\varepsilon<\frac{2\delta(1-x)}{x}$ pour assurer que la série de terme général $P_\alpha(\overline{B}_n)$ est convergente, ce qui montre que pour n assez grand, $\hat{\alpha}_n \notin D_2(\delta)$. Donc

$$D_H^2(\mathcal{L}_\alpha, \mathcal{L}_{\hat{\alpha}_n}) \leq \delta \tag{4.15}$$

Par le lemme 4.4 cette relation implique également

$$1 - e^{-D_0 d_H^2(\alpha,\hat{\alpha}_n)} \leq \delta,$$

soit $e^{-D_0 d_H^2(\alpha,\hat{\alpha}_n)} \geq 1-\delta$, et $-D_0 d_H^2(\alpha,\hat{\alpha}_n) \geq \ln(1-\delta)$. On obtient alors $d_H^2(\alpha,\hat{\alpha}_n) \leq -\frac{1}{D_0}\ln(1-\delta)$. Puisque $\delta<1$, on peut conclure que :

$$d_H^2(\alpha,\hat{\alpha}_n) \leq \frac{\delta}{D_0}.$$

Cette relation étant vraie pour tout $1 > \delta > 0$, l'estimateur de complexité minimale doit donc vérifier P_α-presque-sûrement :

$$d_H^2(\alpha, \hat{\alpha}_n) = o(1) \text{ et } C_n(\hat{\alpha}_n) < n\delta.$$

□

Nous montrons dans la suite que l'estimateur de complexité minimale de la fonction d'intensité est presque-sûrement consistant au sens de la distance de Kullback-Leibler. La preuve nécessite l'hypothèse suivante sur la variance du processus :

H5 : $Var_\alpha(N(T)) < \infty$.

Théorème 4.12. *Si* (Γ_n) *est une famille de fonctions* C^1 *par morceaux, si les hypothèses H3, H4 et H5, les conditions G2, ainsi que G4 avec* $a_n = O(n^{1/4-\mu})$ *(avec* $\mu > 0$ *aussi petit que l'on désire) sont vérifiées, alors l'estimateur de complexité minimale est* P_α *presque-sûrement consistant en distance de Kullback-Leibler : i.e.*

$$d(\alpha \parallel \hat{\alpha}_n) \longrightarrow 0,$$

P_α *presque-sûrement.*

Preuve : La distance de Kullback-Leibler $d(\alpha \parallel \hat{\alpha}_n)$ entre la fonction inconnue α et l'estimateur de complexité minimale $\hat{\alpha}_n$ est $d(\alpha \parallel \hat{\alpha}_n) = H_\alpha(\hat{\alpha}_n) - H_\alpha(\alpha)$ avec

$$H_\alpha(\beta) = \int_0^T [-\alpha(s) \ln \beta(s) + \beta(s) - 1]\, \theta_\alpha(s) ds.$$

La distance entre la fonction inconnue et l'estimateur peut être décomposée de la manière suivante :

$$\begin{aligned} d(\alpha \parallel \hat{\alpha}_n) &= H_\alpha(\hat{\alpha}_n) - \frac{1}{n} \log 1/\mathcal{L}_{\hat{\alpha}_n} - C_n(\hat{\alpha}_n) + \frac{1}{n} \log 1/\mathcal{L}_{\hat{\alpha}_n} + C_n(\hat{\alpha}_n) - H_\alpha(\alpha) \\ &\leq H_\alpha(\hat{\alpha}_n) - \frac{1}{n} \log 1/\mathcal{L}_{\hat{\alpha}_n} + \frac{1}{n} \log 1/\mathcal{L}_{\alpha_n^*} + \frac{1}{n}(C_n(\alpha_n^*) - C_n(\hat{\alpha}_n)) - H_\alpha(\alpha). \end{aligned}$$

La dernière relation provient du fait que

$$\log 1/\mathcal{L}_{\hat{\alpha}_n} + C_n(\hat{\alpha}_n) \leq \log 1/\mathcal{L}_{\alpha_n^*} + C_n(\alpha_n^*).$$

En utilisant les relations $H_\alpha(\alpha) = H_\alpha(\alpha_n^*) - d(\alpha \parallel \alpha_n^*)$, et $(C_n(\alpha_n^*) - C_n(\hat{\alpha}_n)) \leq C([\alpha_n^*])$ on obtient :

$$d(\alpha \parallel \hat{\alpha}_n) \leq \left| H_\alpha(\hat{\alpha}_n) - \frac{1}{n} \log 1/\mathcal{L}_{\hat{\alpha}_n} \right| + \left| \frac{1}{n} \log 1/\mathcal{L}_{\alpha_n^*} - H_\alpha(\alpha_n^*) \right| + \frac{1}{n} C([\alpha_n^*]) + d(\alpha \parallel \alpha_n^*).$$

Par la condition G2 et la proposition (3.12), les deux derniers termes convergent vers zéro. Des deux termes qui restent, on examine uniquement le premier, le second est traité de manière identique. En posant $d\overline{M}_n(s) = d\overline{N}_n(s) - \alpha(s)\overline{Y}_n(s)ds$, on obtient :

$$\begin{aligned}\left|H_\alpha(\hat{\alpha}_n) - \frac{1}{n}\log 1/\mathcal{L}_{\hat{\alpha}_n}\right| &= \left|H_\alpha(\hat{\alpha}_n) - \int_0^T [\hat{\alpha}_n(s) - 1 - \alpha(s)\ln\hat{\alpha}_n(s)]\frac{\overline{Y}_n(s)}{n}\right| ds \\ &+ \frac{1}{n}\left|\int_0^T \ln\hat{\alpha}_n(s)d\overline{M}_n(s)\right|\end{aligned}$$

En décomposant l'expression de l'entropie, on obtient :

$$\begin{aligned}\left|H_\alpha(\hat{\alpha}_n) - \frac{1}{n}\log 1/\mathcal{L}_{\hat{\alpha}_n}\right| &\leq \int_0^T \left|(\alpha(s)\ln\hat{\alpha}_n(s) + 1 - \hat{\alpha}_n(s))\left(\frac{\overline{Y}_n(s)}{n} - \theta_\alpha(s)\right)\right| ds \\ &+ \left|\frac{1}{n}\int_0^T \ln\hat{\alpha}_n(s)d\overline{M}_n(s)\right|.\end{aligned}$$

Par hypothèse sur Γ_n, cette expression est inférieure à :

$$a_n\int_0^T \left|\frac{\overline{Y}_n(s)}{n} - \theta(s)\right| ds + \left|\frac{1}{n}\int_0^T \ln\hat{\alpha}_n(s)d\overline{M}_n(s)\right|.$$

Pour $a_n = n^{1/4-\nu}$ la première intégrale converge vers zéro. Pour le montrer, on applique la loi forte des grands nombres de Marcinkiewicz-Zygmund (comme dans [42], [24]) aux variables aléatoires i.i.d. et centrées $Y_i(s) - \theta(s)$. Par le théorème de Fubini ce résultat est valable presque-partout sur $[0,T]$ par rapport à la mesure de Lebesgue. Puisque Y est supposé borné, le théorème de convergence dominée nous permet de conclure. Pour terminer la preuve, il suffit de borner le dernier terme :

$$\frac{1}{n}\left|\int_0^T \ln\hat{\alpha}_n(s)d\overline{M}_n(s)\right|.$$

En intégrant par parties, on obtient :

$$\begin{aligned}\left|\int_0^T \ln\hat{\alpha}_n(s)d\overline{M}_n(s)\right| &= \left|\overline{M}_n(T)\ln(\hat{\alpha}_n(T)) - \int_0^T \frac{\hat{\alpha}_n'(s)}{\hat{\alpha}_n(s)}\overline{M}_n(s^-)ds\right| \\ &\leq \ln(a_n)\sup_s\left|\overline{M}_n(s)\right| + \left|\int_0^T \frac{\hat{\alpha}_n'(s)}{\hat{\alpha}_n(s)}\overline{M}_n(s)ds\right| \\ &\leq \sup_s\left|\overline{M}_n(s)\right|\left(\ln(a_n) + \int_0^T \left|\frac{\hat{\alpha}_n'(s)}{\hat{\alpha}_n(s)}\right| ds\right).\end{aligned}$$

Pour toute fonction $\beta \in \Gamma_n$, on peut majorer l'intégrale apparaissant dans le membre de droite de l'expression précédente. Soient $E^+ = \{s \in [0,T] : \beta'(s) > 0\}$, et E^- son complémentaire.

Par hypothèse sur Γ_n, il existe une suite d'intervalles $I_k^+ = [a_k, b_k[$ tels que $E^+ = \bigcup_k I_k^+$. On obtient alors

$$\int_0^T \left|\frac{\beta'(s)}{\beta(s)}\right| ds = \int_{E^+} \frac{\beta'(s)}{\beta(s)} ds - \int_{E^-} \frac{\beta'(s)}{\beta(s)} ds.$$

On borne la première intégrale du membre de droite, la seconde se traite de manière identique.

$$\begin{aligned}
\int_{E^+} \frac{\beta'(s)}{\beta(s)} ds &= \sum_k \int_{I_k^+} \frac{\beta'(s)}{\beta(s)} ds \\
&= \sum_k [\ln(\beta(b_k)) - \ln(\beta(a_k))] \\
&\leq \ln(\sup_s \beta(s)) - \ln(\inf_s \beta(s)) \\
&\leq 2\ln(a_n),
\end{aligned}$$

l'avant dernière inégalité dans l'expression précédente, provenant de la monotonie de la fonction $\ln(\beta)$ sur E^+. Finalement, on obtient :

$$\left|\int_0^T \ln \hat{\alpha}_n(s) d\overline{M}_n(s)\right| \leq 5\ln(a_n) \sup_s \left|\overline{M}_n(s)\right|.$$

Pour borner ce dernier terme, on utilise le théorème A.2 de [18] :

Pour tout $p > 0$ il existe des constantes c_p and C_p telles que pour toute martingale M,

$$c_p E_\alpha \left[[M_n]_T^{p/2}\right] \leq E_\alpha \left[\sup_{t\leq T} |M_t|^p\right] \leq C_p E_\alpha \left[[M]_1^{p/2}\right].$$

En appliquant cette inégalité avec $p = 4$ et l'inégalité de Markov on obtient :

$$\begin{aligned}
P_\alpha\left(\frac{a_n}{n} \sup_s |\overline{M}_n(s)| > \varepsilon\right) &= O\left(\left(\frac{n\varepsilon}{a_n}\right)^{-4}\right) E_\alpha \left[\sup_{s\leq T} |\overline{M}_n(s)|^4\right] \\
&= O\left(\left(\frac{n\varepsilon}{a_n}\right)^{-4}) E_\alpha \left[\left[\overline{M}_n\right]_T^2\right]\right).
\end{aligned}$$

La relation $[M]_s = N_s$ permet d'obtenir :

$$\begin{aligned}
E_\alpha \left[\left[\overline{M}_n\right]_T^2\right] &= E_\alpha \left(\overline{N}_n(T)^2\right). \\
&= nVar_\alpha(N(T)) + n^2 E_\alpha(N(T))^2 = O(n^2),
\end{aligned}$$

puisqu'on a supposé que $Var_\alpha(N(T)) < \infty$.

$$P_\alpha\left(\frac{a_n}{n} \sup_s |M_n(s)| > \varepsilon\right) = O\left((a_n)^4 n^{-2}\right).$$

Par le lemme de Borel-Cantelli, pour $a_n \leq n^{1/4-\mu}$ le dernier terme converge vers zéro presque-sûrement par rapport à P_α.

□

Ce théorème est encore vrai pour la norme L_1, et sous des conditions plus fortes sur les familles candidates, pour la norme L_2. C'est l'objet du corollaire suivant.

Corollaire 4.13. *Sous les hypothèses du théorème 4.12 on a :*
a) L'estimateur de complexité minimale converge en norme L_1 P_α presque-sûrement.
b) Si la famille de fonctions candidate est uniformément bornée, alors l'estimateur converge également P_α presque-sûrement en norme L_2.

Preuve : La partie a) découle du lemme 4.9, et la partie b) du lemme 4.8.

□

4.4 Vitesses de convergence.

Le théorème suivant montre que le meilleur compromis entre l'erreur d'estimation en distance de Hellinger, et la complexité de l'estimateur relativement à la taille de l'échantillon est du même ordre que l'indice de résolvabilité, ainsi :

$$d_H^2(\alpha, \hat{\alpha}_n) + \frac{1}{n} C_n(\hat{\alpha}_n) = O(R_n(\alpha)).$$

Une conséquence directe, est que l'erreur d'estimation est presque-sûrement majorée par l'indice de résolvabilité en distance de Hellinger au carré.

Théorème 4.14. *Si les conditions G1 et G2 sont vérifiées, et qu'il existe D_0 tel que $Y(s) > D_0$ P_0 presque-sûrement, alors on a P_α-p.s, pour n suffisamment grand :*

$$d_H^2(\alpha, \hat{\alpha}_n) = O(R_n(\alpha)) \text{ et } C_n(\hat{\alpha}_n) = O(nR_n(\alpha)).$$

Preuve : L'idée centrale de la preuve est de montrer que P_α presque-sûrement il existe des constantes K_2 et K_3 telles que pour n assez grand , $C_n(\hat{\alpha}_n) < nK_2 R_n(\alpha)$ et $d_H^2(\alpha, \hat{\alpha}_n) < \frac{K_3}{D_0} R_n(\alpha)$.

Plus précisément on montre que :

$$P_\alpha(\liminf\{d_H^2(\alpha,\hat{\alpha}_n) < \frac{K_3}{D_0}R_n(\alpha) \text{ et } C_n(\hat{\alpha}_n) < nK_2R_n(\alpha)\}) = 1.$$

Soient $0 < \varepsilon_n$, et A_n défini par $A_n = \{2^{-C_n(\alpha_n^*)}\mathcal{L}_{\alpha_n^*}(N^n,Y^n) \geq \mathcal{L}_\alpha(N^n,Y^n)2^{-n\varepsilon_n}\}$. Par le lemme 3.17, A_n est réalisé P_α presque-sûrement pour n assez grand, avec

$$\varepsilon_n = \frac{1}{n}C_n(\alpha_n^*) + (1+\mu)\frac{\log n}{n} + v_n^2, \tag{4.16}$$

où μ est une constante qui peut être choisie arbitrairement petite, et v_n^2 est l'erreur d'approximation. Soient $A_{n,\beta} = \{2^{-C_n(\beta)}\mathcal{L}_\beta(N^n,Y^n) \geq \mathcal{L}_\alpha(N^n,Y^n)2^{-n\varepsilon_n}\}$, Δ_n le sous-ensemble de Γ_n défini par

$$\Delta_n = \{\beta \in \Gamma_n, \ : \ D_H^2(\mathcal{L}_\alpha,\mathcal{L}_\beta) \geq K_3R_n(\alpha)\} \bigcup \{\beta \in \Gamma_n, \ : \ C_n(\beta) \geq nK_2R_n(\alpha)\}, \tag{4.17}$$

et

$$B_n = \bigcap_{\beta\in\Delta_n} \overline{A}_{n,\beta}.$$

Si on montre que

$$P_\alpha(\liminf B_n) = 1,$$

alors P_α-presque-sûrement, aucune fonction dans Δ_n ne peut maximiser le critère. Pour montrer que $P_\alpha(\liminf B_n) = 1$, il suffit de montrer que $\sum_n P_\alpha(\overline{B}_n) < \infty$, la conclusion découlant alors du lemme de Borel-Cantelli. Par union dénombrable, on a

$$P_\alpha(\overline{B}_n) \leq \sum_{\beta\in\Delta_n} P_\alpha(A_{n,\beta}),$$

et d'après (4.4),

$$P(A_{n,\beta}) \leq 2^{n\varepsilon_n - C_n(\beta)} = 2^{-xC_n(\beta)-(1-x)C_n(\beta)+n\varepsilon_n}.$$

Pour les termes de la somme qui satisfont, $C_n(\beta) > nK_2R_n(\alpha)$ on obtient :

$$P(A_{n,\beta}) \leq 2^{-xC_n(\beta)-(1-x)nK_2R_n(\alpha)+n\varepsilon_n} = 2^{-xC_n(\beta)}2^{-n(K_2R_n(\alpha)(1-x)-\varepsilon_n)},$$

$$\text{et ainsi} \sum_{\beta \,:C_n(\beta)>nK_2R_n(\alpha)} P(A_{n,\beta}) \leq b_2 2^{-n(R_n(\alpha)(K_2(1-x))-\varepsilon_n)}.$$

Cette dernière borne est le terme général d'une série convergente si on peut trouver une suite ε_n telle que :

$$\varepsilon_n \leq K_2(1-x)R_n(\alpha) - (1+\nu)\frac{\log n}{n}. \tag{4.18}$$

Pour les termes de (4.17) satisfaisant $D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta) > K_3 R_n(\alpha)$ on obtient par (4.4) et (4.6) la borne suivante :

$$P_\alpha(\{\mathcal{L}_\alpha(N^n, Y^n)2^{-n\varepsilon_n} \leq \mathcal{L}_\beta(N^n, Y^n)2^{-C_n(\beta)}\}) \leq \min\{2^{n\varepsilon_n - C_n(\beta)}, 2^{\frac{n\varepsilon_n - C_n(\beta)}{2} - nK_3R_n(\alpha)}\}.$$

D'après l'inégalité $\min\{y, z\} \leq y^u z^{1-u}$, vraie pour $u < 1$ appliquée au membre de droite de (4.17) en prenant $u = 2x - 1$, $0 < x < 1$ donne la borne supérieure :

$$2^{(n\varepsilon_n - C_n(\beta))(2x-1)} \cdot 2^{(n\varepsilon_n - C_n(\beta) - 2nK_3R_n(\alpha))(1-x)}. \tag{4.19}$$

L'exposant de (4.19) peut s'écrire :

$$\begin{aligned}
&(n\varepsilon_n - C_n(\beta))(2x-1) + (n\varepsilon_n - C_n(\beta) - 2nK_3R_n(\alpha))(1-x) \\
&= -C_n(\beta)((2x-1) + (1-x)) + n\varepsilon_n((2x-1) + (1-x)) - 2nK_3R_n(\alpha)(1-x) \\
&= -xC_n(\beta) - n(2K_3R_n(\alpha)(1-x) - \varepsilon_n x).
\end{aligned}$$

Si on choisit x tel que $\sum_{\beta \in \Gamma_n} 2^{-x.C_n(\beta)} \leq b_2$,ce qui est possible d'après la condition $G1$, on obtient :

$$\sum_{\beta\, :d(\alpha,\beta) \geq K_3R_n(\alpha)} P(\{\mathcal{L}_\alpha(N^n, Y^n)2^{-n\varepsilon_n} \leq \mathcal{L}_\beta(N^n, Y^n)2^{-C_n(\beta)}\}) \leq b_2 2^{-n((2K_3R_n(\alpha)(1-x)) - x\varepsilon_n)}.$$

Une condition suffisante pour que cette borne soit le terme général d'une série convergente est qu'il existe $\nu > 0$ tel que :

$$\varepsilon_n \leq 2K_3\frac{1-x}{x}R_n(\alpha) - \frac{1+\nu}{x}\frac{\log n}{n}. \tag{4.20}$$

Il reste alors à montrer qu'on peut trouver ε_n satisfaisant (4.20), (4.18) et (4.16). Par (3.25), (3.26) et (3.24), on peut toujours trouver des constantes K_2 et K_3 telles qu'on puisse trouver un tel ε_n. Pour ce choix, les évènements A_n et B_n sont tous les deux réalisés pour n assez grand, donc l'estimateur de complexité minimale n'appartient pas à Δ_n. Nous avons donc établi que

$$D_H^2(\mathcal{L}_\alpha, \mathcal{L}_{\hat{\alpha}_n}) < K_3 R_n(\alpha). \tag{4.21}$$

La relation liant la distance entre les vraisemblances et la distance entre les fonctions d'intensité fournie par le lemme 4.4 permet donc d'affirmer que $1 - e^{-D_0 d_H^2(\alpha, \hat{\alpha}_n)} < K_3 R_n(\alpha)$, soit encore $d_H^2(\alpha, \hat{\alpha}_n) \leq -\frac{1}{D_0}\ln(1 - K_3 R_n(\alpha))$. Puisque $R_n(\alpha) = o(1)$, on peut conclure que :

$$d_H^2(\alpha, \hat{\alpha}_n) \leq \frac{K_3}{D_0} R_n(\alpha).$$

L'estimateur de complexité minimale doit donc satisfaire : P_α-presque-sûrement pour n suffisamment grand,

$$C_n(\hat{\alpha}_n) < nK_2R_n(\alpha) \text{ et } d_H^2(\alpha, \hat{\alpha}_n) < \frac{K_3}{D_0}R_n(\alpha).$$

□

Dans le cadre du théorème 4.14, le prochain résultat présente les vitesses de convergence dans le cadre des familles trigonométriques, polynomiales et splines.

Corollaire 4.15. *Supposons que la fonction inconnue appartienne à l'espace de Sobolev W_2^r des fonctions α sur $[0,1]$, telles que $\alpha^{(r-1)}$ est absolument continue, et $\| \alpha^{(r)} \|_2 < \infty$, avec $r \geq 2$ dans le cas d'une famille Γ_n de fonctions polynomiales, $1 \leq r \leq s$ dans le cas des fonctions splines de degré $s-1$, et $r \geq 1$ dans le cas trigonométrique. L'estimateur de complexité minimale satisfait alors :*

$$d_H^2(\alpha, \hat{\alpha}_n) = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2r}{2r+1}}\right)$$

P_α presque-sûrement.

Preuve : Dans le cas de suites de familles paramétriques, pour des bases de fonctions polynomiales, trigonométriques ou splines, la vitesse d'approximation est liée à la dimension de la famille par la relation :

$$\| \alpha - \tilde{\alpha}_n \|_2^2 = O\left(m(n)\right)^{-2r}).$$

En optimisant μ_n pour cette valeur dans 3.28 on obtient comme vitesse optimale pour la dimension :

$$m_n^* = O\left(\frac{n}{\log n}\right)^{\frac{1}{2r+1}}. \tag{4.22}$$

Pour cette valeur, l'indice de résolvabilité vérifie :

$$R_n(\alpha) = O\left(\frac{\log n}{n}\right)^{\frac{2r}{2r+1}}.$$

□

Remarque : La vitesse $n^{-\frac{2r}{2r+1}}$ est optimale pour l'erreur quadratique intégrée dans le cadre de l'estimation d'une densité de probabilité possédant des normes de Sobolev bornées (voir [31],[16], et [14] chap. 2, p. 38). Barron et Sheu[10] montrent que cette borne est également la vitesse minimax optimale pour la

distance de Kullback-Leibler pour la classe des log-densités possédant une norme de Sobolev bornée. Le résultat du corollaire 4.15 et le corollaire 4.19 montrent qu'on obtient des vitesses de convergence voisines des vitesses optimales, pour la distance de Hellinger, mais également pour la distance de Kullback-Leibler entre les fonctions d'intensité.

◇

Remarque : La méthode d'estimation par des fonctions splines de degré $s-1$ ne peut fournir des vitesses d'approximation supérieures à m^{-2s}, on ne peut atteindre la vitesse $(\frac{n}{\log n})^{-2r/(2r+1)}$ que si $s \geq r$. Par conséquent, pour la méthode usuelle basée sur des splines cubiques, la vitesse optimale est $(\frac{n}{\log n})^{-8/9}$. Le cas de l'histogramme est un cas particulier des splines, pour lequel le degré s est pris égal à 1. Dans ce cas, par la remarque précédente, le corollaire 4.15 fournit une vitesse en $(\frac{n}{\log n})^{-\frac{2}{3}}$.

◇

4.5 Etude de modèles spécifiques.

Les conditions imposées par le théorème précédent sur le processus Y sont relativement fortes, mais elles peuvent être affaiblies en faisant des hypothèses sur le processus N. En outre ces hypothèses vont également conduire à des résultats en termes de distances plus fortes que la distance de Hellinger. On s'intéresse dans cette section à deux applications fréquentes du modèle multiplicatif, les modèles de durées de vie avec censure et des processus de Poisson non-homogènes.

La proposition suivante est obtenue en utilisant l'inégalité 4.6 :

Proposition 4.16.

a) S'il existe une constante K_1 telle que $\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta} < K_1$ P_0 presque-sûrement alors :

$$d(\alpha \parallel \beta) \leq 4K_1 D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta) \tag{4.23}$$

b) S'il existe une constante K_2 telle que $E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right) < K_2$, alors :

$$\frac{d^2(\alpha \parallel \beta)}{4} \leq K_2 D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta). \tag{4.24}$$

Preuve : Afin de montrer le corollaire, il suffit d'établir des relations entre la distance de Hellinger entre les vraisemblances $D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta)$, et la distance Kullback-Leibler relative $d(\alpha \parallel \beta)$.
a) Si on note $p = \mathcal{L}_\alpha$ et $q = \mathcal{L}_\beta$, la relation

$$plog\frac{p}{q} + q - p \leq \psi(p/q)(\sqrt{p} - \sqrt{q})^2 \tag{4.25}$$

permet en posant $u = \frac{p}{q}$ de trouver $\psi(u)$ tel que $h(u) \geq 0$ où $h(u) = \psi(u)(\sqrt{u} - 1)^2 - u\ln(u) + u - 1$. En utilisant l'inégalité $-\ln(u) \geq 1 - u$ on minore h par une fonction qui est nulle pour $\psi(u) = (\sqrt{u} + 1)^2$. En intégrant le membre de gauche de (4.25) on obtient :

$$\int \left[\mathcal{L}_\alpha \log \frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta} + \mathcal{L}_\beta - \mathcal{L}_\alpha\right] dP_0$$

$$= \int \left[\mathcal{L}_\alpha \log \frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right] dP_0 = \int \left[\log \frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right] dP_\alpha = d(\alpha \parallel \beta).$$

En intégrant le membre de droite de (4.25), on obtient :

$$\int \psi(\mathcal{L}_\alpha/\mathcal{L}_\beta)\left(\sqrt{\mathcal{L}_\alpha} - \sqrt{\mathcal{L}_\beta}\right)^2 dP_0. \tag{4.26}$$

Par hypothèse sur β on a $\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta} < K_1\ P_0$ presque-sûrement et on peut choisir $K_1 > 1$, de manière à ce que l'expression (4.26) soit inférieure à

$$(\sqrt{K_1} + \sqrt{K_1})^2 D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta).$$

On obtient ainsi l'inégalité :

$$d(\alpha \parallel \beta) \leq 4K_1 D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta) \tag{4.27}$$

b) Pour la seconde inégalité on a :

$$\log(\sqrt{\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}}) \leq \frac{\sqrt{\mathcal{L}_\alpha} - \sqrt{\mathcal{L}_\beta}}{\sqrt{\mathcal{L}_\beta}}$$

$$E_\alpha\left(\log\sqrt{\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}}\right) \leq \int \left[\frac{\mathcal{L}_\alpha}{\sqrt{\mathcal{L}_\beta}}(\sqrt{\mathcal{L}_\alpha} - \sqrt{\mathcal{L}_\beta})\right] dP_0$$

Par l'inégalité de Schwarz :

$$\leq E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right)^{\frac{1}{2}} \left(\int_\Omega \left(\sqrt{\mathcal{L}_\alpha} - \sqrt{\mathcal{L}_\beta}\right)^2 dP_0\right)^{\frac{1}{2}}$$

Cette inégalité devient :

$$\frac{d^2(\alpha \parallel \beta)}{4} \leq K_2 D_H^2(\mathcal{L}_\alpha, \mathcal{L}_\beta).$$

□

4.5.1 Processus de Poisson non-homogènes.

Les processus de Poisson peuvent être considérés comme des processus de Aalen pour lesquels $Y(s) \equiv 1$. L'importance de ces processus s'explique entre autres par leur simplicité, les propriétés particulières des trajectoires (processus à accroissements indépendants et stationnaires), et le fait que les processus de Poisson permettent de modéliser un grand nombre de phénomènes physiques. De nombreux auteurs traitent des problèmes statistiques des processus de Poisson, on peut citer entre autres [34], [26] plus récemment [69, 70]. Dans cette section, nous présentons plusieurs exemples de processus de Poisson, et nous donnons des résultats d'inférence plus forts que ceux du théorème 4.14 au sens des distances considérées.

Corollaire 4.17. *Supposons les hypothèses du théorème 4.12 vérifiées. Dans le cas des processus de Poisson, pour n suffisamment grand, on obtient P_α presque-sûrement :*

$$d(\alpha \parallel \hat{\alpha}_n) = O\left(\sqrt{R_n(\alpha)}\right).$$

Preuve : Par le théorème 4.12, l'estimateur converge en norme L_1. On peut donc restreindre la recherche de l'estimateur à l'ensemble des fonctions β telles que $\| \alpha - \beta \|_1 < 1$. D'autre part, dans le cas des processus de Poisson, $Y(s) \equiv 1$. On pose $\lambda = \int_0^T \alpha(s)ds$ On a donc :

$$E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right) \leq \exp\left(D_Y \parallel \alpha - \beta \parallel_1\right) \left(\prod_{i=1}^{\overline{N}(T)} \frac{\alpha}{\beta}(\tau_i)\right).$$

Si les familles de fonction candidates sont supposées minorées uniformément par une constante f_0, on obtient en utilisant la proposition 4.16 :

$$\begin{aligned} E_\alpha\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}\right) &\leq e^{D_Y} E_\alpha\left(\prod_{i=1}^{\overline{N}(T)} \frac{\alpha}{\beta}(\tau_i)\right) \\ &\leq E_\alpha\left(\frac{\sup \alpha}{f_0}\right)^{\overline{N}(T)} \\ &\leq \exp\left(\lambda\left(\frac{\sup \alpha}{f_0} - 1\right)\right), \end{aligned}$$

où la dernière inégalité provient du fait que N suit une loi de Poisson. En reprenant la preuve du théorème 4.14 jusqu'à la relation 4.21 et en utilisant d'autre part l'inégalité 4.24, on obtient le résultat annoncé.

□

4.5.2 Durées de vie.

Modèle de durée simple. On suppose que $X_1, \ldots, X_n$ sont des variables aléatoires réelles, non négatives indépendantes et identiquement distribuées, de fonction de répartition F telle que $F(T) < 1$ et de densité $f > 0$ sur $[0, T]$. En analyse des durées de vie, on s'intéresse à la fonction de :hasard définie par $\alpha(t) = f(t)/(1 - F(t))$ pour $0 \leq t \leq T$. La condition $F(T) < 1$ est suffisante pour que α soit intégrable sur $[0, T]$. On suppose de plus que f est continue à gauche et admet des limites à droite. Soit le processus ponctuel multivarié N de composantes $N_i(t) = \mathbf{I}_{\{X_i \leq t\}}$ pour $i = 1, \ldots, n$, et la famille de tribus $\mathcal{F}_t$. F définit une mesure unique sur $\mathcal{F}$, et N_i a une intensité stochastique $\alpha_i(t)\mathbf{I}_{\{X_i \geq t\}}$ par rapport à $\mathcal{F}$ et P. Le lemme suivant est démontré dans [2].

Lemme 4.18.

$$M_i(t) = \mathbf{I}_{\{X_i \leq t\}} - \int_0^t \alpha_i(u)\mathbf{I}_{\{X_i \geq u\}} du \quad i = 1, \ldots, n$$

sont des martingales orthogonales de carré intégrable par rapport à $\mathcal{F}$.

Posons $\overline{N}_n(t) = \sum_{i=1}^n N_i(t)$ et $\overline{\lambda}(t) = \sum_{i=1}^n \lambda_i(t) = \alpha(t)(n - \overline{N}_n(t^-))$. Le processus $\overline{N}_n$ est un processus de comptage d'intensité $\overline{\lambda}(t)$ par rapport à P et $\mathcal{F}_t$. le couple $(\overline{N}_n, \overline{Y}_n)$ est une statistique exhaustive pour l'estimation de α. On a ainsi un modèle multiplicatif composé du processus de comptage $\overline{N}_n$, d'une fonction déterministe α, du processus $\overline{Y}_n(t) = n - \overline{N}_n(t^-)$ et $\mathcal{F}_t$.

Modèle de durées avec censure. On suppose que $X_1, \ldots, X_n$ sont n variables aléatoires positives i.i.d. représentant les durées de vie de n individus, et $C_1, C_2, \ldots, C_n$, n variables aléatoires i.i.d. positives et indépendantes des $X_i, i = 1, \ldots, n$ représentant des dates de censure. Les variables observées sont des triplets i.i.d. (Z_i, δ_i, Y_i) où $Z_i = \min(X_i, C_i)$ représente la durée pendant laquelle le i-ème individu est resté en observation, et $\delta_i = \mathbf{I}_{\{X_i \leq C_i\}}$, δ_i étant l'indicateur de censure. Le processus $Y_i(t) = \mathbf{I}_{\{Z_i \geq t\}}$ indique si l'individu est encore présent (ni mort, ni censuré) à l'instant t. Le processus N_i associé au i-ème individu est alors défini par :

$$N_i = \mathbf{I}_{\{Z_i \leq t, \delta_i = 1\}}.$$

Le processus $\overline{y}_n$ compte le nombre d'individus à risque (ni morts, ni censurés) à l'instant t, et $N_n(t)$ le nombre d'individus dont la mort a été observée dans l'intervalle de temps $[0, t]$.

Corollaire 4.19. *On suppose que les hypothèses du théorème 4.12 sont vérifiées. Dans le cas des modèles de durées de vie,*

a) Si les fonctions de Γ_n *sont uniformément bornées inférieurement, alors* P_α *presque-sûrement*

$$d(\alpha \parallel \hat{\alpha}_n) = O(R_n(\alpha)).$$

b) Pour des familles de fonctions uniformément bornées supérieurement et inférieurement, on a P_α *presque-sûrement :*

$$\parallel \alpha - \hat{\alpha}_n \parallel_2^2 = O(R_n(\alpha)).$$

Preuve : Par le théorème 4.12, on peut se restreindre aux fonctions qui vérifient $\parallel \alpha - \beta \parallel_1 < 1$.

$$\begin{aligned} \frac{\mathcal{L}_\alpha}{\mathcal{L}_\beta}(N,Y) &\leq \exp\left(D_Y \parallel \alpha - \beta \parallel_1 + \sup_{0\leq s\leq T} \ln \frac{\alpha}{\beta}(s) N(T)\right) \\ &\leq \exp\left(D_Y + \ln \frac{\sup \alpha}{f_0} N(T)\right). \end{aligned}$$

La condition de la partie a) de la proposition 4.16 est donc vérifiée. En reprenant la démonstration du théorème 4.14, jusqu'à l'inégalité 4.21, on obtient le résultat annoncé. Le second résultat découle directement du lemme 4.8

□

Chapitre 5

Normalité asymptotique et test exponentiels.

5.1 Normalité asymptotique.

Les estimateurs non paramétriques de la fonction d'intensité sont le plus souvent asymptotiquement normaux. Pour la méthode du noyau par exemple, [60] démontre que si la fonction α est continue, et si $nb_n \to \infty$ et $n^{1/3}b_n$ tend vers 0, on a convergence de $(nb_n)^{1/2}(\hat{\alpha}_n(t) - \alpha(t))$ vers une loi normale dont la variance dépend du noyau. [42] utilise la méthode des tamis, et obtient que $n^{1/2}\int_0^t(\hat{\alpha}_n(s) - \alpha(s))ds$ converge en loi vers une martingale gaussienne de fonction de variance :

$$W_\alpha(t) = \int_0^t \alpha(s)/\theta_\alpha(s)ds, \quad t \in [0,T],$$

où $\theta_\alpha(s) = E_\alpha(Y(s))$. Néanmoins, Karr suppose que la fonction inconnue appartient à l'un des tamis, ce qui serait équivalent dans notre approche à l'existence d'un entier n_0 tel que $\alpha \in \Gamma_{n_0}$. McKeague (1987) montre en utilisant également la méthode des tamis que $n^{1/2}\int_0^t(\hat{\alpha}_n(s)-\alpha(s))ds$ converge dans le sens des lois cylindriques de dimension finie vers une martingale gaussienne centrée de fonction de variance :

$$G(t) = \int_0^t \left[EY^3(s)/(EY^2(s))\right]\alpha(s)ds.$$

Ce résultat permet également d'établir des intervalles de confiance approchés pour l'intensité intégrée, puisqu'on peut estimer $G(t)$ par l'estimateur convergent :

$$\hat{G}(t) = \int_0^t \left[\left(\sum_{i=1}^n Y_i^3(s) \right) / \left(\sum_{i=1}^n Y_i^2(s) \right) \right] \hat{\alpha}_n(s) ds.$$

Nous démontrons un résultat analogue à celui de Karr sans faire d'hypothèse particulière sur la famille contenant α.

Soit Γ_n le sous-ensemble de fonctions positives généré par les $m(n)+1$ fonctions de base $\phi_j, j = 0, \ldots, m(n)$, et dont les coefficients sont tronqués à une précision δ_n. Soit $\alpha^*(t;\cdot)$, la fonction minimisant la fonctionnelle :

$$g(t;\beta) = \int_0^t \beta(s) \overline{Y}_n(s) ds - \int_0^t \ln(\beta(s)) d\overline{N}_n(s),$$

dans l'espace continu $\widetilde{\Gamma}_n$, et $\alpha^*(t;\delta_n;\cdot)$, la fonction obtenue en tronquant les coefficients de $\alpha^*(t;\cdot)$ à la précision δ_n.

Dans ce chapitre nous supposerons pour simplifier les preuves, que les fonctions de complexité sont de la forme $C_n(\beta) = C - m(n) \log \delta_n$ où δ_n représente la précision à laquelle les coefficients sont arrondis, $m(n)$ la dimension de l'espace Γ_n, et C la longueur fixe d'un code pour le vecteur des parties entières des coefficients de β.

Théorème 5.1. *Supposons la condition G4 vérifiée. Soit $\hat{\alpha}_n(t;\cdot)$ l'estimateur de complexité minimale obtenu en minimisant le critère dans l'intervalle $[0,t]$. Pour une précision des coefficients $\delta_n = o(n^{-\frac{1}{2}} a_n^{-\frac{5}{2}})$, le processus $n^{1/2} \int_0^t (\hat{\alpha}_n(t;s) - \alpha(s))\, ds$ converge en distribution vers une martingale gaussienne de fonction de variance :*

$$V_\alpha(t) = \int_0^t \alpha(s)/\theta_{\alpha(s)} ds, \quad t \in [0,T].$$

Preuve : On introduit la fonction $\alpha^*(t;\cdot)$ solution de :

$$Arg \min_{\beta \in \Gamma_n} g(t;\beta) = \int_0^t \beta(s) \overline{Y}_n(s) ds - \int_0^t \ln(\beta(s)) d\overline{N}_n(s).$$

Soient α_j^* ses coefficients. L'optimum $\alpha^*(t;\cdot)$ est caractérisé par la relation :

$$\int_0^t \phi_j(s) \overline{Y}_n(s) ds = \int_0^t \frac{\phi_j(s)}{\alpha^*(t;s)} d\overline{N}_n(s).$$

En effectuant une combinaison linéaire de coefficients α_j^* pour j compris entre 0 et m, puis en intervertissant les signes somme et intégrale, on obtient :

$$\int_0^t \alpha^*(t;s)\overline{Y}_n(s)ds = \int_0^t d\overline{N}_n(s) = \overline{N}_n(t). \tag{5.1}$$

De cette formule, on déduit :

$$\overline{N}_n(t) - \int_0^t \alpha(s)\overline{Y}_n(s)ds = \int_0^t (\alpha^*(t;s) - \alpha(s))\overline{Y}_n(s)ds.$$

D'autre part, on a également (cf. théorème A.3) :

$$\left(n^{-1/2}\left[\overline{N}_n(t) - \int_0^t \alpha(s)\overline{Y}_n(s)ds\right]\right)_{0\le t\le T} \to_d (M_t(\alpha))_{0\le t\le T}, \tag{5.2}$$

où $M_t(\alpha)$ est une martingale gaussienne de variation quadratique :

$$Q_\alpha(t) = \lim_n n^{-1} < M_n >_t = \int_0^t \theta_\alpha(s)\alpha(s)ds \tag{5.3}$$

En utilisant la relation (5.1) on obtient :

$$\begin{aligned} n^{-1/2}\left[\overline{N}_n(t) - \int_0^t \alpha(s)\overline{Y}_n(s)ds\right] &= n^{-1/2}\left[\int_0^t \alpha^*(t;s)\overline{Y}_n(s)ds - \int_0^t \alpha(s)\overline{Y}_n(s)ds\right] \\ &= n^{-1/2}\left[\int_0^t (\hat{\alpha}_n(t;s) - \alpha(s))\overline{Y}_n(s)ds\right] \\ &+ n^{-1/2}\left[\int_0^t (\alpha^*(t;s) - \hat{\alpha}_n(t;s))\overline{Y}_n(s)ds\right]. \end{aligned} \tag{5.4}$$

Sous les hypothèses du théorème, on vérifie que le dernier terme de 5.4 converge vers 0. Considérons la formule des accroissements finis pour les fonctionnelles appliquée au second ordre au point $\alpha^*(t;\cdot)$ à :

$$g(t;\beta) = \int_0^t \beta(s)\overline{Y}_n(s)ds - \int_0^t \ln(\beta(s))d\overline{N}_n(s).$$

En posant $\gamma(t;s) = \alpha^*(t;s) + \theta(t)(\beta(s) - \alpha^*(t;s))$, on obtient :

$$g(t;\beta) = g(t;\alpha^*(t;\cdot)) + \int_0^t \frac{(\beta(s) - \alpha^*(t;s))^2}{\gamma^2(t;s)} d\overline{N}_n(s).$$

Soit $\alpha^*(t;\delta_n;\cdot)$ la fonction (de Γ_n) obtenue en tronquant $\alpha^*(t;\cdot)$ à la précision δ_n. Par définition de $\alpha^*(t;\cdot)$ et de $\hat{\alpha}_n(t;\cdot)$ respectivement, on a les deux inégalités :

$$g(t,\alpha^*(t;\cdot)) \le g(t,\hat{\alpha}_n(t;\cdot)) \le g(t,\alpha^*(t;\delta_n;\cdot)).$$

La première provient du fait que $\alpha^*(t;\cdot)$ réalise le minimum de la fonctionnelle g, la seconde provient du fait que $\hat{\alpha}_n(t;\cdot)$ réalise le minimum de g au sein de Γ_n. Cette dernière inégalité provient du fait que les fonctions de Γ_n ont la même complexité. De ces inégalités, on déduit :

$$g(t,\hat{\alpha}_n(t;\cdot)) - g(t,\alpha^*(t;\cdot)) \leq g(t,\alpha^*(t;\delta_n;\cdot)) - g(t,\alpha^*(t;\cdot)),$$

soit encore :

$$\int_0^t \frac{(\hat{\alpha}_n(t;s)-\alpha^*(t;s))^2}{\gamma_1^2(t,s)} d\overline{N}_n(s) \leq \int_0^t \frac{(\alpha^*(t;\delta_n,s)-\alpha^*(t;s))^2}{\gamma_2^2(t,s)} d\overline{N}_n(s),$$

où γ_1 et γ_2 sont définies respectivement par $\gamma_1(t;\cdot) = \alpha^*(t;\cdot)+\theta_1(t)(\hat{\alpha}_n(t;\cdot)-\alpha^*(t;\cdot))$ et $\gamma_2(t;\cdot) = \alpha^*(t;\cdot)+\theta_2(t)(\hat{\alpha}_n(t;\delta_n;\cdot)-\alpha^*(t;\cdot))$, $\theta_1(t)$ et $\theta_2(t)$ étant des fonctions comprises entre 0 et 1. En notant respectivement $\alpha_j^*(t;\delta_n)$ et $\alpha_j^*(t)$ les coefficients de $\alpha^*(t;\delta_n,\cdot)$ et de $\alpha^*(t;\cdot)$, on obtient :

$$\begin{aligned}\int_0^t \frac{(\alpha^*(t;\delta_n,s)-\alpha^*(t;s))^2}{\gamma_2^2(t,s)} d\overline{N}_n(s) &= \sum_{\tau_i\leq t} \frac{1}{\gamma_2^2(t;\tau_i)} \left(\sum_{j=0}^m (\alpha_j^*(t;\delta_n)-\alpha_j^*(t))\phi_j(\tau_i)\right)^2 \\ &\leq \delta_n^2 \sum_{\tau_i\leq t} \frac{1}{\gamma_2^2(t;\tau_i)} \left(\sum_{j=0}^m \phi_j^2(\tau_i) + 2\sum_{j=0}^m \phi_j(\tau_i)\sum_{l>j}\phi_l(\tau_i)\right).\end{aligned}$$

Les fonctions apparaissant dans la parenthèse sont des fonctions de Γ_n, dont les coefficients de la fonction de base ϕ_j sont respectivement $\phi_j(\tau_i)$ et $\sum_{l>j}\phi_l(\tau_i)$. Compte tenu des hypothèses de domination des fonctions de Γ_n, on obtient finalement pour tout $t\in[0,T]$, qu'il existe une constante K telle que :

$$\int_0^t \frac{(\hat{\alpha}_n(t;s)-\alpha^*(t;s))^2}{\gamma_1^2(t,s)} d\overline{N}_n(s) \leq K\delta_n^2 {a_n}^3 \overline{N}_n(t).$$

En multipliant les deux membres par n^{-1} et en utilisant la loi forte des grands nombres, on obtient :

$$\int_0^t \frac{(\hat{\alpha}_n(t;s)-\alpha^*(t;s))^2}{\gamma_1^2(t,s)} \alpha(s)\theta_\alpha(s) ds = O\left(\delta_n^2 {a_n}^3\right).$$

Soit en tenant compte encore une fois de la domination des fonctions de Γ_n :

$$\int_0^t (\hat{\alpha}_n(t;s)-\alpha^*(t;s))^2 ds = O\left(\delta_n^2 {a_n}^5\right). \tag{5.5}$$

On obtient finalement :

$$\begin{aligned} n^{-\frac{1}{2}} \int_0^t (\alpha^*(t;s)-\hat{\alpha}_n(t;s))\overline{Y}_n(s) ds &\leq n^{\frac{1}{2}} \parallel \alpha^*(t;\cdot)-\hat{\alpha}_n(t;\cdot) \parallel_2 \\ &= O\left(n^{\frac{1}{2}}\delta_n {a_n}^{\frac{5}{2}}\right)\end{aligned}$$

Il suffit alors de prendre $\delta_n = o(n^{-\frac{1}{2}} a_n^{-\frac{5}{2}})$ pour que ce terme converge vers 0. Ainsi par (5.2) on obtient :

$$n^{1/2} \int_0^t (\hat{\alpha}_n(t;s) - \alpha(s)) \frac{\overline{Y}_n(s)}{n} ds \to_d M_\alpha(t), \ \ 0 \leq t \leq T. \tag{5.6}$$

Par la loi des grands nombres, pour n suffisamment grand, le premier terme est de l'ordre de :

$$n^{1/2} \int_0^t (\hat{\alpha}_n(t;s) - \alpha(s)) \theta_\alpha(s) ds,$$

et ainsi

$$n^{1/2} \int_0^t (\hat{\alpha}_n(t;s) - \alpha(s)) \theta_\alpha(s) ds \to_d M_\alpha(t), \ \ 0 \leq t \leq T,$$

où $M_\alpha(t)$ est la martingale de variance quadratique définie dans (5.3). On en déduit que

$$n^{1/2} \int_0^{\cdot} (\hat{\alpha}_n(t;s) - \alpha(s)) ds \to_d \int_0^{\cdot} \theta_\alpha(s)^{-1} dM_\alpha(s),$$

ces convergences étant au sens des lois cylindriques. Cette dernière martingale a pour variation quadratique :

$$\int_0^t \frac{1}{\theta_\alpha^2} \theta_\alpha(s) \alpha(s) ds = V_\alpha(t).$$

□

Le théorème précédent est basé sur l'estimateur obtenu en minimisant la vraisemblance sur l'intervalle de temps $[0,t]$. La normalité asymptotique est obtenue pour l'estimateur séquentiel $\hat{\alpha}_n(t;\cdot)$. Ce théorème est valable pour tout type de famille candidate, et ne fait aucune hypothèse particulière sur la fonction inconnue. Néanmoins, toute application pratique nécessiterait le calcul de l'estimateur à chaque instant. Le prochain résultat permet de s'affranchir de cette limitation. Il est basé sur l'estimateur obtenu dans une famille d'histogrammes et exploite une propriété particulière de cette famille : l'estimateur obtenu sur un intervalle de temps donné ne dépend pas des évènements extérieurs à l'intervalle. En considérant une famille candidate d'histogrammes à intervalles de longueurs identiques dont le nombre est choisi de manière à optimiser l'indice de résolvabilité, on obtient un résultat de convergence gaussienne pour l'estimateur défini sur tout l'intervalle. Un résultat comparable pour l'intensité, a été obtenu par [48]. Les auteurs montrent que pour toute suite de points $s_1, \ldots, s_p$, $p \in \mathrm{N}^*$ de l'intervalle $[0,1]$, pour

une vitesse $m(n)$ inférieure ou égale à $n^{1/2}$, et sous des conditions appropriées, la distribution du vecteur

$$(n/m(n))^{1/2}(\hat{\alpha}_n(s_1) - \alpha(s_1), \ldots, \hat{\alpha}_n(s_p) - \alpha(s_p)),$$

est asymptotiquement gaussienne, de variance $I\sigma$, où I est la matrice unité et σ le vecteur de composantes $\sigma_k = \alpha(s_k)/\theta_\alpha(s_k)$, $k = 1, \ldots, p$.

Théorème 5.2. *Sous les conditions du théorème précédent, soit $\hat{\alpha}_n$ l'estimateur de α dans la famille des histogrammes définis sur des classes de longueurs identiques dont le nombre est choisi de manière à optimiser l'indice de résolvabilité, i.e. $m^*(n) = O\,(n/\log n)^{\frac{1}{3}}$. Le processus*

$$n^{1/2} \int_0^t (\hat{\alpha}_n(s) - \alpha(s))ds$$

converge en distribution vers une martingale gaussienne de fonction de variance définie pour tout $t \in [0, T]$ par :

$$V_\alpha(t) = \int_0^t \frac{\alpha(s)}{\theta_\alpha(s)} ds.$$

Preuve : Soit pour tout $t \in [0, T]$, la fonction $\alpha^*(t; \cdot)$ solution de :

$$Arg \min_{\beta \in \Gamma_n} g(t; \beta) = \int_0^t \beta(s)\overline{Y}_n(s)ds - \int_0^t \ln(\beta(s))d\overline{N}_n(s).$$

Les fonctions $\alpha^*(t; \cdot)$ et $\alpha^*(T; \cdot)$ permettent d'obtenir la décomposition :

$$\begin{aligned}\int_0^t (\hat{\alpha}_n(t; s) - \alpha(s))\overline{Y}_n(s)ds &= \int_0^t (\hat{\alpha}_n(t; s) - \alpha^*(t; s))\overline{Y}_n(s)ds + \int_0^t (\alpha^*(t; s) - \alpha^*(T; s))\overline{Y}_n(s)ds \\ &+ \int_0^t (\alpha^*(T; s) - \hat{\alpha}_n(s))\overline{Y}_n(s)ds + \int_0^t (\hat{\alpha}_n(s) - \alpha(s))\overline{Y}_n(s)ds.\end{aligned}$$

D'après (5.5), on a pour tout $t \in [0, T]$,

$$n^{-\frac{1}{2}} \int_0^t (\hat{\alpha}_n(t; s) - \alpha^*(t; s))\overline{Y}_n(s)ds = O\left(\delta_n n^{\frac{1}{2}+2q}\right).$$

On obtient ainsi :

$$\begin{aligned}n^{-\frac{1}{2}} \int_0^t (\hat{\alpha}_n(t; s) - \alpha(s))\overline{Y}_n(s)ds &= n^{-\frac{1}{2}} \int_0^t (\alpha^*(t; s) - \alpha^*(T; s))\overline{Y}_n(s)ds \\ &+ n^{-\frac{1}{2}} \int_0^t (\hat{\alpha}_n(s) - \alpha(s))\overline{Y}_n(s)ds + O\left(\delta_n n^{\frac{1}{2}+5q}\right).\end{aligned}$$

Dans la preuve du théorème précédent, on a obtenu en (5.6) que

$$n^{\frac{1}{2}} \int_0^t (\hat{\alpha}_n(t;s) - \alpha(s)) \frac{\overline{Y}_n(s)}{n} ds \to_d M_\alpha(t), \quad 0 \leq t \leq T.$$

Pour établir le résultat, il suffit alors de montrer que sous les hypothèses du théorème, le dernier terme vérifie :

$$n^{-\frac{1}{2}} \int_0^t (\alpha^*(t;s) - \alpha^*(T;s)) \overline{Y}_n(s) ds \to 0.$$

Par la relation 5.1 on a $\int_0^t (\alpha^*(t;s) \overline{Y}_n(s) ds = \overline{N}_n(t)$ pour tout $t \in [0,T]$. On obtient en tenant compte de cette relation :

$$\int_0^t (\alpha^*(t;s) - \alpha^*(T;s)) \overline{Y}_n(s) ds = \overline{N}_n(t) - \overline{N}_n(T) + \int_t^T \alpha^*(T;s) \overline{Y}_n(s) ds. \tag{5.7}$$

Soient $t_1^n, \ldots, t_{m^*(n)}^n$, les noeuds de l'histogramme. Soient I_j^n, les $m^*(n)$ intervalles définissant l'histogramme, et $I_{j_0}^n = [t_{j_0^n - 1}, t_{j_0}^n]$, l'intervalle contenant t. On note dans la suite $\nu(I_j)$ le nombre de points de sauts observés dans l'intervalle I_j, et $\psi(I_j) = \int_{I_j} \overline{Y}_n(s) ds$. De la relation :

$$\alpha^*(T;s) = \sum_{j=1}^{m^*(n)} \frac{\nu_j}{\psi_j} \mathbf{I}_{I_j}(s),$$

on déduit aisément :

$$\begin{aligned} \int_t^T \alpha^*(T;s) \overline{Y}_n(s) ds &= \int_t^{t_{j_0^n}} \alpha^*(T;s) \overline{Y}_n(s) ds + \int_{t_{j_0}}^T \alpha^*(T;s) \overline{Y}_n(s) ds \\ &= \int_t^{t_{j_0}^n} \alpha^*(T;s) \overline{Y}_n(s) ds + \overline{N}_n(T) - \overline{N}_n(t_{j_0}^n) \end{aligned}$$

Par (5.7) on obtient :

$$\begin{aligned} \int_0^t (\alpha^*(t;s) - \alpha^*(T;s)) \overline{Y}_n(s) ds &= \int_t^{t_{j_0}^n} \alpha^*(T;s) \overline{Y}_n(s) ds + \overline{N}_n(t_{j_0}^n) - \overline{N}_n(t) \\ &= \frac{\nu(I_{j_0}^n)}{\psi(I_{j_0}^n)} \int_t^{t_{j_0}^n} \overline{Y}_n(s) ds + \overline{N}_n(t_{j_0}^n) - \overline{N}_n(t) \\ &= \frac{\nu(I_{j_0}^n)}{\psi(I_{j_0}^n)} \psi\left([t, t_{j_0}^n]\right) - \nu\left([t, t_{j_0}^n]\right). \end{aligned}$$

On obtient alors la borne supérieure :

$$\begin{aligned} \int_0^t (\alpha^*(t;s) - \alpha^*(T;s)) \overline{Y}_n(s) ds &\leq \nu(I_{j_0}^n) - \nu\left([t, t_{j_0}^n]\right) \\ &\leq \nu(I_{j_0}^n) \end{aligned}$$

Pour la borne inférieure, on obtient :

$$\begin{aligned}\int_0^t (\alpha^*(t;s) - \alpha^*(T;s))\overline{Y}_n(s)ds &\geq \nu\left([t, t_{j_0}^n]\right)\left(\frac{\nu(I_{j_0}^n)}{\psi(I_{j_0}^n)} - 1\right) \\ &\geq -\nu\left([t, t_{j_0}^n]\right) \\ &\geq -\nu(I_{j_0}^n).\end{aligned}$$

L'histogramme est un cas particulier de fonction spline. La vitesse de croissance optimale (au sens de l'indice de résolvabilité) de la dimension des espaces dans le cas d'une famille d'histogrammes (voir 4.22 dans la preuve du corollaire 4.15) à $m(n)$ noeuds équidistants est :

$$m^*(n) = \left(\frac{n}{\log n}\right)^{\frac{1}{3}}.$$

Pour conclure, on utilise cette vitesse pour montrer que $\lim_{n\to\infty} n^{-\frac{1}{2}}\nu(I_{j_0}^n) = 0$ en probabilité. Par l'inégalité de Markov, on a :

$$\begin{aligned}P_\alpha\left(n^{-\frac{1}{2}}\nu\left(I_{j_0}^n\right) > \varepsilon\right) &\leq \frac{E_\alpha\left(\nu\left(I_{j_0}^n\right)\right)}{n^{\frac{1}{2}}\varepsilon} \\ &\leq n^{-\frac{1}{2}}\varepsilon^{-1}\int_{I_{j_0}^n} \alpha(s)\theta_\alpha(s)ds \\ &= O\left(n^{-\frac{1}{2}}\varepsilon^{-1}(t_{j_0}^n - t_{j_0-1}^n)\right) \\ &= O\left(n^{-\frac{1}{2}}\varepsilon^{-1}\left(\frac{n}{\log n}\right)^{-\frac{1}{3}}\right) \\ &= O\left(n^{-\frac{5}{6}}\varepsilon^{-1}(\log n)^{\frac{1}{3}}\right).\end{aligned}$$

On peut conclure que $n^{-\frac{1}{2}}\nu\left(I_{j_0}^n\right)$ converge vers 0 en probabilité, et par conséquent, la suite de processus : $n^{1/2}\int_0^t(\hat{\alpha}_n(s) - \alpha(s))ds$ converge en distribution vers une martingale gaussienne de fonction de variance

$$V_\alpha(t) = \int_0^t \frac{\alpha(s)}{\theta_\alpha(s)}ds,$$

pour tout $t \in [0, T]$.

□

Remarque : Le résultat précédent permet de construire des procédures de tests et des intervalles de confiance pour l'intensité cumulée. En effet,

$$\hat{V}_\alpha(t) = \int_0^t \frac{\hat{\alpha}_n(s)}{\theta_\alpha(s)}ds,$$

est un estimateur convergent de $V_\alpha(t)$. Pour une taille d'échantillon suffisante

$$\frac{n^{1/2}}{[\hat{V}_\alpha(t)]^{\frac{1}{2}}} \int_0^t (\hat{\alpha}_n(s) - \alpha(s))ds$$

est approximativement de loi normale centrée réduite. Soit C_p défini par :

$$P(|N(0,1)| > C_p) \leq 1 - p,$$

les inégalités suivantes sont satisfaites avec une probabilité voisine de p :

$$\int_0^t \hat{\alpha}_n(s)ds - n^{-1/2}C_p\ [\hat{V}_\alpha(t)]^{\frac{1}{2}} \leq \int_0^t \alpha(s)ds \leq \int_0^t \hat{\alpha}_n(s)ds + n^{-1/2}C_p\ [\hat{V}_\alpha(t)]^{\frac{1}{2}}.$$

◇

5.2 Tests exponentiels.

La définition de l'estimateur de complexité minimale se prête très bien à la formulation de tests exponentiels consistants. Un test est dit consistant si les probabilités d'erreur convergent vers 0. Un test est dit exponentiellement consistant si les probabilités d'erreur sont bornées par 2^{-nc} pour une constante c donnée. On connaît des bornes exponentielles pour de nombreux tests dans le cas d'hypothèses simples pour des lois de probabilité H_0 : "$P = P_0$" contre H_1 : "$P = P_1$" (voir par exemple [27]). Dans le cas de familles paramétriques pour des lois absolument continues, et densités lisses, on pourra consulter [36], [7] et [17]. Pour des ensembles convexes avec alternatives définies par des capacités [38].

Dans cette section, nous étudions les propriétés de deux types de tests. Les risques de première et de seconde espèce seront notés respectivement r_1 et r_2, α et β désignant des fonctions candidates. Nous proposons des tests du type H_0^n :"$\alpha = \alpha_0^n$" contre H_1^n :"$\alpha \in A^n$", où A^n est un sous-ensemble de Γ_n. Le premier test que nous considérons est le test de Neyman-Pearson :

$$\begin{cases} H_0^n & : "\alpha = \alpha_0^n" \\ H_1^n & : "\alpha = \alpha_1^n" \end{cases}$$

Dans le cas de deux fonctions de Γ_n, on montre qu'on peut définir des tests consistants et même exponentiellement consistants, si certaines conditions liant la différence des complexités entre les deux fonctions $C_n(\alpha_0^n) - C_n(\alpha_1^n)$, et la distance $d_H^2(\alpha_1^n, \alpha_0^n)$, sont satisfaites. Si l'une des deux

fonctions n'est pas dans Γ_n, on montre qu'on obtient des résultats similaires en substituant au terme de complexité associée à la fonction de Γ, une suite croissante qui satisfait certaines contraintes. Le test classique de Neyman-Pearson est uniformément le plus puissant pour un niveau donné. Le premier test que nous étudions peut être assimilé à un test de Neyman-Pearson dans lequel le niveau ne serait pas arbitraire, mais égal à la différence de complexité entre les deux fonctions.

La seconde famille de test que nous étudions, est de la forme :

$$\begin{cases} H_0 : \text{“}\alpha = \alpha_0^n\text{”} \\ H_1 : \text{“}d_H^2(\alpha, \alpha_0^n) > r_n\text{”} \end{cases}$$

où r_n est une suite dépendant de la taille de l'échantillon. Dans toute la suite, on considère que (N, Y) est un processus de Aalen de fonction d'intensité α et on suppose en outre qu'il existe une constante D_0 telle que : P_0 presque-sûrement, pour tout $s \in [0, T]$, $Y(s) \geq D_0 > 0$.

Théorème 5.3. *Un test de l'hypothèse H_0 : “$\alpha = \alpha_0^n$” contre H_1 : “$\alpha = \alpha_1^n$” de région d'acceptation définie par $W = \{\mathcal{L}_{\alpha_0^n} 2^{-\epsilon_0^n} \geq \mathcal{L}_{\alpha_1} 2^{-\epsilon_1^n}\}$ possède des risques de première et de seconde espèce qui vérifient :*

$$\begin{aligned} -\frac{1}{n}\log r_1 &\geq \frac{1}{2n}(\epsilon_1^n - \epsilon_0^n) + 1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_1^n)} \\ -\frac{1}{n}\log r_2 &\geq \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n) + 1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_1^n)}. \end{aligned} \tag{5.8}$$

Preuve : La zone de rejet est définie par $\overline{W} = \{\mathcal{L}_{\alpha_0^n} 2^{-\epsilon_0^n} < \mathcal{L}_{\alpha_1^n} 2^{-\epsilon_1^n}\}$. Le risque de première espèce est donné par :

$$\begin{aligned} r_1 = P_{\alpha_0^n}\left(\overline{W}\right) &\leq P_{\alpha_0^n}\left(\mathcal{L}_{\alpha_0^n} 2^{-\epsilon_0^n} \leq \mathcal{L}_{\alpha_1^n} 2^{-\epsilon_1^n}\right) \\ &\leq 2^{\frac{1}{2}(\epsilon_0^n - \epsilon_1^n)} 2^{-n D_H^2(\mathcal{L}_{\alpha_0^n}, \mathcal{L}_{\alpha_1^n})} \\ &= 2^{-n\left(D_H^2(\mathcal{L}_{\alpha_0^n}, \mathcal{L}_{\alpha_1^n}) - \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n)\right)} \\ &\leq 2^{-n\left(1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_1^n)} - \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n)\right)} \end{aligned}$$

où la dernière inégalité provient du lemme 4.4. Le risque de seconde espèce est défini par :

$$r_2 = P_{\alpha_1^n}(W) = P_{\alpha_1^n}\left(\mathcal{L}_{\alpha_1^n} 2^{-\epsilon_1^n} \leq \mathcal{L}_{\alpha_0^n} 2^{-\epsilon_0^n}\right).$$

On obtient de manière symétrique :

$$\begin{aligned} r_2 = P_{\alpha_1^n}(W) &\leq 2^{\frac{1}{2}(\epsilon_1^n - \epsilon_0^n)} 2^{-nD_H^2(\mathcal{L}_{\alpha_1^n}, \mathcal{L}_{\alpha_0^n})} \\ &= 2^{-n\left(D_H^2(\mathcal{L}_{\alpha_1^n}, \mathcal{L}_{\alpha_0^n}) - \frac{1}{2n}(\epsilon_1^n - \epsilon_0^n)\right)} \\ &\leq 2^{-n\left(1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_1^n)} - \frac{1}{2n}(\epsilon_1^n - \epsilon_0^n)\right)} \end{aligned}$$

□

La proposition suivante fournit un exemple de test.

Proposition 5.4. *On suppose la condition G2 vérifiée.*

1) Pour tout n, soient α_0^n et α_1^n deux fonctions de Γ_n, et soient $C_n(\alpha_0^n)$ et $C_n(\alpha_1^n)$ leurs complexités respectives. Si les deux suites (α_0^n) et (α_1^n) vérifient pour tout n :

$$d_H^2(\alpha_0^n, \alpha_1^n) > \frac{1}{nD_0} \left| C([\alpha_0^n]) - C([\alpha_1^n]) \right|,$$

alors $W = \{\mathcal{L}_{\alpha_0^n} 2^{-C([\alpha_0^n])} > \mathcal{L}_{\alpha_1^n} 2^{-C([\alpha_1^n])}\}$ est la région de rejet d'un test consistant de H_0 :"$\alpha = \alpha_0^n$" contre H_1 :"$\alpha = \alpha_1^n$" dont les risques de première et de seconde espèce vérifient respectivement :

$$-\frac{1}{n} \log r_1 \geq \frac{1}{2n} \left(C([\alpha_1^n]) - C([\alpha_0^n]) \right) + 1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_1^n)},$$

$$-\frac{1}{n} \log r_2 \geq \frac{1}{2n} \left(C([\alpha_0^n]) - C([\alpha_1^n]) \right) + 1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_1^n)}.$$

2) Soit $\alpha_0^n \in \Gamma$ et $\alpha_1^n \in \Gamma_n$, de complexité $C_n(\alpha_1)$. Soit ϵ_0^n, une suite vérifiant $\epsilon_0^n = o(n)$ et :

$$|\epsilon_0^n - C_n(\alpha_1^n)| < nD_0 d_H^2(\alpha_0^n, \alpha_1^n),$$

alors $W = \{\mathcal{L}_{\alpha_0^n} 2^{-\epsilon_0^n} > \mathcal{L}_{\alpha_1^n} 2^{-C_n(\alpha_1^n)}\}$ est la région de rejet d'un test consistant de H_0 :"$\alpha = \alpha_0^n$" contre H_1 :"$\alpha = \alpha_1^n$" dont les risques de première et de seconde espèce vérifient 5.8.

Preuve : On détermine une condition suffisante de convergence pour le risque de première espèce, la condition pour le risque de seconde espèce s'en déduit par symétrie. Pour que le risque de première espèce converge vers 0, il suffit que :

$$e^{-D_0 d_H^2(\alpha_0^n,\alpha_1^n)} < 1 - \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n),$$

soit encore

$$-\ln(1 - \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n)) < D_0 d_H^2(\alpha_0^n, \alpha_1^n). \tag{5.9}$$

La suite $\varepsilon_0^n - \varepsilon_1^n$ étant en $o(n)$, en utilisant le dévelopement limité du logarithme, on obtient :

$$-\ln(1 - \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n)) < \frac{1}{2n}(\varepsilon_0^n - \varepsilon_1^n),$$

une condition suffisante pour que 5.9 soit vraie, est qu'elle soit vérifiée par un majorant du membre de gauche de (5.9), d'où la condition suffisante :

$$D_0 d_H^2(\alpha_0^n, \alpha_1^n) \geq \frac{1}{2n}(\epsilon_0^n - \epsilon_1^n).$$

En utilisant la symétrie des deux hypothèses, les deux conditions peuvent se résumer par :

$$D_0 d_H^2(\alpha_0^n, \alpha_1^n) > \frac{1}{2n}\left|\epsilon_1^n - \epsilon_0^n\right|.$$

Dans le cas de deux fonctions de Γ_n, la différence $|\epsilon_0^n - \epsilon_1^n|$ représente la valeur absolue de la différence des complexités des deux fonctions, et se réduit à celle des longueurs de description de leurs parties entières, car elles appartiennent au même espace, d'où le résultat du 1). Si l'une des deux fonctions, α_0^n par exemple, n'appartient pas à Γ_n, une condition suffisante pour que le test soit convergent est qu'on puisse choisir la suite ϵ_0^n telle que

$$C_n(\alpha_1^n) - 2nD_0 d_H^2(\alpha_0^n, \alpha_1^n) < \epsilon_0^n < C_n(\alpha_1^n) + 2nD_0 d_H^2(\alpha_0^n, \alpha_1^n).$$

□

Remarque : Dans le cas de deux fonctions de Γ_n, la condition de validité du test est que le carré de la distance de Hellinger entre les deux fonctions soit plus grand que la valeur absolue de la différence entre les deux complexités, relativement à la taille de l'échantillon. Intuitivement, la distance entre les deux fonctions apparaissant dans les hypothèses doit être suffisamment grande par rapport à la différence des complexités pour permettre de les séparer. Dans le cas de suites constantes $\alpha_0^n \equiv \alpha_0$ et $\alpha_1^n \equiv \alpha_1$, les test obtenus sont exponentiellement consistants, compte tenu de l'hypothèse de croissance sur les fonctions de complexités $C_n = o(n)$.

◇

Exemple 5.1 : Test sur la dimension.

Etant donné un échantillon de réalisations de processus ponctuel dont on connaît la loi $\alpha \in \Gamma$, on peut désirer déterminer si la meilleure approximation de α au vu des données observées, est dans un sous-ensemble de Γ_n de dimension m_0 ou m_1, et pour une précision fixée. Pour cela, on peut appliquer le test précédent avec pour α_0^n et α_1^n respectivement, la projection au sens de la norme L_2, de la fonction α dans les sous-espaces de dimension m_0 et m_1. La réponse à un tel test permet alors sans recourir à l'estimateur, d'obtenir la fonction qui réalise la meilleure compression en moyenne.

Remarque : Le test précédent peut servir à réaliser un codage variable de réalisations de processus ponctuels, ou encore comme critère de sélection dans un algorithme adaptatif d'estimation de la fonction d'intensité. A partir d'une solution initiale obtenue par la minimisation sans contrainte de la partie vraisemblance du critère de complexité minimale, on peut utiliser de tels tests pour sélectionner de manière adaptative le sous-ensemble des fonctions de base réalisant le meilleur compromis entre approximation et longueur de description. A chaque étape, il suffit de comparer la restriction de la solution continue au sous-espace considéré à la solution obtenue à l'étape précédente.

◇

Le résultat suivant montre qu'on peut construire des tests exponentiels dans le cas d'hypothèses du type H_0 :"$\alpha \equiv \alpha_0^n$" contre H_1 : "$d_H^2(\alpha, \alpha_0^n) > r_n$". La preuve de ce résultat nécessite le lemme suivant :

Lemme 5.5. *Soit (α_n), une suite de fonctions telle que $v_n = \| \alpha - \alpha_n^* \|_2$ converge vers 0. Soit également r_n, et β telles que : $r_n \geq 4\sqrt{2K} v_n$, et $D_H(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta) > r_n$. Si la condition G3 et l'hypothèse H3 sont vérifiées, alors pour toute suite positive u_n, on a :*

$$P_\alpha \left(\mathcal{L}_{\alpha_n^*} \leq \mathcal{L}_\beta 2^{-u_n} \right) \leq 2^{\frac{u_n}{2}} 2^{-\frac{n}{2 \ln 2} D_H^2(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta)},$$

pour n suffisamment grand.

Preuve : Soient $U_1, \ldots, U_n$, n variables aléatoires i.i.d et $E_n = \{\sum_{i=1}^n U_i \geq n u_n\}$. Par l'inégalité exponentielle, on a :

$$P(E_n) \leq 2^{\left(\inf_{y \geq 0} \left(-nyu_n + \sum_{i=1}^n \log E\left(e^{yU_i}\right)\right)\right)}.$$

Appliquée à $U_i = \log \frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}$, l'inégalité donne :

$$P_\alpha\left[\sum_{i=1}^n \log\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}(N_i,Y_i)\right) \geq nu_n\right] \leq 2^{\left[\inf_{y\geq 0}\left(-nyu_n+\sum_{i=1}^n \log E_\alpha\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}(N_i,Y_i)\right)^y\right)\right]}.$$

Pour $y = \frac{1}{2}$, on obtient :

$$\begin{aligned}
\frac{1}{n}\log P_\alpha(E_n) &\leq -\frac{u_n}{2} + \frac{1}{n}\sum_{i=1}^n \log E_\alpha\left(\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}\right)^{\frac{1}{2}}\right) \\
&\leq -\frac{u_n}{2} + \log E_\alpha\left(\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}\right)^{\frac{1}{2}}\right) \\
&= -\frac{u_n}{2} + \log \int \mathcal{L}_\alpha\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}\right)^{\frac{1}{2}} dP_0 \\
&= -\frac{u_n}{2} + \log \int \mathcal{L}_{\alpha_n^*}(1+\nu_n^*)\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}\right)^{\frac{1}{2}} dP_0, \qquad (5.10)
\end{aligned}$$

où $\nu_n^* = \frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}} - 1$. Analysons le second terme de cette dernière borne :

$$\begin{aligned}
\log \int \mathcal{L}_{\alpha_n^*}(1+\nu_n^*)\left(\left(\frac{\mathcal{L}_\beta}{\mathcal{L}_{\alpha_n^*}}\right)^{\frac{1}{2}}\right) dP_0 &= \log \int (1+\nu_n^*)\left(\mathcal{L}_{\alpha_n^*}\mathcal{L}_\beta\right)^{\frac{1}{2}} dP_0 \\
&= \log\left[\int \left(\mathcal{L}_{\alpha_n^*}\mathcal{L}_\beta\right)^{\frac{1}{2}} dP_0 + \int \nu_n^*\left(\mathcal{L}_{\alpha_n^*}\mathcal{L}_\beta\right)^{\frac{1}{2}} dP_0\right] \\
&= \log\left[1 - D_H^2(\mathcal{L}_{\alpha_n^*},\mathcal{L}_\beta) + \int \nu_n^*\left(\mathcal{L}_{\alpha_n^*}\mathcal{L}_\beta\right)^{\frac{1}{2}} dP_0\right] \qquad (5.11)
\end{aligned}$$

On développe l'intégrale apparaissant dans 5.11 :

$$\begin{aligned}
\int \nu_n^*\left(\mathcal{L}_{\alpha_n^*}\mathcal{L}_\beta\right)^{\frac{1}{2}} dP_0 &= \int \nu_n^*\sqrt{\mathcal{L}_{\alpha_n^*}}\left(\sqrt{\mathcal{L}_\beta} - \sqrt{\mathcal{L}_{\alpha_n^*}}\right) dP_0 + \int \nu_n^*\mathcal{L}_{\alpha_n^*} dP_0 \\
&= \int \nu_n^*\sqrt{\mathcal{L}_{\alpha_n^*}}\left(\sqrt{\mathcal{L}_\beta} - \sqrt{\mathcal{L}_{\alpha_n^*}}\right) dP_0 \\
&\leq \left[\int \left(\sqrt{\mathcal{L}_{\alpha_n^*}} - \sqrt{\mathcal{L}_\beta}\right)^2 dP_0\right]^{\frac{1}{2}}\left[\int \nu_n^{*2}\mathcal{L}_{\alpha_n^*} dP_0\right]^{\frac{1}{2}} \\
&= 2D_H(\mathcal{L}_{\alpha_n^*},\mathcal{L}_\beta)\left[E_{\alpha_n^*}(\nu_n^{*2})\right]^{\frac{1}{2}} \\
&= 2D_H(\mathcal{L}_{\alpha_n^*},\mathcal{L}_\beta)\left[E_{\alpha_n^*}\left(\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right)^2\right) - 2E_{\alpha_n^*}\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right) + 1\right]^{\frac{1}{2}} \\
&= 2D_H(\mathcal{L}_{\alpha_n^*},\mathcal{L}_\beta)\left[E_{\alpha_n^*}\left(\left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}\right)^2\right) - 1\right]^{\frac{1}{2}}
\end{aligned}$$

$$= 2D_H(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta) \left[\int \mathcal{L}_{\alpha_n^*} \left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}} \right)^2 dP_0 - 1 \right]^{\frac{1}{2}}$$

$$= 2D_H(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta) \left(E_\alpha \left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}} \right) - 1 \right)^{\frac{1}{2}} \tag{5.12}$$

Par le lemme 3.15, on a pour n suffisamment grand :

$$E_\alpha \left(\frac{\mathcal{L}_\alpha}{\mathcal{L}_{\alpha_n^*}}(N, Y) \right) \leq \exp\left(K v_n^2 \right),$$

où $v_n = \| \alpha - \alpha_n^* \|_2$. Cette borne permet d'obtenir par 5.12 et pour n suffisamment grand :

$$\int \nu_n^* \left(\mathcal{L}_{\alpha_n^*} \mathcal{L}_\beta \right)^{\frac{1}{2}} dP_0 \leq 2D_H(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta) \sqrt{2K} v_n.$$

Les hypothèses $D_H(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta) > r_n \geq 4\sqrt{2K} v_n$ impliquent

$$2\sqrt{2K} v_n \leq \frac{1}{2} D_H(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta).$$

On obtient ainsi la majoration :

$$\int \nu_n^* \left(\mathcal{L}_{\alpha_n^*} \mathcal{L}_\beta \right)^{\frac{1}{2}} dP_0 \leq \frac{1}{2} D_H^2(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta).$$

De 5.10 et 5.11, on peut déduire :

$$\begin{aligned} \frac{1}{n} \log P_\alpha(E_n) &\leq -\frac{u_n}{2} + \log\left(1 - \frac{1}{2} D_H^2(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta) \right) \\ &\leq -\frac{u_n}{2} + -\frac{1}{2\ln(2)} D_H^2(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta). \end{aligned}$$

On a donc

$$P_\alpha\left(\mathcal{L}_{\alpha_n^*} \leq 2^{-n u_n} \mathcal{L}_\beta \right) \leq 2^{-\frac{n u_n}{2}} 2^{-\frac{n}{2\ln 2} D_H^2(\mathcal{L}_{\alpha_n^*}, \mathcal{L}_\beta)},$$

pour n suffisamment grand.

□

Théorème 5.6. *Sous les mêmes hypothèses que dans le théorème 5.3 et en supposant de plus les conditions G1 et G3 vérifiées, soit* $\alpha_n = Arg \inf_{\gamma \in \Gamma_n} d_H^2(\alpha, \gamma)$ *et* $r_n \geq \max(D_0 d_H^2(\alpha, \alpha_n), 4\sqrt{2K} v_n)$. *La suite de tests définie par* H_0 *:"*$\alpha = \alpha_0^n$*" contre* H_1 *:"*$d_H^2(\alpha, \alpha_0^n) > 2\frac{r_n}{D_0}$*", de région de rejet* $W = \{\mathcal{L}_{\alpha_0^n} 2^{-n\varepsilon_n} < \mathcal{L}_{\hat{\alpha}_n} 2^{-C_n(\hat{\alpha}_n)}\}$ *possède des erreurs de première et de seconde espèce vérifiant :*

$$\begin{cases} -\frac{1}{n} \log r_1 \geq 1 - e^{-D_0 d_H^2(\alpha, \alpha_n)} - \varepsilon_n/2 \\ -\frac{1}{n} \log r_2 \geq \frac{1}{2} \left(\varepsilon_n + 1 - e^{-r_n} - \frac{1}{n} C_n(\alpha_n) \right). \end{cases} \tag{5.13}$$

Preuve : Soit le test de zone de rejet définie par $W = \{\mathcal{L}_{\alpha_0^n} 2^{-n\varepsilon_n} < \mathcal{L}_{\hat{\alpha}_n} 2^{-C_n(\hat{\alpha}_n)}\}$. En utilisant l'inégalité (4.4) on obtient pour le risque de première espèce :

$$\begin{aligned} r_1 &= P_\alpha\left(\mathcal{L}_\alpha 2^{-n\varepsilon_n} < \mathcal{L}_{\hat{\alpha}_n} 2^{-C_n(\hat{\alpha}_n)}\right) \\ &\leq P_\alpha\left(\bigcup_{\gamma\in\Gamma_n}\left(\mathcal{L}_\alpha 2^{-n\varepsilon_n} < \mathcal{L}_\gamma 2^{-C_n(\gamma)}\right)\right) \\ &\leq \sum_{\gamma\in\Gamma_n} 2^{\frac{n\varepsilon_n - C_n(\gamma)}{2}} 2^{-nD_H^2(\mathcal{L}_\alpha,\mathcal{L}_\gamma)} \\ &\leq 2^{-n\left(1-e^{-D_0 d_H^2(\alpha,\alpha_n)}-\varepsilon_n/2\right)} \sum_{\gamma\in\Gamma_n} 2^{-\frac{C_n(\gamma)}{2}}. \end{aligned}$$

Par l'inégalité de Kraft on peut conclure que :

$$\log r_1 \leq -n\left(1 - e^{-D_0 d_H^2(\alpha,\alpha_n)} - \varepsilon_n/2\right)$$

L'erreur de seconde espèce est définie par :

$$\begin{aligned} r_2 &= P_\alpha\left(\mathcal{L}_{\hat{\alpha}_n} 2^{-C_n(\hat{\alpha}_n)} \leq \mathcal{L}_{\alpha_0^n} 2^{-n\varepsilon_n}\right) \\ &= P_\alpha\left(\sup_{\gamma\in\Gamma_n} \mathcal{L}_\gamma 2^{-C_n(\gamma)} \leq \mathcal{L}_{\alpha_0^n} 2^{-n\varepsilon_n}\right) \\ &\leq P_\alpha\left(\mathcal{L}_{\alpha_n} 2^{-C_n(\alpha_n)} \leq \mathcal{L}_{\alpha_0^n} 2^{-n\varepsilon_n}\right) \end{aligned}$$

Afin de majorer cette probabilité, on applique le lemme 5.5 avec $u_n = C_n(\alpha_n) - \varepsilon_n$, et $\beta = \alpha_0^n$. Pour cela, on vérifie que les hypothèses du lemme sont satisfaites. L'hypothèse $r_n \geq 4\sqrt{2K}v_n$ est satisfaite d'après l'énoncé du théorème. Il reste à vérifier que

$$D_H(\mathcal{L}_{\alpha_n}, \mathcal{L}_{\alpha_0^n}) > r_n. \tag{5.14}$$

Par le lemme 4.4,

$$D_H(\mathcal{L}_{\alpha_n}, \mathcal{L}_{\alpha_0^n}) \geq 1 - e^{-D_0 d_H^2(\alpha_n,\alpha_0^n)},$$

donc une condition suffisante pour que l'inégalité 5.14 soit vérifiée, est que :

$$1 - e^{-D_0 d_H^2(\alpha_n,\alpha_0^n)} > r_n.$$

Une condition nécessaire pour que la seconde inégalité soit vérifiée, est que :

$$d_H^2(\alpha_n, \alpha_0^n) > \frac{r_n}{D_0}.$$

Il reste à montrer que sous l'alternative H1, et les conditions du théorème, cette dernière inégalité est vérifiée.

Sous H1, et compte tenu de l'inégalité triangulaire, on a :

$$\frac{2r_n}{D_0} < d_H^2(\alpha, \alpha_0^n) \leq d_H^2(\alpha, \alpha_n) + d_H^2(\alpha_n, \alpha_0^n).$$

Par hypothèse, $d_H^2(\alpha, \alpha_n) < \frac{r_n}{D_0}$, d'où $\frac{2r_n}{D_0} < \frac{r_n}{D_0} + d_H^2(\alpha_n, \alpha_0^n)$, et

$$d_H^2(\alpha_n, \alpha_0^n) \geq \frac{r_n}{D_0}.$$

L'inégalité 5.14 est donc vérifiée, et on peut appliquer le lemme 5.5, pour obtenir :

$$\begin{aligned} r_2 &\leq 2^{-\frac{n\varepsilon_n - C_n(\alpha_n)}{2}} 2^{-\frac{n}{2} D_H^2(\mathcal{L}_{\alpha_0^n}, \mathcal{L}_{\alpha_n})} \\ &\leq 2^{-\frac{n\varepsilon_n - C_n(\alpha_n)}{2}} 2^{-\frac{n}{2}(1 - e^{-D_0 d_H^2(\alpha_n, \alpha_0^n)})} \\ &\leq 2^{-\frac{n\varepsilon_n - C_n(\alpha_n)}{2}} 2^{-\frac{n}{2}(1 - e^{-r_n})}. \end{aligned}$$

Cette minoration fournit pour r_2, la borne :

$$-\frac{1}{n} \log r_2 \geq \frac{1}{2} \left(\varepsilon_n + 1 - e^{-r_n} - \frac{1}{n} C_n(\alpha_n) \right)$$

.

□

Corollaire 5.7. *Soit r_n une suite vérifiant $r_n \geq \max(D_0 d_H^2(\alpha, \alpha_n), 4D_Y v_n)$, et ε_n une suite positive telle que :*

$$\frac{1}{n} C_n(\alpha_n) + r_n \leq \varepsilon_n \leq 2D_0 d_H^2(\alpha_0^n, \alpha_n),$$

alors $W = \{\mathcal{L}_{\alpha_0} 2^{-n\varepsilon_n} < \mathcal{L}_{\hat{\alpha}_n} 2^{-C_n(\hat{\alpha}_n)}\}$ définit la région de rejet d'un test consistant de H_0 : "$\alpha = \alpha_0^n$" contre H_1 : "$d_H^2(\alpha, \alpha_0^n) > r_n$" dont les risques de première et de seconde espèce vérifient 5.13.

Preuve : Il suffit de vérifier que les bornes données par 5.13 sont strictement positives. La première borne

$$1 - e^{-D_0 d_H^2(\alpha_0^n, \alpha_n)} - \varepsilon_n/2,$$

est strictement positive si :

$$e^{-D_0 d_H^2(\alpha_0^n, \alpha_n)} < 1 - \frac{\varepsilon_n}{2},$$

soit encore

$$\begin{aligned} d_H^2(\alpha_0^n, \alpha_n) &\geq -\frac{1}{D_0}\ln(1-\frac{\varepsilon_n}{2}) \\ &= -\frac{1}{D_0}(-\frac{\varepsilon_n}{2}+\frac{\varepsilon_n^2}{4}+\ldots) \end{aligned}$$

Le terme $\varepsilon_n/2D_0$ est un majorant du membre de droite de l'expression précédente. Une condition suffisante pour que l'inégalité soit vraie, est qu'elle soit valable pour le majorant, soit :

$$d_H^2(\alpha_0^n, \alpha_n) \geq \frac{\varepsilon_n}{2D_0}.$$

Pour la seconde borne

$$\varepsilon_n + 1 - e^{-r_n} - \frac{1}{n}C_n(\alpha_n) > 0,$$

si

$$\varepsilon_n > \frac{1}{n}C_n(\alpha_n) + 1 - e^{-r_n},$$

une condition suffisante est obtenue pour un majorant de $1 - e^{-r_n}$, soit

$$\varepsilon_n > \frac{1}{n}C_n(\alpha_n) + r_n.$$

□

Remarque : Sous les hypothèses de la proposition 4.16, ce résultat est encore valable pour une distance autre que la distance de Hellinger apparaissant dans l'hypothèse alternative H1.

◇

Chapitre 6

Algorithmes de minimisation et simulations.

Rappelons qu'on s'intéresse à des processus (N, Y) définis sur un espace $(\Omega, (\mathcal{F}_t), P_\alpha)$ où P_α est absolument continue par rapport à P_0 et de dérivée de Radon-Nikodym donnée par :

$$\frac{dP_\alpha}{dP_0} = \exp\left(-\int_0^T (\alpha(s)-1)Y(s)ds + \int_0^T \ln(\alpha(s))dN(s)\right) \quad (6.1)$$

$$= \mathcal{L}_\alpha(N, Y). \quad (6.2)$$

Si l'on observe un échantillon $(\xi^n, y^n) = ((\xi_1, y_1), \ldots, (\xi_n, y_n))$, on note $(\overline{\xi}_n, \overline{y}_n)$ le couple défini par les cumulés $\overline{\xi}_n = \sum_{i=1}^n \xi_i$ et $\overline{y}_n = \sum_{i=1}^n y_i$. Pour toute fonction candidate β, la dérivée de Radon-Nikodym calculée en (ξ^n, y^n) a pour expression à un terme multiplicatif près indépendant de β, :

$$\mathcal{L}_\beta(\xi^n, y^n) = \exp\left(\int_0^T \ln(\beta(s))d\overline{\xi}_n(s) - \int_0^T \beta(s)\overline{y}_n(s)ds\right).$$

Le critère pour une fonction candidate β s'écrit :

$$-\log \mathcal{L}_\beta(\xi^n, y^n) + C_n(\beta) = \frac{1}{\ln 2}\left[\int_0^T \ln(\beta(s))d\overline{\xi}_n(s) - \int_0^T \beta(s)\overline{y}_n(s)ds\right] + C_n(\beta) \quad (6.3)$$

$$= \frac{1}{\ln 2}\left[-\int_0^T \beta(s)\overline{y}_n(s)ds + \sum_{i=1}^{\overline{\xi}_n(T)} \ln(\beta(\tau_i))\right] + C_n(\beta), \quad (6.4)$$

où τ_i représente le $i-$ème instant de saut observé du processus $\overline{\xi}_n = \sum_{i=1}^n \delta_{\tau_i}$. Le terme $-\log \mathcal{L}_\beta(\xi^n, y^n)$ représente la longueur (à un bit près) d'un code de Shannon pour les données observées $((\xi_1, y_1), \ldots, (\xi_n, y_n))$. Les fonctions β étant toutes définies sur la même base, la

minimisation s'effectue en réalité sur les coefficients du développement de la fonction sur la base. Cette minimisation est délicate pour deux raisons. La première est la contrainte de positivité sur les fonctions de Γ_n. Afin de ne pas avoir recours à des techniques de minimisation sous contraintes, il est préférable de modifier le critère en exprimant la fonction inconnue comme une fonctionnelle positive d'une autre fonction, qui elle, est autorisée à varier dans tout l'espace. La seconde difficulté provient du fait que la partie complexité de β (i.e. $C_n(\beta)$)dans le critère n'est pas une fonction continue des coefficients du développement de la fonction sur la base. La plupart des techniques de minimisation, basées sur les calculs de dérivées, ne pourront donc pas être employées sans précautions dans ce cadre. Les termes $C_n(\beta)$, pour définir les longueurs d'un code instantanément décodable doivent satisfaire l'inégalité de Kraft $\sum 2^{-C_n(\beta)} \leq 1$. Les longueurs C_n doivent en outre pour des raisons liées à la convergence de l'estimateur, vérifier la condition $C_n = o(n)$. La minimisation de la longueur de description totale s'effectue en deux étapes. La composante continue (comme fonction des coefficients) du critère $(-\log \mathcal{L}_\beta(\xi^n, y^n))$, est minimisée en utilisant une méthode de descente. Pour m fixé, on obtient une solution du maximum de vraisemblance. En tronquant les coefficients de cette solution á diverses précisions, on obtient des solutions $\hat{\alpha}_n(m, \delta)$. Pour chacune de ces fonctions on calcule la longueur totale du code (y compris le terme de complexité), et pour m fixé, on détermine la précision qui minimise la longueur totale de description. En faisant varier m dans l'ensemble des précisions admissibles, on obtient alors la solution $\hat{\alpha}_n$. Dans la section suivante, on analyse la partie lisse du critère.

6.1 Minimisation de la partie lisse du critère.

La fonction critère à optimiser dépend de paramètres qui varient continûment, ne figurant que dans la vraisemblance, et d'un certains nombre de paramètres discrets apparaissant dans la fonction de complexité $C_n(\beta)$. De manière pratique, on est donc conduits à fixer les paramètres discrets et à optimiser la partie vraisemblance du critère puis à changer les valeurs des paramètres discrets et à effectuer à nouveau l'optimisation de la vraisemblance et ainsi de suite.

Dans toute la suite, m étant fixé, on considère une base $\{\phi_j,\ j = 0, \ldots, m-1\}$ de fonctions positives de Γ_n. Pour tenir compte de la contrainte de positivité, nous avons étudié deux méthodes. La première consiste à rechercher des fonctions de la forme $\beta(s) = \left(\sum_{j=0}^{m-1} \beta_j \phi_j(s)\right)^2$, ce qui garantit la positivité, et la seconde consiste à rechercher le minimum dans un voisinage

de l'optimum tel que la contrainte de positivité soit satisfaite.

Contrainte sur la fonction d'intensité.

Pour m fixé, on cherche :

$$\beta^* = Arg \min_{\beta_0,\ldots,\beta_{m-1}} \left\{ \int_0^T \left(\sum_j \beta_j \phi_j(s) \right)^2 \overline{y}_n(s) ds - \sum_{i=1}^{\overline{\xi}_n(T)} \ln \left(\sum_j \beta_j \phi_j(\tau_i) \right)^2 \right\} \tag{6.5}$$

Le vecteur gradient a pour $j-$ème composante :

$$\nabla_j(\mathcal{L}_\beta) = \frac{\partial \mathcal{L}_\beta}{\partial \beta_j} = 2 \int_0^T \beta(s) \phi_j(s) \overline{y}_n(s) ds - 2 \int_0^T \frac{\phi_j(s)}{\beta(s)} d\overline{\xi}_n(s)$$

L'élément $\mathbf{H}_{j,l}(\beta)$ de la matrice hessienne est donné par :

$$\mathbf{H}_{j,l}(\beta) = \frac{\partial^2 \mathcal{L}_\beta}{\partial \beta_j \beta_l} = 2 \int_0^T \phi_j(s) \phi_l(s) \overline{y}_n(s) ds + 2 \int_0^T \frac{\phi_j(s) \phi_l(s)}{\beta^2(s)} d\overline{\xi}_n(s)$$

A l'optimum β^* on a la relation :

$$\int_0^T \beta^*(s) \phi_j(s) \overline{y}_n(s) ds = \int_0^T \frac{\phi_j(s)}{\beta^*(s)} d\overline{\xi}_n(s).$$

Par combinaison linéaire de coefficients β_j^*, on obtient en intervertissant les signes somme et intégrale :

$$\int_0^T \beta^{*2}(s) \overline{y}_n(s) ds = \int_0^T d\overline{\xi}_n(s) = \overline{\xi}_n(T).$$

Le critère optimal a donc pour expression :

$$C_m^* = \overline{\xi}_n(T) - \sum_{i=1}^{\overline{\xi}_n(T)} \ln(\beta^*(\tau_i))^2.$$

C_m^* représente la complexité minimale des données, étant donné la fonction β^*.

Approximation locale.

Pour m fixé, on cherche :

$$\beta^* = Arg \min_{\beta_0,\ldots,\beta_{m-1}} \int_0^T \left(\sum_j \beta_j \phi_j(s) \right) \overline{y}_n(s) ds - \sum_{i=1}^{\overline{\overline{\xi}}_n(T)} \ln \sum_j \beta_j \phi_j(\tau_i). \tag{6.6}$$

On obtient pour vecteur gradient :

$$\nabla_j(\beta) = \frac{\partial \mathcal{L}_\beta}{\partial \beta_j} = \int_0^T \phi_j(s)\overline{y}_n(s)ds - \int_0^T \frac{\phi_j(s)}{\beta(s)} d\overline{\xi}_n(s)$$

L'élément $\mathbf{H}_{j,l}(\beta)$ de la matrice hessienne est donné par :

$$\mathbf{H}_{j,l}(\beta) = \frac{\partial^2 \mathcal{L}_\beta}{\partial \beta_j \beta_l} = \int_0^T \frac{\phi_j(s)\phi_l(s)}{\beta(s)} d\overline{\xi}_n(s).$$

A l'optimum β^* on a la relation :

$$\int_0^T \phi_j(s)\overline{y}_n(s)ds = \int_0^T \frac{\phi_j(s)}{\beta^*(s)} d\overline{\xi}_n(s).$$

Si, de la même manière que précédemment, on multiplie chaque terme par β_j^*, et si on effectue la sommation sur j, en intervertissant les signes sommes et intégrale, on obtient :

$$\int_0^T \beta^*(s)\overline{y}_n(s)ds = \int_0^T d\overline{\xi}_n(s) = \overline{\xi}_n(T).$$

Le critère optimal a donc pour expression :

$$C_m^* = \overline{\xi}_n(T) - \sum_{i=1}^{\overline{\xi}_n(T)} \ln \beta(\tau_i).$$

Résolution numérique.

De nombreuses méthodes d'optimisation peuvent être utilisées pour déterminer la solution du système non-linéaire. Les méthodes du gradient, du gradient conjugué, de Newton-Raphson et du gradient à métrique variable ont été étudiées par [51]. Le gradient a été implémenté par la méthode de la plus profonde descente, le gradient conjugué par la méthode de Fletcher et Reeves [33], l'algorithme à métrique variable selon la méthode de Kelley [43]. Le principe de ces méthodes est le suivant : à partir de la solution β_k à l'étape k, une direction de descente d_k, un coefficient λ_k, déterminent un vecteur :

$$\beta_{k+1} = \beta_k + \lambda_k d_k. \tag{6.7}$$

Le tableau 6.1 donne un aperçu des principales méthodes de descente qui ont été utilisées pour résoudre ce problème.

Les meilleurs résultats ont été obtenus en premier lieu par l'algorithme de variance puis par l'algorithme de Newton-Raphson (voir [51]). L'algorithme de variance nécessite une estimation

Algorithme	Evaluation de d_k (c.f.6.7)
Gradient	$d_k = \nabla(\beta_k)$
Gradient conjugué	$d_k = \nabla(\beta_k) + \lambda_{k-1} d_{k-1}, k \geq 1$ $d_0 = \nabla(\beta_0)$ $\lambda_{k-1} = \frac{[\nabla(\beta_k)]^T \nabla(\beta_k)}{[\nabla(\beta_{k-1})]^T \nabla(\beta_{k-1})}$
Métrique variable	$d_k = M_k \nabla(\beta_k)$ $M_0 > 0$ $M_k = M_{k-1} - \frac{(M_{k-1} y_{k-1})(M_{k-1} y_{k-1})^T}{y_{k-1}^T M_{k-1} y_{k-1}}, k \geq 1, k \neq i+1$ $M_{i+1} = \sum_{i-m+1}^{i} q_j q_j^T (q_j^T q_j)^{-1}$ $q_j = \beta_{j+1} - \beta_j$ $y_j = \nabla(\beta_{j+1}) - \nabla(\beta_j)$
Newton-Raphson	$d_k = -\mathbf{H}^{-1}(\beta_k) \nabla(\beta_k)$
Algorithme de variance	$d_k = -[E(\mathbf{H}(\beta_k))]^{-1}$

Figure 6.1 : Méthodes de descente.

de l'espérance de la matrice hessienne, et ne correspond pas au cadre théorique que nous avons adopté, nous avons donc privilégié l'algorithme de Newton dans l'étude numérique. L'algorithme de Newton-Raphson pour résoudre le système non-linéaire,

$$\int_0^T \frac{\phi_j(s)}{\beta(s)} d\bar{\xi}_n(s) = \int_0^T \beta(s)\phi_j(s)\bar{y}_n(s)ds, \quad 0 \leq j \leq m-1.$$

correspondant à la première méthode, et pour la seconde :

$$\int_0^T \frac{\phi_j(s)}{\beta(s)} d\bar{\xi}_n(s) = \int_0^T \phi_j(s)\bar{y}_n(s)ds, \quad 0 \leq j \leq m-1,$$

Pour obtenir une valeur initiale relativement proche de l'optimum recherché, le principe utilisé consiste à rechercher une approximation constante par morceaux de la fonction initiale. Pour cela on cherche une solution $\beta^0(s) = \sum_{j=0}^{m-1} \beta_j^0 \mathbf{I}_{I_{j,m}}$ où $I_{j,m} = [jT/m, (j+1)T/m]$, $j = 0, \ldots, m-1$, i.e. une fonction de type histogramme définie sur m intervalles de longueurs identiques. On en déduit une fonction $\beta^1(s) = \sum_{j=0}^{m-1} \beta_j^1 \phi_j(s)$ dont $\beta^0(s)$ est une interpolation et qui servira de valeur initiale pour l'algorithme. Pour la recherche de $\beta_0^0, \ldots, \beta_{m-1}^0$, les composantes du gradient s'écrivent

$$\nabla_j(\beta) = \frac{\partial \mathcal{L}_\beta}{\partial \beta_j} = \psi_{j,m} - \frac{\mu(I_{j,m})}{\beta_{j,m}}, \quad j = 0, \ldots, m-1,$$

où $\mu(I_{j,m})$ désigne le nombre de points de sauts du processus $\overline{\xi}_n$ dans l'intervalle $I_{j,m}$ et $\psi_{j,m} = \int_{I_{j,m}} \overline{y}_n(s)ds$. Le système admet alors une solution explicite donnée par :

$$\beta^0_{j,m} = \frac{\mu(I_{j,m})}{\psi_{j,m}}.$$

On détermine les coefficients de la solution initiale β^1 dans la base considérée en résolvant le système linéaire d'ordre m, $A\beta^1 = \beta^0$, la matrice A étant définie par $A_{k,j} = \phi_j(t_k)$, $j, k = 0, \ldots, m-1$, les valeurs t_k représentant les points milieux des intervalles définissant l'histogramme $t_k = (2k+1)T/2m$.

Pour β^1 donné, l'algorithme de Newton-Raphson pour $k \geq 1$ sera :

$$\begin{cases} \mathbf{H}(\beta^k)\Delta^k = -\nabla(\beta^k) \\ \beta^{k+1} = \beta^k + \Delta^k \end{cases}$$

Le point d'arrêt de l'algorithme étant donné pour $\| \Delta^k \|_2 \leq \varepsilon$, pour un ε prédéfini.

6.2 Prise en compte de la complexité.

A partir des procédures de la section précédente, on déduit un l'algorithme 6.2 déterminant la solution de complexité minimale pour P familles de bases indicées par j, chaque famille possédant un nombre maximal de fonctions de base $M(j)$. La procédure $SolInit(m, j)$ permet pour une famille j donnée et une dimension m, de calculer les coefficients de la solution initiale β^1 dans la base considérée. La procédure $Newton$ résout numériquement le système non-linéaire de solution initiale β^1. La procédure $MinPrec$ détermine la précision optimale.

Dans nos simulations, nous avons effectué la recherche de l'estimateur de complexité minimale au sein de trois familles. Les familles, trigonométriques polynomiales et splines. On peut concevoir que cette recherche soit enrichie par l'apport d'autres familles. Le but étant à travers cette démarche, de trouver la fonction la plus simple, présentant la meilleure adéquation aux données observées. Le critère de sélection dans notre méthode, est basé sur la complexité. Les familles candidates devront êtres calculables, au sens où une complexité représentant la difficulté liée à leur description pourra leur être affectée. D'autre part, les procédures de minimisation comportant des méthodes très lentes (gradient, Newton-Raphson), il sera nécessaire de partir d'une solution initiale suffisamment proche de l'optimum recherché. Théoriquement, on sait que si la fonction inconnue appartient à l'une des familles, elle est presque-sûrement découverte

Debut

Repeter

Repeter

SolInit (m, j, u)

Newton (m, j, u)

MinPrec (m, j, u, δ)

Si **LCode** $(m, j, u, \delta) < Lmin$ Alors

Lmin $\leftarrow$ LCode ;$u* \leftarrow u$

jusqu'à $((m = M(j))$ ou $(C(m) + C(j) > Lmin))$

jusqu'à $j = P$.

Fin

Figure 6.2 : Algorithme de recherche.

pour n suffisamment grand, il est donc clair que l'ensemble des fonctions candidates est également l'ensemble des fonctions susceptibles d'être découvertes par l'estimateur. En particulier cette famille peut contenir par exemple des splines cubiques, dont le nombre et la position des noeuds doivent être déterminés par la méthode. Supposant les familles de fonctions candidates indicées par j et $C(\Gamma^j)$ désignant la longueur d'un code pour la famille $(\Gamma_n^j)_{n\geq 0}$, l'estimateur de complexité minimale $\hat{\alpha}_n$ est une fonction réalisant le minimum

$$\min_j \left\{ C(\Gamma^j) + \min_{\beta \in \Gamma_n^j} (\log 1/\mathcal{L}_\beta(\xi^n, y^n) + C_n(\beta)) \right\}.$$

La recherche n'est pas exhaustive, car la première famille fournit une borne inférieure de la longueur du code qui définit un test d'arrêt de la procédure de recherche. En réalité la procédure de minimisation précédente admet la formulation séquentielle suivante :

$$\min_j \left\{ C(\Gamma^j) + \min_{m\ :C(m)<K_j} \left(\min_{(\beta_0,\ldots,\beta_{m-1})} (\log 1/\mathcal{L}_\beta(\xi_n, yn) + C_n(\beta)) \right) \right\},$$

où K_j désigne la borne inférieure de la longueur totale du code atteinte avant la famille j et $C(m)$ la complexité liée à la dimension. Si la recherche s'effectue au sein de chaque famille par ordre de complexité croissante, il est inutile de poursuivre l'algorithme pour des dimensions de complexité supérieure à K_j.

6.3 Exemples de bases de fonctions.

6.3.1 Cas de l'histogramme.

On peut considérer par exemple, la base de fonctions $\{\phi_j,\ j = 1,\ldots,m-1\}$ formée des fonctions indicatrices d'intervalles. L'estimateur de complexité minimale s'obtient alors en minimisant le nombre de classes m de l'histogramme, la position des noeuds, et la précision à laquelle les valeurs de la fonction dans les classes sont arrondies. On peut également se placer dans le contexte de noeuds équidistants, ou de coefficients arrondis à une précision fixée à l'avance et fonction de la taille de l'échantillon. Nous examinons tous ces cas, et nous étudions essentiellement trois estimateurs. Le premier, que nous noterons $\hat{\alpha}_n^{(e)}$ est basé sur des noeuds équidistants, dont le nombre est déterminé par le critère. Le second, $\hat{\alpha}_n^{(v)}$ est basé sur des intervalles dyadiques entre noeuds, de longueurs variables. Dans chaque cas, on peut choisir de minimiser par rapport à la précision à laquelle les coefficients sont arrondis. Le troisième estimateur $\hat{\alpha}_n^{(m)}$ est défini à partir du premier comme la moyenne des estimations obtenues à nombre d'intervalles fixé, pour une quantité M de valeurs de ce nombre d'intervalles. Soit $\hat{\alpha}_{n,m}^{(e)}$ (resp. $\hat{\alpha}_{n,m}^{(v)}$) défini sur m intervalles de même longueur (resp. sur m intervalles de longueur quelconque) par :

$$\hat{\alpha}_{n,m}^{(\cdot)} = Arg \min_{\beta\equiv(\beta_0,\ldots,\beta_{m-1})} (\log 1/\mathcal{L}_\beta(\xi^n, y^n) + C_n(\beta))$$

Par définition, $\hat{\alpha}_n^{(\cdot)}$ réalise le minimum :

$$\min_m \left(\log 1/\mathcal{L}_{\hat{\alpha}_{n,m}^{(\cdot)}}(\xi^n, y^n) + C_n(\hat{\alpha}_{n,m}^{(\cdot)})\right),$$

et

$$\hat{\alpha}_n^{(m)} = \frac{1}{M}\sum_{j=1}^{M} \hat{\alpha}_{n,j}^{(e)}.$$

Ces trois estimateurs étant basés sur $\hat{\alpha}_{n,m}^{(\cdot)}$, on commence le calcul en déterminant (sans minimiser par rapport à la précision) les valeurs des coefficients de $\hat{\alpha}_{n,m}$ sur les intervalles de l'histogramme. Le critère à minimiser aura ainsi pour expression :

$$\mathcal{L}_\beta(\xi^n, y^n) = \int_0^T \left(\sum_{j=0}^{m-1} \beta_j\phi_j(s)\right) \overline{y}_n(s)ds - \int_0^T \ln(\sum_{j=0}^{m-1} \beta_j\phi_j(s)) d\overline{\xi}_n(s)) + \frac{m}{2}\log n \qquad (6.8)$$

Si on pose

$$\psi_j = \int_0^T \phi_j(s)\overline{y}_n(s)ds = \int_{I_j} \overline{y}_n(s)ds \quad j = 0,\ldots,m-1,$$

on obtient :

$$\mathcal{L}_\beta = \sum_{j=0}^{m-1} \beta_j \psi_j - \int_0^T \ln \left(\sum_{j=0}^{m-1} \beta_j \phi_j(s) \right) d\bar{\xi}_n(s) + \frac{m}{2} \frac{\log n}{n}.$$

Pour m fixé, le gradient aura pour expression :

$$\nabla_j(\beta) = \psi_{j,m} - \int_0^T \frac{\phi_{j,m}(s)}{\beta_{j,m}} d\bar{\xi}_n(s) = \psi_{j,m} - \int_{I_{j,m}} \frac{1}{\beta_{j,m}} d\bar{\xi}_n(s)$$

En utilisant $\bar{\xi}_n = \sum_{i=1}^{\bar{\xi}_n(T)} \delta_{\tau_i}$, on obtient :

$$\nabla_j(\beta) = \psi_{j,m} - \frac{\mu(I_{j,m})}{\beta_{j,m}},$$

où $\mu(I_{j,m})$ est le nombre de sauts du processus $\bar{\xi}_n$ sur l'intervalle $I_{j,m}$. Ainsi on obtient $\hat{\beta}_{j,m} = \frac{\mu(I_{j,m})}{\psi_{j,m}}$. Les estimateurs $\hat{\alpha}_n^{(e)}$ et $\hat{\alpha}_n^{(v)}$ sont obtenus en choisissant des intervalles $I_{j,m}$ variables ou équidistants, et en minimisant par rapport à m, le critère (6.8) appliqué à $\hat{\beta}^{(m)}(s) = \sum_{j=0}^{m-1} \beta_{j,m} \phi_j(s)$, et dans le cas des intervalles de longueurs variables, en minimisant la position des noeuds. Afin de déterminer la position optimale des noeuds, on construit une partition de l'intervalle $[0,T]$ en $m(n)$ dyadiques de longueurs $\delta_n = 2^{-m(n)}$, précision à laquelle on arrondit les coefficients de l'histogramme [1] Pour toute dimension p, le choix de p positions parmi $m(n)$ nécessite $C_{m(n)}^p$ comparaisons. Au total, la minimisation par rapport à p nécessitera de l'ordre de $\sum_p C_{m(n)}^p$ comparaisons, soit $2^{m(n)}$. Par exemple, pour une discrétisation de $[0,1]$ en 16 intervalles, on devra effectuer de l'ordre de 65536 comparaisons; on ne peut donc pas procéder par recherche exhaustive. On peut néanmoins résoudre le problème en utilisant la théorie des graphes. Soit A l'arbre binaire dont la racine R représente l'intervalle $[0,T]$, et pour tout sommet $[a,b]$, les fils gauches et droit représentent respectivement les intervalles $[a,c]$ et $[c,b]$ où $c = (b-a)/2$ par exemple. Pour un niveau m, cet arbre contient $2^m - 1$ éléments. Chaque niveau i de l'arbre contient la partition de $[0,T]$ en i intervalles dyadiques. Pour deux éléments (i,j) de l'arbre, on notera $c(i,j) = 0$ s'il n'existe pas d'arc de i vers j, et $c(i,j) = 1$ sinon. Soient $[a_i,b_i]$ l'intervalle correspondant au sommet i, et $[a_j,b_j]$ l'intervalle correspondant au sommet j. On posera $c(i,j) = 1$ si $b_i = a_j$ et $c(i,j) = 0$ sinon. De plus, on définit deux sommets E et F, représentant des intervalles fictifs tels que $c(E,i) = 1$ si $a_i = 0$ et $c(i,F) = 1$ si $b_i = 1$. Le problème revient alors à déterminer le plus court chemin dans un graphe dont les sommets représentent des intervalles, et les arcs, les longueurs de codes associées à ces intervalles (la longueur

[1] Les positions des noeuds doivent également être codées dans ce cas de figure, afin que le décodeur puisse reconstituer l'histogramme.

d'un code pour les données comprises dans cet intervalle). Cette méthode a été implémentée avec l'algorithme de recherche de plus court chemin de Dykstra. La section suivante présente l'application de la méthode à trois familles usuelles.

6.3.2 Fonctions trigonométriques, polynomiales et splines.

Dans cette section, nous formulons les différents résultats obtenus pour trois bases de fonctions usuelles, les fonctions splines, polynomiales et trigonométriques. Nous donnons les vitesses de convergence dans ces cas particuliers, et nous présentons quelques résultats de simulations. Les simulations ont été effectuées à partir des trois familles de fonctions citées : les fonctions splines, trigonométriques et polynomiales. On suppose que la fonction d'intensité admet l'écriture :

$$\beta = \left(\sum_{j=0}^{m-1} \beta_j \phi_j(s) \right)^2 .$$

Chaque fonction est codée par un préfixe désignant la famille, un ensemble de paramètres (nombre de fonctions de base, degré pour la spline etc.), un vecteur des coefficients tronqués à une précision donnée. L'algorithme est néanmoins conçu pour toute base de fonctions positives, et peut être enrichi à volonté par l'ajout de nouvelles bases de fonctions. Dans chaque cas, on doit calculer les intégrales $\psi_{j,l} = \int_0^T \phi_j(s)\phi_l(s)\overline{y}_n(s)ds$ et $\xi_j = \int_0^T \phi_j(s)\overline{y}_n(s)ds$ apparaissant dans les expressions du gradient et de la hessienne (voir section 6.1.1). Comme nous nous sommes concentrés, en ce qui concerne les expérimentations numériques, sur la situation des durées de vie avec possibilité de censure, nous avons choisi de nous limiter dans les calculs ci-dessous à des réalisations $\overline{y}_n$ de la forme $\overline{y}_n(s) = \sum_{i=1}^n \mathbf{I}_{[0,A_i[}(s)$.

Fonctions polynomiales. Les fonctions de base sont données par $\phi_j(s) = s^j, j = 0, \ldots, m-1$. On obtient :

$$\xi_j = \sum_{i=1}^n \int_0^{A_i} s^j ds = \frac{1}{j+1} \sum_{i=1}^n A_i^{j+1}.$$

$$\psi_{j,l} = \sum_{i=1}^n \int_0^{A_i} s^{j+l} ds = \frac{1}{j+l+1} \sum_{i=1}^n A_i^{j+l+1}.$$

Fonctions trigonométriques. Pour m pair, on peut obtenir une base de m fonctions trigonométriques en posant pour $k = 0, \ldots, m-1$:

$$\phi_k(s) = \begin{cases} \cos(2\pi k s) \text{ si } k \text{ est pair.} \\ \sin(2\pi k s) \text{ si } k \text{ est impair.} \end{cases}$$

Pour le calcul des fonctions ξ_j, on distingue les cas pair et impair.

j pair

$$\xi_j = \int_0^T \cos(2\pi j s)\overline{y}_n(s)ds = \frac{1}{2\pi j}\sum_{i=1}^n \sin(2\pi j A_i).$$

j impair

$$\xi_j = \int_0^T \sin(2\pi j s)\overline{y}_n(s)ds = \frac{1}{2\pi j}\sum_{i=1}^n (1 - \cos(2\pi j A_i)).$$

Pour le calcul des fonctions $\psi_{j,l}$, on distinguera donc les cas j, l pairs ou impairs.

j et l pairs.

$$\begin{aligned} \psi_{j,l} &= \int_0^T \cos(2\pi j s)\cos(2\pi l s)\overline{y}_n(s)ds \\ &= \begin{cases} \frac{1}{4\pi(j+l)}\sum_{i=1}^n \sin(2\pi(j+l)A_i) + \frac{1}{4\pi(j-l)}\sum_{i=1}^n \sin(2\pi(j-l)A_i) \text{ si } j \neq l \\ \frac{1}{8\pi j}\sum_{i=1}^n \sin(4\pi j)A_i) + \frac{1}{2}\sum_{i=1}^n A_i \text{ si } j = l \end{cases} \end{aligned}$$

j pair et l impair.

$$\begin{aligned} \psi_{j,l} &= \int_0^T \cos(2\pi j s)\sin(2\pi l s)\overline{y}_n(s)ds \\ &= \frac{1}{4\pi(j+l)}\sum_{i=1}^n (1 - \cos(2\pi(j+l)A_i)) + \frac{1}{4\pi(j-l)}\sum_{i=1}^n (\cos(2\pi(j-l)A_i) - 1). \end{aligned}$$

j et l impairs.

$$\begin{aligned} \psi_{j,l} &= \int_0^T \sin(2\pi l s)\sin(2\pi j s)\overline{y}_n(s)ds \\ &= \begin{cases} \frac{1}{4\pi(j+l)}\sum_{i=1}^n (-\sin(2\pi(j+l)A_i)) + \frac{1}{4\pi(j-l)}\sum_{i=1}^n (\sin(2\pi(j-l)A_i)) \text{ si } j \neq l \\ \frac{1}{8\pi j}\sum_{i=1}^n ((-\sin(4\pi j)A_i) - A_i) \text{ si } j = l \end{cases} \end{aligned}$$

Fonctions splines. Une base de fonctions splines de degré q à noeuds équidistants est donnée par :

$$\phi_0 \equiv 1, \ \phi_1(s) = s, \ldots, \phi_q(s) = s^q, \ \phi_{q+1} = [(s-\Delta)^+]^q, \ldots, \phi_{m-1}(s) = [(s-(m-1-q)\Delta)^+]^q.$$

Avec $\Delta = 1/(m-q)$. On a donc $m-(q+1)$ noeuds intérieurs et m fonctions de base. Pour le calcul des termes $\psi_{j,l}$, on distingue pour $q > 0$ les cas $j, l \leq q$ et $j, l > q$.

Histogramme

$q = 0$

$\psi_{j,l} = 0$ si $j \neq l$, et si ν désigne la mesure de Lebesgue, on obtient :

$$\begin{aligned} \xi_j &= \psi_{j,j} \\ &= \sum_{i=1}^{n} \int_0^{A_i} \mathbf{I}_{I_j}(s) ds \\ &= \sum_{i=1}^{n} \nu([0, A_i] \cap I_j). \end{aligned}$$

Splines affines

$q = 1$

Dans le cas de la deuxième méthode, on distingue les cas $j > q$ et $j \leq q$.

$j \leq q$: on obtient $\xi_j = \frac{1}{j+1} \sum_{i=1}^{n} A_i^{j+1}$.

$j > q$: $\xi_j = \frac{1}{2} \sum_{i=1}^{n} (A_i - t_{j-1})^2$.

Pour la méthode 1, on distingue trois cas :

$j \leq q$ et $l \leq q$

$\psi_{j,l} = \frac{1}{j+l+1} \sum_{i=1}^{n} A_i^{j+l+1}$.

$j \leq q$ et $l > q$

On obtient en posant $t_{l-1} = ((l-1)\Delta)$:

$$\begin{aligned} \psi_{j,l} &= \sum_{i=1}^{n} \int_0^{A_i} s^j [(s - t_{l-1}) +] ds \\ &= \int_{t_{l-1}}^{A_i} s^j (s - t_{l-1}) ds \\ &= \frac{1}{j+2} \sum_{i=1}^{n} \left(A_i^{j+2} - t_{l-1}^{j+2} \right) - \frac{t_{l-1}}{j+1} \sum_{i=1}^{n} \left(A_i^{j+1} - t_{l-1}^{j+1} \right). \end{aligned}$$

$j > q, l > q,$

En posant $v = \max(t_{l-1}, t_{j-1})$ on obtient :

$$\begin{aligned} \psi_{j,l} &= \sum_{i=1}^{n} \int_{v=\max(t_{l-1}, t_{j-1})}^{A_i} (s - t_{l-1})(s - t_{j-1}) ds \\ &= \sum_{i=1}^{n} (A_i - v) \left[\frac{1}{3}(A_i^2 + A_i v + v^2) - \frac{t_{l-1} + t_{j-1}}{2}(A_i + v) + t_{l-1} t_{j-1} \right]. \end{aligned}$$

Splines cubiques

On distingue les cas $j > q$ et $j \leq q$.

$\underline{j > q}$

On obtient $\xi_j = \frac{1}{4}\sum_{i=1}^{n}(A_i - t_{j-3})^4$.

$\underline{j \leq q}$ On obtient $\xi_j = \sum_{i=1}^{n} A_i^{j+1}$.

Pour la méthode 2 et $q = 3$ on distinguera deux cas, le cas $j \leq q$ et $l \leq q$ étant identique au premier cas de $q = 1$:

$\underline{j \leq q \text{ et } l > q}$

$$\begin{aligned}
\psi_{j,l} &= \sum_{i=1}^{n}\int_0^{A_i} s^j[(s-t_{l-1})+]^3 ds \\
&= \int_{t_{l-3}}^{A_i} s^j(s-t_{l-3})^3 ds \\
&= \frac{1}{j+4}\sum_{i=1}^{n}(A_i^{j+4} - t_{l-3}^{j+4}) - \frac{3t_{l-3}}{j+3}\sum_{i=1}^{n}(A_i^{j+3} - t_{l-3}^{j+3}) \\
&+ \frac{3t_{l-3}^2}{j+2}\sum_{i=1}^{n}(A_i^{j+2} - t_{l-3}^{j+2}) - \frac{t_{l-3}^2}{j+1}\sum_{i=1}^{n}(A_i^{j+1} - t_{l-3}^{j+1}).
\end{aligned}$$

$\underline{j > q \text{ et } l > q}$

En posant $v = \max(t_{l-3}, t_{j-3})$ on a :

$$\begin{aligned}
\psi_{j,l} &= \sum_{i=1}^{n}\int_v^{A_i} (s-t_{l-3})^3(s-t_{j-3})^3 ds \\
&= \frac{1}{7}\sum_{i=1}^{n}(A_i^7 - v^7) - \frac{1}{2}(t_{l-3}+t_{j-3})\sum_{i=1}^{n}(A_i^6 - v^6) \\
&+ \frac{3}{5}(t_{l-3}^2 + 3t_{l-3}t_{j-3} + t_{j-3}^2)\sum_{i=1}^{n}(A_i^5 - v^5) \\
&- \frac{1}{4}(t_{l-3}^3 + 9t_{l-3}^2 t_{j-3} + 9t_{j-3}^2 t_{l-3} + t_{j-3}^3)\sum_{i=1}^{n}(A_i^4 - v^4) \\
&+ (t_{j-3}t_{l-3}^3 + 3t_{l-3}^2 t_{j-3}^2 + t_{j-3}^3 t_{l-3})\sum_{i=1}^{n}(A_i^3 - v^3) \\
&- \frac{3}{2}(t_{l-3}t_{j-3})^2(t_{j-3}+t_{l-3})\sum_{i=1}^{n}(A_i^2 - v^2) + t_{j-3}^3 t_{l-3}^3(A_i - v).
\end{aligned}$$

6.4 Estimation dans des modèles de durées de vies avec censure.

6.4.1 Cas de l'histogramme

Les trois estimateurs basés sur l'histogramme, introduits à la section sont comparés dans le cas des modèles de durées avec censure. Les figures présentent des résultats d'estimation pour respectivement 200, 400 et 800 données. On peut remarquer que dans tous les cas, les meilleurs résultats sont obtenus pour le mélange d'histogrammes à intervalles de mêmes longueurs.

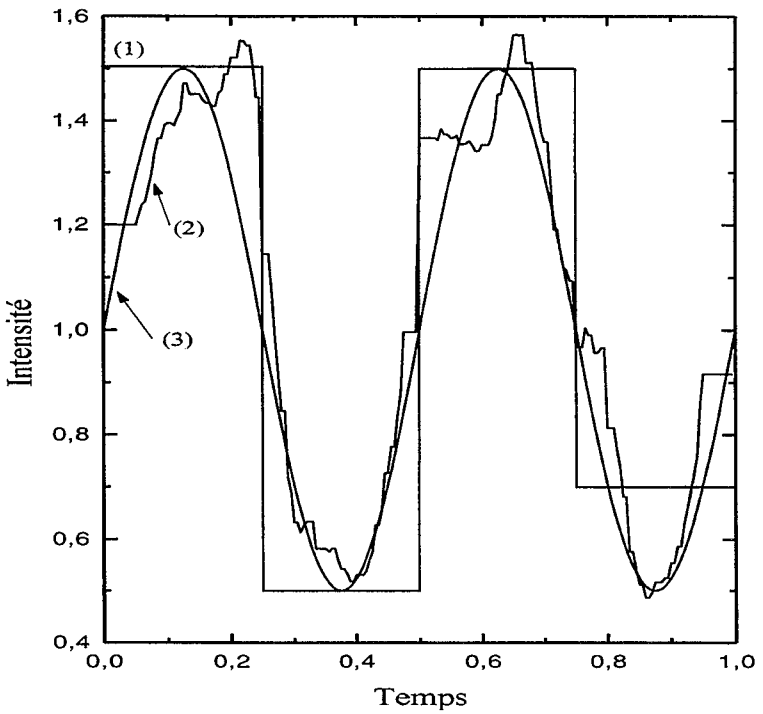

Figure 6.3 : 200 processus. 130 morts observées. (1) Estimateurs à intervalles de même longueurs et à intervalles de longueurs variables, confondus. (2) Melange. (3) Fonction théorique.

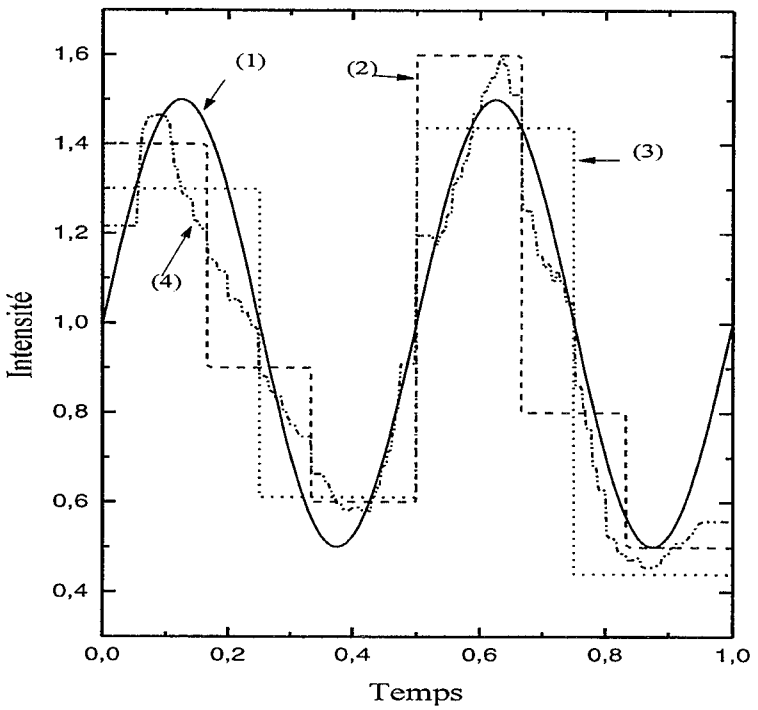

Figure 6.4 : 400 processus. 245 morts observées. (1) Fonction théorique. (2) Estimateur à intervalles de même longueurs. (3) Estimateur à intervalles de longueurs variables. (4) Melange.

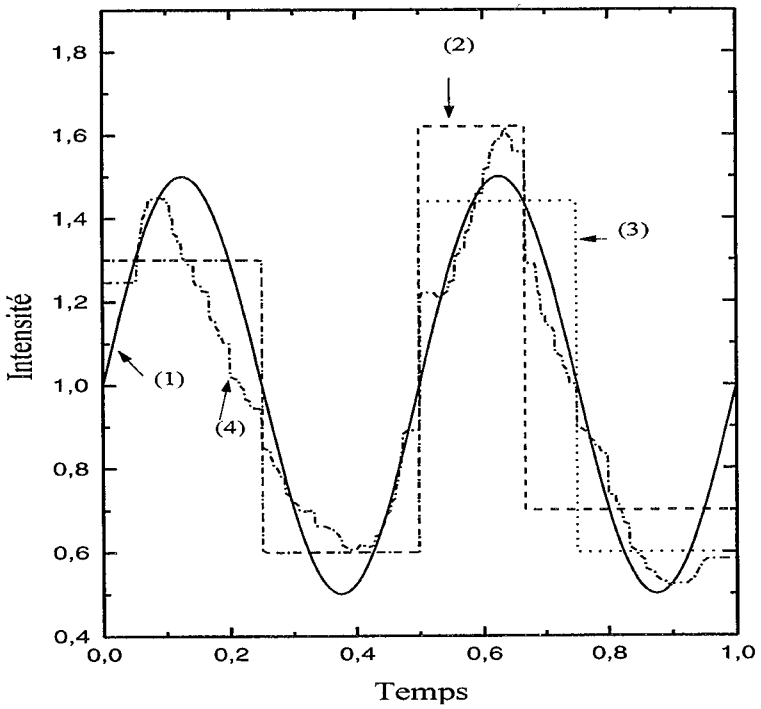

Figure 6.5 : 800 processus. 496 morts observées. (1) Fonction théorique. (2) Estimateur à intervalles de longueurs variables. (3) Estimateur à intervalles de même longueurs. (4) Melange.

6.4.2 Recherche simultanée dans plusieurs familles.

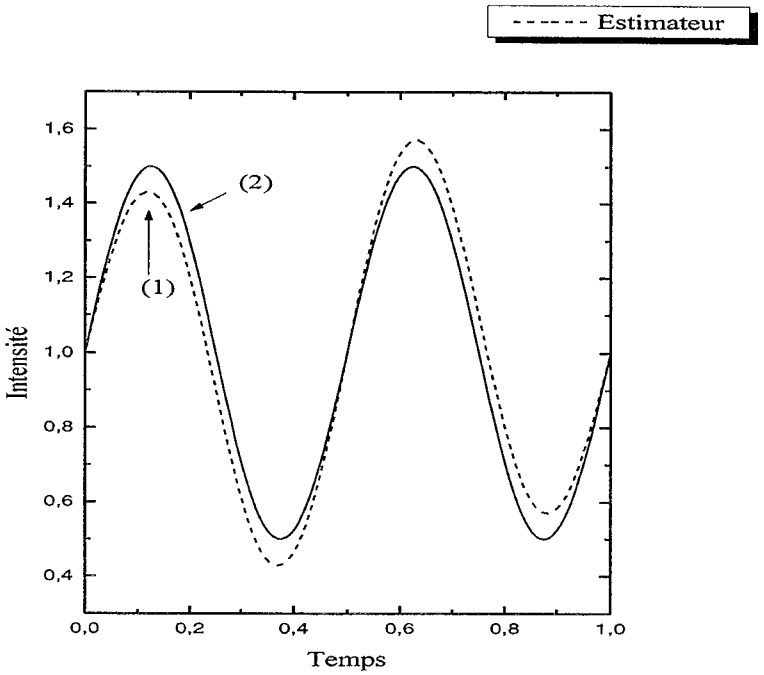

Figure 6.6 : 400 processus. 252 morts observées. (1) Estimateur (2) Fonction théorique.

N	Erreur	Ecart-type	Famille	Fonction
200	0.27	0.11	40%	1%
400	0.19	0.1	64%	2%
600	0.16	0.08	64%	4%
800	0.13	0.06	76%	5%

Figure 6.7 : Comportement asymptotique.

N	Erreur	Ecart-type	Famille	Fonction
200	0.28	0.29	61%	11%
400	0.16	0.14	62%	13%
600	0.15	0.1	63%	16%
800	0.13	0.07	65%	23%

Figure 6.8 : Comportement asymptotique dans le cas d'une fonction exponentielle de paramètre 2.

6.4.3 Commentaires.

Dans toutes les simulations, nous avons considéré comme censure, la fonction indicatrice de l'intervalle $[0, T]$, soit pour la durée X observée, $Y(s) = \mathbf{I}_{\{\min(X,T)>s\}}$.
Les trois estimateurs basés sur l'histogramme, ont pour avantage d'être très faciles à calculer, (leurs coefficients sont solution d'une équation linéaire), par contre ils présentent un grand inconvénient, le manque de lissitude. Les résultats obtenus pour le mélange d'histogrammes, montrent qu'on peut obtenir de très bonnes estimations d'une fonction continue. Ces estimateurs semblent fournir un bon compromis entre qualité d'estimation et temps de calcul. Il serait intéressant d'étudier plus en détail la méthode d'estimation basée sur des mélanges dont la dimension optimale doit être déterminée par le critère de complexité minimale et d'inclure le cas des mélanges basés sur des intervalles de longueurs variables.
La recherche simultanée dans plusieurs familles fournit une bien meilleure qualité d'estimation, compte tenu du fait que l'histogramme en est un cas particulier, et que l'ensemble des familles de fonctions candidates contient des fonctions lisses. Les données du tableau de la figure 6.8 ont été obtenues en simulant pour chaque taille d'échantillon 100 processus de même loi, de fonction d'intensité :

$$y = 0.5 \sin(4\pi x) + 1.$$

La fonction estimée fait parti de l'une des familles (famille de fonctions trigonométriques), et on s'attendrait d'après les résultats théoriques, à pouvoir la découvrir exactement. Nous avons compté dans les colonnes "famille" et "fonction", les pourcentages respectifs d'identification de la bonne famille, et de découverte de la fonction estimée. Dans la figure 6.8, on remarque dans la colonne "fonction", que le taux de découverte de la fonction estimée augmente régulièrement. En ce qui concerne le taux d'identification de la bonne famille, on remarque également une progression, mais moins marquée entre les tailles d'échantillons 400 et 600. L'erreur d'estimation (colonne "erreur") décroît dans un premier temps assez rapidement de 200 à 400 données, puis plus lentement pour des échantillons de taille 600. Nous pensons que cette irrégularité dans la progression de la qualité des résultats est due à une forte variance. La colonne "ecart-type" montre de grandes fluctuations dans l'erreur d'estimation. Cette forte variance est due au fait que la seconde partie du critère, la partie complexité aura tendance à favoriser les fonctions les plus simples du point de vue de la description (histogrammes à intervalles équidistants). Une manière

naturelle d'améliorer la variance, serait d'exclure pour les petits échantillons, des bases qui ne possèdent pas de bonnes qualités d'approximation, et de ne permettre la recherche dans celles-ci qu'à partir d'une taille suffisamment grande pour que l'influence de la partie "complexité" du critère soit faible. Enfin, il serait intéressant d'étudier le comportement de l'estimateur dans le cas de durées de vies non-censurées, afin de mesurer l'influence de la censure, et dans le cas des processus de Poisson, pour mesurer l'influence du nombre moyen de sauts sur la qualité de l'estimation.

Annexes

Annexe A

Annexe.

A.0.4 Un complément sur les machines de Turing.

On peut en donner la définition formelle suivante :

Définition A.1. (Machine de Turing) *Une machine de Turing est formellement décrite par un septuple* $M = (Q, \Gamma, \Sigma, \delta, s, B, F)$ *où :*

- Q *est un ensemble fini d'états.*
- Γ *est l'alphabet du ruban.*
- $\Sigma \subseteq \Gamma$ *est l'alphabet d'entrée (alphabet utilisé pour le mot d'entrée).*
- $s \in Q$ *est l'état initial.*
- $F \subseteq Q$ *est l'ensemble des états accepteurs.*
- $B \in \Gamma - \Sigma$ *est le symbole "Blanc".*
- $\delta : Q \times \Gamma \longrightarrow Q \times \Gamma \times (L, R)$ *est la fonction de transition (L et R désignant le déplacement de la tête de lecture vers la gauche et vers la droite).*

Une machine de Turing fonctionne à temps discret. Le temps t est à valeurs dans l'ensemble des entiers naturels, la machine commence à fonctionner à $t = 0$. A cet instant, la tête de lecture est située sur la case de départ et l'unité de contrôle est dans un état particulier q_0. D'autre part, toutes les cases contiennent des B sauf une suite finie de cases adjacentes s'étendant de la case

de départ vers la droite et contenant des 0 et des 1. Cette chaîne binaire est appelée l'entrée. A chaque instant elle est dans un état interne (en mémoire) particulier; la tête de lecture-écriture examine une à une les lettres inscrites sur le ruban. A chaque étape de l'exécution, la machine est dans une configuration qui peut s'exprimer par le quadruplet (p, t, a, q) où p est l'état courant de l'unité de contrôle, t est le symbole analysé, a est l'opération suivante de type écriture d'un symbole ou déplacement de la tête, soit un élément de $S = (0, 1, B, L, R)$, et q est l'état de l'unité de contrôle à exécuter après cette étape.

La valeur lue t et l'état p déterminent une exécution a et un état suivant q de l'unité de contrôle déterminé par la table de transition d'états. On peut résumer ce fonctionnement séquentiel, si p_n, t_n représentent l'état et le symbole lu à l'instant n, et a_n l'opération effectuée à l'instant n, on aura

$$p_{n+1} = f(t_n, p_n), a_{n+1} = g(t_{n+1}, p_{n+1})$$

Si le couple (t_{n+1}, p_{n+1}) ne correspond pas à une entrée de la fonction g, alors la machine ne fait rien, on lui permet ainsi de s'arrêter. Si à l'arrêt $p_n \in F$ alors la fin d'exécution sera considérée valable, ces états sont dits "accepteurs". Si la machine se trouve à la fin de l'exécution dans un état q de $F \subseteq Q$, le résultat sera considéré comme la sortie de l'exécution, sinon la sortie sera indéfinie. On peut associer à chaque machine de Turing une fonction partielle de la manière suivante. L'entrée de la machine est un n-uplet $(x_1, , x_n)$ de chaînes binaires concaténées en une seule chaîne binaire constituée de versions distinguables des x_i[1]. L' entier représenté par la chaîne binaire maximale bordée de symboles "Blanc" dans laquelle un élément est analysé quand la machine s'arrête dans un état accepteur est appelé résultat de l'exécution.

A.0.5 Deux résultats relatifs aux martingales.

Le premier résultat présenté ici sert à établir le théorème 4.12. Il est tiré de [18], et permet d'obtenir des majorations presques-sûres pour les moments d'une martingale.

Théorème A.2. (Burkholder (1973)) *Pour tout $p \geq 1$ il existe des constantes c_p et C_p telles que pour toute martingale M,*

$$c_p E\left[[M]_1^{p/2}\right] \leq E\left[\sup_{t\leq 1} |M_t|^p\right] \leq C_p E\left[[M]_1^{p/2}\right] \tag{A.1}$$

[1] On peut séparer les mots constituant l'entrée

En particulier pour $p = 2$

$$E\left[\sup_{t\leq 1} M_t^2\right] \leq K_2 E\left[[M]_1\right] = K_2 E\left[< M >_1\right].$$

Le théorème suivant est un théorème central limite pour des suites de martingales. Il permet d'établir le théorème 5.1.

Théorème A.3. Rebolledo (1978) *Soit* V *une fonction continue sur* $[0,T]$, *non-décroissante, telle que* $V_0 = 0$, *et supposons que*

(a) pour tout t, $< M_n >_t \to V_t$ *en probabilité ;*

(b) il existe des constantes $b_n \downarrow 0$ *telles que*

$$P\left(\sup_{t\leq T} |\Delta M_t^n| \leq b_n\right) \to 1.$$

Alors, il existe une martingale M *continue, centrée telle que* $< M >_t = V_t$, *telle que* $M^n \to_d M$ *sur* $D[0,T]$, *où* $\to_d$ *désigne la convergence en loi. La martingale* M *est un processus gaussien centré à accroissements indépendants et* $E[M_t M_s] = V_{t\wedge s}$.

Index

Bibliographie

[1] R. N. Moll A. J. Kfoury and M. A. Arbib. *A programming approach to computability.* Springer Verlag, New-York, 1982.

[2] O. Aalen . *Statistical inference for a family of counting processes.* PhD thesis, University of California, Berkeley, Berkeley, CA, December 1975.

[3] O. Aalen. Nonparametric inference for a family of counting processes. *Ann. Statist.*, 6:701–726, 1991.

[4] A. Anderson and A. Senthiselvan. Smooth estimates for the hazard function. *Journal of the Royal Statistical Society*, B 42:322–327, 1980.

[5] A. Antoniadis and G. Grégoire . Penalized likelihood estimation for rates for censored survival data. *Scand. J. Statist.*, 17:43–63, 1990.

[6] A. Antoniadis, G. Grégoire, and I.W. McKeague. Wavelet methods for curve estimation. *Journal of the American Statistical Association*, 89(428):1340–1353, 1994.

[7] R. R. Bahadur. An optimal property of the likelihood ratio statistic. volume 1 of *Proc. Fifth Berkeley Symp. Math. Statist. Probab.*, pages 13–26, 1966.

[8] A. R. Barron. *Logically smooth density estimation.* PhD thesis, Stanford University, California, Stanford, CA 94305, August 1985.

[9] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *Transactions on Information Theory*, 37:1034–1054, 1991.

[10] A. R. Barron and C. Sheu. Approximation of density functions by sequences of exponential families. *Ann. Statist*, 6, 1991.

[11] R. BARTOSZYNSKI, B. W. BROWN, C. M. MCBRIDE, and j. R. THOMPSON. Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *Annals of Statistics*, 9:1050–1060, 1981.

[12] M. S. BIRMAN and M. Z. SOLOMJAK. Piecewize-polynomial approximations of functions of the classes W_p^α. *Mat. USSR-Sbornik*, 2:295–317, 1967.

[13] R. BOEL, P. VARAYIA, and E. WONG. Martingales on jump processes I:Representation results. *SIAM J. Control*, 13:999–1021, 1975.

[14] D. BOSQ and J. LECOUTRE. *Theorie de l'estimation fonctionnelle.* Economica, Paris, 1987.

[15] N. BRESLOW. Contribution to the discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society*, B 34:187–220, 1972.

[16] J. BRETAGNOLLE and C. HUBER. Estimation des densités:Risque minimax. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete.*, 47:119–137, 1979.

[17] L. D. BROWN. Non local optimality of appropriate likelihodd ratio tests. *Ann. Math. Statist.*, 42:1206–1240, 1971.

[18] D. BURKHOLDER. Distribution function inequalities for martingales . *Ann. Prob.*, 1:19–42, 1973.

[19] G. CHAITIN. On the length of programs for computing finite binary sequences:statistical considerations. *Journal of the ACM*, 22:329–340, 1944.

[20] G. CHAITIN. Journal of the ACM. *On the length of programs for computing finite binary sequences*, 13:547–569, 1966.

[21] G. CHAITIN. A theory of program size formally identical to information theory. *Journal of the ACM*, 22:329–340, 1975.

[22] G. CHAITIN. Algorithmic entropy of sets. *Computers and mathematics with Applications*, 2:233–245, 1976.

[23] N. CHOMSKY. On certain formal properties of grammers. *Information and control*, 2:137–161, 1959.

[24] Y. S. CHOW and H. TEICHER. *Probability theory : Independence, Interchangeability, Martingales.* Springer Verlag, New-York, 1978.

[25] T. M. COVER. Enumerative Source Encoding. *IEEE Transactions on Information Theory*, 29:73–77, 1973.

[26] D. R COX and P. A. W. LEWIS. *The statistical analysis of series of events.* Methuen, London, 1966.

[27] I. CSISZAR and G. LONGO. On the error exponent for source coding for testing simple statistical hypotheses. *Studia Sci. Math. Hungar.*, 6:181–191, 1971.

[28] L. D. DAVISSON. Universal noiseless coding. *Transactions on information Theory*, 19:783–785, 1973.

[29] L. D. DAVISSON. Minimax noiseless universal coding for Markov sources. *IEEE Transactions on Information Theory*, 29:211–215, 1983.

[30] L. D. DAVISSON and A. LEON-GARCIA. A source matching approach to finding minimax codes. *IEEE Transactions on Information Theory*, 26:166–174, 1980.

[31] S.Y. EFROIMOVICH and M.S. PINSKER. Estimation of square-integrable probability density of a random variable. *Problems Inform. Transmission*, 18:175–189, 1983.

[32] P. ELIAS. Universal codeword sets and representation of the integers. *IEEE Transactions on Information Theory*, 21:194–203, 1975.

[33] R. FLETCHER and C. M. REEVES. Function minimization by conjugate gradients. *Computing. J.*, 7:149–154, 1964.

[34] F. HAIGHT. *Handbook of the Poisson distribution.* Wiley, New-York, 1967.

[35] P. HALL and E. J. HANNAN. Stochastic complexity and nonparametric density estimation. *Biometrika*, 75:705–714, 1988.

[36] W. HOEFFDING and J. WOLFOWITZ. Distinguishability of sets of distributions. *Ann. Math. Statist.*, 29:700–718, 1958.

[37] J. E. HOPCROFT and J.E. ULLMAN. *Formal langages and their relations to automata.* Addison-Wesley, Reading, Massachusetts, 1969.

[38] P. J. HUBER and V. STRASSEN. Minimax tests and the Neyman-Pearson lemma for capacities. *Annals of Statistics*, 1:251–263, 1973.

[39] J. JACOD. Multivariate point processes :Predictable projection, Radon-Nikodym derivatives, representations of martingales. *Z.Wahrscheinlichkeitstheorie und Verw.Gebiete*, 31:235–253, 1975.

[40] J.D. KALBFLEISCH and R.L. PRENTICE. Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60:267–278, 1973.

[41] A. F. KARR. *Point processes and their statistical inference.* Marcel Dekker, New-York, 1986.

[42] A. F. KARR. Maximum likelihood estimation in the multiplicative intensity model via sieves. *Ann. Statist*, 15:473–490, 1987.

[43] H. J. KELEY and G. E. MYERS. Conjuguate direction methods for parameter optimization. *18th Congr. of the Int. Astronaut. Fed.*, 1967.

[44] J. C. KIEFFER. A unified approach to weak universal coding. *IEEE Transactions on Information Theory*, 26:674–682, 1978.

[45] A. N. KOLMOGOROV. Three approaches to the quantitative definition of information. *Problemy Peradachi Information*, 1:3–11, 1965.

[46] L.G. KRAFT. A device for Quantizing, Grouping, and Coding Amplitude Modulated Pulses. Master's thesis. Master's thesis, Dept. of Electrical Engineering, 1949.

[47] R. E. KRICHEVSKY and V. K. TROFIMOV. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27:199–207, 1981.

[48] J. LEŚKOW and R. RÓŻAŃSKI. Histogram maximum likelihood estimator in the multiplicative intensity model. *Stoch. Processes. Appl.*, 17:151–159, 1989.

[49] S. K. LEUNG-YAN-CHEONG and T. M. COVER. Some equivalences between Shannon entropy and Kolmogorov complexity. *Trans. Inform. Theory*, 24:331–339, 1978.

[50] G. LONGO and A. SGARRO. The source coding theorem revisited: a combinatorial approach. *IEEE Transactions on Information Theory*, 25:544–548, 1979.

[51] J. MARKHAM, D. L. SNYDER, and J. R. COX. A numerical implementation of the maximum likelihood method of parameter estimation for Tracer-kinetic data. *J. Mathematical Biosciences*, 28:275–300, 1976.

[52] A. A. MARKOV. Theory of Algorithms. *Trudy Mathematicheskego Instituta imeni V.A Steklova*, 42, 1954. Translation by Israel Program for Scientific Translations, Jerusalem , 1961.

[53] I. MCKEAGUE. Estimation for a semimartingale regression model using the method of sieves. *Ann. Statist*, 14:579–589, 1986.

[54] I. MCKEAGUE. Asymptotic theory for weighted least squares estimators in Aalen's additive risk model. *Contemporary Math*, 80:139–152, 1988.

[55] B. MCMILLAN. Two inequalities implied by unique decipherability. *IEEE Transactions on Information Theory*, 2:115–116, 1956.

[56] P. ODIFREDDI. *Classical recursion theory.* North-Holland, 1989.

[57] E.L. POST. Finite combinatory processes-formulation 1. *Journal of Symbolic Logic*, 1:103–105, 1936.

[58] E.L. POST. Formal reductions of the general combinatorial decision problem. *American Journal of Mathematics*, 65:197–215, 1943.

[59] G. QIAN, G. GABOR, and R. P. GUPTA. Minimum Complexity estimation: A decision -theoretic approach. *IEEE Transactions on Information Theory*, 40:1081–1191, 1994.

[60] H. RAMLAU-HANSEN. Smoothing counting process intensities by means of kernel functions. *Ann. Statist*, 11:453–466, 1983.

[61] J. RICE and M. ROSENBLATT. Estimation of the log survivor function and hazard function. *Sankhya*, A 38:60–78, 1976.

[62] J. RISSANEN. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[63] J. RISSANEN. A universal prior for integers and estimation by minimum description length. *Ann. Statist*, 11:416–431, 1983.

[64] J. RISSANEN, T. P. SPEED, and B. Yu. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38:315–323, 1992.

[65] J. ROBINSON. General recursive functions. *Proceedings of the American Mathematical Society*, 1:703–718, 1950.

[66] H. ROGERS. *Theory of recursive functions and effective computability.* McGraw-Hill, New-York, 1967.

[67] C. E. SHANNON. A Mathematical Theory of Communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[68] C. E. SHANNON. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana Ilinois, 1949.

[69] D. L. SNYDER. *Random point processes.* Wiley, New-York, 1975.

[70] D. L. SNYDER and M. I. MILLER. *Random point processes in time and space.* Springer, New-York, 1991.

[71] R. J. SOLOMONOFF. A formal theory of inductive inference. *Information and Control*, pages 1–22,224–254, 1964.

[72] M. A. TANNER and W. H. WONG. The estimation of the hazard function by the kernel method. *Annals of Statistics*, 11:989–993, 1983.

[73] V. K. TROFIMOV. Redundancy of universal coding of arbitrary Markov Sources. *Problemy Peredachi Informatsii*, 10:16–24, 1974.

[74] A. TURING. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42, 1936.

[75] G. S. WATSON and M. R. LEADBETTER. Hazard Analysis I. *Biometrika*, 51:175–184, 1964.

Résumé

Soit un processus ponctuel observé sur un intervalle de temps fini, et admettant une intensité stochastique conforme au modèle de Aalen. La fonction d'intensité du processus est estimée à partir d'un échantillon indépendant et identiquement distribué de paires constituées par la réalisation du processus ponctuel et du processus prévisible associé, par la minimisation d'un critère qui représente la longueur d'un code variable pour les données observées. L'estimateur de complexité minimale est la fonction minimisant ce critère dans une famille de fonctions candidates. Un choix judicieux des fonctions de complexité permet de définir ainsi des codes universels pour des réalisations de processus ponctuels. Les estimateurs de la fonction d'intensité obtenus par minimisation de ce critère sont presque-sûrement consistants au sens de l'entropie, et au sens de la distance de Hellinger pour des fonctions de complexité satisfaisant l'inégalité de Kraft. L'étude des vitesses de convergence pour la distance de Hellinger montre qu'elles sont majorées par celle de la redondance du code. Ces vitesses, sont précisées dans le cas des familles de fonctions trigonométriques, polynomiales et splines. Dans le cas particulier des processus de Poisson et des modèles de durées de vie avec censure, les mêmes vitesses de convergence sont obtenues pour des distances plus fortes. D'autres propriétés de l'estimateur sont présentées, notamment la découverte exacte de la fonction inconnue dans certains cas, et la normalité asymptotique. Des suites de tests exponentiels consistants sont également étudiées. Le comportement numérique de l'estimateur est analysé à travers des simulations dans le cas des modèles de durées de vie avec censure.

Mots clés : estimation non-paramétrique, processus ponctuels, fonction d'intensité,modèle de Aalen, complexité, codage universel, longueur de description minimale, entropie.

Abstract

A counting process with stochastic intensity in the Aalen model is observed over a finite time interval. The intensity function of the process is estimated from an independant and identically distributed sample of pairs consisting of the point process and the related predictable process, by minimizing a criteria which represents the length of a two-stage variable code for the observed data. The minimum complexity estimator is the function which minimizes this criteria in a set of candidate functions. Universal codes for the outcomes of a point process can be built by a suitable choice of the complexity functions. For complexity functions which satisfy Kraft inequality, the estimators of the intensity function obtained by the minimization of this criteria are almost-surely consistant for the Hellinger distance and the Kullback-Leibler distance. The convergence rates in Hellinger distance are showned to be bounded by the code redundancy. These rates are specified in the case of polynomial, trigonometric and spline candidate basis families. In the particular case of Poisson processes and censored survival lifetimes, the same rates are seen to hold for stronger distances. Other properties of the estimator are studied, mainly the discovery of the true function in some cases and the asymptotic normality. Consistant sequences of exponential tests are also studied. The numerical behavior of the estimator is analysed throughout simulations in the case of censored lifetimes.

Keywords : nonparametric estimation, point processes, intensity function, Aalen model, complexity, universal coding, minimum description length, entropy