Table des matières

Introduction

Un véritable engouement dans divers domaines est observé depuis ces vingt dernières années pour les méthodes de modélisation bayésienne. Leur approche intégrant la prise en compte d'informations *a priori* constitue leur principale attractivité dans les domaines de l'environnement, de l'agronomie, de la médecine et bien d'autres. Une modélisation des connaissances *a priori* au travers d'une loi de probabilité permet de combiner différentes sources d'informations. Cette loi est mise à jour grâce aux observations pour obtenir la loi *a posteriori*. Cependant les solutions analytiques obtenues par ces approches sont complexes. Des méthodes d'approximation ont donc été mises au point afin d'approximer la loi *a posteriori* lorsque l'on ne peut pas le faire analytiquement. Ces approximations sont basées sur le principe de simulation de Monte Carlo. Plusieurs algorithmes existent comme par exemple les méthodes de Monte Carlo par chaîne de Markov ou le filtrage particulaire.

Toute recherche fondamentale peut aboutir à un développement pratique. La recherche clinique est un aboutissement naturel des développements des méthodes statistiques. Cette recherche portant sur l'être humain a des spécificités propres (coûts élevés, problèmes d'éthique, faible nombre de données). Ainsi, l'approche bayésienne fournit un cadre naturel pour résoudre des problèmes d'inférence statistique en recherche clinique. En effet, c'est un domaine où l'on dispose de nombreux jugements, informations de la part des experts, des médecins ou des biologistes. L'interprétation de ces informations au travers d'une loi de probabilité *a priori* apparait comme une chose naturelle. Cette thèse permet de montrer l'utilité des méthodes de modélisation bayésienne dans le cadre de la recherche clinique.

Dans le premier chapitre, nous introduisons les concepts de base de la statistique bayésienne ainsi que les notations et la terminologie appropriée. Une partie est consacrée aux méthodes d'approximations comme les méthodes de Monte Carlo par chaîne de Markov (MCMC) et les méthodes de filtrage particulaire. Ces méthodes ont été décrites car elles sont utilisées et développées dans les paragraphes suivants. Une section est également consacrée au stage doctoral réalisé à l'Université de Sherbrooke. Ce travail en collaboration avec Éric Marchand se place dans le cadre de la théorie de la décision et traite d'intervalles de confiance bayésiens pour un paramètre contraint. Les premiers résultats basés sur des approximations sont intéressants, une des perspectives est de confirmer certaines propriétés théoriques.

Dans le deuxième chapitre, nous traitons de deux modèles couramment util-

isés en recherche clinique : le modèle de Cox et le modèle logistique. Le premier modèle est un modèle courant en analyse de la survie pour étudier la dépendance entre le temps de survie et les covariables. Cependant, la relation entre la fonction de risque associée et les covariables est log-linéaire. Cette hypothèse est remise en question lorsque les effets des covariables sont mieux représentés par des fonctions lisses non linéaires. On introduit donc des fonctions B-splines, où le nombre et la position des nœuds sont considérés comme des variables libres pour modéliser ces effets et améliorer l'ajustement. L'algorithme du Reversible Jump MCMC (RJMCMC) est utilisé pour choisir le nombre et la position des nœuds. Ce problème fait l'objet du premier article accepté dans "Communication in Statistics-Theory and Methods".

Le deuxième modèle est utilisé dans le cadre de l'analyse de la dépendance entre une variable réponse et une ou plusieurs variables explicatives. Ce modèle suppose une relation linéaire entre le *log-odds* et les variables, ce type de modélisation est restrictif. Comme pour le modèle de Cox, on utilise une représentation B-spline pour modéliser cette relation et on met en oeuvre l'algorithme RJMCMC pour sélectionner le nombre et la position des nœuds. Le deuxième article y est consacré.

Le troisième chapitre traite également de la représentation B-spline et du choix du nombre et de la position des nœuds dans un contexte plus général. Une méthode de sélection de modèle via une procédure de pénalisation pour estimer ces paramètres inconnus est proposée. Le modèle choisi est celui minimisant le critère des moindres carrés pénalisé. La fonction de pénalité est estimée à partir des données en utilisant la méthode de Birgé et Massart. Cette méthode est basée sur un mélange de théorie et d'idées heuristiques, l'heuristique des pentes. On obtient ainsi un estimateur réalisant le risque quadratique minimal. Un algorithme de calibration de pénalités reposant sur une généralisation de cette heuristique est également utilisé pour estimer la pénalité.

Le dernier chapitre est consacré à la modélisation bayésienne non paramétrique et notamment aux modèles de mélange suivant un processus de Dirichlet notés modèles MDP. Ces modèles ont été développés pour de nombreuses applications, telles que l'estimation de la densité, l'analyse de la survie, la classification... Ce sont des modèles hiérarchiques non paramétriques. L'algorithme que l'on a utilisé pour les approcher est le Blocked Gibbs Sampler (BGS) qui utilise l'échantillonneur de Gibbs. Plus spécialement, le BGS a été mis en place pour classer des observations issues d'un mélange de lois et pour déterminer le nombre de composantes, leurs poids et leurs paramètres. Une application a été réalisée dans le cadre d'une collaboration avec des biologistes. Une deuxième partie s'intéressse aux processus de Dirichlet hiérarchiques. Ce type d'approche est utilisée dans le cas où l'on a plusieurs groupes de données et que l'on souhaite "lier" ces groupes. L'algorithme du BGS a été adapté pour ce type de problème.

Chapitre 1

L'analyse bayésienne

Ce premier chapitre commence par une introduction à la statistique bayésienne en exposant les concepts de base. Puis, une section sera consacrée à la théorie de la décision dans un cadre bayésien où l'on présentera le début d'un travail réalisé, en collaboration avec Éric Marchand de l'Université de Sherbrooke. Enfin, nous aborderons une présentation de différentes méthodes d'appoximation numérique.

1.1 L'approche bayésienne

L'approche bayésienne fournit un cadre naturel pour résoudre des problèmes d'inférence statistique. Elle se distingue de la statistique classique parce qu'elle considère le(s) paramètre(s) du modèle comme des variables aléatoires. Soit Θ l'espace des paramètres et $\mathcal{X}$ l'espace des observations.

On considère un modèle statistique de loi de probabilité P_θ de densité $p(x|\theta)$ dépendant d'un paramètre inconnu de dimension k : $\theta \in \mathbb{R}^k$. On dispose d'un échantillon aléatoire de n observations $x = (x_1, \ldots, x_n)$ issues de cette distribution. La fonction de vraisemblance associée est donnée par :

$$l(\theta, x) = p(x|\theta). \tag{1.1}$$

La connaissance *a priori* sur le paramètre θ s'exprime, à travers une loi de probabilité nommée loi *a priori* notée π de densité $p(\theta)$. Cette loi peut permettre de traduire les connaissances que l'on a avant l'expérience, avant d'avoir des informations sur les valeurs des observations. On l'interprète comme la représentation formelle sous forme probabiliste de la connaissance sur les paramètres.

On distingue deux types de lois *a priori* : les lois informatives et les lois non informatives. Nous analyserons ces différentes notions dans le paragraphe suivant. L'information sur θ est mise à jour grâce au théorème de Bayes qui permet de prendre en compte l'information apportée par les observations. Ainsi, on obtient la loi de θ conditionnellement aux observations, appelée loi *a posteriori* de θ, notée π_x de densité $p(\theta|x)$. La densité jointe de x et θ s'écrit :

$$p(x, \theta) = p(x|\theta)\, p(\theta). \tag{1.2}$$

La formule de Bayes est basée sur la décomposition inverse, ce qui permet d'obtenir la densité *a posteriori* de θ :

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{m(x)}, \tag{1.3}$$

où $m(x)$ est la constante d'intégration, appelée également densité marginale ou densité prédictive de x donnée par :

$$m(x) = \int_{\Theta} p(x|\theta)\, p(\theta)\, d\theta. \tag{1.4}$$

D'autre part, comme le dénominateur de (1.3) ne dépend pas de θ, on peut écrire :

$$p(\theta|x) \propto p(x|\theta)\, p(\theta). \tag{1.5}$$

Lorsque la distribution *a priori* provient d'une mesure non finie mais σ-finie, au lieu d'une mesure de probabilité, c'est-à-dire qu'elle vérifie :

$$\int_{\Theta} p(\theta)\, d\theta = +\infty,$$

on parle de distribution *a priori* impropre. Les concepts de l'approche bayésienne semblent naturels, simples et flexibles. Les difficultés résident dans :

- la spécification de la loi *a priori* en prenant en compte les informations avant l'observation des données,
- la détermination du modèle d'échantillonnage,
- le calcul de la loi *a posteriori*.

Les calculs menant à la distribution *a posteriori* peuvent être complexes. En effet, la résolution de l'intégrale (1.4) peut s'avérer difficile à résoudre d'un point de vue analytique. Différentes méthodes existent afin de faciliter ce calcul ou d'en donner une approximation. Nous verrons dans la section (1.2) un aperçu de ces quelques méthodes.

L'approche bayésienne a donc permis un changement majeur en statistique en passant de la notion de paramètre inconnu à la notion de paramètre aléatoire [57]. Le paragraphe suivant met en évidence le choix des lois *a priori* comme point sensible dans l'analyse bayésienne.

1.1.1 Les distributions a priori

L'analyse bayésienne se heurte à deux choix :

(*i*) Le premier concerne le choix de la famille de la probabilité d'échantillonnage, il n'est pas spécifique à l'approche bayésienne.

(*ii*) Le second, qui est l'essence de l'analyse bayésienne, est le choix de la distribution *a priori* sur les paramètres.

En ce qui concerne la distribution *a priori*, deux approches existent, le choix d'un prior informatif ou non. Ces deux approches sont expliquées brièvement dans les paragraphes suivants. Pour plus de détails on pourra consulter les livres de Droesbeke, Fine et Saporta [23] et de Robert [57].

Priors informatifs

Un des interêts de l'approche Bayésienne est de combiner les connaissances que l'on a sur les paramètres, avant l'observation des données. Ceci est fait à travers une loi *a priori* informative, avec l'information venant des données. Ce mode de pensée est subjectiviste, i.e. que l'on prend en compte l'information ne résultant pas des données.

Dans ce contexte, les distributions *a priori* sont utilisées pour traduire les connaissances avant l'observation des données. Elles aident à exprimer l'opinion des experts. L'un des critères de choix pour ces lois est de simplifier les calculs, d'où l'utilisation répandue de distributions naturelles conjuguées à un modèle d'échantillonnage. Ces lois facilitent le calcul de la loi *a posteriori* (1.3). Cependant, grâce à l'évolution des méthodes d'approximation numérique, le choix de lois *a priori* simplifiant les calculs est délaissé au profit de lois *a priori* plus pertinentes.

Définissons la notion de distributions naturelles conjuguées à une famille de probabilités d'échantillonnage. Notons $\mathcal{M}$ une famille de lois de probabilité sur Θ.

Définition 1.1.1. *On dira que $\mathcal{M}$ est fermée si, pour toute probabilité de $\mathcal{M}$ choisie comme loi a priori et pour tout échantillon observé, la loi a posteriori déduite est encore un élèment de $\mathcal{M}$.*

ou bien,

Définition 1.1.2. *Une famille $\mathcal{M}$ de distributions de probabilité sur Θ est dite conjuguée (ou fermée par échantillonnage) par une fonction de vraisemblance $p(x|\theta)$ si, pour tout $\pi \in \mathcal{M}$ la distribution a posteriori π_x appartient aussi à $\mathcal{M}$.*

Définissons ensuite les familles exponentielles, elles sont constituées de distributions d'échantillonnage qui permettent toujours la dérivation des distributions *a priori* qui y sont conjuguées. Elles sont étudiées en détail par Brown [14].

Définition 1.1.3. *Posons μ une mesure σ-finie sur $\mathcal{X}$. Posons C et n des fonctions respectivement de $\mathcal{X}$ et Θ dans $\mathbb{R}^+$,et posons R et N des fonctions de $\mathcal{X}$ et Θ dans $\mathbb{R}^k$. La famille de distributions de densités*

$$p(x|\theta) = C(\theta)\, n(x)\, exp\{R(\theta).T(x)\} \tag{1.6}$$

est appelée une famille exponentielle de dimension k. En particulier, quand $\Theta \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^k$, et

$$p(x|\theta) = C(\theta)\, n(x)\, exp\{\theta.x\},$$

la famille est dite naturelle.

De nombreuses distributions usuelles continues et discrètes appartiennent à des familles exponentielles comme par exemple la loi de Dirichlet, la distribution normale, la loi Poisson, la loi Gamma, la loi binomiale, ...[57].

Considérons $p(x|\theta) = n(x)\, exp\{\theta.x - \psi(\theta)\}$ une distribution générique issue d'une famille exponentielle, la proposition suivante détermine une famille conjuguée :

Proposition 1.1.4. *Une famille conjuguée pour $p(x|\theta)$ est donnée par*

$$p(\theta|\mu,\,\lambda) = K(\mu,\,\lambda)\, e^{\theta.\mu - \lambda\,\psi(\theta)},$$

avec $K(\mu,\,\lambda)$ la constante de normalisation de la densité. La distribution a posteriori correspondante est $p(\theta|\mu + x,\,\lambda + 1)$

Certaines lois n'appartiennent pas à une famille exponentielle [57], comme par exemple la loi de Student qui ne peut pas s'exprimer sous la forme (1.6). Aussi, la définition (1.1.3) exclut toutes les lois avec un support non constant alors que certaines d'entre elles admettent des lois *a priori* conjuguées avec un nombre fini de paramètres. C'est le cas de la loi de Pareto et des distributions uniforme $\mathcal{U}_{[-\theta,\,\theta]}$ et $\mathcal{U}_{[0,\,\theta]}$.

Dans le cas où l'on a beaucoup d'informations avant l'observation des données, la spécification d'une loi *a priori* adéquate s'avère difficile. Une méthode est de déterminer une loi de probabilité subjectiviste découlant d'une spécification de la distance marginale de l'échantillon. Pour résumé, la connaissance de l'expert va s'exprimer à travers la distribution marginale de l'échantillon plutôt qu'en terme d'*a priori* [23].

Une autre approche, en terme de prior informatif, concerne les priors hiérarchiques. On parle, dans la littérature, de spécification hiérarchique de l'*a priori* dans un modèle d'échantillonnage de fonction de vraisemblance $p(x|\,\theta)$ quand la distribution *a priori* sur le paramètre θ est fonction d'un hyperparamètre θ_i sur lequel une distribution *a priori* est spécifiée.

On appelle hyperparamètres les paramètres des distributions *a priori* qui sont mises sur les paramètres de la loi *a priori* $p(\theta)$.

Définition 1.1.5 (Modèle bayésien hiérarchique)**.** *Un modèle bayésien hiérarchique est un modèle statistique bayésien, $(p(x|\,\theta),\, p(\theta))$, où la distribution a priori est décomposée en distributions conditionnelles de densités*

$$p_1(\theta|\,\theta_1),\, p_2(\theta_1|\,\theta_2),\, \ldots,\, p_n(\theta_{n-1}|\,\theta_n)$$

et une densité marginale $p_{n+1}(\theta_n)$ telle que

$$p(\theta) = \int_{\Theta_1\times\cdots\times\Theta_n} p_1(\theta|\,\theta_1)\, p_2(\theta_1|\,\theta_2)\, \ldots\, p_{n+1}(\theta_n)\, d\theta_1 \ldots d\theta_n.$$

On appelle θ_i l'hyperparamètre de niveau i.

Plaçons nous dans le cas où l'on a un hyperparamètre de niveau 1. On construit une probabilité sur $(\theta_1,\,\theta,\,x)$ décomposée en $p(\theta_1)\, p(\theta|\theta_1)\, p(x|\theta)$ où x ne dépend pas de θ_1 conditionnellement à θ. De plus, l'introduction de ce type de prior permet une spécification *a priori* permettant, par exemple, de mélanger des *a priori* informatifs et non informatifs Nous verrons dans le chapitre 4 l'utilité d'une telle modélisation dont les principales avancées reposent sur les travaux de Good ([35], [36]).

Priors non informatifs

Le paragraphe précédent a montré que les priors conjugués sont très utiles. Cependant dans le cas où l'on n'a pas ou peu d'information *a priori*, leur utilisation est justifiée par le seul fait que ces lois *a priori* facilitent les calculs. On recherche donc à défaut des distributions *a priori* non informatives permettant d'exprimer l'ignorance *a priori* tout en gardant un cadre bayésien, ce qui permet de conserver les paramètres aléatoires. Le choix de tels *a priori* se caractérise par la spécification d'une mesure sur l'espace paramétrique Θ à partir du mécanisme d'échantillonnage et non d'une probabilité. Trois approches sont couramment utilisées : la recherche d'*a priori* invariantes, l'approche de Jeffrey [43] et l'approche dite des *a priori* de référence [7].

La première approche est assez naturelle, elle consiste à traduire l'ignorance *a priori* par l'invariance par rapport à une famille Φ de transformations de Θ en lui-même. En effet, dire que la distribution *a priori* pour le paramètre θ est la même que pour le paramètre $\phi(\theta)$, ϕ appartenant à la famille Φ exprime certainement l'ignorance sur le paramètre inconnu. Les mesures utilisées dans ce cadre sont les mesures Haar invariantes à droite. Cette approche est limitée parce qu'elle fait appel à une structure d'invariance qui peut, ne pas exister ou ne pas être intéressante, pour la prise de décisions. L'introduction des distributions *a priori* non informatives de Jeffreys permet une approche plus intrinsèque ne nécessitant pas de structure d'invariance. Elles sont basées sur l'information de Fisher qui est donnée, dans le cas unidimensionnel, par :

$$I(\theta) = \mathbb{E}_\theta\Big(\frac{\partial log\, p(X|\theta)}{\partial\theta}\Big)^2.$$

La distribution *a priori* correspondante est

$$p^*(\theta) \propto I^{1/2}(\theta),$$

modulo une constante de normalisation quand p^* est propre. Bernardo [7] propose une modification de l'approche de Jeffreys appelée approche des lois *a priori* de référence, qui permet de distinguer les paramètres d'intérêts des paramètres de nuisance. Elle permet la spécification d'une loi *a priori* contenant aussi peu d'information que possible et ce, à travers une comparaison permettant de mesurer l'information apportée par le modèle statistique. On définira alors une loi *a priori* de référence qui minimise cette information ([23], [57]).

Pour conclure cette présentation très brève des différentes distributions *a priori*, on peut noter le manque de méthode "générale" pour se repérer dans la multiplicité des approches.

1.1.2 La théorie de la décision en analyse bayésienne

Nous commencerons par un rappel des différentes notions que l'on rencontre dans la théorie de la décision dans un cadre bayésien. Puis, nous explorerons, à travers le travail réalisé avec Éric Marchand de l'Université de Sherbrooke, une piste sur la théorie de la décision.

Soit $\mathcal{D}$ l'espace des décisions i.e. l'ensemble de toutes les décisions possibles, $\mathcal{X}$ l'espace des observations et Θ l'espace des paramètres. Notons $\delta : \mathcal{X} \to \mathcal{D}$ un

estimateur de θ, qui à chaque observation x, associe une décision $\delta(x)$. La plupart du temps, l'espace des décisions pour l'estimation ponctuelle correspond à l'espace des paramètres, ou à une tranformation de cet espace, $h(\Theta)$. L'inférence statistique consiste à prendre une décision d permettant d'estimer le paramètre θ ou une fonction du paramètre θ, $h(\theta)$.

Définition 1.1.6. *Une fonction de coût est une fonction définie de la façon suivante :*

$$L : \Theta \times \mathcal{D} \to [0, +\infty) \tag{1.7}$$

La fonction de coût $L(\theta, d)$ évalue la pénalité (ou l'erreur) associée à la décision d quand le paramètre prend la valeur θ.

Dans le cadre de la théorie de la décision, on est en mesure d'évaluer cette décision.

La théorie de la décision bayésienne est fondée sur le coût moyen *a posteriori*. Ainsi, au lieu d'utiliser le critère de comparaison fréquentiste :

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta[L(\theta, \delta(x))] \\ &= \int_{\mathcal{X}} L(\theta, \delta(x))\, P_\theta(dx), \end{aligned} \tag{1.8}$$

on utilise le coût moyen *a posteriori* ou coût espéré

$$\begin{aligned} \rho(\pi, d|x) &= \mathbb{E}^\pi[L(\theta, d)|\, x] \\ &= \int_\Theta L(\theta, d)\, \pi_x(d\theta) \\ &= \mathbb{E}_{\pi_x}[L(\theta, d)]. \end{aligned} \tag{1.9}$$

Observons une différence importante entre ces deux solutions : le coût moyen *a posteriori* (1.9) est fonction de x, contrairement au risque (1.8) qui dépend de θ.

On définit le risque intégré comme le risque fréquentiste moyenné sur les valeurs de θ en fonction de leur distribution *a posteriori* :

$$r(\pi, \delta) = \mathbb{E}^\pi[R(\theta, \delta)] \tag{1.10}$$

On obtient pour chaque estimateur un nombre réel fournit par $r(\pi, \delta)$, et non une fonction de θ comme dans l'approche fréquentiste. De plus, on est en mesure d'effectuer une comparaison directe entre les estimateurs car un ordre total est instauré, à condition que $r(\pi, \delta) < \infty$.

Nous pouvons, maintenant, définir la notion d'estimateur de Bayes :

Définition 1.1.7. *Un estimateur de Bayes associé à :*
- *une distribution a priori π*
- *une fonction de coût L*

est un estimateur δ^π minimisant $r(\pi, \delta)$. Pour chaque $x \in \mathcal{X}$:

$$\delta^\pi(x) = \arg\min_d \rho(\pi, d|x).$$

La valeur $r(\pi) = r(\pi, \delta^\pi)$ est appelée le risque de Bayes.

Théorème 1.1.8. *Un estimateur minimisant le risque intégré (1.10) peut être obtenu en sélectionnant, pour chaque $x \in \mathcal{X}$, la valeur $\delta(x)$ qui minimise le coût moyen a posteriori, $\rho(\pi, d|x)$.*

Démonstration. Les égalités suivantes sont obtenues grâce au théorème de Fubini, en effet $L(\theta, \delta) \geq 0$.

$$\begin{aligned} r(\pi, \delta) &= \mathbb{E}^{\pi}[R(\theta, \delta)] \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x))\, P_{\theta}(dx)\, \pi(d\theta) \\ &= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta(x))\, p(x|\theta)\, dx\, p(\theta)\, d\theta \\ &= \int_{\mathcal{X}} \int_{\Theta} L(\theta, \delta(x))\, p(x|\theta)\, p(\theta)\, d\theta\, dx \\ &= \int_{\mathcal{X}} \underbrace{\int_{\Theta} L(\theta, \delta(x))\, p(\theta|\, x)\, d\theta}_{\rho(\pi, d|x)}\, m(x)\, dx. \end{aligned}$$

□

Beaucoup d'autres notions sont importantes en théorie de la décision. Nous allons définir, ici, celles qui nous sont utiles. La notion de minimaxité permet de minimiser le coût moyen dans le cas le moins favorable.

Définition 1.1.9. *Le risque minimax associé à une fonction de perte L est la valeur :*

$$\bar{R} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} r(\theta, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} \mathbb{E}_{\theta}[L(\theta, \delta(x))]$$

où D^* est l'espace des distributions de probabilité sur $\mathcal{D}$ et un estimateur minimax est un estimateur δ_0 tel que :

$$\sup_{\theta} r(\theta, \delta_0) = \bar{R}$$

Plusieurs estimateurs minimax peuvent exister en même temps, nous devons donc déterminer l'estimateur optimal. La définition d'admissibilité permet une comparaison de ces estimateurs minimax.

Définition 1.1.10. *δ_0 est un estimateur inadmissible s'il existe un estimateur δ_1 qui domine δ_0 i.e. que pour tout θ*

$$r(\theta, \delta_0) \geq r(\theta, \delta_1)$$

et pour au moins une valeur θ_0 du paramètre $r(\theta_0, \delta_0) > r(\theta_0, \delta_1)$. Sinon δ_0 est dit admissible.

Proposition 1. *S'il existe un unique estimateur minimax, il est admissible.*

D'autres propriétes, sont étudiées dans le livre [57] concernant l'estimateur de Bayes. Nous nous intéressons à l'estimation paramétrique par intervalles de confiance bayésiens dans le cas où l'espace des paramètres est restreint.

Cette sous-section est nécessaire pour la suite. En effet, nous nous intéresserons à la probabilité de recouvrement qui est une fonction de risque pour une perte du type :

$$L(\theta,\, I(X)) = \begin{cases} 1 & \text{si } \theta \notin I(X)\ , \\ 0 & \text{si } \theta \in I(X), \end{cases}$$

où l'espace des décisions correspond aux ensembles mesurables ou à la classe des intervalles sur $\mathbb{R}$ [6].

1.1.3 Estimation bayésienne par intervalle sur un espace paramétrique restreint

Plusieurs méthodes permettent de construire un intervalle de confiance, citons entre autres : l'inversion d'un test d'hypothèses, la construction à l'aide de pivot et les intervalles de confiances bayésiens. A ce sujet, plusieurs références sont à rappeler comme les livres de Casella et Berger [15], de Fourdrinier [29]. Intéressons nous à l'inférence bayésienne développée par Marchand et Strawderman [49].

Supposons qu'il y ait des contraintes sur le paramètre θ, une des méthodes intuitives pour construire un intervalle pour θ est de tronquer l'intervalle de confiance "classique" sur les valeurs possibles du paramètre. Cependant, il peut arriver que ce nouvel intervalle soit vide ou jugé trop court par l'expérimentateur. La méthode de Marchand et Strawderman est une alternative à ce type de problème. Leurs objectifs sont de déterminer un intervalle de confiance bayésien ayant de bonnes propriétes fréquentistes, plus précisement une bonne probabilité de recouvrement fréquentiste. Pour cela, ils déterminent une borne inférieure de la probabilité de recouvrement fréquentiste. Cette borne est établie grâce aux propriétés des densités étudiées et au choix de lois *a priori* non informatives. Une revue des différents résultats est présentée dans le mémoire de Kevin Bosa [11], [50].

Pour une revue de tels problèmes avec une perspective d'estimation ponctuelle lire l'article de Marchand et Strawderman [48].

Dans la suite de notre thèse nous nous sommes intéressés à un couple $(\theta_1,\, \theta_2)$ de paramètres. On désire estimer θ_1 avec l'information additionnelle ou sous la contrainte que $(\theta_1,\, \theta_2) \in A$, où A est un sous-ensemble strict de $\mathbb{R}^2$. La question est de savoir si l'information additionnelle améliore la précision de l'intervalle de confiance bayésien.

Ainsi, nous commençerons cette section par quelques définitions sur les intervalles de confiance bayésien. Ensuite, nous exposerons des résultats préliminaires sur la construction d'intervalles de confiances bayésiens dans ce cas particulier.

Les différents résultats présentés sont nouveaux et prometteurs.

Préliminaires

Avant d'exposer les résultats, voici quelques définitions. Soit le modèle $X \sim F_\theta(.)$, où $F_\theta(.)$ est la fonction de répartition et $\theta \in \Theta$.

Définition 1.1.11. *Soit $I(X) = [l(X), u(X)]$ un intervalle de confiance pour θ. La probabilité de recouvrement fréquentiste de θ associée à $I(X)$ est donnée par $\mathbb{P}_\theta(\theta \in [l(X), u(X)])$, i.e. la probabilité, sous le paramètre θ fixé, pour que l'intervalle $I(X)$ contienne θ.*

Plaçons nous dans le cadre de l'inférence bayésienne.

Définition 1.1.12. *Soit $p(.|x)$ la densité de la loi a posteriori associée à l'observation x et à la loi a priori π. Un intervalle de confiance bayésien de niveau $1-\alpha$ est un intervalle $I_\pi(x) = [l(x), u(x)])$ tel que $\int_{l(x)}^{u(x)} p(\theta|x)\, d\theta = 1-\alpha$.*

Remarque 1.1.13. *La probabilité de recouvrement fréquentiste associée à $I_\pi(x)$ n'est généralement pas égale à $1-\alpha$.*

Définition 1.1.14. *Soit $\mathcal{R}(x)$ telle que :*

$$\mathcal{R}(x) = \{\theta \in \Theta \,|\, p(\theta|x) \geq c\}, \quad c \geq 0.$$

Si c est choisi tel que $\int_{\mathcal{R}(x)} p(\theta|x)d\theta = 1-\alpha$ pour tout x, alors on dit que $\mathcal{R}(.)$ est la région de confiance bayésienne de niveau $1-\alpha$ ayant la plus grande densité a posteriori ou que c'est la région "HPD" pour Highest Posterior Density.

Notons que dans le cas unidimensionnel où la densité *a posteriori* est unimodale, la région correspond à un intervalle.

Les résultats

Décrivons le contexte dans lequel nous nous plaçons. Posons $\theta = (\theta_1, \theta_2) \in \Theta^2$ et $X = (X_1, X_2) \in \mathcal{X}^2$. On souhaite estimer le paramètre θ_1 sachant que :

$$X_1 \sim \mathcal{N}(\theta_1, 1) \;\; et \;\; X_2 \sim \mathcal{N}(\theta_2, 1)$$

avec X_1 indépendante de X_2 et sous la contrainte $\theta_1 - \theta_2 \geq 0$. Notons $\phi_{\theta_1}(.)$ la densité de la loi normale réduite de moyenne θ_1 et $\Phi_{\theta_1}(.)$ sa fonction de répartition.

L'objectif est de déterminer un intervalle de confiance HPD pour θ_1 sachant l'information additionnelle apportée par l'observation X_2, puis de comparer cet intervalle à l'intervalle "classique".

Dans un premier temps, déterminons la loi *a posteriori* de θ_1 sachant les observations X_1, X_2.

Soit la loi "uniforme" sur $\{(\theta_1, \theta_2) : \theta_1 \geq \theta_2\}$, la loi *a priori* obtenue pour θ est donc donnée par :

$$\pi(\theta_1, \theta_2) = I_{[0, \infty[}(\theta_1 - \theta_2).$$

Les variables X_1 et X_2 étant indépendantes, la loi jointe s'écrit :

$$\pi(x_1, x_2|\, \theta_1, \theta_2) = \phi_{\theta_1}(x_1)\, \phi_{\theta_2}(x_2) \tag{1.11}$$

Ces deux variables suivent des lois à paramètres de position. Rapellons cette notion :

Définition 1.1.15. *Un modèle multidimensionnel $X \sim F_\theta(.)$ est dit à paramètre de position si, pour tout $\theta \in \Theta$, la loi $F_\theta(.)$ est l'image de la loi $F_0(.)$ par la translation $x \to x+\theta$, i.e. $F_0(x) = F_\theta(x+\theta)$ pour tout x, θ. On dit que θ est un paramètre de position ou que la famille $\{F_\theta : \theta \in \Theta\}$ est à paramètre de position*

D'après cette définition, on peut écrire (1.11) :

$$\pi(x_1,\, x_2|\, \theta_1,\, \theta_2) = \phi_0(x_1 - \theta_1)\, \phi_0(x_2 - \theta_2)$$

Ainsi la loi de X_1, X_2 sachant θ_1, θ_2 est donnée par :

$$\pi(x_1,\, x_2|\, \theta_1,\, \theta_2) \propto \phi_0(x_1 - \theta_1)\, \phi_0(x_2 - \theta_2))\, I_{[0,\,\infty[}(\theta_1 - \theta_2) \tag{1.12}$$

On en déduit la loi *a posteriori* de θ_1 sachant X_1, X_2 :

$$\begin{aligned}
\pi(\theta_1|X_1,\, X_2) &\propto \int_{\mathbb{R}} \phi_0(X_1 - \theta_1)\, \phi_0(X_2 - \theta_2)\, I_{[0,\,\infty[}(\theta_1 - \theta_2)\, d\theta_2 \\
\pi(\theta_1|X_1,\, X_2) &\propto \int_{-\infty}^{\theta_1} \phi_0(X_1 - \theta_1)\, \phi_0(X_2 - \theta_2)\, d\theta_2
\end{aligned}$$

Par le changement de variables $V = X_2 - \theta_2$ et la symétrie de ϕ_0, on est en mesure de calculer l'intégrale, ce qui donne :

$$\pi(\theta_1|X_1,\, X_2) = \frac{\phi_0(X_1 - \theta_1)\, \Phi_0(\theta_1 - X_2)}{N} \tag{1.13}$$

avec $N = \int_{\mathbb{R}} \phi_0(X_1 - \theta_1)\, \Phi_0(\theta_1 - X_2)\, d\theta_1$.

On introduit la variable $U = \theta_1 - X_1$. Elle facilite les calculs et ne change pas l'interprétation des résultats. Sa loi *a posteriori* s'écrit :

$$\pi(U|X_1,\, X_2) = \frac{\phi_0(U)\, \Phi_0(U + \Delta)}{N} \tag{1.14}$$

où $\Delta = X_1 - X_2$.

Notons que cette densité est un cas particulier de la classe des densités décrite par Arnold et Beaver [5], appelée densité "skew-normal" de paramètres α_0, α_1 dont la forme générale est donnée par :

$$\frac{\phi_0(u)\, \Phi_0(\alpha_0 + \alpha_1\, u)}{\Phi_0(\frac{\alpha_0}{\sqrt{1+\alpha_1^2}})}$$

Dans notre situation, on a que $\alpha_1 = 1$ et $\alpha_0 = \Delta$.

Le calcul de la constante N se fait assez facilement :

$$\begin{aligned}
N &= \int_{\mathbb{R}} \phi_0(X_1 - \theta_1)\, \Phi_0(\theta_1 - X_2)\, d\theta_1 \\
&= \int_{\mathbb{R}} \phi_0(u)\, \mathbb{P}(T \leqslant u + \Delta)\, du \;\text{ avec }\; T \sim \mathcal{N}(0,\, 1) \\
&= \Phi_0(\frac{\Delta}{\sqrt{2}}).
\end{aligned}$$

Ainsi, la densité *a posteriori* (1.14) de U devient :

$$\pi(u|X_1,\, X_2) = \frac{1}{\Phi_0(\frac{\Delta}{\sqrt{2}})}\,\phi_0(u)\,\Phi_0(u+\Delta) \tag{1.15}$$

Une forme explicite de la loi *a posteriori* de $U = \theta_1 - X_1$ est obtenue. De plus, à X_1 près l'interprétation de cette loi est identique à la loi *a posteriori* de θ_1.

Propriétés

Après calculs, on obtient des formes explicites pour la variance et l'espérance de la loi *a posteriori* de U.

Ainsi l'espérance est égale à :

$$\mathbb{E}[U] = \frac{1}{N}\,\frac{1}{\sqrt{2}}\,\phi_0(\frac{\Delta}{\sqrt{2}}) = \frac{1}{\sqrt{2}}\,R(\frac{\Delta}{\sqrt{2}})$$

et la variance :

$$\mathbb{V}ar[U] = 1 - \frac{\Delta}{2\sqrt{2}}\,R(\frac{\Delta}{\sqrt{2}}) - \frac{1}{2}\,R^2(\frac{\Delta}{\sqrt{2}}).$$

Avant de décrire analytiquement les résultats concernant la variance et l'espérance, observons quelques graphiques. La Figure (1.1) représente l'espérance de U en fonction de Δ. On observe une décroissance de l'espérance en Δ et une convergence vers 0 lorsque Δ tend vers l'infini. En ce qui concerne la représentation de la variance de U en fonction de Δ (Fig. (1.2)), on observe une fonction croissante dont la limite est 1 lorsque Δ tend vers l'infini.

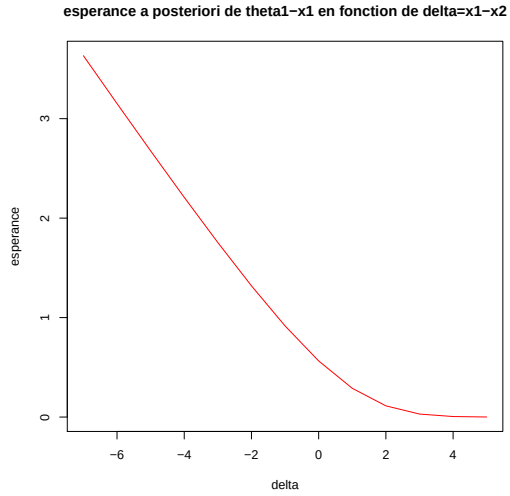

FIG. 1.1 – Espérance a posteriori de $U = \theta_1 - X_1$ en fonction de $\Delta = X_1 - X_2$

Intéressons nous aux résultats analytiques.

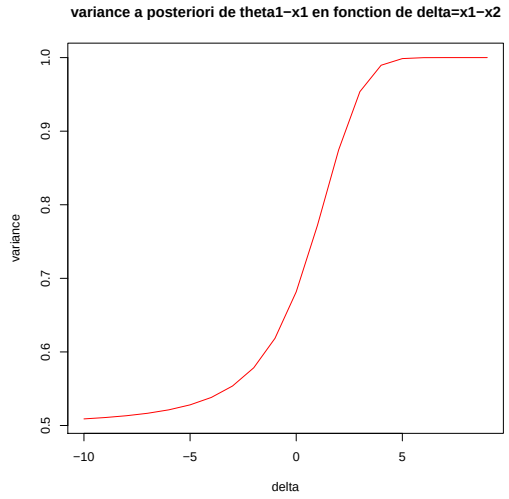

FIG. 1.2 – Variance a posteriori de $U = \theta_1 - X_1$ en fonction de $\Delta = X_1 - X_2$

Remarquons que ces deux termes dépendent de l'inverse du rapport de Mill qui a été étudié par Ruben [64] et Boyd [12]. Ce rapport inverse est défini de la façon suivante :

$$R(x) = \frac{\phi_0(x)}{\Phi_0(x)}. \tag{1.16}$$

Afin de déterminer des propriétés pour la variance et l'espérance *a posteriori* de U, il faut rappeler les propriétés de l'inverse du rapport de Mill (1.16).

Commençons par ces trois propriétés :

(*i*) $R(x) = \frac{\phi_0(x)}{\Phi_0(x)}$ est une fonction décroissante en x
(*ii*) $\lim_{x\to+\infty} R(x) = 0$
(*iii*) $R'(x) = -R(x)\,(x + R(x))$ et est croissante en x

Cette définition permet d'obtenir un Lemme mettant en avant des propriétés intéressantes :

Lemme 1.1.16.

(*i*) $\sqrt{x^2+2} \geq R(x) \geq -x \quad \forall x \in \mathbb{R}$
(*ii*) $\lim_{x\to-\infty} R(x)(x + R(x)) = 1$

La démonstration de ce Lemme se trouve dans le livre de DasGuptas [18]
Ce premier corollaire donne des propriétés sur la variance de U.

Corollaire 1.1.17.

(*i*) $\lim_{\Delta\to-\infty} \mathbb{V}ar_\Delta(u) = \frac{1}{2}$,
(*ii*) *La variance* $Var_\Delta(.)$ *est une fonction croissante en* Δ,
(*iii*) $\lim_{x\to+\infty} Var_\Delta(u) = 1$.

La preuve de ce corollaire est obtenue grâce aux propriétés de l'inverse du rapport de Mill :

Démonstration.

(i) Par définition, on a

$$\begin{aligned}\mathbb{V}ar[U] &= 1 - \frac{\Delta}{2\sqrt{2}}\, R(\frac{\Delta}{\sqrt{2}}) - \frac{1}{2}\, R^2(\frac{\Delta}{\sqrt{2}}) \\ &= 1 - \frac{1}{2}\, R(\frac{\Delta}{\sqrt{2}})\,(\frac{\Delta}{\sqrt{2}} + R(\frac{\Delta}{\sqrt{2}})).\end{aligned}$$

D'après la propriéte (ii) du lemme (1.1.16), on obtient le résultat suivant :

$$\lim_{\Delta\to-\infty} \mathbb{V}ar_\Delta(u) = 1 - \frac{1}{2} = \frac{1}{2}$$

(ii) Par la définition de l'inverse du rapport de Mill, la variance est croissante en Δ.

(iii) On sait que $\lim_{\Delta\to+\infty} \Delta\,\phi_0(\Delta) = 0$ et par définition de l'inverse du rapport de Mill, on obtient la limite suivante :

$$\lim_{\Delta\to+\infty} \mathbb{V}ar_\Delta(u) = 1$$

□

Ce corollaire permet de confirmer d'un point de vue théorique ce que l'on a observé sur les deux figures précédentes. Afin de déterminer de nouvelles propriétés pour la famille de densités sur $\mathbb{R}$ données en 1.15

$$\frac{\phi_0(u)\,\Phi_0(u+\Delta)}{\Phi_0(\frac{\Delta}{\sqrt{2}})},$$

on présente deux lemmes utiles pour la suite des résultats :

Lemme 1.1.18. *Soit* $Z \sim \mathcal{N}(0,\,1)$ $\mathbb{E}[\Phi_0(Z+\Delta)] = \Phi_0(\Delta/\sqrt{2})$

Lemme 1.1.19. *Soit* $Z \sim \mathcal{N}(0,\,1)$ *alors* $\mathbb{E}[\phi_0(Z+c)] = \dfrac{\phi_0(c/\sqrt{2})}{\sqrt{2}}$ *; avec* $c \in \mathbb{R}$

Démonstration.

$$\begin{aligned}\mathbb{E}[\phi_0(Z+c)] &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(z+c)^2}\, \frac{1}{\sqrt{2\pi}}\, e^{-z^2}\, dz \\ &= \int_{\mathbb{R}} \frac{1}{2\pi}\, e^{-z^2-\frac{c^2}{2}-cz}\, dz \\ &= \frac{e^{-c^2/4}}{\sqrt{2\pi}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\, e^{-(z+\frac{c}{2})^2}\, dz \\ &= \frac{1}{\sqrt{2}}\, \phi_0(c/\sqrt{2}).\end{aligned}$$

□

Le Lemme suivant permet d'établir de nombreuses propriétés :

Lemme 1.1.20. *Pour U une famille de densités sur $\mathbb{R}$ données par $\dfrac{\phi_0(u)\,\Phi(u+\Delta)}{\Phi(\frac{\Delta}{\sqrt{2}})}$ de paramètre $\Delta \in \mathbb{R}$, nous avons*

(i) *un rapport de vraisemblance monotone décroissant en u avec paramètre Δ.*

(ii) $\mathbb{E}_\Delta[e^{tU}] = \frac{e^{t^2/2}}{N}\,\Phi_0(\frac{t+\Delta}{\sqrt{2}})$; $t \in \mathbb{R}$

(iii) $\mathbb{E}_\Delta[U] = \frac{1}{\sqrt{2}}\,R(\frac{\Delta}{\sqrt{2}})$

(iv) $\mathbb{V}ar_\Delta[U] = 1 - \frac{\Delta}{2\sqrt{2}}\,R(\frac{\Delta}{\sqrt{2}}) - \frac{1}{2}\,R^2(\frac{\Delta}{\sqrt{2}})$

(v) $U \xrightarrow{\Delta \to +\infty} \mathcal{N}(0,1)$ *(en loi)*

(vi) $\frac{U+\Delta/2}{\sqrt{1/2}} \xrightarrow{\Delta \to -\infty} \mathcal{N}(0,1)$ *(en loi)*

Démonstration.

(i) Découle de la log concavité de Φ

(ii)

$$\begin{aligned}
\mathbb{E}_\Delta[e^{tU}] &= \int_{\mathbb{R}} \frac{e^{tu}}{N}\,\phi_0(u)\,\Phi_0(u+\Delta)\,du \\
&= \frac{1}{N}\int_{\mathbb{R}} e^{tu}\,\frac{1}{\sqrt{2\pi}}\,e^{\frac{1}{2}u^2}\,\Phi_0(u+\Delta)\,du \\
&= \frac{e^{t^2/2}}{N}\int_{\mathbb{R}} \phi_0(y)\,\Phi_0(t+\Delta+y)\,dy \\
&= \frac{e^{t^2/2}}{N}\,\mathbb{E}[\Phi_0(t+\Delta+Y)]
\end{aligned}$$

Et d'après le lemme (1.1.18) on a que $\mathbb{E}_\Delta[e^{tU}] = \frac{e^{t^2/2}}{N}\,\Phi_0(\frac{t+\Delta}{\sqrt{2}})$.

□

Les dessins de la Figure 1.3 représentent la densité de la variable aléatoire U pour Δ positif et pour Δ négatif. De plus ils permettent d'illustrer les propriétés (v), (vi) du Lemme (1.1.20).

L'intervalle de confiance HPD

Notre but est de déterminer l'intervalle de confiance HPD de θ_1 afin d'étudier sa probabilité de recouvrement fréquentiste. On étudie donc un intervalle très proche de l'intervalle HPD et pour lequel les propriétés analytiques de la probabilité de recouvrement sont calculables plus aisément. Ainsi, une alternative plus explicite est l'approximation $\mathbb{E}_\Delta[U] \pm z_{\alpha/2}\sqrt{\mathbb{V}ar_\Delta(U)}$ que nous baptisons l'intervalle "ad hoc". On pourra avoir des idées sur le comportement de l'intervalle HPD et sur sa probabilité de recouvrement.

Soit l'intervalle ad hoc défini par :

$$\begin{aligned}
IC_{adhoc}(X) &= [l(\Delta),\, u(\Delta)] \\
&= [\mathbb{E}_\Delta[U] \pm z_{\alpha/2}\sqrt{\mathbb{V}ar_\Delta(U)}]
\end{aligned}$$

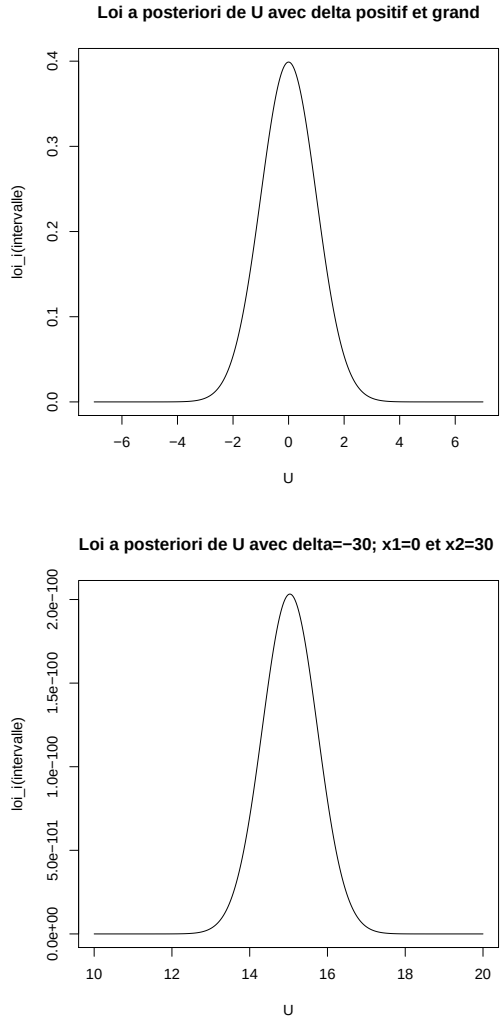

FIG. 1.3 – Densités de la variable U avec Δ positif (à gauche) et Δ négatif (à droite)

ou, en remplaçant Δ par sa valeur :

$$IC_{adhoc}(X) = [\frac{1}{\sqrt{2}} R(\frac{X_1 - X_2}{\sqrt{2}}) \pm z_{\alpha/2} \sqrt{\mathbb{V}ar_{X_1 - X_2}(U)}] \tag{1.17}$$

Il est symétrique, centré en la moyenne et l'on a une forme élégante et relativement explicite pour sa probablité de recouvrement fréquentiste.

Lemme 1.1.21. *Soient $X_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, 2$ indépendants et l'intervalle de confiance*

$$IC(X) = [X_1 + l(\Delta),\ X_1 + u(\Delta)].$$

La probablité de recouvrement fréquentiste associée est donnée par :

$$\begin{aligned} C(\theta_1, \theta_2) &= \mathbb{P}_\theta(\theta_1 \in IC(X)) \\ &= \mathbb{E}[\Phi_0(-\sqrt{2}\, l(\sqrt{2}\, Z + \beta) - Z)] - \mathbb{E}[\Phi_0(-\sqrt{2}\, u(\sqrt{2}\, Z + \beta) - Z)] \end{aligned}$$

où $\beta = \theta_1 - \theta_2$ *et* $Z \sim \mathcal{N}(0,\, 1)$

Démonstration.

$$\begin{aligned} C(\theta_1, \theta_2) &= \mathbb{P}_\theta(\theta_1 \in IC(X)) \\ &= \mathbb{P}_\theta(l(X_1 - X_2) \leq \theta_1 - X_1 \leq u(X_1 - X_2)) \\ &= \mathbb{P}_\theta(-u(X_1 - X_2) \leq X_1 - \theta_1 \leq -l(X_1 - X_2)) \\ &= \mathbb{P}_\theta(-u(X_1 - \theta_1 - (X_2 - \theta_2) + (\theta_1 - \theta_2)) \leq X_1 - \theta_1 \\ &\quad \leq -l(X_1 - \theta_1 - (X_2 - \theta_2) + (\theta_1 - \theta_2))) \\ &= \mathbb{P}_{(0,\,0)}(-u(X_1 - X_2 + \beta) \leq X_1 \leq -l(X_1 - X_2 + \beta)) \end{aligned}$$

où $\beta = \theta_1 - \theta_2$.

Les bornes dépendent de $X_1 - X_2$, on regarde donc la loi de $X_1 | X_1 - X_2$. Posons $Y = X_1 - X_2$ comme $X_1 \sim \mathcal{N}(0,\, 1)$ et $X_2 \sim \mathcal{N}(0,\, 1)$;

$$Y = X_1 - X_2 \sim \mathcal{N}(0,\, \sqrt{2}),$$

ainsi

$$X_1 | Y \sim \mathcal{N}(Y/2,\, 1/2).$$

Rappelons que les espérances conditionnelles vérifient $\mathbb{E}\big[\mathbb{E}[Y|X]\big] = \mathbb{E}[Y]$. *Ici, nous souhaitons avoir* $\mathbb{P}_{(0,\,0)}(X_1 \in IC(Y))$, *ce qui se décompose de la façon suivante :*

$$\begin{aligned} \mathbb{E}[\mathbb{P}_{(0,\,0)}(X_1 \in IC(Y)|Y)] &= \mathbb{E}\big[\mathbb{E}[1_{IC(Y)}(X_1)|Y)]\big] \\ &= \mathbb{E}[1_{IC(Y)}(X_1)] \\ &= \mathbb{P}_{(0,\,0)}(X_1 \in IC(Y)) \end{aligned}$$

Donc

$$\begin{aligned} C(\theta_1, \theta_2) &= \mathbb{P}_{(0,\,0)}(-u(X_1 - X_2 + \beta) \leq X_1 \leq -l(X_1 - X_2 + \beta)) \\ &= \mathbb{P}_{(0,\,0)}\Big(\frac{-u(Y + \beta) - (Y/2)}{\sqrt{1/2}} \leq Z' \leq \frac{-l(Y + \beta) - (Y/2)}{\sqrt{1/2}}\Big) \\ &= \mathbb{P}_{(0,\,0)}(\sqrt{2}\,[-u(Y + \beta) - (Y/2)] \leq Z' \leq \sqrt{2}\,[-l(Y + \beta) - (Y/2)]) \end{aligned}$$

avec $Z' = \frac{X_1 - (Y/2)}{\sqrt{1/2}}$.
De plus $Y \sim \mathcal{N}(0,\, \sqrt{2})$ donc $Z = Y/\sqrt{2} \sim \mathcal{N}(0,\, 1)$.

Ainsi on a que

$$\begin{aligned} &\mathbb{E}^Z\Big[\mathbb{P}_{(0,\,0)}\big(\sqrt{2}\,[-u(\sqrt{2}Z + \beta) - (Z/\sqrt{2})] \leq Z' \leq \sqrt{2}\,[-l(\sqrt{2}Z + \beta) - (Z/\sqrt{2})]\big)\Big] \\ &\quad = \mathbb{E}\Big[\Phi(-\sqrt{2}\, l(\sqrt{2}\, Z + \beta) - Z)\Big] - \mathbb{E}\Big[\Phi(-\sqrt{2}\, u(\sqrt{2}\, Z + \beta) - Z)\Big] \end{aligned}$$

□

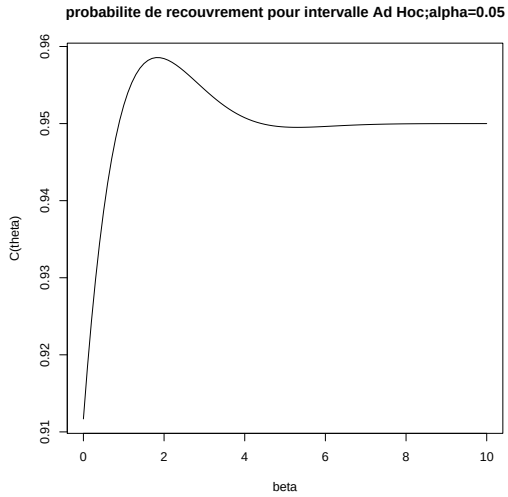

FIG. 1.4 – Probabilité de recouvrement pour l'intervalle ad hoc et alpha=0.05

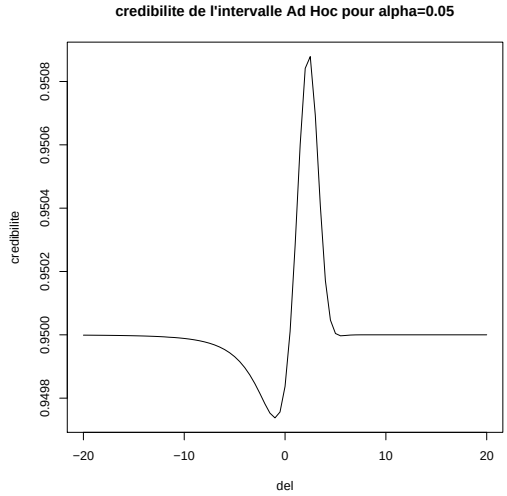

FIG. 1.5 – Crédibilité de l'intervalle ad hoc et alpha=0.05

On obtient une forme explicite de la probabibilité de recouvrement fréquentiste de l'intervalle ad hoc représentée sur la Figure (1.4). La limite semble être proche de 0.95 de plus elle semble atteindre un minimum en $t\theta_1 = \theta_2$. Cette probabilité est satisfaisante. De plus, on observe une bonne couverture pour β compris entre, approximativement, 2 et 3. La crédibilité associée à cet intervalle

est définie par

$$\begin{aligned}\mathbb{P}(U \in IC_{AH}|X_1,\, X_2) &= \mathbb{P}\Big(\frac{1}{\sqrt{2}}\, R(\frac{X_1 - X_2}{\sqrt{2}}) - z_{\alpha/2}\,\sqrt{Var_{X_1-X_2}(u)} \leq U \\ &\leq \frac{1}{\sqrt{2}}\, R(\frac{X_1 - X_2}{\sqrt{2}}) + z_{\alpha/2}\,\sqrt{Var_{X_1-X_2}(u)}\Big)\end{aligned} \tag{1.18}$$

et sa représentation graphique est donnée par la Figure (1.5). De même, la crédibilité de l'intervalle ad hoc est bonne. De plus, comparativement à l'intervalle "classique" la longueur de l'intervalle ad hoc est plus petite, avec une meilleure couverture pour β compris en 2 et 3.

Ainsi, en supposant, que l'intervalle ad hoc et l'intervalle HPD sont très proches, on peut conclure que les résultats semblent prometteurs. L'information apportée par X_2 et se traduisant par une contrainte sur le paramètre θ_1 permettrait d'obtenir une meilleure précision.

Un approfondissement intéressant consisterait d'une part à établir toutes les propriétés de la probabilité de recouvrement de l'intervalle ad hoc. On pourra ainsi mieux connaître son comportement, déterminer les intervalles de croissance et de décroissance, et en déduire une borne inférieure pour la probabilité de recouvrement.

D'autre part, l'intervalle d'interêt étant l'intervalle HPD, un des principaux objectifs est de déterminer les expressions analytiques représentant l'intervalle HPD et sa probabibité de recouvrement. Ce qui permettra d'étudier le comportement de $C(\theta_1, \theta_2)$ et de déterminer une borne inférieure. Enfin l'étude d'un cas plus général, comme par exemple un paramètre θ_1 multidimensionnel, est à envisager.

Comme précisé au début de cette section, nous sommes au stade des préliminaires.

1.2 Les méthodes d'approximation

Dans cette section, nous nous intéresserons à quelques techniques d'approximation. Nous commencerons d'abord par un bref rappel sur les méthodes de Monte Carlo, puis nous nous intéresserons aux méthodes de Monte Carlo par chaînes de Markov, appelées méthodes MCMC. Les deux principaux algorithmes, à savoir, l'algorithme de Gibbs et l'algorithme de Metropolis-Hasting seront étudiés. Nous finirons par le filtrage particulaire et les améliorations apportées. Pour une revue plus détaillée, il faut consulter Robert et Casella [61], Robert [59], Gamerman [30].

Les méthodes MCMC et de filtrage particulaire étant toutes les deux basées sur le même principe de simulation de Monte Carlo, nous commençons cette section par un rappel sur ces méthodes.

Les méthodes de Monte Carlo sont des méthodes numériques permettent de calculer une quantité déterministe en utilisant des procédés aléatoires. Intéressons-

nous à l'évaluation de l'intégrale suivante :

$$\mathbb{E}_{\pi_x}[g(\Theta)] = \int_{\Theta} g(\theta)\,\pi_x(d\theta). \tag{1.19}$$

L'idée est de générer des variables aléatoires selon la loi *a posteriori* π_x. Ainsi, par la loi des grands nombres :

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} g(\theta_i) = \int_{\Theta} g(\theta)\,\pi_x(d\theta),$$

ce qui donne :

$$\mathbb{E}_{\pi_x}[g(\Theta)] \simeq \frac{1}{m}\sum_{i=1}^{m} g(\theta_i).$$

Dans le cas où l'on ne sait pas simuler selon π_x, il nous faut utiliser une autre méthode, appelée échantillonnage d'importance, qui permet également d'évaluer l'intégrale (1.19). Elle fait intervenir une fonction d'importance h.

L'intégrale (1.19) s'écrit alors de la façon suivante :

$$\begin{aligned} \mathbb{E}_{\pi_x}[g(\Theta)] &= \int_{\Theta} \frac{g(\theta)\,h(\theta)}{h(\theta)}\,\pi_x(d\theta) \\ &= \int_{\Theta} \frac{g(\theta)\,p(\theta|x)}{h(\theta)}\,h(\theta)\,d\theta \end{aligned} \tag{1.20}$$

On génére $\theta_1, \ldots, \theta_m$ suivant h ainsi, on a, par la loi des grands nombres :

$$\mathbb{E}_{\pi_x}[g(\Theta)] \simeq \frac{1}{m}\sum_{i=1}^{m} \frac{g(\theta_i)\,p(\theta_i|\,x)}{h(\theta_i)}.$$

Par exemple, l'utilisation de la loi *a priori* π comme fonction d'importance fournit l'approximation :

$$\frac{\sum_{i=1}^{m} g(\theta_i)p(x|\theta_i)}{m},$$

avec les θ_i générés selon π.

Le choix de la loi d'importance est crucial, elle doit être simple à simuler. Si elle est bien choisie elle permet de réduire la variance. De plus, il faut éviter les lois d'importance telles que

$$\sup_{\theta\in\Theta} \frac{p(\theta|x)}{h(\theta)} = \infty,$$

il faut en effet que la fonction $h(\theta)$ soit assez proche de $p(\theta|x)$.

1.2.1 État de l'art des méthodes de Monte Carlo par Chaînes de Markov

Supposons que l'on souhaite échantillonner selon une loi de probabilité π de forme complexe. Les méthodes dites de Monte Carlo par chaînes de Markov (MCMC) vont être utilisées dans le cas où l'on n'est pas en mesure de calculer analytiquement ou par des méthodes d'approximation numérique classiques la densité p associée à cette distribution.

Le principe de ces méthodes repose sur la génération d'une chaîne de Markov ergodique (généralement homogène) dont la loi stationnaire est π. La distribution π est appelée "loi cible".

Ces méthodes sont devenues un important outil de calcul en statistique bayésienne.

A) Les méthodes

Pour leur mise en place, on fait appel aux chaînes de Markov ainsi qu'à leurs propriétés, voici ci-après, un rappel sur les chaînes de Markov.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ un espace d'état avec $d \geq 1$. On appelle *processus stochastique* $X = (X_k)_{k\geq 0}$ à valeurs dans $\mathbb{R}^d$ une suite de variables aléatoires à valeurs dans $\mathbb{R}^d$. On utilisera la notation $X_{0:k} = (X_0, \ldots, X_k)$. La loi de ce processus est la famille de mesures de probabilité $\mathbb{P}(X_{0:k} \in \mathrm{d}x_{0:k})$ pour tout $k \geq 0$.

Définition 1.2.1 (Chaîne de Markov)**.** *Un processus $X = (X_k)_{k\geq 0}$ à valeurs dans $\mathbb{R}^d$ est appelé* chaîne de Markov *(ou processus de Markov) s'il satisfait à la propriété de Markov suivante :*

$$\mathbb{P}(X_{k+1} \in \mathrm{d}x | X_{0:k}) = \mathbb{P}(X_{k+1} \in \mathrm{d}x | X_k), \qquad \forall k \geq 0.$$

Ce qui signifie que le futur conditionné par le passé est équivalent au futur conditionné par le présent.

On obtient grâce à la définition 1.2.1 la factorisation de la loi du processus :

$$\begin{aligned}\mathbb{P}(X_{0:k} \in \mathrm{d}x_{0:k}) =& \mathbb{P}(X_k \in \mathrm{d}x_k | X_{k-1} = x_{k-1})\, \mathbb{P}(X_{k-1} \in \mathrm{d}x_{k-1} | X_{k-2} = x_{k-2}) \cdots \\ & \cdots \mathbb{P}(X_1 \in \mathrm{d}x_1 | X_0 = x_0)\, \mathbb{P}(X_0 \in \mathrm{d}x_0) \\ =& \mathbb{P}(X_0 \in \mathrm{d}x_0) \prod_{i=1}^{k} \mathbb{P}(X_i \in \mathrm{d}x_i | X_{i-1} = x_{i-1})\,.\end{aligned}$$

La loi d'un processus de Markov est complètement déterminée par sa loi initiale :

$$\mu_0(\mathrm{d}x) \stackrel{def}{=} \mathbb{P}(X_0 \in \mathrm{d}x),$$

et par son noyau de transition :

$$P_k(x, \mathrm{d}x') \stackrel{def}{=} \mathbb{P}(X_k \in \mathrm{d}x' | X_{k-1} = x)\,.$$

La loi de transition détermine l'évolution de la loi de la variable X_k.

Hypothèse 1.2.2 (Homogénéité en temps)**.** *Un processus est homogène en temps, c'est à dire que le noyau de transition ne dépend pas de k, ce qui s'écrit :*

$$P(x, \mathrm{d}x') = \mathbb{P}(X_k \in \mathrm{d}x' | X_{k-1} = x)\,, \qquad \forall k \geq 0\,.$$

Un noyau de transition $P(x,\, dx')$ sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ est dit **réversible** par rapport à la mesure $\pi(dx)$ lorsque :

$$\int_A \pi(\mathrm{d}x)\, P(x,\, B) = \int_B \pi(\mathrm{d}x)\, P(x,\, A)\,, \quad \forall A,\, B \in \mathcal{B}(\mathbb{R}^d) \tag{1.21}$$

lorsque $A = \mathbb{R}^d$: $\int \pi(\mathrm{d}x)\, P(x,\, B) = \int_B \pi(\mathrm{d}x) = \pi(B) \quad \forall B \in \mathcal{B}(\mathbb{R}^d)$ c'est-à-dire que $\pi P = \pi$, on dit que «P préserve π» ou que «π est invariante par P».

L'équation de bilan détaillée (1.21) s'écrit, pour tout $A,\, B \in \mathcal{B}(\mathbb{R}^d)$,

$$\int\int \mathbf{1}_A(x)\, \mathbf{1}_B(x')\, \pi(\mathrm{d}x)\, P(x, \mathrm{d}x') = \int\int \mathbf{1}_A(x)\, \mathbf{1}_B(x')\, \pi(\mathrm{d}x')\, P(x', \mathrm{d}x)$$

$$\iff \pi(\mathrm{d}x)\, P(x, \mathrm{d}x') = \pi(\mathrm{d}x')\, P(x', \mathrm{d}x)$$

Cette égalité entre mesures est définie sur $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d))$

Définition 1.2.3 (réversibilité)**.** *Une chaîne de Markov $\{X_n\}_{n\in\mathbb{N}}$ définie sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ de loi initiale $\pi(\mathrm{d}x)$ et de noyau de transition $P(x,\, dx')$ est dite* réversible *si P est réversible par rappport à π :*

$$\mathbb{P}(X_n \in A\,,\ X_{n+1} \in B) = \mathbb{P}(X_{n+1} \in A\,,\ X_n \in B)\,, \quad \forall A,\, B \in \mathbb{R}^d\,.$$

Remarquons que vérifier la réversibilité (1.21) est équivalent à vérifier l'équation de bilan détaillé. Ainsi π est invariante par P pour tout $A \in \mathbb{R}^d$.

Définition 1.2.4 (irréductibilité et périodicité)**.**

– *une chaîne est π-**irréductible**, avec π mesure de probabilité, si*

$$\forall x \in \mathbb{R}^d,\, \forall A \in \mathcal{B}(\mathbb{R}^d) \text{ avec } \pi(A) > 0\,,\ \exists n \in \mathbb{N} \text{ tel que } P^n(x, A) > 0$$

– *une chaîne π-irréductible est **périodique** s'il existe une partition disjointe $\mathbb{R}^d = A_0 + \cdots + A_n$ (avec $n+1 \geq 3$) telle que $\pi(A_n) = 0$ et*

$$z \in A_{i-1} \Rightarrow P(x,\, A_i) = 1\,,\ 1 \leq i \leq n-1\,,$$
$$z \in A_{n-1} \Rightarrow P(x,\, A_0) = 1\,.$$

*Dans le cas contraire elle est dite **apériodique**.*

La méthode de Métropolis-Hastings

L'objectif de la méthode de Metropolis-Hastings est de construire un noyau de transition réversible $P(x, dx')$ dont π est la mesure invariante. Cette méthode se décompose en deux étapes. On commence par générer des échantillons aléatoires à partir d'une distribution Q appelée noyau de proposition. Ces échantillons sont ensuite "corrigés" afin qu'ils se comportent asymptotiquement comme des observations aléatoires de la distribution cible ou de la distribution stationnaire. La méthode incorpore donc des étapes de proposition et d'acceptation.

Soit $X^{(t)} = x$ l'état t de la chaîne de Markov. Pour le prochain état $X^{(t+1)}$, on propose un candidat x' généré selon un noyau de proposition Q de densité $q(x, .)$. Ce noyau peut dépendre, ou non, de x. Si c'est le cas sa densité sera celle

de x' sachant x que l'on notera $p(x'|x)$. L'état candidat est accepté avec une probabilité $a(x, x')$, ainsi si x' est accepté $X^{(t+1)} = x'$, sinon $X^{(t+1)} = x$.

Pour schématiser, la méthode de Metropolis-Hastings simule de façon itérative la chaîne de Markov $(X^{(t)})_{t\geq 1}$ grâce à l'algorithme suivant :

Initialisation : choix arbitraire de $x^{(0)}$

Itération t :

- Conditionnellement à x, on génère $x' \sim q(x, .)$
- On calcule $a(x, x') = \min\{1, \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')}\}$
-

$$x^{(t+1)} = \begin{cases} x' & \text{avec probabilité } a(x, x'), \\ x & \text{avec probabilité } 1 - a(x, x'), \end{cases}$$

La quantité $a(x, x')$ représente la probabilité d'acceptation de la nouvelle valeur. C'est cette fonction qui permet d'obtenir un noyau de transition réversible $P(x, dx')$ dont la loi cible est la mesure invariante.

En effet, pour $Q(x, dx')$ le noyau de proposition, le noyau de transition de la chaîne de Metropolis-Hasting s'écrit de la façon suivante :

$$P(x, dx') \stackrel{def}{=} a(x, x')\, Q(x, dx') + \underbrace{\{\int (1 - a(x, x'))\, , Q(x, dx')\}}_{r(x)}\, \delta_x(dx'), \quad (1.22)$$

où $\delta_x(dx')$ est la mesure de dirac et $r(x)$ la probabilité de rejet. L'algorithme a donc perturbé un noyau de proposition $Q(x, dx')$ pour obtenir un noyau de transition $P(x, dx')$ laissant invariant la distribution cible π. La probabilité d'acceptation est donc construite pour que le noyau de transition (1.22) préserve la mesure cible. Pour ce faire, le noyau doit vérifier la condition suffisante de réversibilité qui, d'après le paragraphe précédent, est équivalente à l'équation de bilan détaillé :

$$\pi(dx)\, P(x, dx') = \pi(dx')\, P(x', dx). \quad (1.23)$$

Ainsi le noyau défini par (1.22) vérifie l'équation de bilan détaillée (1.23) si et seulement si :

$$\pi(dx)\, Q(x, dx')\, a(x, x') = \pi(dx')\, Q(x', dx)\, a(x', x). \quad (1.24)$$

Dans le cas où le noyau de proposition et la loi cible sont dominés, il existe une mesure ν telle que : $Q(x, dx') = q(x, x')\, \nu(dx')$ et $\pi(dx) = p(x)\, \nu(dx)$. On définit :

$$R = \{(x, x') \in \mathbb{R}^d \times \mathbb{R}^d : p(x)\, q(x, x') > 0 \; et \; p(x')\, q(x', x) > 0\}$$

et

$$r(x, x') = \frac{p(x)\, q(x, x')}{p(x')\, q(x', x)}$$

Ainsi l'équation de bilan détaillée (1.24) est satisfaite si et seulement si :

(*i*) $p(x)\, q(x, x')\, a(x, x') = 0$ pour tout $(x, x') \in R^c$

(*ii*) $a(x, x')\, r(x, x') = a(x', x)$ pour tout $(x, x') \in R$

La probabilité d'acceptation est donc :

$$a(x,x') = \begin{cases} 1 \wedge r(x',x), & \text{si } (x,x') \in R \\ 0 & \text{si } (x,x') \in R^c \end{cases}$$

D'autre part, l'irréductibilité découle de la condition sur le support de q, qui n'est cependant pas nécessaire pour assurer la validité de l'algorithme.

Lemme 1.2.5. *Lorsque $supp(p) \subset supp(q)$, la chaîne de Markov $(X^{(t)})_{t\geq 0}$ est irréductible et de distribution stationnaire π.*

Cette méthode permet d'obtenir une chaîne de Markov $(X^{(t)})_{t\geq 0}$ irréductible ergodique et de distribution stationnaire la loi cible π. La paramétrisation et le choix de la distribution *a priori* joue un rôle fondamental. Deux cas de figures se distinguent, quand le ratio d'acceptation est proche de 1, les valeurs proposées entre l'état précédent et l'état proposé seront similaires. Une lente convergence est attendue. A contrario, si l'état proposé a des valeurs très éloignées de l'état précédent, nous obtiendrons un haut taux de rejet. Il n'y a pas de conditions générales, seulement des lignes conductrices générales à suivre ([62], [63]).

En pratique, le choix du noyau de proposition et l'initialisation sont assez difficiles, le temps de chauffe peut être assez long et enfin il y a une difficulté à diagnostiquer la convergence.

L'échantillonnage de Gibbs

L'algorithme de Gibbs est très populaire parmi les méthodes MCMC grâce à la simplicité de ses calculs. Son principe est de construire une chaîne de Markov $(X^{(t)})_{t>0}$ admettant π comme distribution stationnaire en effectuant une mise à jour séquentielle des composantes de $X_i^{(t)}$ de la chaîne.

Soit $\pi(dx)$ la loi cible d'une variable aléatoire $X_{1:d}$ définie sur l'espace mesurable produit

$$(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) = (\mathbb{R} \times \cdots \times \mathbb{R}, \mathcal{B}(\mathbb{R}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R})).$$

On définit les lois marginales conditionnelles :

$$\pi_i(dx_i|x_{-i}) \stackrel{def}{=} \mathbb{P}(X_i \in dx_i | X_{-i} = x_{-i}),\ i = (1, \ldots, d)$$

où x_{-i} est égal à x privé de la $i$ème composante, i.e. $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

On part d'un état initial $X^{(0)}$ et l'itération qui permet de passer de $X^{(t)}$ à $X^{(t+1)}$ se déroule de la façon suivante :

$$\begin{aligned} X_1^{(t+1)} &\sim \pi_1(.\,|X_{2:d}^{(t)}) \\ &\vdots \\ X_l^{(t+1)} &\sim \pi_l(.\,|X_{1:(l-1)}^{(t+1)}, X_{(l+1):d}^{(t)}) \\ &\vdots \\ X_d^{(t+1)} &\sim \pi_d(.\,|X_{1:(d-1)}^{(t+1)}). \end{aligned}$$

Ainsi le noyau de transition de la chaîne de Gibbs s'écrit :

$$P(x,\, dx') \stackrel{def}{=} \pi_1(dx'_1 \,|\, x_{2:d}) \,\ldots\, \pi_l(dx'_l \,|\, x'_{1:(l-1)}, x_{(l+1):d}) \,\ldots\, \pi_d(dx'_d \,|\, x_{1:(d-1)}).$$

π est invariante par P.

D'autre part, l'échantillonneur de Gibbs est un cas particulier de l'algorithme de Metropolis-Hasting. Supposons qu'il existe une mesure ν telle que :

$$\pi(dx) = p(x)\,\nu(dx) \;\text{ et }\; \pi_l(dx'_l \,|\, x'_{1:(l-1)}, x_{(l+1):d}) = p_l(x_l \,|\, x_{1:(l-1)}, x'_{(l+1):d})\,\nu(dx').$$

Soit

$$q(x,y) = p_1(x_1 \,|\, x'_{2:d}) \,\ldots\, p_l(x_l \,|\, x_{1:(l-1)}, x'_{(l+1):d}) \,\ldots\, p_d(x_d \,|\, x_{1:(d-1)}).$$

La forme du noyau de transition est donnée par l'équation (1.22). Afin de déterminer la probabilité d'acceptation calculons :

$$r(x',x) = \frac{p(x')\,q(x',x)}{p(x)\,q(x,x')}.$$

$$r(x',x) = \frac{p(x')\,p_d(x_d \,|\, x_{1:(d-1)}) \,\ldots\, p_l(x'_l \,|\, x'_{1:(l-1)}, x_{(l+1):d}) \,\ldots\, p_1(x'_1 \,|\, x_{2:d})}{p(x)\,p_d(x'_d \,|\, x_{1:(d-1)}) \,\ldots\, p_l(x'_l \,|\, x'_{1:(l-1)}, x_{(l+1):d}) \,\ldots\, p_1(x'_1 \,|\, x_{2:d})}$$

La probabilité d'acceptation est toujours égale à 1 et la proposition est toujours acceptée. Cette méthode est en général plus efficace que l'algorithme de Metropolis-Hasting. De plus, elle a l'avantage de ne pas avoir à choisir une distribution de proposition et ne rejette jamais ces propositions. Cependant, celle-ci nécessite de savoir simuler selon les lois marginales conditionnelles et explore mal les diagonales.

Ces algorithmes permettent d'obtenir des réalisations approximatives de la loi cible π et des estimateurs convergents de quantités $\mathbb{E}^{\pi}[g(\theta)]$. Notons que nous exposons ici les deux principaux algorithmes, de nombreux autres existent (c.f. [33]).

B) Les critères de convergence

On vient d'étudier, d'un point de vue théorique, la garantie de la convergence des algorithmes MCMC, on souhaite cependant vérifier, contrôler, cette convergence lors du déroulement de l'algorithme. Ce contrôle est assez difficile, deux approches existent :

- (*i*) les méthodes "on line" : elles ne modifient pas l'algorithme
- (*ii*) les méthodes "customisées" : elles modifient l'algorithme en le réécrivant pour y inclure des variables de contrôle, des chaînes parallèles ou encore des modifications par renouvellement des transitions.

Les premières méthodes sont plus intéressantes pour un contrôle automatique mais restent moins robustes. Les secondes demandent plus d'efforts. En effet, elles s'accompagnent d'un temps de calcul et de programmation bien plus important. L'un des problèmes qui se pose concerne le nombre d'itérations nécessaire

pour obtenir des échantillons selon la loi cible. Il existe depuis 10 ans dans la littérature beaucoup d'éléments sur les diagnostics de convergence.

Il nous faut décrire ici les diagnostics les plus souvent utilisés : les méthodes graphiques et les méthodes un peu plus formelles. On obtiendra alors des réalisations approximatives de cette distribution, ainsi que des estimateurs convergents de quantités *a posteriori* $\mathbb{E}[g(\theta)|x]$. Pour un approfondissement, consulter les ouvrages suivant : Cowles et Carlin [16], Brooks et Roberts [13], Robert [60], Robert et Casella [61], Mengersen et al. [51].

Méthodes graphiques

Les premières méthodes graphiques viennent de Gelfand et Smith [31] et notamment de leur "thick pen" qui consiste à stopper l'algorithme dès que les variations de la moyenne empirique ne dépassent pas l'épaisseur d'un gros crayon. Depuis, de nombreux critères ont pris forme, notamment la comparaison d'estimateurs multiples de Robert [58], mais également la méthode fondée sur les Cusums.

L'autocorrélation

L'examen des autocorrélations est important, en effet de fortes autocorrélations entraînent une convergence très lente, l'exploration de la loi *a posteriori* en entier est longue. En général, on représente les autocorrélations en fonctions des pas (appelés lag en anglais) afin d'obtenir le corrélogramme. Ce graphique permet de détecter le "lag" pour lequel la corrélation est très faible et donc d'estimer la longueur effective de la chaîne (Sorensen et al., [66]).

Méthodes de comparaison inter-chaînes

Un diagnostic de convergence populaire a été mis en place par Gelman et Rubin [32]. Cette méthode génère m chaînes indépendantes chacune de longueur $2 * n$. Chaque chaîne possède des valeurs initiales différentes échantillonnées selon une distribution plus dispersée que la distribution cible. Ce dernier point est important car il permet d'oberver, ou pas, un manque de convergence. On écarte les n premières valeurs de la chaîne afin de garder les n dernières. Chaque caractéristique de la loi *a posteriori* est étudiée séparément.

L'étape qui va suivre est une analyse de la variance qui consiste à diagnostiquer la convergence quand la variabililité entre les chaînes n'est pas plus grande que celle intra chaîne.

Expliquons brièvement cette méthode d'un point de vue analytique. Soit la chaîne i $(i = 1, \ldots, m)$, on considère que les valeurs simulées θ_{ij}, $j = 1, \ldots, n$ selon la loi *a posteriori* π_x sont utilisables. En prenant en compte les m chaînes, on organise nos observations comme suit : m classes avec n observations par classe. Soit B la variance inter groupe et W la variance intra groupe telle que :

$$B = \frac{n \sum_{i=1}^{m} (\bar{\theta_{i.}} - \bar{\theta_{..}})}{m - 1}$$

et

$$W = \frac{\sum_{i=1}^{m} S_i^2}{m},$$

où

$$S_i^2 = \frac{\sum_{j=1}^{n} (\theta_{ij} - \bar{\theta_{i.}})}{n-1}, \; \bar{\theta_{i.}} = \frac{\sum_{j=1} n\theta_{ij}}{n}, \;\; \bar{\theta_{..}} = \frac{\sum_{i=1}^{m} \bar{\theta_{i.}}}{m},$$

avec S_i^2 l'estimation de la variance des valeurs échantillonnées de la chaîne i.

Posons

$$\mu = \int \theta \, p(\theta|y) \, d\theta$$

et

$$\sigma^2 = \int (\theta - \mu)^2 \, p(\theta|y) \, d\theta,$$

la moyenne et la variance, respectivement, de la distribution cible.

Si les valeurs simulées sont échantillonnées selon la loi *a posteriori*, on peut vérifier rapidement que les espérances suivantes reposent sur :

$$\mathbb{E}[B] = \sigma^2 \tag{1.25}$$

et

$$\mathbb{E}[S_i^2] = \mathbb{E}[W] = \sigma^2. \tag{1.26}$$

Gelman et Rubin [32] suggèrent l'estimateur de la variance *a posteriori*

$$\hat{\sigma^2} = \frac{n-1}{n} W + \frac{1}{n} B, \tag{1.27}$$

qui est clairement non biaisé pour σ^2, à condition que tous les échantillons soient générés selon la distribution cible.

Pour conclure, nous ne sommes pas en mesure de garantir la convergence pour un temps d'arrêt donné. Par conséquent la recommandation générale est d'utiliser une combinaison des différents outils de diagnostic (y compris les méthodes graphiques) et d'apprendre le plus possible sur la loi cible.

1.2.2 Le filtrage particulaire

L'utilisation des méthodes MCMC nécessite une attention particulière sur deux points : le diagnostic de convergence et l'accélération de la convergence. Ces deux aspects causent une certaine incertitude sur les résultats obtenus. Aussi les utilisateurs des méthodes MCMC, en particulier dans l'environnement, sont intéressés par d'autres techniques qui permettraient une validation de leurs résultats. Les techniques de filtrage particulaire, développées dans le milieu des années 1990, en font partie. Ces méthodes ont été, à l'origine, développées lorsque de nombreuses observations sont à traiter en temps réel, comme dans le domaine de la robotique ou bien encore de la poursuite de cible.

Les techniques particulaires ont le même objectif que les MCMC, à savoir l'approximation de la loi *a posteriori* et sont basées sur le même principe de simulation de Monte Carlo. Cependant, elles sont séquentielles, i.e. elles traitent les données en ligne et sont ainsi plus rapides.

A) Le filtrage

Commençons par expliquer ce concept. Il consiste à calculer à chaque pas de temps la loi conditionnelle π_k, appelée *filtre*, de l'état X_k sachant une réalisation des observations $Y_{1:k} = y_{1:k}$ jusqu'à l'instant courant k, i.e. :

$$\pi_k(\mathrm{d}x) = \mathbb{P}(X_k \in \mathrm{d}x | Y_{1:k} = y_{1:k})\,. \tag{1.28}$$

La suite de probabilités conditionnelles $(\pi_k)_{k\geq 0}$, avec la convention $\pi_0 = \mu_0$ constitue *le filtre optimal bayésien*. Pour des contraintes de temps réel, nous souhaitons calculer ce filtre de façon récursive, i.e. de telle façon que le calcul de π_k ne soit fonction que de la loi conditionnelle précédente π_{k-1} et de l'observation Y_k. L'évolution du filtre optimal dans le temps est relativement facile à décrire, elle se décompose en une étape de prédiction et une étape de correction faisant intervenir la nouvelle observation.

Proposition 1.2.6 (Filtre optimal)**.** *Le filtre se calcule de façon récursive en deux étapes. Le filtre prédit* $\pi_{k^-}(\mathrm{d}x) = \mathbb{P}(X_k \in \mathrm{d}x | Y_{1:k-1})$ *à l'étape* k *se déduit du filtre* π_{k-1} *selon l'équation de prédiction :*

$$\pi_{k^-}(\mathrm{d}x) = \int_{\mathbb{R}^n} \pi_{k-1}(\mathrm{d}x')\, Q_k(x', \mathrm{d}x) = (\pi_{k-1}\, Q_k)(\mathrm{d}x)\,. \tag{1.29}$$

Le filtre π_k *se déduit à l'étape* k *du filtre prédit* π_{k^-} *selon l'équation de correction :*

$$\pi_k(\mathrm{d}x) = \frac{\psi_k(x, y_k)\, \pi_{k^-}(\mathrm{d}x)}{\int_{\mathbb{R}^n} \psi_k(x', y_k)\, \pi_{k^-}(\mathrm{d}x')} = (\psi_k \cdot \pi_{k^-})(\mathrm{d}x)\,. \tag{1.30}$$

L'évolution du filtre peut se résumer par le schéma suivant :

$$\pi_{k-1} \xrightarrow{\text{prédiction}} \pi_{k^-} = (Q_k\, \pi_{k-1}) \xrightarrow[\text{correction}]{Y_k \downarrow} \pi_k = (\psi_k \cdot \pi_{k^-})$$

TAB. 1.1 – Décomposition d'une itération du filtre

Ce filtre est une équation récursive dans un espace infini-dimensionnel et ne peut donc pas être utilisé en pratique. Dans certains cas, comme le cas linéaire-gaussien, ce filtre admet une formulation explicite. Dans le cas général, il est nécessaire de faire appel à des méthodes d'approximation.

B) Les méthodes particulaires

Les méthodes particulaires sont des méthodes d'approximation fondées sur le principe de Monte Carlo pour estimer et prédire en ligne des systèmes dynamiques. L'idée est d'approcher la loi π_k à l'aide d'une mesure discrète égale à une somme finie de mesures de Dirac concentrées en des points dénommés "particules" et pondérées par des cœfficients appelés "poids des particules". L'ap-

proximation consiste donc à écrire le filtre prédit et le filtre sous la forme :

$$\pi_{k^-}(\mathrm{d}x) \simeq \pi^N_{k^-}(\mathrm{d}x) = \sum_{i=1}^{N} w^i_{k^-}\, \delta_{\xi^i_{k^-}}(\mathrm{d}x)\,,$$
$$\pi_k(\mathrm{d}x) \simeq \pi^N_k(\mathrm{d}x) = \sum_{i=1}^{N} w^i_k\, \delta_{\xi^i_k}(\mathrm{d}x)\,.$$

Ces méthodes séquentielles de Monte Carlo (ou filtres particulaires) font appel, de manière récursive, au principe d'échantillonnage d'importance et de rééchantillonnage. Nous conseillons la lecture du livre de Doucet [22] (édité par Doucet, Freitas et Gordon) pour faire le point sur les connaissances dans le domaine ainsi que l'article de Gordon [37] et le livre de Godsill, Cappé et Moulines [34].

Filtre de Monte Carlo

Comme nous l'avons vu précédemment, on cherche à approximer le filtre de façon récursive, supposons que :

$$\pi_{k-1}(\mathrm{d}x) = \sum_{i=1}^{N} w^i_{k-1}\, \delta_{\xi^i_{k-1}}(\mathrm{d}x)\,.$$

Algorithme 1.2.7. *On obtient alors une approximation du filtre prédit et du filtre grâce aux équations (1.29) et (1.30) :*

Prédiction :

$$\pi_{k^-}(\mathrm{d}x) \simeq \pi^N_{k^-}(\mathrm{d}x) = \sum_{i=1}^{N} w^i_{k^-}\, \delta_{\xi^i_{k^-}}(\mathrm{d}x) \quad \textit{où} \quad \left\{ \begin{array}{ll} \xi^i_{k^-} & \sim Q_k(\xi^i_{k-1}, dx) \\ w^i_{k^-} & = w^i_{k-1} \end{array} \right.$$

Correction :

$$\pi_k(dx) \simeq \pi^N_k(\mathrm{d}x) = \sum_{i=1}^{N} w^i_k\, \delta_{\xi^i_k}(\mathrm{d}x) \quad \textit{où} \quad \left\{ \begin{array}{ll} \xi^i_k & = \xi^i_{k^-} \\ w^i_k & \propto w^i_{k-1}\, \psi_k(\xi^i_{k^-}, y_k) \end{array} \right.$$

Démonstration. :

Prédiction :

En utilisant l'équation de prédiction (1.29), on obtient :

$$\begin{aligned} \pi_{k^-}(\mathrm{d}x) &= \int_{\mathbb{R}^n} Q_k(x', \mathrm{d}x) \sum_{i=1}^{N} w^i_{k-1}\, \delta_{\xi^i_{k-1}}(\mathrm{d}x') \\ &= \sum_{i=1}^{N} w^i_{k-1}\, Q_k(\xi^i_{k-1}, \mathrm{d}x) \\ &\simeq \pi^N_{k^-}(\mathrm{d}x) = \sum_{i=1}^{N} w^i_{k^-}\, \delta_{\xi^i_{k^-}}(\mathrm{d}x)\,. \end{aligned}$$

où $\xi^i_{k^-} \sim Q_k(\xi^i_{k-1}, \mathrm{d}x)$ et $w^i_{k^-} = w^i_{k-1}$.

Cette étape est également appelée étape de propagation des particules ; dans le cas précis du filtre de Monte Carlo, la propagation se fait selon la loi de transition de la chaîne.

Correction :
L'équation de correction (1.30) donne :

$$\begin{aligned}
\pi_k(\mathrm{d}x) &= \frac{\psi_k(x, y_k)\, \pi_{k^-}(\mathrm{d}x)}{\int_{\mathbb{R}^n} \psi_k(x', y_k)\, \pi_{k^-}(\mathrm{d}x')} \\
&\simeq \pi^N_k(\mathrm{d}x) = \frac{\psi_k(x, y_k) \sum_{i=1}^N w^i_{k-1}\, \delta_{\xi^i_{k^-}}(\mathrm{d}x)}{\int_{\mathbb{R}^n} \psi_k(x', y_k) \sum_{i=1}^N w^i_{k-1}\, \delta_{\xi^i_{k^-}}(\mathrm{d}x')} \\
&= \frac{\sum_{i=1}^N w^i_{k-1}\, \psi_k(\xi^i_{k^-}, y_k)\, \delta_{\xi^i_{k^-}}(\mathrm{d}x)}{\sum_{i=1}^N \psi_k(\xi^i_{k^-}, y_k)\, w^i_{k-1}} \\
&= \sum_{i=1}^N w^i_k\, \delta_{\xi^i_k}(\mathrm{d}x)\,.
\end{aligned}$$

où $w^i_k = \frac{w^i_{k-1}\, \psi_k(\xi^i_{k^-}, y_k)}{\sum_{i=1}^N w^i_{k-1}\, \psi_k(\xi^i_{k^-}, y_k)}$ et $\xi^i_k = \xi^i_{k^-}$.

Cette étape effectue une mise à jour des particules.

Algorithme 1.2.8. ***(Filtre de Monte Carlo)***
$\xi_{1:N} \sim \mu_0$
$w_{1:N} \leftarrow 1/N$
pour $k = 2 : T$ ***faire***
$\tilde{\xi}_i \sim Q_k(\xi_i, .)$ *pour* $i = 1 : N$ *%propagation*
$\tilde{w}_i \leftarrow w_i\, \psi_k(\tilde{\xi}_i, y_k)$ *pour* $i = 1 : N$ *%mise à jour des poids*
$w_i \leftarrow \tilde{w}_i / \sum_{i=1}^N \tilde{w}_i$ *pour* $i = 1 : N$ *%normalisation*
$\xi_{1:N} \leftarrow \tilde{\xi}_{1:N}$
sortie $(\xi_{1:N}, w_{1:N})$
fin

Ce filtre n'est pas viable, dans la pratique, car il s'avère que la majorité des poids dégénère très rapidement vers zéro or idéalement tous les poids devraient rester proches de $1/N$, ce qui signifierait que les particules sont d'égales importance dans l'approximation.

Ce filtre ne devient donc utilisable que lorsqu'une étape de rééchantillonnage est ajoutée permettant ainsi de limiter la dégénérescence des poids. En effet, cette étape va favoriser les particules dites "importantes" par rapport aux particules ayant un poids négligeable. Nous dupliquons alors les particules à fort poids au détriment de celles qu'on souhaite défavoriser. On obtient un des premiers algorithmes proposé en 1993 par Gordon, Salmond et Smith [37], ce filtre est connu sous le nom de *filtre Bootstrap*. Nous conserverons dans la suite la dénomination de filtre Bootstrap pour désigner cette classe initiale de filtres particulaires, pour laquelle les particules sont propagées selon la loi de transition et

rééchantillonnées à chaque étape selon un schéma classique de rééchantillonnage.

On présente ici la technique de rééchantillonnage multinomial. D'autres techniques sont exposées dans mon mémoire [20]. L'algorithme consiste à tirer N particules parmi $\{\xi_k^1, \cdots, \xi_k^N\}$ avec les probabilités $\{w_k^1, \ldots, w_k^N\}$, on obtient ainsi j_1 fois la particule ξ_k^1, j_2 fois la particule ξ_k^2, $\cdots$. Le vecteur $(j_1, \cdots, j_N)$ suit la loi multinomiale suivante $\mathcal{M}(N, w_k^1, \cdots, w_k^N)$.

Algorithme 1.2.9. ***(Filtre Bootstrap)***
$\xi_{1:N} \sim \mu_0$
$w_{1:N} \leftarrow 1/N$
pour $k = 2 : T$ ***faire***
$\tilde{\xi}_i \sim Q_k(\xi_i, .)$ *pour* $i = 1 : N$ *%propagation*
$\tilde{w}_i \leftarrow \psi_k(\tilde{\xi}_i, y_k)$ *pour* $i = 1 : N$ *%mise à jour des poids*
$w_i \leftarrow \tilde{w}_i / \sum_{i=1}^N \tilde{w}_i$ *pour* $i = 1 : N$ *%normalisation*
$\xi_{1:N} \leftarrow resample\_multi(w_{1:N}, \tilde{\xi}_{1:N})$
sortie $(\xi_{1:N})$
fin

Comme nous venons de le voir, l'algorithme du bootstrap est identique à l'algorithme du filtre de Monte Carlo et seule une étape de rééchantillonnage est ajoutée. D'autres filtres existent, notamment les filtres SIS et SIR. Ils correspondent respectivement au filtre de Monte Carlo et au filtre Bootstrap dans lesquels l'étape de propagation des particules s'effectue non plus au travers de la loi de transition de la chaîne, mais grâce à une loi de proposition [20].

Améliorations et Applications

Différentes améliorations existent dans la littérature, cependant ces méthodes n'ont pas été utilisées par la communauté du «pistage» : elles ne sont pas assez rapides pour faire des calculs en temps réel.

Dans le domaine de l'environnement, le contexte est un peu différent car il n'y a pas de contrainte de temps réel pour les calculs. Par conséquent, il était intéressant de les adapter. Ainsi une application à la pêcherie d'un algorithme d'amélioration a été réalisée afin de comparer la performance des méthodes particulaires aux méthodes MCMC. Les comparaisons ont été faites grâce au filtre particulaire auxiliaire : ASIR.

Ce filtre introduit des variables auxiliaires représentant les composantes d'un mélange et couple cette idée avec le principe d'échantillonnage pondéré séquentiel [56]. Voici l'algorithme :

Algorithme 1.2.10. ***(ASIR)***

$\xi_{1:N} \sim \mu_0$
$w_{1:N} \leftarrow 1/N$
pour $k = 1 : T$ ***faire***
$\mu_i \leftarrow \mathbb{E}[X_k | X_{k-1} = \xi_i]$ *pour* $i = 1 : N$
$\tilde{\beta}_i \leftarrow \psi_k(\mu_i, y_k)\, w_i$ *pour* $i = 1 : N$
$\beta_i \leftarrow \tilde{\beta}_i / \sum_{i=1}^N \tilde{\beta}_i$ *pour* $i = 1 : N$ *%normalisation*
$i^j \sim multi(\beta_i)$ *pour* $j = i : N$

$\xi_j \sim Q_k(\xi_{i^j}, .)$ *pour* $j = 1 : N$ *%propagation*
$\tilde{w}_j \leftarrow \frac{\psi_k(\xi_j, y_k)}{\psi_k(\mu^k_{i^j}, y_k)}$ *pour* $j = 1 : N$ *%mise à jour des poids*
$w_j \leftarrow \tilde{w}_j / \sum_{j=1}^N \tilde{w}_j$ *pour* $j = 1 : N$ *%normalisation*
sortie $(\xi_{1:N}, w_{1:N})$

fin

Nous avons mis en œuvre le filtre ASIR et le filtre bootstrap afin d'évaluer le stock de poulpe à l'échelle de la ZEE mauritanienne. Les données récoltées entre 1971 et 2005 proviennent de la Délégation à la Surveillance Pêche et au Contrôle en Mer (DSPCM). Elles ont été transmises par Étienne Rivot dans le cadre d'une collaboration avec l'Agrocampus de Rennes.

Les résultats des estimations de l'évolution de la biomasse obtenus avec ces différents algorithmes et les méthodes MCMC sont comparables [20]. Ces méthodes mettent ainsi à disposition des écologues une approche alternative aux méthodes MCMC, de plus leur mise en œuvre est aussi simple que les méthodes MCMC mais le temps de calcul est beaucoup plus faible.
Les liens entre filtrage particulaire et MCMC peuvent donc être des points intéressants à étudier.

1.3 Conclusion

Ce premier chapitre constitue une introduction aux méthodes statistiques bayésiennes, il permet de mettre en avant plusieurs concepts de base utiles pour le développement de cette thèse.

La première section explique les fondamentaux des méthodes statistiques bayésiennes, notamment grâce à des travaux réalisés dans le cadre de la théorie de la décision. Ces travaux préléminaires ont permis d'obtenir des résultats théoriques ainsi que des approximations intéressantes concernant le problème d'intervalle de confiance bayésien pour un paramètre contraint.

Ensuite, par le biais d'un état de l'art des méthodes d'approximation, nous avons présenté deux méthodes basées sur le même principe de simulation de Monte Carlo, les méthodes MCMC et les méthodes particulaires. Les méthodes particulaires sont utilisées lorsque l'on souhaite traiter les données en ligne, ce sont des méthodes séquentielles. A contrario les méthodes MCMC traitent les données *a posteriori*. Un travail a été effectué afin de comparer ces deux types de méthodes dans le cas particulier d'un modèle d'évolution de la biomasse. Les résultats obtenus sont similaires et suggèrent de s'intéresser au contrôle du nombre de particules dans l'approximation particulaire. D'autre part, l'aspect MCMC se développe aussi considérablement au travers d'approche de chaînes en parallèles.

Le deuxième chapitre met en exergue les méthodes MCMC dans le cadre des modèles de Cox et logistique.

Chapitre 2

RJMCMC et fonctions splines pour des données cliniques

Ce deuxième chapitre traite de deux modèles bien connus en biostatistique : le modèle de Cox et le modèle logistique. Dans ces deux modèles une hypothèse de relation linéaire ou log-linéaire est posée : le modèle de Cox suppose une relation log-linéaire entre la fonction de risque et les covariables, tandis que le modèle logistique fait apparaître une relation linéaire entre l'*odds ratio* et les variables.

Ces deux types de modélisation sont trop restrictifs pour de nombreuses raisons. L'idée a donc été d'utiliser une représentation B-spline pour les modéliser. En effet, le lissage par B-spline constitue un type de régression populaire grâce à ses bonnes propriétés numériques. Cependant une difficulté réside dans le choix du nombre de nœuds intérieurs nécessaire à une bonne approximation. Nous avons donc mis en œuvre l'algorithme du Reversible Jump Markov Chain Monte Carlo (RJMCMC) afin de sélectionner le nombre de nœuds et de déterminer leur position.

Ce chapitre s'organise de la façon suivante : une première partie présentera le modèle de Cox, le modèle logistique et les splines, ensuite nous présenterons l'article traitant du modèle de Cox et accepté dans le journal *Communication in Statistics - Theory and methods*, enfin nous finirons par un deuxième article soumis à *Journal of Biostatistics* et s'intéressant au modèle logistique.

2.1 Introduction

En recherche clinique nous sommes amenés à définir la relation entre une maladie et des facteurs prédictifs. Les principaux modèles utilisés sont la régression linéaire multiple, le modèle de Cox et la régression logistique. Le choix du modèle se fait en fonction de la nature des variables modélisées.

Ainsi le modèle de Cox sera privilégié dans le cas où la variable à expliquer est dichotomique et que l'on cherche à exprimer le risque instantané de survenue d'un évènement en fonction des facteurs explicatifs.

La régression logistique pourra être utilisée lorsque l'on cherche à déterminer la probabilité de survenue d'un évènement en fonction des variables explicatives

(qualitatives ou quantitatives). Différents évènements peuvent être considérés : la récidive d'une maladie, le décès, la réponse à un traitement, ... Les deux sous-sections suivantes explicitent plus en détail ces deux modèles.

2.1.1 Le modèle de Cox

Le modèle de Cox est un modèle d'analyse de survie semi-paramétrique qui permet d'étudier le délai de survenue d'un évènement. Commençons par définir quelques notions d'analyse de survie

Définitions et notations

Les données de survie comportent deux notions :

(i) une notion qualitative : la survenue de l'évènement (oui/non),

(ii) une notion quantitative : le moment de la survenue de l'évènement T.

Afin de différencier plusieurs types d'observations, plusieurs dates importantes sont à définir :

- La **date d'origine ou date d'inclusion** correspond à la date où l'on débute l'observation,
- La **date de dernières nouvelles (ddn)** est la date la plus récente où l'on est renseigné sur le sujet. Si le sujet est décédé, cette date correspond à la date de décès. Notons que si la ddn se produit avant la date de point et que le sujet n'est pas décédé, les sujets sont des perdus de vue.
- La **durée de surveillance** représente la durée entre la date d'origine et la date de dernières nouvelles,
- La **date de point** est la date de fin d'observation (on ne s'occupe pas de ce qu'il se passe après), c'est également la date à laquelle on souhaite connaître l'état du patient. Cette date correspond au gel de la base, c'est une forme de censure administrative.

Un temps de participation t_i est associé à chaque sujet i, il est défini selon les trois cas de figures suivants :

- Si la date de dernières nouvelles est antérieure à la date de point, le temps de participation correspond à la durée entre la date d'origine et la date de dernières nouvelles,
- Si la date de dernières nouvelles est postérieure à la date de point, le temps de participation correspond à la durée entre la date d'origine et la date de point,
- Si la date de dernières nouvelles est égale à la date de point, le temps de participation est égal à la durée de surveillance.

Les sujets pour lesquels on ne connait pas l'état à la date de point ou pour lesquels l'évènement n'est pas survenu constituent **les données censurées**. Ainsi, on associe à chaque sujet i une variable binaire c_i indiquant l'état du sujet au temps t_i. On va donc se servir des données sous la forme (t_i, c_i) pour

chaque sujet i.

$$\begin{cases} c_i = 1 & \text{si l'évènement est observé avant la date de point} \\ c_i = 0 & \text{sinon} \end{cases}$$

Si l'on a observé le décès du sujet i avant la date de point on aura que $c_i = 1$. A contrario $c_i = 0$ si le sujet est encore vivant au temps t_i. Un sujet dont on ne connait pas l'état à la date de point est un sujet **perdu de vue** et $c_i = 0$. De façon générale, on parlera de donnée censurée lorsque $c_i = 0$.

Soit $T \geq 0$ la variable aléatoire de survie (ou de durée de vie) représentant le délai entre la date d'origine et la date de survenue de l'évènement étudié. La densité associée à cette variable est donnée pour $t \in \mathcal{R}^+$ par :

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

La fonction de répartition associée est définie par :

$$F(t) = \mathbb{P}(T \leq t) = \int_0^t f(u)\, du.$$

La fonction de survie (ou courbe de survie ou survie en t) :

$$S(t) = \mathbb{P}(T > t) = 1 - F(t),$$

représente la probabilité de survivre au delà du temps t. C'est une fonction monotone décroissante et continue telle que $S(0) = 1$ et $\lim_{t \to \infty} S(t) = 0$. Le risque instantané de décès (ou taux instantané de décès) est défini par :

$$h(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T < t + dt | T > t)}{dt} = -\frac{d}{dt} Ln\big(S(t)\big),$$

$h(t)\, dt$ est la probabilité de décéder entre l'instant t et l'instant $t + dt$ pour un sujet sachant que ce sujet est encore vivant en t. La fonction $h(t)$ est également appelée fonction de risque (ou force de mortalité).

On peut écrire la fonction de survie en fonction de la fonction de risque :

$$S(t) = exp\big(-\int_0^t h(u)\, du\big).$$

On note $H(t)$ la fonction de risque cumulé de $h(u)$ entre 0 et t :

$$H(t) = \int_0^t h(u)\, du.$$

Grâce à ces différentes égalités on peut écrire $H(t) = -Ln\big(S(t)\big)$ et

$$f(t) = h(t)\, exp\big(-H(t)\big) \tag{2.1}$$

Plusieurs modèles paramétriques de survie existent : le modèle exponentiel, le modèle de Weibull, le modèle log-normal, ... Dans ces modèles la fonction de risque est une fonction mathématique qui dépend d'un ou plusieurs paramètres. Pour plus de détails sur ces différents modèles, nous recommandons le livre de

Hill et al. [38] qui traite de l'analyse des données de survie.

On s'intéresse ici au modèle de Cox (ou modèle à hasard proportionnel) qui est un modèle semi-paramétrique de survie. Il permet de prendre en compte simultanément plusieurs variables sans donner de formes paramétriques précises aux fonctions de survie. On définit la fonction de risque en fonction des covariables de la façon suivante :

$$h(t,\, X) = h_0(t)\, exp(\beta' X), \tag{2.2}$$

où $X' = [x_1, \ldots, x_p]$ correspond au vecteur des p variables explicatives et $\beta' = [\beta_1, \ldots, \beta_p]$ aux coefficients associés. On ne donne pas de forme paramétrique à la fonction $h_0(t)$.

Cox [17] a suggéré de considérer $h_0(t)$ comme une fonction inconnue que l'on ne cherche pas à estimer. Dans ce modèle, la fonction de risque de chaque individu est le produit de deux fonctions : une dependante du temps et l'autre non. Ainsi le rapport des risques instantannés pour deux sujets de caractéristiques X^1 et X^2 ne dépend pas du temps :

$$\frac{h(t,\, X^1)}{h(t,\, X^2)} = \frac{exp(\beta' X^1)}{exp(\beta' X^2)}. \tag{2.3}$$

Ce rapport est appelé risque relatif à l'instant t des sujets de caractéristiques X^1 par rapport aux sujets de caractéristiques X^2. Il est donc indépendant du temps t.

Le modèle de Cox est appelé modèle à risque proportionnel. La vraisemblance associée au modèle de Cox est une vraisemblance partielle qui est définie par :

$$\mathcal{L}^*(\beta) = \prod_{i \in \text{décès}} \frac{exp(\beta' x_i)}{\sum_{j \in R_i} exp(\beta' x_j)}, \tag{2.4}$$

où l'ensemble R_i correspond à l'ensemble des personnes à risque au temps t_i. Les coefficients β vont être estimés en maximisant $\mathcal{L}^*$.

Une des spécificités du modèle de Cox est la relation log-linéaire entre la fonction de risque et les covariables, c'est à dire que $Ln(h(t,\, X))$ est une fonction linéaire de X. On a donc de (2.2) :

$$Ln(h(t,\, X)) = Ln(h_0(t)) + \beta' X.$$

Cette hypothèse de linéarité est remise en question lorsque les effets des covariables sont mieux représentés par des fonctions lisses non linéaires. On introduit donc des fonctions B-splines, où le nombre et la position des nœuds sont considérés comme des variables libres, pour modéliser ces effets et améliorer l'ajustement. L'algorithme du Reversible Jump MCMC (RJMCMC) est utilisé pour choisir le nombre et la position des nœuds.

2.1.2 La régression logistique

Le modèle de régression logistique permet de représenter sous la forme d'un risque (ou d'une probabilité) la relation entre une variable à expliquer Y dichotomique et une ou plusieurs variables explicatives $X = (X_1, \ldots, X_p)$ (qualitatives ou quantitatives). La probabilité ou le risque de survenue de l'évènement

lorsque les valeurs des variables X_i sont connues et s'écrit :

$$\mathbb{P}(Y = 1|X = x) = \frac{e^{\alpha_0+\alpha' x}}{1 + e^{\alpha_0+\alpha' x}},$$

où $\alpha' = [\alpha_1, \ldots, \alpha_p]$ est le vecteur des coefficients associés à chaque variable. On note $f(x)$ la fonction définie par

$$f(x) = \frac{e^{\alpha_0+\alpha' x}}{1 + e^{\alpha_0+\alpha' x}},$$

et $g(x)$ la fonction logit qui s'écrit de la façon suivante :

$$g(x) = Ln\big(\frac{f(x)}{(1 - f(x))}\big) = \alpha_0 + \alpha' x.$$

Le rapport $\frac{f(x)}{(1-f(x))}$ est appelé *odds* (ou chance). Le modèle logistique suppose donc une relation linéaire entre le *log-odds* et les variables, ce type de modélisation est restrictif. Ainsi, comme pour le modèle de Cox, une représentation B-spline est utilisée pour modéliser cette relation. L'algorithme du Reversible Jump Markov Chain Monte Carlo (RJMCMC) est également mis en œuvre pour déterminer le nombre et sélectionner la position des nœuds.

Les deux articles qui suivent présentent donc l'adaptation du RJMCMC dans le cadre du modèle de Cox et du modèle logistiquue.

Rappelons tout d'abord la notion de fonction spline.

2.1.3 Les fonctions splines

Commençons par définir cette notion.

Définition 2.1.1 (Fonction spline)**.** *Soient un intervalle $[a, b]$, un entier $d \geq 0$, un entier k et une suite de k points $m_1, \ldots, m_k$ dans $[a, b]$. On appelle spline polynomiale de degré d (ou d'ordre $d+1$) ayant pour nœuds intérieurs les points $m_1, \ldots, m_k$ toute fonction s de $[a, b]$ dans $\mathbb{R}$ telle que :*

- *s est continuement dérivable jusqu'à l'ordre $d-1$ (si $d \geq 1$),*
- *les restrictions de s aux intervalles inter-nœuds $[a, m_1], \ldots, [m_i, m_{i+1}], \ldots, [m_k, b]$ coïncident avec des polynômes de degré d.*

L'espace des fonctions splines de degré d aux nœuds $(m_1, \ldots, m_k)$ est un espace fonctionnel linéaire noté $\mathbf{S}^d(m)$ de dimension $k + d + 1$.

Par la suite, nous choisirons les fonctions splines de degré $d = 1$, ce sont des fonctions continues et linéaires par morceaux. La régression spline est utilisée comme une alternative à la régression polynômiale. Ainsi l'utilisation de polynômes par morceaux constitue une méthode flexible et intéressante.

Le choix des fonctions de base de l'espace des fonctions splines constitue une étape importante. Soient $m_1 < \cdots < m_k$ les k nœuds intérieurs, m_0 et m_{k+1} les deux nœuds limites. Un premier choix simple est d'utiliser la base des puissances tronquées.

On définit la puissance tronquée de la façon suivante.

$$x_+ = \begin{cases} x & \text{si } x \geq 0, \\ 0 & \text{sinon.} \end{cases}$$

Une base de l'espace des fonctions splines est donnée par :

$$\{1, x, \ldots, x^d, (x - m_1)_+^d, \ldots, (x - m_k)_+^d\}.$$

Soit s une fonction spline associée à cette base, les propriétés requises sont vérifiées :

- s est un polynôme de degré d sur chaque intervalle $[m_j, m_{j+1})$,
- s a deux dérivées inconnues,
- s a une troisième dérivée qui est une fonction étagée (en étage) avec des sauts aux nœuds.

Bien que cette base soit correcte, elle n'est pas recommandée pour le calcul de la régression spline car elle nécessite de nombreux calculs. Une alternative qui est supérieure d'un point de vue numérique est donnée par la base des fonctions B-splines [19]. Chaque fonction de base $B_j^d(x)$ est non nulle sur un ensemble d'au plus $d + 2$ nœuds distincts, c'est ce qui constitue leur principal avantage. Ainsi d'un point de vue pratique leur évaluation se fait rarement au delà de ces nœuds ce qui permet d'obtenir une matrice de régression "bande". La définition des fonctions B-splines est assez simple, elle est construite à partir des différences divisées.

Commençons par définir cette notion.

Définition 2.1.2. *Soit f une fonction réelle définie sur l'intervalle $[a, b]$ et définie au moins sur les $(n + 1)$ valeurs $x_0, \ldots, x_n$ toutes distinctes.*

La différence divisée d'ordre 0 pour la valeur x_0 notée $[x_0]f$ est définie par :

$$[x_0]f = f(x_0).$$

La différence divisée d'ordre 1 pour la valeur x_0, x_1 notée $[x_0\, x_1]f$ est définie par :

$$[x_0, x_1]f = \frac{[x_1]f - [x_0]f}{x_1 - x_0}.$$

La différence divisée d'ordre n pour la valeur $x_0, \ldots, x_n$ notée $[x_0, \ldots, x_n]f$ est définie par :

$$[x_0, \ldots, x_n]f = \frac{[x_1, \ldots, x_n]f - [x_0, \ldots, x_{n-1}]f}{x_n - x_0}$$

Soit la partition $\tilde{\Delta} = \{\tau_i\}_{i=1,\ldots,k+2\,(d+1)}$ avec $m_0 = \tau_1 = \cdots = \tau_{d+1}$, $m_{k+1} = \tau_{k+d+2} = \cdots = \tau_{k+2\,(d+1)}$ et $\tau_{i+d+1} = m_i$ pour $i = 1$ à k.

On obtient la définition des B-splines.

Définition 2.1.3. *On appelle B-splines polynomiales normalisées de degré d associées à la partition $\tilde{\Delta}$, les fonctions $(B_j^d(x))_{j=1,\ldots,r}$ définies sur l'intervalle $[a, b]$ par :*

$$B_j^d(x) = (-1)^{(d+1)}(\tau_{j+d+1} - \tau_j)\,[\tau_j, \ldots, \tau_{j+d+1}](x - \tau)_+^d$$

Voici les principales propriétés vérifiées par les fonctions B-splines :

(*i*) **Propriété** 1 : le support de $B_j^d(x)$ est $[\tau_j, \tau_{j+d+1}]$ et $B_j^{d+1}(x) = 0$ si $x \notin [\tau_j, \tau_{j+d+1}]$

(*ii*) **Proprieté** 2 : les B-splines de degré 0 correspondent aux indicatrices de support, on a que :

$$B_j^0(x) = \begin{cases} 1 & \text{si } x \in [\tau_j, \tau_{j+1}] \\ 0 & \text{sinon} \end{cases}$$

(*iii*) **Propriété** 3 : les valeurs des fonctions B-splines en un point de l'intervalle $[a, b]$ forment une partition de l'unité. Soit $x \in [a, b]$, $\{B_j^d(x)\}_{j=1,\ldots,k+d+1}$ forment une partition de l'unité. C'est-à-dire :

$$\begin{cases} B_j^d(x) \geq 0 \\ \sum_{j=1}^{k+d+1} B_j^d(x) = 1 \end{cases}$$

(*iv*) **Propriété** 4 : les fonctions $\{B_j^d(x)\}_{j=1,\ldots,k+d+1}$ forment une base de l'espace des fonctions splines.

L'ensemble des graphiques (Fig 2.1) représentent les fonctions B-splines de degré $d = 1$ définies sur l'intervalle [0, 5] avec les nœuds intérieurs (1, 2, 3). Ces graphiques permettent de visualiser les différentes propriétés mentionnées ci-dessus.

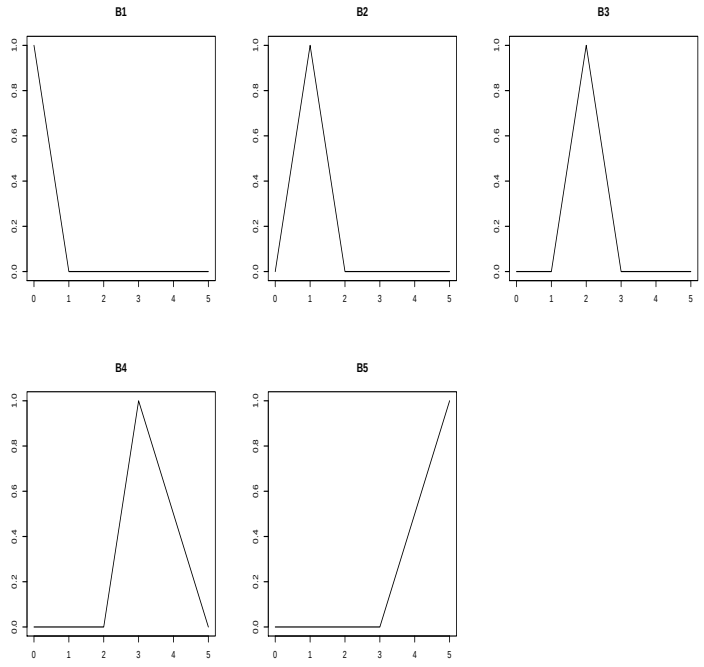

FIG. 2.1 – Les fonctions de base B-splines de degré 2 définies sur l'intervalle [0,5] avec les nœuds intérieurs (1, 2, 3)

D'après la propriété 4 les B-splines constituent une base de l'espace des fonctions splines, ainsi toute fonction spline s se décompose de façon unique sous la forme suivante :

$$s(x) = \sum_{j=1}^{k+d+1} \beta_j \, B_j^d(x). \tag{2.5}$$

Un autre aspect intéressant des fonctions B-splines est donné par la relation de récurrence suivante :

$$B_i^d(x) = \frac{(x - \tau_i)}{(\tau_{i+d} - \tau_i)} \, B_i^{d-1}(x) + \frac{(\tau_{i+d+1} - x)}{(\tau_{i+d+1} - \tau_{i+1})} \, B_{i+1}^{d-1}(x) \tag{2.6}$$

Les splines de régression sont attractives grâce à leurs propriétés calculatoires. La principale difficulté de cette approche concerne le choix du nombre et de la position des nœuds intérieurs. De nombreuses méthodes existent : la plus simple, appelée *cardinal splines*, requiert un seul paramètre : le nombre de nœuds. Les positions sont choisies uniformément sur la rangée des données. Une légère amélioration consiste à placer les nœuds aux quantiles appropriés de la variable prédictive, i.e. trois nœuds intérieurs placés aux trois quantiles. D'autres approches considérant les nœuds comme des variables existent et sont présentées dans les deux articles qui suivent.
Dans ces deux articles une approche basée sur les méthodes MCMC et qui considére les nœuds comme des variables est proposée. Le premier traite du modèle de Cox, le second du modèle logistique.

2.2 Article : "Free knot splines with RJMCMC in survival data analysis"

Ce premier article accepté dans *Communication and Statistics : Theory and Methods* traite du modèle de Cox.

Free knot splines with RJMCMC in survival data analysis

M. Denis [a], N. Molinari [a,b]

[a] *Laboratoire de Biostatistique, Institut Universitaire de Recherche Clinique, UFR médecine, Université montpellier 1, 641, avenue Gaston Giraud, 34093 Montpellier, France*

[b] *Hopital Caremeau, CHU Nîmes, Place du Pr. R. Debré, 30029 Nîmes cedex 9, France*

Abstract

The B-spline representation is a common tool to improve the fitting of smooth non-linear functions, it offers a fitting as a piecewise polynomial. The regions that define the pieces are separated by a sequence of knots. The main difficulty in this type of modeling is the choice of the number and the locations of these knots. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm provides a solution to simultaneously select these two parameters by considering the knots as free parameters. This algorithm belongs to the MCMC techniques that allow simulations from target distributions on spaces of varying dimension. The aim of the present investigation is to use this algorithm in the framework of the analysis of survival time, for the Cox model in particular. In fact, the relation between the hazard ratio function and the covariates being assumed to be log-linear, this assumption is too restrictive. Thus, we propose to use the RJMCMC algorithm to model the log hazard ratio function by a B-spline representation with an unknown number of knots at unknown locations. This method is illustrated with two real data sets: the Stanford heart transplant data and lung cancer survival data. Another application of the RJMCMC is selecting the significant covariates, and a simulation study is performed.

Preprint submitted to Communications in Statitics - Theory and Methods

Key words: Reversible Jump Markov Chain Monte Carlo method, Splines, B-splines, Cox model

1 Introduction

In medical statistics, the analysis of the dependence of survival time on independent variables or covariates X has received considerable attention. The Cox model (Cox (1972)) is a popular choice for this analysis. It is a semiparametric model with a hazard function that is assumed to take the form

$$h(t,\, X) \;=\; h_0(t)\, e^{\beta' X}\,, \tag{1}$$

where $\beta' = (\beta_1,\, \ldots,\, \beta_m)$ is an unknown vector of parameters reflecting the effects of covariates $X' = (X^1,\, \ldots,\, X^m)$ on survival and h_0 denotes the baseline hazard function. Note that no particular shape is assumed for the baseline hazard function. In this model, the relation between the hazard ratio function and covariates X is assumed to be log-linear, i.e., $ln(\frac{h(t,\, X)}{h_0(t)})$ is a linear function of X. More precisely, we have:

$$ln(\frac{h(t,\, X)}{h_0(t)}) \;=\; \beta' X.$$

This assumption is questioned when covariate effects are best represented by smooth non-linear functions. If we take this remark into account, we can write the hazard model under the following form:

$$h(t,\, X) \;=\; h_0(t)\, e^{g(X)},$$

where g, the log hazard ratio function (LHR), is an unspecified smooth function of X. O'Sullivan (1988) uses smoothing splines to estimate non-linear covariate effects in the Cox model. Kooperberg et al. (1995) use linear splines

Email address: marie.denis@inserm.fr (M. Denis).

and their tensor products to estimate the LHR function. In these two methods, the knots are fixed. However, Molinari et al. (2001) use a B-spline representation of degree one for modeling the effect of quantitative covariates. Unlike previous methods, knots are seen as free parameters. The knot location thus corresponds to a break point for the function g and can be interpreted as a threshold value. Furthermore, as the linear model is nested in spline models, a likelihood ratio test is used to select the model and determine the number of knots.

Over the last few years, MCMC simulation techniques have become very important computational tools in Bayesian statistics. These methods belong to a larger class of algorithms which aim at sampling from target distributions on a space of fixed dimension. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm introduced in Green (1995), allows simulations from target distribution on spaces of variable dimension. One application is the comparison of models: the "true" model is unknown but is assumed to come from a specified class of parametric models $\{\mathcal{M}_0, \mathcal{M}_1, \dots\}$.

One of the main purposes of the present investigation is to use these RJMCMC methods to determine the covariates, called significant, which have a possible effect on survival, and to model the log hazard ratio function with a spline function.

In our approach, we provide a Bayesian version that models g by a B-spline representation with an unknown number of knots at unknown locations. To this end, the RJMCMC algorithm is used.

The paper is organized as follows. In Section 2, a short review of the Cox model and B-splines is given. In Section 3, we introduce the Reversible Jump MCMC algorithm, and we give three applications in Section 4, namely simulations, the Stanford heart data presented by Miller and al. (1982), and a real lung cancer survival data set.

2 The model

2.1 The Cox regression model

The Cox model is a well-recognized statistical technique to analyze survival. Survival data set are given, for the ith patient, by (t_i, c_i, X_i), where t_i is the observed survival time, c_i is a binary variable that indicates if the failure, like death, relapse or infection is observed ($c_i = 1$ corresponding to uncensored failure time), and X_i is the covariate value. Let us consider n randomly selected individuals. Let $t_1 < t_2 < \cdots < t_l$ denote the ordered uncensored failure times with corresponding values $X_1, X_2, \ldots, X_l$. We denote by R_{t_i} the collection of individuals with censored or uncensored failure times $\geq t_i$. Cox's partial likelihood is given by:

$$L(\beta) = \prod_{i=1}^{l} \frac{exp(\beta' X_i)}{\sum_{j \in R_{t_i}} exp(\beta' X_j)}. \tag{2}$$

2.2 Splines and Cox model

The use of splines introduces a nonparametric character in the regression. The tuning parameters for regression splines are the number of interior knots k, the degree d, and the location of the knots.

A spline function belongs to a linear functional space of dimension $d + 1 + k$. Let $r_1 < r_2 < \cdots < r_k$ be the k interior knots, where $r_j \in]X_{min}, X_{max}[$ and $r_0 = X_{min} = \min_{i=1,\ldots,n}(X_i)$, and $r_{k+1} = X_{max} = \max_{i=1,\ldots,n}(X_i)$ the boundary knots. The most popular basis function for this linear space is called B-splines (de Boor C (1978)) and is denoted by

$$s(X, \beta, r) = \sum_{i=1}^{d+1+k} \beta_i B_i(X),$$

where $\beta = (\beta_1, \ldots, \beta_{d+1+k})$ are the spline coefficients. In the case where we

assume that covariate effects are best represented by smooth non-linear functions, we can approximate the log hazard ratio function in the Cox model by a spline function s of degree d:

$$LHR(X) = ln(\frac{h(t,\, X)}{h_0(t)}) \;=\; s(X,\, \beta, r).$$

The spline partial likelihood function is defined by

$$L(\beta) \;=\; \prod_{i=1}^{l} \frac{exp(s(X_i,\, \beta, r))}{\sum_{j \in R_{t_i}} exp(s(X_j,\, \beta, r))}. \tag{3}$$

In this paper, we estimate the log hazard ratio function with a linear spline $d = 1$. This implies easy interpretations: in fact, knots are points where the slope changes to the shape of the piecewise linear function. Therefore, a quick change of slope can be interpreted as a point separating the variable range into two parts and the knot location corresponds to a threshold value. From a clinical point of view, this corresponds to a change in the risk function.

3 Estimation

This section presents essential background on the Reversible Jump MCMC proposed by Green (1995). The adaptation of this algorithm for covariate selection and the spline regression is also given.

3.1 RJMCMC

The reversible jump MCMC algorithm allows simulation from target distributions on spaces of variable dimension, and it can be considered as a general framework for Metropolis-Hastings algorithms (Hastings (1970), Metropolis and al. (1953)).

Consider the following hierarchical model: let k be an indicator from a countable set $\mathcal{K}$. $\theta^{(k)} \in \Theta^{(k)}$ denotes the parameter vector. Each k determines a model $\mathcal{M}_k$ defined by the parameter $\theta^{(k)}$, with a dimension of parameter space $\Theta^{(k)}$, which can vary with k. The joint distribution of $(k,\, \theta^{(k)},\, y)$, where y is the data vector, is modeled as:

$$p(k,\, \theta^{(k)},\, y) \;=\; p(k)\, p(\theta^{(k)} \mid k)\, p(y \mid k,\, \theta^{(k)}),$$

i.e., the product of model probability, parameter prior and likelihood. Inference about k and $\theta^{(k)}$ will be based on the joint posterior $p(k,\, \theta^{(k)}|y)$, which is known as the *target* distribution. For convenience, we abbreviate $(k,\, \theta^{(k)})$ as z and we note $\pi(dz)$ this *target* distribution. Given k, z lies in $C_k \;=\; \{k\}\times\Theta^{(k)}$, while generally $z \;\in\; C = \bigcup_{k\,\in\,\mathcal{K}}\, C_k$.

In the Markov Chain Monte Carlo computation, an aperiodic and irreductible Markov transition kernel $P(z,\, dz')$ is constructed. It satisfies detailed balance:

$$\int_A \int_B \pi(dz)\, P(z,\, dz') = \int_B \int_A \pi(dz')\, P(z',\, dz), \tag{4}$$

where $A,\, B \in C$. We simulate this chain to obtain a dependent, approximate sample from $\pi(dz)$.

In our case, we have multiple parameter subspaces $\{C_k\}$ of different dimension. Therefore, a method that switches between these subspaces is needed. To this end, different types of move m between the subspaces can be defined. If the current state is z, a move of type m to state dz' with probability $q_m(z,\, dz')$ is defined and is accepted with probability

$$\alpha_m(z,\, z') = min\{1,\, \frac{\pi(dz')\, q_m(z',\, dz)}{\pi(dz)\, q_m(z,\, dz')}\}. \tag{5}$$

For moving z to z', we must generate random numbers u and set z' as a determinist function of z and u : $z' \;=\; z'(z,\, u)$. The reverse move from z' to z has to be defined symmetrically by generating random numbers u' and

setting $z = z(z',\, u')$. The vectors of Markov chain states and proposal random variables $(z,\, u)$ and $(z',\, u')$ must be of equal dimension, that is, the crucial dimension matching condition:

$$n_1 + n_1' = n_2 + n_2',$$

where n_1, n_2 are the dimensions of z, z', respectively, and n_1', n_2' are the dimensions of u, u', respectively.

The ratio (5) becomes

$$\begin{aligned}\alpha_m(z,\, z') =& min\{1,\, \text{likelihood ratio}\times \text{ prior ratio } \times \text{ proposal ratio}\} \\ =& min\{1,\, \frac{p(y|z')}{p(y|z)}\, \frac{p(z')}{p(z)}\, \frac{p(k|k')\, q_2(u')}{p(k'|k)\, q_1(u)}\ |\ \frac{\partial(z',\, u')}{\partial(z,\, u)}\ |\}.\end{aligned}$$

where q_1, q_2 are the distribution of u, u' and $|\ \frac{\partial(z',\, u')}{\partial(z,\, u)}\ |$ the jacobian. Often in practice $n_1 + m_1 = n_2$; consequently, a random u is necessary only for the birth step and we omit the terms $q_2(u')$ and u' in $\alpha_m(z,\, z')$.

3.2 RJMCMC for covariate selection

In this section, we consider the problem of covariate selection in the Cox regression model. We want to select the clinically significant covariates. Let us suppose that we have m covariates. $\mathcal{M}_k$ denotes the model with k covariates. The parameter $\theta^{(k)}$ represents the covariate vector. For convenience, we associate a number with each covariate, for example, the covariate X^1 is associated with 1 and X^2 is associated with 2, ... Thus $\theta^{(k)}$ is a subset of $\{1,\, 2,\, \ldots,\, m\}$ of length k.

We shall generate samples from the joint posterior of $(k,\, \theta^{(k)})$. The different moves are :

(1) the addition of a covariate,

(2) the deletion of a covariate,

(3) the change of a covariate.

Steps (1) and (2) change the dimension of the model.

The model indicator k is assumed to lie in a set $\mathcal{K} = \{1, \ldots, m\}$. Consequently, a prior for k corresponds to a discrete distribution. Several alternatives are plausible: a Poisson distribution with parameter λ restricted to the set $\mathcal{K}$, a discrete uniform distribution on $\mathcal{K}$, or a negative binomial. In our problem after a simulation study on a simple example, we observe that a discrete uniform distribution systematically overestimates the number of significant covariates. The other distributions lead to the same posterior mode of the number of significant covariates (after a burn-in period). Therefore, we choose the Poisson distribution truncated to $k \leq m$ and $k > 0$ to specify the prior probability of k; this choice is somewhat arbitrary (Denison et al. (1998)). The parameter $\theta^{(k)}$ is taken uniformly from the state space $\{1, 2, \ldots, m\}$:

$$p(\theta^{(k)}|k) = (C_m^k)^{-1} = \frac{(m-k)!\,k!}{m!}.$$

The proposed covariate X^i to add in step (1) is found by uniformly choosing one of the covariates that is not in the current model $p(X^i) = 1/(m-k)$. The proposed covariate X^i to delete in step (2) is chosen uniformly from the covariates of the current model with probability $p(X^i) = 1/k$. In step (3), we choose a covariate X^j uniformly from the covariates of the current model and we choose another covariate X^l $(l \neq j)$ uniformly from the other covariates with probability $p(X^l|X^j)\,p(X^j) = (1/(m-k))\,(1/k)$.

According to Green, the acceptance ratio for each of the types of move is given by:

$$\alpha = min(1, \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio})$$

where the likelihood is the partial likelihood (2). An application of this algo-

rithm is given in Section 4.

3.3 RJMCMC for a Cox spline regression model

The Cox B-spline regression model is defined by

$$h(t,\, X) \;=\; h_0(t)\, e^{g(X)}$$

with

$$g(X) = LHR(X) = ln(\frac{h(t,\, X)}{h_0(t)}) \;=\; s(X,\, \beta, r).$$

s is a B-spline function of degree d= 1

$$s(X,\, \beta, r) = \sum_{i=1}^{k+2} \beta_i\, B_i(X,\, r).$$

The reversible Jump Markov Chain Monte Carlo method allows the selection of the number and position of the knots in order to obtain the best adjustment. For a Bayesian approach, let us formulate the hierarchical model: we take the number of interiors knots k random, from some countable set $\mathcal{K}$. $\mathcal{M}_k$ denotes the model with exactly k interior knots and $r^{(k)} \;=\; (r_1,\, \ldots,\, r_k)$ denotes the k interior knot locations, with $r_0 = X_{min}$ and $r_{k+1} = X_{max}$ the boundary knots.

Note that the same notation k is used to indicate the covariate number (Section 3.2) and the number of knots because these parameters define the space dimension of the model $\mathcal{M}_k$.

The vector of spline coefficients $\beta = (\beta_i)_{1 \leq i \leq k+2}$ is to be estimated from the data. We use the spline partial likelihood function (3). We shall generate samples from the joint posterior of $(k,\, r^{(k)})$. To take the varying dimensionality into account, we have to develop an appropriate reversible jump move. For this problem, possible transitions are

(1) the addition of a knot (a birth step),
(2) the deletion of a knot (a death step),

(3) the movement of a knot.

These independent move types are randomly chosen with probability b_k for move k to $k+1$ (i.e., birth step), d_k for move k to $k-1$ (i.e., death step), and η_k for the move step. These probabilities satisfy $b_k + d_k + \eta_k = 1$ for all k.

3.3.1 Prior specifications

Let $k \in \mathcal{K} = \{0, \ldots, k_{max}\}$; we use a Poisson distribution, with parameter λ restricted to the countable set $\mathcal{K}$, to specify the prior for k:

$$p(k) = \frac{\lambda^k \, exp(-\lambda)}{k\,!}.$$

The r_i are taken to be the order statistics from a uniform random variable with the state space of the candidate knot sites $\mathcal{R} = \{r_{01}, \ldots, r_{0K}\}$, where $r_{01}, \ldots, r_{0K}$ are distributed equidistantly over interval $]X_{min}, X_{max}[$, i.e., the knots are equally spaced by $\frac{X_{max}-X_{min}}{K}$. Then the prior distribution for $r^{(k)} = (r_1, \ldots, r_k)$ is

$$p(r^{(k)}|k) = \frac{k\,!}{K^k},$$

where K is the number of possible emplacements.

3.3.2 Move step

The move step consists of uniformly choosing a knot, say r_j, from the set of moveable knots and proposing this knot to be moved to another position $r'_j \in \mathcal{R}$. A knot $r_j \in \{r_1, \ldots, r_k\}$ is called moveable (Biller (1998)) if the number m_j of vacant candidate knots $r_{0i} \in \mathcal{R}$ with $r_{j-1} < r_{0i} < r_{j+1}$ is at least 1. Let $r = r^{(k)}$; the number $n(r)$ of moveable knots is then defined as

$$n(r) = card\{r_j \mid m_j \geq 1,\, j \in \{1, \ldots, k\}\}.$$

So, first, we draw a knot r_j uniformly from $n(r)$ moveable knots with proba-

bility $p(r_j) = \frac{1}{n(r)}$ and, given r_j, we draw uniformly r'_j (the new position) from the set of m_j of vacant candidate knots, with probability $p(r'_j \,|\, r_j) = \frac{1}{m_j}$. The corresponding proposal ratio for move step is given by

$$\text{proposal ratio} = \frac{p(r_j|r'_j)\, p(r'_j)}{p(r'_j|r_j)\, p(r_j)} = \frac{n(r)\, m_j}{n(r')\, m'_j} = \frac{n(r)}{n(r')}.$$

The prior ratio is 1 because all collections of the same number of knots have the same prior probability. The acceptance probability for the move step is

$$\alpha = min\{1, \frac{p(y \,|\, (k,\, r'))\, n(r)}{p(y \,|\, (k,\, r))\, n(r')}\},$$

where $p(y \,|\, (k,\, r'))$ is the Cox partial likelihood.

3.3.3 Changing dimension

Let $z = (k,\, r)$, where $r = (r_1,\, \ldots,\, r_k)$. We define $b_0 = d_{k_{max}} = 1$, $b_{k_{max}} = d_0 = 0$ and otherwise $b_k = d_k = 1/3$.

In the birth step, given k, we add a new knot $r'_j \in (r_j,\, r_{j+1})$. r'_j is drawn uniformly with probability $p(r'_j) = 1/(K-k)$ from the set of the $(K-k)$ vacant candidate knots $r_{0i} \in \mathcal{R}$. We have $z' = (k+1,\, r')$ where $r' = (r_1,\, \ldots,\, r_j,\, r'_j,\, r_{j+1},\, \ldots,\, r_k)$. For the birth step, the prior ratio is given by:

$$\begin{aligned}
\text{prior ratio} =& \frac{\text{prior for } k+1 \text{ knots}}{\text{prior for } k \text{ knots}} \frac{\text{prior for location of } k+1 \text{ knots}}{\text{prior for location of } k \text{ knots}} \\
=& \frac{p(k+1)}{p(k)} \frac{p(r'|k+1)}{p(r|k)} \\
=& \frac{p(k+1)}{p(k)} \frac{k+1}{K}.
\end{aligned}$$

The corresponding proposal ratio is given by

$$\begin{aligned}
\text{proposal ratio} =& \frac{d_{k+1}\,(1/k+1)}{b_k\,(1/K-k)} \\
=& \frac{d_{k+1}\,(K-k)}{b_k\,(k+1)}.
\end{aligned}$$

In the death step, the proposed knot to delete is simply chosen uniformly from the knots of the current model, so it is drawn with probability $p(r_{j+1}) = 1/(k+1)$.

The acceptance probability for the birth step is

$$\alpha(z,\, z') \,=\, min\{1,\, \frac{p(y \,|\, z')}{p(y \,|\, z)} \times \text{prior ratio} \times \text{proposal ratio}\}.$$

For the death step, it is the same except that the fraction is inverted. The coefficients β are estimated at each step with the method of maximum likelihood.

4 Applications

In this section, we illustrate the reversible jump algorithm with three examples: a basic problem of variable selection with simulations and a Cox spline regression with two sets of data: the Stanford heart transplant data and lung cancer data.

4.1 Covariate selection

In our first application, we use the reversible jump algorithm for the problem of variable selection in the Cox model. The RJMCMC allows the determination of the number k of significant covariates (section 3.2). We compare this method with classic alternatives, like the AIC and the BIC criteria.

We have simulated data according to a Cox model, where the Weibull function is used as the baseline hazard function. The parameters of the Weibull function are the shape $\alpha = 2$ and the scale $\lambda = 15/2$. Seven covariates $X' = (X^1, \ldots, X^7)$ are included in the model with the coefficient vector

$\beta = (1, 0, 0.8, 0.4, 1.5, 0, 0.2)$. We take $\lambda = 2$ for the parameter of the Poisson distribution, $m = 7$ corresponds to the number of covariates, and the size of the simulated sample of survival times is $n = 200$. Regarding the choice of the parameter $\lambda = 2$, it depends first on the prior beliefs that the researcher may have about the number of significant covariates. We have tested different values of λ in a simulation study; $\lambda = 1$ seems to be too restrictive and the values $\lambda = 2, 3, 4$ lead to the same results, so we take the smaller value $\lambda = 2$. Furthermore, these observations show us that the method is robust. The goal is to select the significant covariates, i.e., X^1, X^3, X^4, X^5, X^7. The difficulty lies in the fact that for small coefficients, like β_4, β_7, the selection problem becomes harder. In order to compare the performance of the RJMCMC algorithm with the classic alternatives, we have run them with a wide range of censoring percentages and for each percentage $N = 200$ samples are simulated. For each sample a stepwise procedure with AIC and BIC is applied as well as the RJMCMC algorithm to select the significant covariates. In Table 4.1, we give the percentage of times when the covariates were selected during the N simulations for a censoring percentage equal to 30%. Similar results are obtained for different values, so the amount of censoring does not seem to influe on the results.

We have seen in Section 3.2 the different reversible jump moves of the RJMCMC algorithm. In each of those three moves, we have used the maximum likelihood method to estimate the vector β. The estimates are obtained with 10000 iterations. The posterior distribution for k allows an estimation of the number of significant covariates. With resped to the posterior distribution of θ_k given k, it provides the significant covariates. Indeed, the k modes of this distribution correspond to the k significant covariates.

From Table 1, we can see that the BIC criterion tends to underestimate the number of significant covariates, unlike the AIC criterion, which overestimates

censoring percentage: 30%	X^1	X^2	X^3	X^4	X^5	X^6	X^7
BIC	100%	0%	100%	64%	100%	0%	33%
AIC	100%	7 %	100%	79%	100 %	10 %	60%
RJMCMC	100%	0%	100 %	69 %	100 %	2 %	50%

Table 1

Percentage of times that covariates are selected by the different methods during $N = 200$ simulations.

it. This is due to the penalty term, and it is a common observation. The RJMCMC algorithm takes place in the middle; it is neither better nor worse. In fact, the RJMCMC algorithm is not a tool for covariate selection although it seems as successful as the classic methods. We use it instead for the difficult problem of the choice of the number and location of interior knots in the B-spline representation. In this context, the RJMCMC is of real interest. In fact, for the procedures using the BIC or AIC criterion, we must first determine for each number of knots k the spline coefficients and the position. We then apply a criterion for model selection for each possibility to determine the knot number. Unlike this procedure, the RJMCMC method allows to select simultaneously the number and the position of knots in a single algorithm. Furthermore, the B-spline representation for the log hazard ratio function allows us to remove the linear assumption if the covariate effects are best represented by a smooth non-linear function, but it is also suitable if the link is linear.

4.2 Stanford heart transplant data

Miller and al. (1982) provided a number of analyses of the Stanford heart transplant data. The program began in October 1967. By February 1980, 157 patients had received heart transplants. Of these 157 patients, 55 were still alive, i.e., were censored, as of February 1980 and 102 were deceased, i.e., were

uncensored. The data vector contains 157 observations of the status indicator (censored or uncensored), the survival time (months), and one covariable; age (years).

The reversible jump algorithm is used to determine the number and location of knots to have the best adjustment. The estimates are obtained with 10000 iterations.

We take $k \in \mathcal{K} = \{0, \ldots, k_{max}\}$ with a maximal number of knots $k_{max} = 5$ and a Poisson distribution with $\lambda = 2$ for the prior of k. The candidate knot sites $\mathcal{R}$ are of length $K = 20$. The left part of Figure 1 shows the posterior distribution of k, the mode is at 1. The right part illustrates the posterior distribution of r given $k = 1$; the mode is in the interval [45, 50] which corresponds to the knot $r_{0i} = 46$.

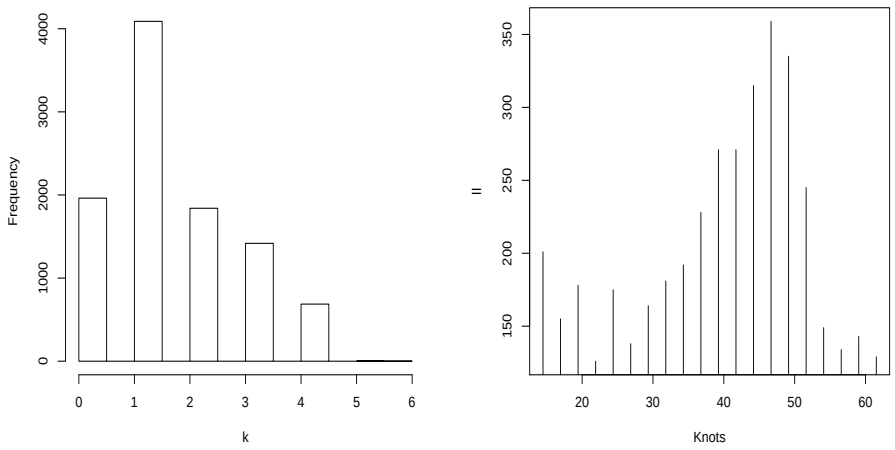

Fig. 1. The posterior distribution of k and of r given k

The result is shown in Figure 2. In fact, there is an obvious break for the log hazard ratio function. This knot is truly meaningful: we can assume the age of 46 years as a threshold value for heart transplantation. Similar results were obtained by Durrleman et al. (1989) with restricted cubic splines and by Hastie and al. (1986) with a local likelihood and local scoring introduction.

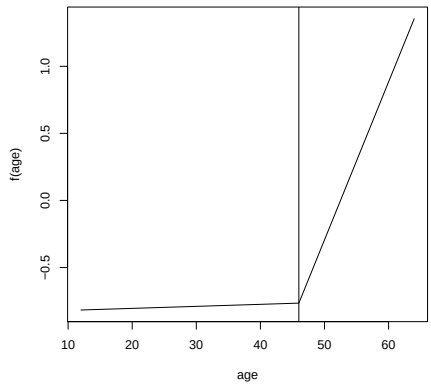

Fig. 2. The log-hazard ratio function computed with linear spline and one optimized knot for the Stanford heart transplant data.

4.3 Lung cancer data

Serum markers have been proposed to help in the management of small-cell lung cancer (SCLC) during chemotherapy. In this setting, the most established serum marker is the gamma-gamma isomer of a glycolytic enzyme referred to as neuron specific enolase (NSE). Recently, a tumor marker, named CYFRA, was proposed to detect cytokeratins in the serum. In this application, we have 124 patients. The sample is (X_i, t_i, c_i), $i = 1, \ldots, 124$, where $X_i'=(X_i^1, X_i^2)=(\text{cyfra}_i, \text{nse}_i)$ are two predictive variable values, t_i is the survival time, and c_i is the final vital status indicator for th ith patient. We want to establish the relationship between risk of death and marker level during treatment. We have to use the reversible jump algorithm, first to find the best B-spline representations for NSE and CYFRA, and second to determine the significant marker.

The B-spline representation for NSE and CYFRA.

We proceed in the same way as in Section 4.2. With two variables, an additive model is assumed and the log hazard ratio function is given by

$$\begin{aligned} g(X) &= g_1(nse) + g_2(cyfra) \\ &= \sum_{i=1}^{k_1+2} \beta_i^1 \, B_i(nse) \; + \; \sum_{i=1}^{k_2+2} \beta_i^2 \, B_i(cyfra). \end{aligned}$$

In this bivariate case, we let $k = \sum_{i=1}^{2} k_i$ where k_i is the spline degree s_i and $r^{(k)} = (r^{(k_1)}, r^{(k_2)})$ is the parameter vector for each spline. The same movements as in the previous algorithm are used: addition, deletion or movement of a knot. At each iteration, we randomly choose the spline which we are going to modify. The prior for k is a Poisson distribution truncated with parameter λ. The choice of this parameter reflects, in this context, the parsimony of the model. Here, let (k_1, k_2) and $(r^{(k_1)}, r^{(k_2)})$ represent respectively, the number of interior knots and the knot locations for NSE and CYFRA. For CYFRA, the site of candidate knots is the space of 20-quantiles. For NSE, the candidate knots are distributed uniformly over the interval $]X_{min}, X_{max}[$.

The estimates are obtained with 10000 iterations. Figure 3 shows the posterior distribution of k_1 and k_2. For NSE, we retain any interior knot, which means that the NSE effect is linear: a patient with a high NSE level has a high risk of death. Concerning CYFRA, the posterior distribution indicates a mode at 1. From Figure 4, the posterior distribution of r_2 given $k_2 = 1$ admits a mode at 35.1. These results correspond to those obtained in Molinari et al. (2001).

Figure 5 represents the optimized adjustment.

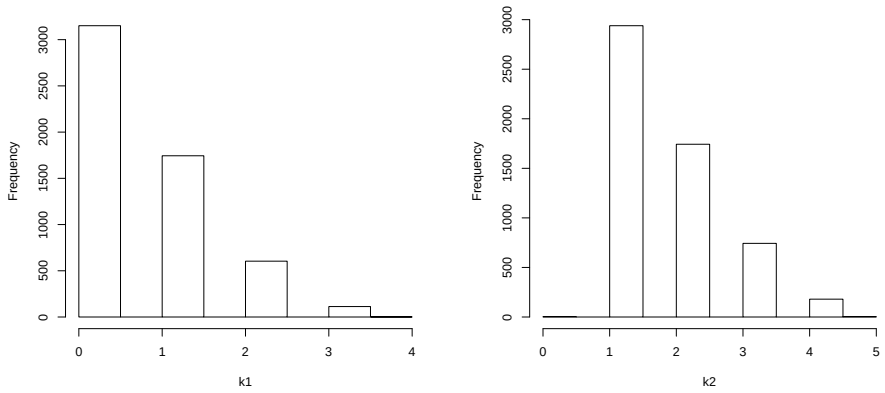

Fig. 3. The posterior distribution of k_1 and k_2.

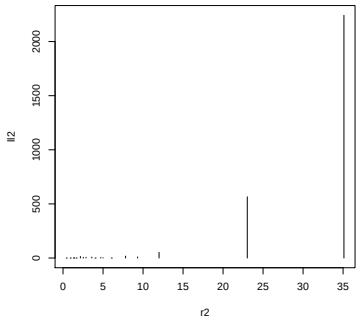

Fig. 4. The posterior distribution of r_2 given $k_2 = 1$.

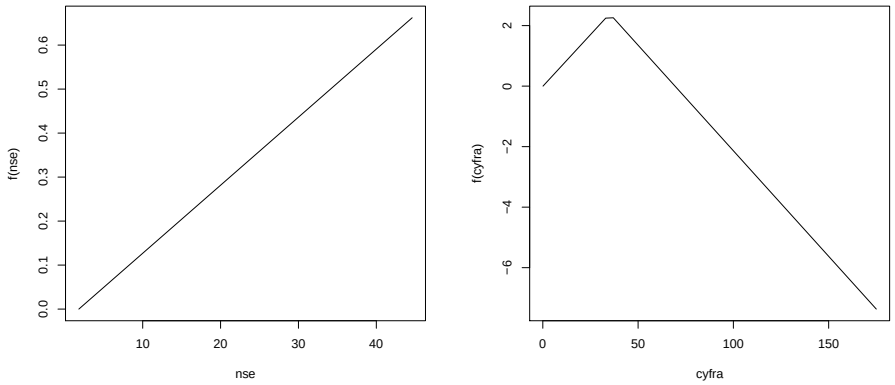

Fig. 5. The log hazard ratio function modeled by B-spline with NSE (left) and by one knot B-spline with CYFRA (right).

Covariate selection

Let $X_i' = (nse_i,\, cyfra_i)$; we use the same process seen in Section 3.2 to select the significant covariates. Unlike in Section 3.2 where the relation between the log hazard ratio function and the covariates was linear, we have here an additive model with B-spline representations :

$$log(\frac{h(t,\, X)}{h_0(t)}) \,=\, \sum_{i=1}^{2} \beta_i^1 \, B_i(nse) + \sum_{i=1}^{3} \beta_i^2 \, B_i(cyfra).$$

The algorithm selects one significant covariate: NSE, with a proportion of appearance of 95%.

5 Conclusion

In this paper, a derivation of the RJMCMC algorithm is used to analyse survival data with the Cox model and non-linear spline covariate effect. Variable selection and free knot spline fitting are performed with this procedure. An interesting perspective would be to construct an algorithm that simultaneously allows the selection of the significant covariates, the best adjustment B-spline and the degree of splines.

References

Biller, C. (1998). Adaptive Bayesian regression splines in semiparametric generalized linear models. Journal of Computational and Graphical Statistics 9:122–140

Cox, DR. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, Series B 34:187-220.

de Boor, C. (1978). A Practical Guide to Splines. New-York: Springer-Verlag.

Denison D.G.T, Mallick B.K. and Smith A.F.M. (1998). Automatic Bayesian curve fitting. Journal of the Royal Satistical Society (B) 60:333-350.

Durrleman, S. Simon, R. (1989). Flexible regression models with cubic splines. Statistics in Medicine 8:551-561.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711-732.

Hastie, T. Tibshhirani, R. (1986). Generalized additive models. Statistical Science 1:297-318.

Hasting, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Kooperberg, C. Stone, C.J. Truong, Y.K. (1995). Hazard regression. Journal of the American Statistical Association 90:78-94.

Metropolis, N. Rosenbluth, A.W. Rosenbluth, M.N. Teller, A.H. Teller, E. (1953). Equations of state calculations by fast computing machines. J. Chem. Phys 21:1087-1091.

Miller, R.G. Hallpern, J. (1982). Regression with censored data. Biometrika 69:521-531.

Molinari, N. Daurès, J.P. Durand, J.F. (2001). Regression splines for threshold selection in survival data analysis. Statistics in Medicine 20:237-247.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. SIAM Journal of Scientific and Statistical Computing 9:531-542.

2.3 Article : "Free knot splines with RJMCMC for logistic models and threshold selection"

Ce deuxième article soumis à *Journal of Biostatistics* traite du modèle logistique.

Free-knot splines with RJMCMC for logistic models and threshold selection

M. Denis[1], N. Molinari[2]

[1]

Abstract

In medical statistics, the logistic model is a popular choice for the analysis of the dependence between a response variable and one or more explanatory variables. The response variable is the log odds and it is a linear function of explanatory variables. This type of modeling is restrictive, as the behaviour of the log odds can be best represented by a smooth non-linear function. Thus, we use a representation B-spline, where the number and location of knots are seen as free variables, is used to improve the fitting. For a piecewise linear spline, knots are points where the slope is changing in the shape of the function. Therefore, a quick change of slope allows to interpret the knot location as a threshold value. The use of MCMC simulation techniques is a very important computational tool in Bayesian statistics. These methods belong to a class of algorithms for sampling from target distributions on a space of fixed dimension. The Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm, allows simulations from target distributions on spaces of varying dimension. One of the main purposes of the present investigation is to use this RJMCMC method for modeling the log odds by a B-spline representation with an unknown number of knots at unknown locations. The method is illustrated with simulations and a real data set from an in vitro fertilization program.

1 INTRODUCTION

Logistic regression is a powerful and flexible means to analyze the relationship between a dependent dichotomous variable (e.g. which only takes two possible values) and one or more risk factors (e.g. explanotary variables). It is a method very used in applied research, but it assumes that these explanotary variables have a linear effect on the model. This assumption is restrictive, in fact in most of problems the underlying processes are complex and not well understood. Using spline functions seems to be an interesting alternative to study this relationship. It permits to detect the possibility of non-linear effects of the explanotary variables. The name spline function was introduced by Schöenberg (1) in 1946. The real explosion in the theory, and in practical applications, began in the early 1960s. Spline functions are used in many applications such as

[1]Institut Universitaire de Recherche Clinique (IURC), University of Montpellier 1, 641, avenue Gaston Giraud, 34093 Montpellier, France. E-mail: marie.denis@inserm.fr.
[2] Hôpital Carremeau, CHU Nîmes, Place du Pr. R. Debré, 30029 Nîmes cedex 9, France.

interpolation, data fitting, solving numerically ordinary and partial differential equations (finite element method), and in curve and surface fitting. For survival data analysis, Sleeper and Harrington (2) introduced spline function into the Cox model. Kooperberg et al. (3) developed the hazard regression (HARE) method which uses piecewise linear regression splines to model the hazard function. The diversity of applications exists due to the great flexibility of splines. But, the main difficulty of splines is the selection of the number and location of knots. In this paper, we utilize the Reversible Jump Markov Chain Monte Carlo (RJMCMC) technique introduced by Green (4) to handle this difficulty.

In recent years the use of MCMC simulation techniques has been a very important computational tool in Bayesian statistics. These methods belong to a class of algorithms for sampling from target distributions on a space of fixed dimension. The RJMCMC algorithm allows simulations from target distribution on spaces of varying dimension. One application is the comparaison of models: the "true" model is unknown but is assumed to come from a specified class of parametric model $\{\mathcal{M}_0, \mathcal{M}_1, \dots\}$.

One of the main purposes of the present investigation is to use this RJMCMC method for modeling the logit function by a B-spline representation with an unknown number of knots at unknown locations. Considering the spline knots as free parameters implies more flexibility and improves data approximation. Moreover, the use of spline allows to both define threshold values and remove the linearity assumption of the logit function. If the estimation of the logit function is based on piecewise linear splines, the knot location corresponds to a break point in the linearity, so a quick change of slope can be interpreted as a point separating the variable range in two parts and the knot location corresponds to a threshold value. Finally, the RJMCMC algotrithm gives directly the knot number without using a model selection criterion and it allows to estimate a wide range of features for the function of the interest. This approach has been introduced by (9) and developped by severals authors ((1)).

The paper is organized as follows. In section 2, a short review of the spline functions and the logistic model is given. In section 3, we shall introduce the Reversible Jump MCMC algorithm, and we give two applications in section 4 with simulations and a real data set from an in vitro fertilization program.

2 THE MODEL

2.1 Spline functions

Let $(r_0 =)a < r_1 < r_2 < \dots < r_k < b(= r_{k+1})$ be a subdivision of k distinct points on the interval $[a, b]$ on which the x variable is valued. We denote the points $(r_1, \dots, r_k)$ as the k interior knots, r_0 and r_{k+1} as the boundary knots. The spline function $s(x)$ used to transform the x variable is a polynomial of degree d (or order $d + 1$) on any intervals $[r_{i-1}, r_i]$, and has $d - 1$ continuous derivatives on the open interval $[a, b]$. These functions provide great flexibility for fitting data, which is controlled by the number of knots. Spline functions belong to a linear functionnal space of dimension $d + 1 + k$. The most popular basis function for this linear space is given by Schoenberg's B-splines, or Basic-splines, and is denoted by $\{B_1^d(., r), \dots, B_{d+1+k}^d(., r)\}$ for a fixed sequence of knots $r = (r_1, \dots, r_k)'$. Their structure is advantageous as it requires less

computation as compared to other basis functions such as the truncated power basis (Eubank (15); Ramsay and Silverman (16)). De Boor (5) proposes a recursive algorithm to compute B-splines of any degree from B-splines of a lower degree.

We can define B-spline basis functions by:

$$B_j^1(x,\, r) \quad = \quad \begin{cases} 1 & \text{si } r_j \leq x \leq r_{j+1} \\ 0 & \text{sinon} \end{cases}$$

$$B_j^s(x,\, r) \quad = \quad \frac{r - r_j}{r_{j+s-1} - r_j}\, B_j^{s-1}(x,\, r) + \frac{r_{j+s} - r}{r_{j+s} - r_{j+1}}\, B_{j+1}^{s-1}(x,\, r),$$

where $j = 1, \ldots, k + d + 1$ and $s = 2, \ldots, d + 1$. Thus, each basis function is non zero in a limited interval spanned by a $d + 1$ adjacent knots which leads to stable estimates and reduces computation. These are piecewise polynomials with continuity constraints on the polynomial and its first $d - 1$ derivaties at the interior knots.

So, a spline function can be written

$$s(x,\, \beta,\, r) \;=\; \sum_{i=1}^{d+k+1} \beta_i\, B_i^d(x,\, r), \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_{d+1+k})'$ is the vector of the spline coefficients and $r = (r_1, \ldots, r_k)'$ is the vector of the interior knots. We can extend to the multivariate case by using additive models. With additive modeling (6) one can to decompose a function of the form $h(X) = h(X_1, \ldots, X_p)$ by a sum of functions of the individual components of X, where $Y = h(X)$ is the response variable and $X = (X_1, \ldots, X_p)$ the explanotary variables.

Let $(y^i,\, x^i)$ the observations, where each x^i is a p-vector $(x_1^i, \ldots, x_p^i)$.

So, h is defined by:

$$Y = h(X) = h(X_1, \ldots, X_p) = \sum_{i=1}^{p} h_j(X_j), \tag{2}$$

and, an estimator s of f can be given by:

$$s(x) = s(x_1, \ldots, x_p) = \sum_{j=1}^{p} s_j(x_j,\, \beta^j,\, r^j), \tag{3}$$

where $\beta^j = (\beta_1^j, \ldots, \beta_{k_j+d_j+1}^j)'$, $r^j = (r_1^j, \ldots, r_{k_j+d_j+1}^j)'$ for $j = 1, \ldots, p$ and each function s_j is defined according to the equation (1). If X_i and X_j are two variables, and the response variable depends on the combination of levels of X_i and X_j, then X_i and X_j are said to interact. We incorporate a term to model this interaction; thus the model can be represented by an additive model including multiplicative interaction of order 1, as follows

$$s(x) = s(x_1, \ldots, x_p) = \sum_{j=1}^{p} s_j(x_j,\, \beta^j,\, r^j) + \sum_{i<j} s_{ij}(x_i \times x_j,\, \beta^{ij},\, r^{ij}). \tag{4}$$

2.2 Spline logistic regression model with free-knots

The logistic model is used to study the relationship between a dichotomous variable (or response variable) and one or more explanotary variables. This

model estimates the probability of a certain event occurring. The specific form of the logistic regression is

$$f(x) = \frac{exp(\alpha_0 + \alpha' x)}{1 + exp(\alpha_0 + \alpha' x)'}, \tag{5}$$

where $f(x)$ is the expected value of a randomly obtained proportion of the subpopulation corresponding to the vector $x = (x_1, x_2, \ldots, x_p)'$ where α^0 and $\alpha = (\alpha^1, \ldots, \alpha^p)'$ are the regression coefficients which have to be estimated from the data. We can define the logit function g as follows

$$g(x) = ln\frac{f(x)}{1 - f(x)} = \alpha_0 + \alpha' x. \tag{6}$$

This equation shows a linear relation between the logit function and the explanotary variables. This type of modeling is too restrictive, in fact the behavior of the logit function can be non-linear. The use of splines in this regressive model allows the investigation of non-linear effects with continuous covariates and introduces a nonparametric character. The tuning parameters for regression splines are the number k and the location of knots. In this work, we model the logit function (6) with B-splines with free-knots in order to allow maximum flexibility and improve the fit.

This approach has been used by Denison (9), Lindstrom (14). More precisely, a Markov chain Monte Carlo algorithm is used to estimate a Bayesian version of the B-spline model. Unlike, Dimatteo (13) and Johnson (17) which use a prior on the coefficients β, we estimate these parameters with least-squares estimator. Thus, we avoid the "delicate re-balancing of the coefficients" like mentionned by Dimatteo.

Other approach exists concerning the spline approximation notably P-splines (Eilers and Marx, 1996; Brezger and Lang, 2006). These methods use a relatively large number of knots and to prevent overfitting, a penalty on the second derivative restricts the flexibilty of the fitted curve. In our work, we use the knot location to interpret the results, and in this context the P-splines are not adapted.

With respect to the spline logistic regression model, it is defined by

$$f(x) = \frac{exp(s(x, \beta, r))}{1 + exp(s(x, \beta, r))}. \tag{7}$$

Thus, the logit function (6) can be written as a spline function

$$g(x) = ln\frac{f(x)}{1 - f(x)} = s(x, \beta, r). \tag{8}$$

Let (y^i, x^i), where $i = 1, \ldots, n$, the n observed independent pairs. We approxime the logit of the conditional probability of success by a B-spline model.

Using (8), we obtain:

$$\begin{aligned} logit(P(Y=1\,|\,X_1,\ldots,X_p)) &= ln\frac{P(Y=1\,|\,X_1,\ldots,X_p)}{1-P(Y=1\,|\,X_1,\ldots,X_p)} \\ &= \sum_{j=1}^{p} s_j(x_j,\,\beta^j,\,r^j) \qquad (9) \\ &= \sum_{l=1}^{k_1+d_1+1} \beta_l^1\,B_l^1(x_1,\,r^1)+\cdots+\sum_{l=1}^{k_p+d_p+1}\beta_l^p\,B_l^p(x_p,\,r^p), \end{aligned}$$

where for $j=1,\ldots,p$, $(B_l^j(.,\,r^j))_{l=1,\ldots,k_j+d_j+1}$ is the B-spline matrix, r^j is the knots vector, β_l^j are the spline coefficients, k_j is the fixed number of knots and d_j the fixed degree of the spline function. The associated likelihood is defined by

$$\mathcal{L}(\beta,\,r)=\prod_{i=1}^{n}\left(\frac{exp(\sum_{j=1}^{p}\sum_{l=1}^{k_j+d_j+1}\beta_l^j\,B_l^j(x_j^i,\,r_j))}{1+exp(\sum_{j=1}^{p}\sum_{l=1}^{k_j+d_j+1}\beta_l^j\,B_l^j(x_j^i,\,r_j))}\right)^{y^i} \left(\frac{1}{1+exp(\sum_{j=1}^{p}\sum_{l=1}^{k_j+d_j+1}\beta_l^j\,B_l^j(x_j^i,\,r_j))}\right)^{1-y^i}. \qquad (10)$$

The estimate of the logit function with a linear spline $d=1$ implies easy interpretation. So, we let $d_1=\cdots=d_p=1$. In fact, knots are points where the slope is changing in the shape of the piecewise linear function. So, a quick change of slope can be interpreted as a point separating the variable range into two parts and the knot location corresponds to a threshold value.

From a clinical point of view, knot location represents the threshold value of the risk factor for which the probability of a disease occuring suddenly changes. Moreover, we can define the notion of odds ratio on each interval. In practice, only a small number of threshold values are of clinical interest. A good working model provides one or two threshold values which allow the classification of the patients into two or three groups for differentiation of treatment.

3 BAYESIAN ESTIMATION OF THE LOGIT FUNCTION

This section presents essential background on Reversible Jump MCMC, proposed by Green (4). The adaptation of this algorithm for the spline regression is given.

3.1 RJMCMC

The reversible jump MCMC algorithm allows simulation from target distributions on spaces of varying dimension, it can be considered as a general framework for Metropolis-Hastings algorithms ((7), (8)).

Consider the following hierarchical model: let k be a indicator from a coutable set $\mathcal{K}$ and $\theta^{(k)}$ be the parameter vector. Each k determines a model $\mathcal{M}_k$ defined by $\theta^{(k)}$, with dimension of the parameter space $\Theta^{(k)}$ allowed to vary with k.

The joint distribution of $(k,\, \theta^{(k)},\, y)$, where y is the data vector, is modelled as:

$$p(k,\, \theta^{(k)},\, y) \;=\; p(k)\, p(\theta^{(k)} \mid k)\, p(y \mid k,\, \theta^{(k)}),$$

i.e. the product of model probability, parameter prior and likelihood. Inference about k and $\theta^{(k)}$ will be based on the joint posterior $p(k,\, \theta^{(k)}|y)$, which is known as the *target* distribution. For convenience, we abbreviate $(k,\, \theta^{(k)})$ as z and we note $\pi(dz)$ this *target* distribution. Given k, z lies in $C_k \;=\; \{k\} \times \Theta^{(k)}$, while generally $z \;\in\; C = \bigcup_{k \,\in\, \mathcal{K}} C_k$.

In Markov Chain Monte Carlo computation, an aperiodic and irreductible Markov transition kernel $P(z,\, dz')$ is constructed and it satisfies detailed balance:

$$\int_A \int_B \pi(dz)\, P(z,\, dz') = \int_B \int_A \pi(dz')\, P(z',\, dz), \tag{11}$$

where $A,\, B \in C$. We simulate this chain to obtain a dependent, approximate, sample from $\pi(dz)$.

In our case, we have multiple parameter subspaces $\{C_k\}$ of different dimension. A method that switches between these subspaces is needed. For all that, different types of move between the subspaces can be defined. If the current state is z, a move of type m to state dz' with probability $q_m(z,\, dz')$ is defined and is accepted with probability

$$\alpha_m(z,\, z') = min\{1,\, \frac{\pi(dz')\, q_m(z',\, dz)}{\pi(dz)\, q_m(z,\, dz')}\}. \tag{12}$$

For move z to z', we must generate ramdom numbers u and set z' as a determinist function of z and u : $z' = z'(z,\, u)$. The reverse move from z' to z has to be defined symmetrically by generating random numbers u' and setting $z = z(z',\, u')$. The vectors of Markov chain states and proposal random variables $(z,\, u)$ and $(z',\, u')$ must be of equal dimension, that is, the crucial dimension matching condition:

$$n_1 + n_1' = n_2 + n_2',$$

where $n_1,\, n_2$ are the dimensions of $z,\, z'$ respectively, and $n_1',\, n_2'$ are the dimensions of $u,\, u'$ respectively.

The ratio (12) becomes

$$\begin{aligned} \alpha_m(z,\, z') =& min\{1,\, \text{likelihood ratio}\times \text{ prior ratio } \times \text{ proposal ratio}\} \\ =& min\{1,\, \frac{p(y|z')}{p(y|z)}\, \frac{p(z')}{p(z)}\, \frac{p(k|k')\, q_2(u')}{p(k'|k)\, q_1(u)}\; |\; \frac{\partial(z',\, u')}{\partial(z,\, u)}\; |\}. \end{aligned}$$

where $q_1,\, q_2$ are the distributions of $u,\, u'$ and $|\, \frac{\partial(z',\, u')}{\partial(z,\, u)}\, |$ is the jacobian. Often in practice $n_1 + m_1 = n_2$. Consequently only for the birth step a random u is necessary and we omit in the $\alpha_m(z,\, z')$ the terms $q_2(u')$ and u'.

3.2 RJMCMC for B-spline logistic regression

We have defined in 2.2 the spline logistic regression model with free-knots. We use the RJMCMC algorithm of previous section to select the number and the position of knots to have the best ajustment. For a Bayesian approach, let us

formulate the hierarchical model: we take the number of interior knots k as random, from some countable set $\mathcal{K}$. $\mathcal{M}_k$ denotes the model with exactly k interior knots and $r^{(k)} = (r_1, \ldots, r_k)$ denotes the interior knot locations, with $r_0 = X_{min}$ and $r_{k+1} = X_{max}$ as the boundary knots.

As concerns the vector of spline coefficients $\beta = (\beta_i)_{1 \leq i \leq k+d+1}$ is to be estimed from the data by means of the standard least squares regression theory. A complete Bayesian approach would include these coefficients in the vector of parameters (see Dimatteo (13), Johnson (17)). However, Denison and al. (9) seem shown that the least squares estimation approach leads to no significant deterioration in the performance of the algorithm and avoids an additional computational burden.

We shall generate samples from the joint posterior of $(k, r^{(k)})$. By taking into account the varying dimensionality, we have to develop appropriate reversible jump moves.

For this problem, possible transitions (9) are

1. the addition of a knot (a birth step),
2. the deletion of a knot (a death step),
3. the movement of a knot.

These independent move types are randomly chosen with probability b_k for move k to $k+1$ (i.e. birth step), d_k for move k to $k-1$ (i.e. death step) and η_k for the move step. These probabilities satisfy $b_k + d_k + \eta_k = 1$ for all k.

3.2.1 Prior specifications

Let $k \in \mathcal{K} = \{0, \ldots, k_{max}\}$. We use a truncated Poisson distribution, with parameter λ restricted to the countable set $\mathcal{K}$, to specify the prior for k:

$$p(k) \propto \frac{\lambda^k \, exp(-\lambda)}{k\,!} \, 1_{\{0, \ldots, kmax\}}(k)$$

The r_i are taken to be the order statistics from a uniform random variable with state space the candidate knot locations $\mathcal{R} = \{r_{01}, \ldots, r_{0K}\}$, where $r_{01}, \ldots, r_{0K}$ are disributed equidistantly over the interval $]X_{min}, X_{max}[$, i.e. the knots are equally spaced. Then the prior distribution for $r^{(k)} = (r_1, \ldots, r_k)$ is

$$p(r^{(k)}|k) = \frac{k\,!}{K^k},$$

where K is the number of possible emplacements. As concerns the parameter λ, it could be altered depending on the prior beliefs the researchers may have about the smoothess of the logit function. Small values of λ reflects a stong insistence on smoothness.

3.2.2 Move step

The move step consists in choosing a knot uniformly, say r_j, among the set of moveable knots and proposing this knot to be moved to another position $r'_j \in \mathcal{R}$. A knot $r_j \in \{r_1, \ldots, r_k\}$ is called moveable ((1)), if the number m_j

of vacant candidate knots $r_{0i} \in \mathcal{R}$ with $r_{j-1} < r_{0i} < r_{j+1}$ is at least 1. Let $r = r^{(k)}$. The number $n(r)$ of moveable knots then is defined as

$$n(r) = card\{r_j \mid m_j \geq 1,\, j \in \{1, \ldots, k\}\}.$$

So, firstly, we draw a knot r_j uniformly among $n(r)$ moveable knots with probability $p(r_j) = \frac{1}{n(r)}$ and, given r_j, we draw uniformly r'_j (the new position) among the set of m_j vacant candidate knots, with probability $p(r'_j \mid r_j) = \frac{1}{m_j}$. The corresponding proposal ratio is given by

$$\text{proposal ratio} = \frac{p(r_j|r'_j)\, p(r'_j)}{p(r'_j|r_j)\, p(r_j)} = \frac{n(r)\, m_j}{n(r')\, m'_j} = \frac{n(r)}{n(r')}.$$

The prior ratio is 1 because all collections of the same number of knots have the same prior probability. The acceptance probability for such a move step is

$$\alpha = min\{1, \frac{p(y \mid (k, r'))\, n(r)}{p(y \mid (k, r))\, n(r')}\},$$

where $p(y \mid (k, r'))$ is the spline likelihood.

3.2.3 Changing dimension

Let $z = (k, r^{(k)})$ the current state of parameters. We define $b_0 = d_{k_{max}} = 1$, $b_{k_{max}} = d_0 = 0$ and otherwise $b_k = d_k = 1/3$.

In the birth step, given k, we add a new knot $r'_j \in (r_j, r_{j+1})$. r'_j is drawn uniformly with probability $p(r'_j) = 1/(K-k)$ from the set of the $(K-k)$ vacant candidate knots $r_{0i} \in \mathcal{R}$. We have $z' = (k+1, r')$ where $r' = (r_1, \ldots, r_j, r'_j, r_{j+1}, \ldots, r_k)$. For the birth step, the prior ratio is given by:

$$\begin{aligned}
\text{prior ratio} =& \frac{\text{prior for } k+1 \text{ knots}}{\text{prior for } k \text{ knots}} \frac{\text{prior for location of } k+1 \text{ knots}}{\text{prior for location of } k \text{ knots}} \\
=& \frac{p(k+1)}{p(k)} \frac{p(r'|k+1)}{p(r|k)} \\
=& \frac{p(k+1)}{p(k)} \frac{k+1}{K}.
\end{aligned}$$

The corresponding proposal ratio is given by

$$\begin{aligned}
\text{proposal ratio} &= \frac{d_{k+1}\,(1/k+1)}{b_k\,(1/K-k)} \\
&= \frac{d_{k+1}\,(K-k)}{b_k\,(k+1)}.
\end{aligned}$$

In the death step, the proposed knot to delete is simply chosen uniformly from the knots of the current model, so it is drawn with probability $p(r_{j+1}) = 1/(k+1)$.

The acceptance probability for the birth step is

$$\alpha(z, z') = min\{1, \frac{p(y \mid z')}{p(y \mid z)} \times \text{prior ratio} \times \text{proposal ratio}\}.$$

For the death step, it is the same except that the fraction is inverted. The coefficients β are estimated at each step through the function *glm.fit* available in the R package.

In the multivariate case, we let $k = \sum_{i=1}^{p} k_i$ where k_i is the spline degree s_i and $r^{(k)} = (r^{(k_1)}, \ldots, r^{(k_p)})$ the parameter vector for each spline. The same movements are used that in previous algorithm: addition, deletion or movement of a knot. At each iteration, we choose randomly the spline which we are going to modify. The prior for k is a truncated Poisson distribution with parameter λ. The choice of this parameter reflects, in this context, the parsimony of the model.

4 DATA ANALYSIS

In this section, we illustrate the reversible jump algorithm with two examples: a simulation study and an analysis of a real data set from an in vitro fertilization program.

4.1 Simulation settings

We have simulated 2000 data according to a logistic model defined by:

$$g(x) = ln\frac{f(x)}{1-f(x)} = cos(x/8), \qquad x \in [15,\, 65],$$

where x is genereted from a uniform distribution on [15, 65]. We use the RJMCMC algorithm with splines of degree $d = 1$ and $d = 2$, to estimate this function. Let $\lambda = 1$ be the parameter of Poisson distribution and $k_{max} = 5$, in fact a large number of knots is unlikely to be required. In Figure 1, we display the true function along with corresponding spline model estimates with degree $d = 1,\, 2$. The MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\{\hat{f}(x_i) - f(x_i)\}^2,$$

where f is the true function and $\hat{f}$ is our estimate to the true function. The MSE are : 0.03 for the spline of degree 2 and 0.04 for the linear spline. Thus, we find a slightly lower MSE for the estimate when using a spline of degre 2 instead of linear spline.

4.2 Analysis of FIV data

Many couples resort to in vitro fertilization (IVF), when they have difficulties conceiving children. The principal advantage of IVF is to control follicular growth, ovulation, sperm quality and the early development of fertilized eggs. The study carried out by Roseboom et al. performed a multiple logistic regression analysis in order to evaluate the relationship between various factors and pregnancy. The study led by Demouzon et al. (2) leading even results: the probablity of pregnancy for each cycle is affected by the age of the patients. We want to validate this result by the new method proposed in the previous sections.

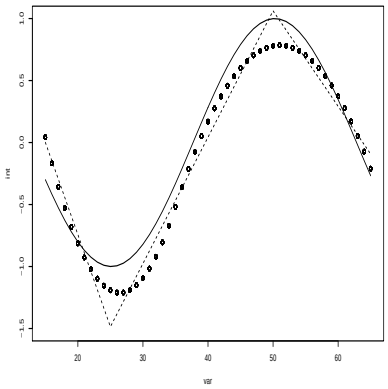

Figure 1: The true curve: -, the estimated logit function by a spline model of degree $d = 1$: $--$ and by a spline model of degree $d = 2$: ...

The french national register of in vitro fertilization (Fivnat) records all of the IVFs carried out in France. This population-based study is a cohort of $23,520$ couples which underwent IVF for the first time between 1994 and 1996. Couples were followed up until they obtained a possible clinical pregnancy or until the 31st of December 1998. A total of 7892 pregnancies were recorded. For each couple, the age of the woman and the age of the man at the first attempt are available. Generally, we take the degree of the spline $d_i > 0$ and the knot number $k_i \geq 0$. However, in epidemiology a smaller number of groups is preferred, so $k_i \in \{0, \ldots, 5\}$, and to allow the interpretation of the results, more precisely to separate the patients in different groups, we let $d_i = 1$.

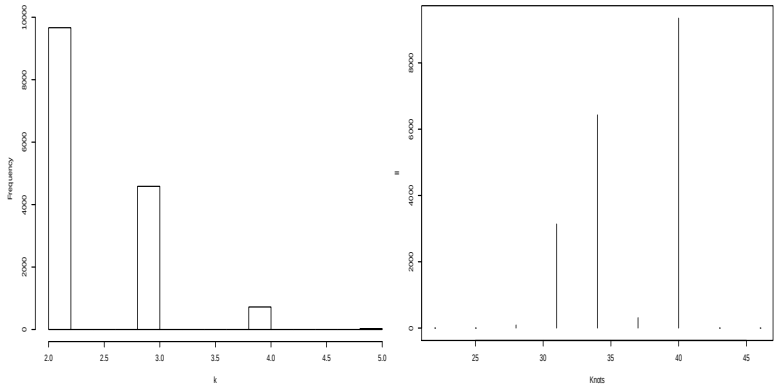

Figure 2: The posterior distribution of k (left) and of r given k (right)

First, we consider the univariate spline model for the age of women. The RJMCMC is used to select the number and location of knots. We let $\lambda = 1$ for the parameter of the prior distribution of k and $k_{max} = 5$. We choose these values because we wish have a smooth function (i.e. with few knots) and few groups of patients.

As concerns the parameter λ, we have tested others values; the results are

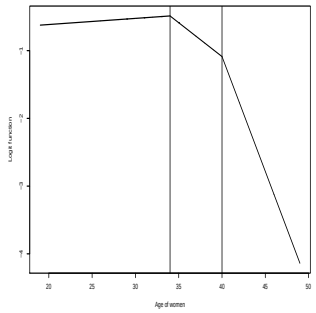

Figure 3: The logit function (i.e. the IVF success rate) according to age of women approximed by a spline model with $k = 2$ and $d = 1$.

the same thus the method seems robust. For the candidate knot location $\mathcal{R}$ the knots are equally spaced of 3 years. The different reversible jump moves have seen in 3.2, the vector β is approximated at each iteration. The estimates are obtained with 20000 iterations and a burn-in time of 5000.

The posterior distribution for k is shown at left in Figure 2, it indicates a mode at $k = 2$. From the right part of this figure, we see the posterior distribution of r given $k = 2$. The knot locations selected are 34 and 40. The figure 3 shows the corresponding logit function estimated by a spline of degree one and with two knots located at age of 34 and 40 years. We have fixed the spline degree at $d = 1$ to be able to interpret the results. From figure 3, the knot locations correspond to break points of the logit function. Indeed, before the first knot, the function seems constant, between the two knots it decreases, and after the second knot, it decreases sharply. Thus, the ages of 34, 40 can be interpreted as threshold values for IVF success. These results are consistent with the results found in previous studies ((11), (2)) using the classical criterion BIC.

Secondly, we model the bivariate spline model for the age of women and men. Let $k_{max} = 10$, $\lambda = 1$ and $d_1 = 1$, $d_2 = 1$. For each variable, we define a candidate knot site where the knots are equally spaced.

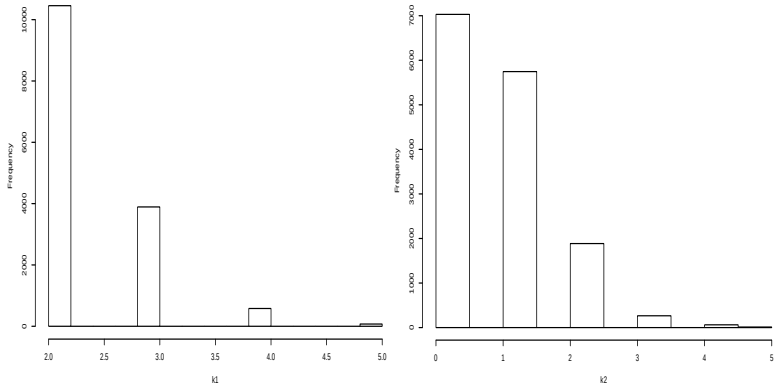

Figure 4: The posterior distribution of k_1 and k_2.

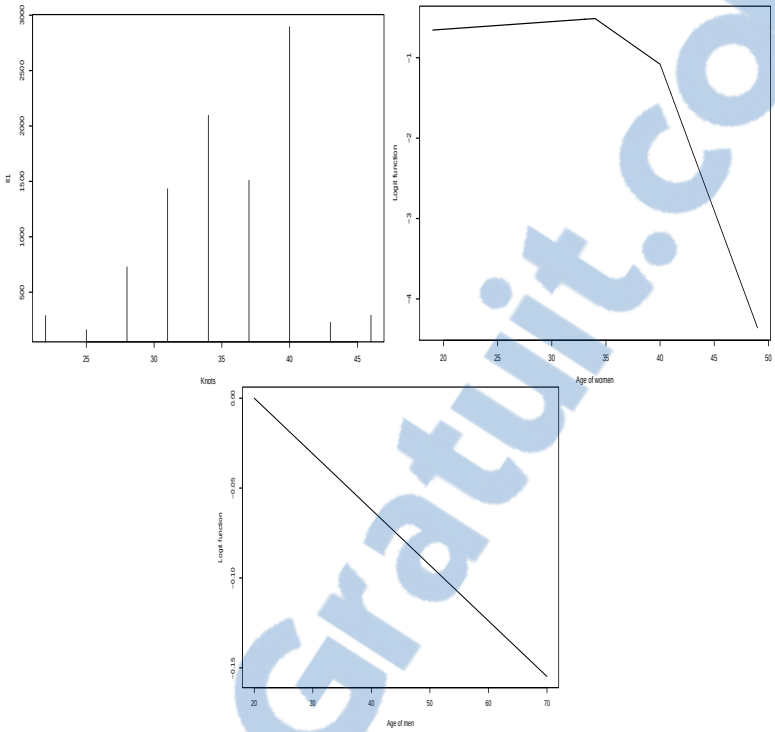

Figure 5: The posterior distribution of r_1 given $k_1 = 2$ (left), the logit function (i.e. the IVF success rate) according to age of women approximed by a spline model with $k_1 = 2$ and $d_1 = 1$ (right) and the logit function according to age of men approximated by a spline model with $k_2 = 0$ and $d_2 = 1$ (down).

Figure 4 shows the posterior distribution of k_1 and k_2. For the age of women, the posterior distribution indicates a mode at 2. Concerning the age of men, we retain any interior knot, the figure 5 shows a linear effect of age of men in IVF success. The left part of the figure 5 illustrates the posterior distribution of r_1 given $k_1 = 2$ (i.e. for the age of women); it indicates two knot locations at 34 and 40 years. These knots are full meaningful and according to the right part of the figure 5: we can assume the ages of 34 and 40 as threshold values for the IVF success. These results are consistent with the previous study using the univariate spline model. These results show the important role played by the age of women in IVF success.

5 DISCUSSION AND FUTURE PLANS

In summary, the use of B-spline to model the logit function helps explain the relationship between response and explanotary variables without imposing a linear link between these variables. In fact, B-spline modeling is more flexible. Furthermore, the linear spline model reconsiders the knots as threshold values. Thus we can classify the patients into groups for treatment differentiation. Finally, the advantage of the RJMCMC algorithm is demonstrated by the direct identification of the number of knot without resorting to model selection criterion such as the BIC or AIC.

References

[1] Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math.* 1946; **4**:45–99;112–141.

[2] Sleeper LA and Harrington DP. Rgression splines in the Cox Model with application to covariate effects in liver disease. *Journal of the American Statistical Association* 1990; **85**:941–949.

[3] Kooperberg C, Stone CJ and Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**:78–94.

[4] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**:711–732.

[5] De Boor C. *A Practical Guide to Splines.* Springer-Verlag: New-York, 1978.

[6] Hastie T, Tibshirani R.. *Generalized Additive Models.* Chapman and Hall: London, 1990.

[7] Hasting WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**:97–109.

[8] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E. Equations of state calculations by fast computing machines. *J. Chem. Phys* 1953; **21**:1087–1091.

[9] Denison DGT, Mallick BK and Smith AFM. Automatic Bayesian curve fitting. *J. R. Statist. Soc. B* 1998; **60**:333–350.

[1] Biller C. Adaptive Bayesian regression splines in semiparametric generalized linear models. *Sonderforschungsbereich* 1998; **51**:4178–4192.

[11] Molinari N, Daurès JP, Durand JF. Regression splines for threshold selection in survival data analysis. *Statistics in Medicine* 2001; **20**:237–247.

[2] Demouzon J, Rossin-Amar B, Bachelot A, Renon C and Devecchi A. Influence du rang de la tentative en FIV, *Contraception, Fertilité, et Sexologie* 1998; **26**:466–472.

[13] DiMatteo, I., Genovese, C.R. and Kass, R.E Bayesian curve-fitting with free-knot splines, *Biometrika* 2001; **88**:1055–1071.

[14] Lindstrom, M.J. Penalized estimation of free-knot splines, *J. Comp. Graph. Statist* 1999; **8**:333–52.

[15] Eubank, R. Spline Smoothing and Nonparametric Regression, *Dekker, New York.* 1988;

[16] Ramsay, J. and Silverman, B. Functionnal Data Analysis, *Springer, New York.* 1997;

[17] Johnson, M.S. Modeling dichotomous item responses with free-knot splines, *Computational statistics and Data Analysis* 2007; **51**:4178-4192.

[18] Li, C.S and Hunt, D. Regression splines for threshold selection with application to a random-effect logistic dose-response model, *Computational statistics and Data Analysis* 2004; **46**:1–9.

[19] Zhou, S. and Shen, X. Spatially adaptive regression splines and accurate knot selection schemes, *Journal of the American Statistical Association* 2001; **96**:247–259.

2.4 Conclusion

Ce deuxième chapitre a permis de répondre à des questions rencontrées en recherche clinique.

En effet, l'objectif était de remettre en question la modélisation linéaire ou log-linéaire entre l'*odds ratio* ou le risque instantané et les covariables et, d'autre part, d'avoir une interprétation simple des résultats. L'idée a donc été de se tourner vers la représentation B-spline qui permet un ajustement flexible sous la forme de polynômes par morceaux. Dans les deux articles qui ont été présentés nous nous sommes limités aux polynômes de degré un, en effet ce choix permet de parler de valeur "seuil" et, ainsi, de pouvoir déterminer des "groupes" de sujets.

Les difficultés rencontrées avec les fonctions B-splines sont le choix du nombre de nœuds et de leur position. De nombreuses méthodes font appel à un critère de choix de modèles pour déterminer le nombre de nœuds. L'algorithme mis en œuvre dans ce deuxième chapitre offre une méthode alternative dans laquelle le nombre et la position des nœuds sont déterminés en même temps. La méthode proposée et implémentée avec le logiciel R est une méthode MCMC qui permet de simuler selon des distributions cibles sur des espaces de dimensions variables.

Les résultats obtenus sur différents jeux de données ont permis une interprétation simple et sont en adéquation avec les études précédentes. Ce chapitre a donc permis de mettre en place une méthode MCMC dans le cadre de modèles couramment utilisés en recherche clinique.

Le chapitre suivant traite également de la représentation B-spline au travers d'une méthode de sélection de modèle.

Chapitre 3

L'heuristique des pentes

Le chapitre précédent propose une approche bayésienne dans le cadre de la régression B-spline permettant de determiner le nombre et de sélectionner la position des nœuds. Ce troisième chapitre expose une approche différente basée sur une méthode de sélection de modèles via une procédure de pénalisation. L'approche considérée a été développée par Birgé et Massart [8], elle est basée sur un mélange de théories et d'idées heuristiques : l'heuristique des pentes. Nous commencerons par une introduction aux méthodes de sélection de modèles via une procédure de pénalisation. Pour étayer nos propos, nous présenterons un article en révision au journal de la Société Française De Statistique (SFDS).

3.1 La sélection de modèles via une procédure de pénalisation

La sélection de modèles via une procédure de pénalisation est une méthode très utilisée depuis plusieurs années. Elle consiste à choisir un modèle minimisant un critère défini comme la somme d'un risque empirique et d'un terme mesurant la complexité du modèle.

Considérons le modèle de régression gaussien suivant où les coordonnées de $Y \in \mathbb{R}^n$ sont définies par :

$$Y_i = s(x_i) + \varepsilon_i, \;\; \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\, \sigma^2), \text{ pour } i = 1, \ldots, n. \tag{3.1}$$

On utilise le contraste des moindres carrés $\gamma(t, (x, y)) = (t(x) - y)^2$, la fonction de contraste empirique associée est définie pour tout $t(.)$ par :

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} (y_i - t(x_i))^2.$$

La fonction de perte correspondante est donnée, pour tout s, t, par

$$l(s, t) = \mathbb{E}_s[\gamma_n(t) - \gamma_n(s)],$$

où $\mathbb{E}_s$ représente l'espérance sous la distribution des y_i.

Soit $\{\mathcal{S}_m,\, m \in \mathcal{M}\}$ une collection de modèles au plus dénombrable et $\{\hat{s}_m,\, m \in \mathcal{M}\}$ les estimateurs des moindres carrés associés. On appelle "modèle" un sous-espace linéaire de $\mathbb{R}^n$, notons $\mathcal{S} = \bigcup_{m\in\mathcal{M}} \mathcal{S}_m$. La qualité d'un modèle sera jugée au travers de son risque quadratique :

$$\mathbb{E}_s[l(s,\hat{s}_m)] = \underbrace{l(s,\overline{s}_m)}_{\text{terme de biais}} + \underbrace{\mathbb{E}_s[l(\overline{s}_m,\hat{s}_m)]}_{\text{terme de variance}},$$

où $\overline{s}_m$ est la projection orthogonale de s sur $\mathcal{S}_m$. Le terme de biais mesure la qualité d'approximation de la réalité que procure le modèle $\mathcal{S}_m$, le terme de variance est proportionnel au nombre de paramètres à estimer.

Un bon modèle doit être le reflet convenable de la réalité tout en restant d'une compléxité raisonnable.

Deux types de méthodes existent : les méthodes *efficaces* et *consistentes*. La première approche se base sur l'idée de l'estimation non-biaisée du risque. Le but est de choisir le modèle minimisant ce risque. Les critères du C_p de Mallows et de l'AIC d'Akaike appartiennent aux méthodes efficaces. A contrario la deuxième approche suppose l'existence d'un vrai modèle de taille minimale, l'objectif est de trouver ce modèle. Le BIC fait partie des méthodes consistentes.

Par la suite plaçons nous dans le cadre des procédures efficaces, le "meilleur" modèle sera celui qui minimise le risque quadratique quand $m \in \mathcal{M}$. L'estimateur associé à ce modèle est noté $\hat{s}_{m*}$ et s'appelle *l'oracle*.

On définit m^* de la manière suivante :

$$m^* = \inf_{m\in\mathcal{M}} \mathbb{E}_s[l(s,\hat{s}_m)] = \inf_{m\in\mathcal{M}} \{l(s,\overline{s}_m) + \mathbb{E}_s[l(\overline{s}_m,\hat{s}_m)]\}. \tag{3.2}$$

Cette expression dépend de s qui est inconnue, il est donc impossible de choisir ce modèle. Le but est de construire une procédure de sélection de modèles qui ne dépend que des observations. Ainsi l'idée va être de choisir $\hat{m}$ tel que le risque de l'estimateur associé $\tilde{s} = \hat{s}_{\hat{m}}$ soit aussi proche que possible du risque de l'oracle :

$$\mathbb{E}_s[l(s,\tilde{s})] \leq C \inf_{m\in\mathcal{M}} \{\mathbb{E}_s[l(s,\hat{s}_m)]\}, \tag{3.3}$$

avec $C > 1$ une constante positive indépendante de s.

L'approche utilisée par la sélection de modèle via une procédure de pénalisation consiste à définir un critère des moindres carrés pénalisé sur $\mathcal{M}$:

$$crit(m) = \gamma_n(\hat{s}_m) + pen(m), \tag{3.4}$$

où $pen : \mathcal{M} \to \mathbb{R}^+$ est une fonction de pénalité. Ce type de méthode remonte au début des années 70 avec le critère du C_p de Mallows et d'Akaike. Le principal problème est de déterminer la fonction de pénalité qu'il convient de choisir. Ces deux critères souffrent du même inconvénient qui est leur nature asymptotique. Quelques améliorations ont été apportées dans ce sens pour l'AIC par Hurvich et Tsai [39]. L'approche qui va suivre adopte un point de vue non-asymptotique permettant de ne faire aucune hypothèse sur l'appartenance ou non de la vraie fonction s à un des modèles.

Dans ce chapitre la méthode utilisée est celle proposée par Birgé et Massart ([8], [9]) reposant sur un mélange de théories et d'idées heuristiques. Elle permet d'estimer la fonction de pénalité à partir des données dans le cadre de la régression gaussienne homoscédastique.

Comme les critères de Mallows [47] ou de Akaike ([1], [2]), le critère de Birgé et Massart repose sur une estimation non biaisée du risque quadratique. Leur objectif est d'obtenir une procédure d'estimation fournissant un critère dicté par les données sélectionnant un estimateur $\tilde{s}$ avec un risque aussi proche que possible du risque oracle. Le concept fondamental pour comprendre et valider cette méthode est le concept de pénalité minimale.

Éclaircissons cette notion. Supposons que la pénalité est proportionnelle à une fonction de la dimension $f(D_m)$ (i.e. $pen(m) = K\, f(D_m)$), Birgé et Massart détermine une constante minimale K_{min} tel que si $K < K_{min}$ le rapport des risques est asymptotiquement grand et est fini si $K > K_{min}$. Ils montrent également que lorsque $K = 2\, K_{min}$ on obtient une procédure de sélection de modèles efficaces. Ils arrivent à la conclusion que la pénalité optimale est égale à deux fois la pénalité minimale.

Cette relation caractérise l'heuristique des "pentes". Arlot et Massart [4] ont récemment developpé, dans le cadre de la régression hétéroscedastique, un algorithme basé sur une généralisation de l'heuristique des pentes. Cet algorithme permet d'estimer K_{min} à partir des données. De nombreux résultats de concentrations sont également donnés. Bien que ces résultats soient prouvés dans le cadre particulier du régressogramme nous supposons qu'ils restent valides au moins dans le cadre de la régression gaussienne.

En résumé, on peut considérer deux approches pour estimer la pénalité minimale à partir des données : soit en utilisant l'algorithme développé par Arlot et Massart et mis en œuvre par Lebarbier [24], soit en estimant K_{min} par la pente de la partie linéaire de la fonction $-\gamma_n(t)$.

L'article présenté dans la section suivante est une application de la méthode de l'heuristique des pentes dans le cadre de la régression spline afin de déterminer le nombre et la position des nœuds.

3.2 L'heuristique des pentes pour la régression spline

Comme expliqué dans l'article suivant le problème du choix du nombre de nœuds dans la régression spline est équivalent à un problème de sélection de modèles. En ce qui concerne la position des nœuds, notre approche considère un ensemble de taille N de nœuds initial $\{m_1, \ldots, m_N\}$ placés aux N-quantiles sur l'ensemble des données $\{x_1, \ldots, x_n\}$. L'ensemble de toutes les combinaisons possibles est testé pour chaque dimension.

Cet article se décompose de la façon suivante : dans un premier temps, un rappel sur la représentation B-spline et les méthodes de sélection de modèles via une procédure de pénalisation est effectué.

Puis, une seconde partie est dédiée à la méthode de l'heuristique des pentes dans le cadre de la régression spline. L'application des résultats de Birgé et

Massart [9] permet d'obtenir une fonction de pénalité dans le cadre de la régression spline et une borne supérieure pour le risque quadratique de l'estimateur correspondant. La fonction de pénalité dépend de deux constantes inaccessibles théoriquement et dépendant de la variance σ^2 qui est inconnue en pratique. Ainsi, plusieurs simulations sont effectuées afin d'estimer ces deux constantes en supposant σ^2 connue.

Puis, dans le cas général où σ^2 est inconnue, nous appliquons deux approches développées par Birgé et Massart [9], Arlot et Massart [4]. Ces deux méthodes reposent sur l'heuristique des pentes, elles sont expliquées et utilisées afin d'estimer la fonction de pénalité à partir des données.

Dans une dernière partie, plusieurs situations sont simulées afin d'accéder à la performance de ces deux méthodes en les comparant aux critères classiques : le BIC et le C_p de Mallows.

Voici l'article qui est en révision au journal de la SFDS.

CHOICE OF KNOT NUMBERS IN SPLINE REGRESSION BY SLOPE HEURISTICS

Marie Denis & Nicolas Molinari

Abstract. — This paper deals with the choice of number and locations of interior knots in the B-spline regression. This type of representation is a very useful tool in medical statistics to detect threshold values. This formulation can be seen as a model selection problem, indeed each knot sequence can be associated with a model. From this point of view, we used a method of model selection via a penalization procedure: the so-called "slope heuristics" introduced by Birgé and Massart. First, a penalty function is proposed in the context of the B-spline regression, some simulation studies are done to estimate the penalty shape. Secondly, two approaches developed by Birgé and Massart (2007) and Arlot and Massart (2008) are applied to estimate the penalty function from the data. Some simulation experiments are performed to assess the performance of these data driven penalties by comparing them to the "classical" penalties.

1. Introduction

In this paper, we deal with the problem of model selection via a penalization procedure in spline regression with an additive gaussian noise. This introduction presents essential background on spline regression and model selection via a penalization procedure. We observe a gaussian vector $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ such that the coordinates are defined by :

$$Y_i = s(x_i) + \varepsilon_i \text{ where } \varepsilon_1, \ldots, \varepsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

with $(x_1, \ldots, x_n)$ n fixed points. The regression function s is a spline function, it represents the fit as a piecewise polynomial. The regions that define the pieces

Mots clefs. — B-splines, Data-driven calibration, Model selection by penalization, Slope heuristics.

are separated by a sequence of interior knots $m_1, \ldots, m_k$. Spline functions are characterized by the spline degree d, the number k and the location of the interior knots ; they belong to a linear functional space of dimension $d+k+1$. The most popular basis function for this linear space is given by Schoenberg's B-splines, or Basic-splines, and is denoted by $\{B_1^d(\cdot, m), \ldots, B_{d+k+1}^d(\cdot, m)\}$ for a fixed sequence of knots $m = (m_1, \ldots, m_k)'$. This structure has the advantage of requiring less computation as compared to other basis functions, such as the truncated power basis. A spline function can be written as

$$s(x, \beta, m) = \sum_{l=1}^{k+d+1} \beta_l B_l^d(x, m), \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_{k+d+1})'$ is the vector of spline coefficients.

In this setting, we can write (1) in the following form :

$$Y_i = \sum_{l=1}^{k+d+1} \beta_l B_l^d(x_i, m) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n. \tag{3}$$

The use of B-splines in medical statistics is very interesting for modelling the logistic model as well as the Cox model, it is more flexible compared to classical models.

Moreover, this representation allows detecting some threshold values necessary for the interpretation of results. This interpretation is possible thanks to the use of a piecewise linear spline where the knot locations can be interpreted as thresholds. Consequently, we use in the sequel the spline degree $d = 1$.

However, the main difficulty with spline regression is the choice of the location $(m_1, \ldots, m_k)$ and the number k of interior knots. Different approaches are used for selecting the number of knots. Martin-Magniette (2005) used the BIC criterion. Several methods consider the knots as free parameters and estimate them by using the maximum likelihood method (Molinari et al (2001)), a cross-validation procedure (Wensin and Zhao (2003)) or in a bayesian context (DiMatteo and al. (2001), Lindstrom (2002), Brezger and Lang (2006), Holmes and Mallick (2003)). In this paper, our aim is to present a model selection procedure via a penalization approach to determine the number and the location of knots.

This approach is an old idea which consists in minimizing the sum of the empirical risk and of some measure of model complexity, called the penalty.

We consider in this work the "slope heuristics" method developed by Birgé and Massart (2007) and a generalization of this method by Arlot and Massart (2008).

The paper is organized as follows. In Section 2, we explain the slope heuristics method and define the penalty shape in the context of spline regression. An upper bound of the quadratic risk for the corresponding penalized estimator is given. Since the penalty function depends on two unknown constants, first we estimate both constants thanks to simulation procedures by assuming a known variance. In Section 3, we applied two approaches, developped by Birgé and Massart (2007) and Arlot and Massart (2008), to estimate the penalty from the data when the variance σ^2 is unknown. In Section 4, we study different simulation experiments to assess the performance of the methods compared to the "classical" penalties. An application on real data is performed to test the prediction performance of this method. Finally, we discuss future plans in Section 5.

2. The slope heuristics for the spline regression

In this section we explain the so-called "slope heuristics" method introduced by Birgé and Massart (2001, 2007) and recently generalized by Arlot and Massart (2008). These methods allow to estimate the penalty function from the data. First, we introduce some functions for the model selection, then we explain the main concept for understanding the method : the minimal penalty.

2.1. Introduction

In order to estimate the unknown function s given by (2), we consider a subset of N knots $\{m_1, \ldots, m_N\}$ defined on the subset $\{x_1, \ldots, x_n\}$. We define by $\mathcal{M}$ the set of all subsets m of $\{m_1, \ldots, m_N\}$. To each subset m of $\mathcal{M}$ corresponds a linear subspace of $\mathbb{R}^n$ defined by

(4)
$$\mathcal{S}_m = \Big\{ \Big(\sum_{l=1}^{|m|+d+1} \beta_l \, B_l^d(x_i, m) \Big)_{1 \le i \le n} \,|\, \beta = (\beta_1, \ldots, \beta_{|m|+d+1}) \in \mathbb{R}^{|m|+d+1} \Big\}.$$

The associated dimension at each subspace is given by $D_m = |m| + d + 1$ where $|m|$ corresponds to the cardinality of the subset of knots m. We call $\mathcal{S}_m$ a model. The estimation of the spline function s can be viewed as a model selection problem among the collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$. We denote $\mathcal{S} = \bigcup_{m \in \mathcal{M}} \mathcal{S}_m$.

Next, we explain the different functions involved in the model selection procedure.

First, the contrast used for the spline regression is the least-squares contrast $\gamma(t,(x,y)) = (t(x)-y)^2$. The associated empirical contrast function is defined for any $t \in \mathcal{S}$ by

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}(y_i - t(x_i))^2 = \|y-t\|_n^2, \tag{5}$$

where $\|\cdot\|_n$ corresponds to the normalized Euclidean norm on $\mathbb{R}^n$. We construct the collection of least-squares estimators $\{\hat{s}_m, m \in \mathcal{M}\}$ associated to the collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$, where

$$\hat{s}_m = \arg\min_{t\in\mathcal{S}_m}\{\gamma_n(t)\}.$$

$\hat{s}_m$ denotes the projection of Y onto $\mathcal{S}_m$. The corresponding loss function is defined, for any $s,t \in \mathcal{S}$, by

$$l(s,t) = \mathbb{E}_s[\gamma_n(t) - \gamma_n(s)] = \|t-s\|_n^2,$$

where $\mathbb{E}_s$ denotes the expectation under the distribution of the y_i.

The quality of a model $\mathcal{S}_m$ (or an estimator $\hat{s}_m$) is given by its corresponding quadratic risk : the loss mean of the least-squares estimator $\hat{s}_m$. It can be decomposed under the following form :

$$\begin{aligned}\mathbb{E}_s[l(s,\hat{s}_m)] = &\quad l(s,\overline{s}_m) + \quad \mathbb{E}_s[l(\overline{s}_m,\hat{s}_m)] \\ = &\quad l(s,\overline{s}_m) + \quad \frac{\sigma^2}{n}D_m.\end{aligned} \tag{6}$$

$l(s,\overline{s}_m)$ is the bias term measuring the quality of approximation of s by $\mathcal{S}_m$, $\mathbb{E}_s[l(\overline{s}_m,\hat{s}_m)]$ the variance term mesuring the estimation error in $\mathcal{S}_m$ and $\overline{s}_m$ is the orthogonal projection of s on $\mathcal{S}_m$.

An ideal model for s is the one with the smallest risk, which satisfies :

$$m^* = \arg\min_{m\in\mathcal{M}}\mathbb{E}_s[l(s,\hat{s}_m)] = \arg\min_{m\in\mathcal{M}}\{l(s,\overline{s}_m) + \mathbb{E}_s[l(\overline{s}_m,\hat{s}_m)]\}. \tag{7}$$

Nevertheless, it is impossible to choose the associated estimator $\hat{s}_{m^*}$, called the oracle, because it depends on the true function s which is unknown.

So, model selection consists in finding a data driven criterion to select an estimator, $\tilde{s}$, such that its risk is as close as possible to the oracle risk :

$$\mathbb{E}_s[l(s,\tilde{s})] \leq C \inf_{m\in\mathcal{M}}\{\mathbb{E}_s[l(s,\hat{s}_m)]\}, \tag{8}$$

for a nonnegative constant $C > 1$ independent of s. This inequality is called the oracle inequality. Furthermore, the aim will be to choose a model leading

to stability between the biais and variance terms. In fact, when the values of D_m increase, the biais term decreases and the variance term increases.

We use a model selection approach via a penalization procedure, which consists in defining a penalized least-squares criterion over $\mathcal{M}$ by

$$crit(m) = \gamma_n(\hat{s}_m) + pen(m), \tag{9}$$

with $pen : \mathcal{M} \to R^+$ being some penalty function. The selected model $\hat{m}$ is a minimizer of the penalized criterion (9) and $\hat{s}_{\hat{m}}$ is the associated estimator.

The final problem is also to choose a convenient penalty function which selects an estimator verifying the oracle inequality (8). Traditionally, penalization procedures are approached from two angles : *efficiency* and *consistency*. Efficiency is based on unbiased risk estimation and aims at choosing a model which minimizes risk. Consistency assumes the existence of a true model of minimal size and aims at finding it. The BIC criterion has been designed for this purpose. For more details about the distinction between these points of view, see the first chapter of McQuarrie and Tsai (1998). Mallows'C_p (Mallows (1973)), FPE and AIC of Akaike (Akaike (1973), Akaike (1974)) are designed for efficient procedures.

These two criteria present serious practical drawbacks. For example, the AIC relies on a strong asymptotic assumption. The optimal calibration of Mallows' C_p requires the knowledge of the noise level σ^2, assumed to be constant. This calibration is difficult because the variance is unknown in practice and must be separately estimated by a plug-in method.

The method proposed by Birgé and Massart (2007) avoids this difficulty : rather than estimating σ^2, they try to estimate the penalty itself, or calibrate it using the data at hand. This method is called the "*slope heuristics*", it is based on a mixture of theoretical and heuristic ideas for defining efficient penalty functions from the data in the Gaussian homoscedastic regression on a fixed design framework. They focus on a non asymptotic point of view meaning that assume we don't that the true s belongs to one of the models. Recently, Arlot and Massart (2008) have generalized these results to the heteroscedastic random design case.

The next subsection applies the slope heuristics to give a general shape of the penalty function in the spline regression.

2.2. Minimal penalties

The existence of a minimal penalty is essential to understand and validate the slope heuristics from the theorical point of view. We note $\hat{m}(K)$ the selected

model by minimizing the empirical risk penalized by $pen(m) = K\, pen_{min}(m)$, with $K > 0$ a positive constant needing calibration.

The term "minimal" means that if we choose $K < 1$, the dimension $D_{\hat{m}(K)}$ of the selected model will be close to the dimension of the largest models. In other words, penalties below this minimal penalty lead to procedures that systematically select models of large dimension. On the other hand, if $K > 1$ the selected model $\hat{m}(K)$ will have a smaller dimension compared to the largest models. A reasonnable choice for K seems to be 2 (or close to two). In fact, it provides an optimal strategy in some cases. Thus, doubling the minimal penalty would lead to an efficient procedure : the penalized estimator satisfies the oracle inequality with a constant C approximately equal to one. We can write :

$$pen_{opt}(m) = 2 * pen_{min}(m). \tag{10}$$

This relationship characterizes the Birgé and Massart's heuristics.

Now, the main issue is to estimate the minimal penalty. Birgé and Massart (2007) prove the existence of a minimal penalty in the Gaussian linear process framework. The following theorem gives a general non-asymptotic risk bound based on this minimal penalty. Firstly, we define a family of nonnegative weights $\{L_m\}_{m\in\mathcal{M}}$ satisfying the condition

$$\Sigma = \sum_{\{m\in\mathcal{M}|\, D_m>0\}} \exp^{-L_m D_m} < +\infty. \tag{11}$$

***Théorème 2.1* (Birgé and Massart (2007)).** — *Given the collection of models $\{\mathcal{S}_m\}_{m\in\mathcal{M}}$, let us consider a family of nonnegative weights $\{L_m\}_{m\in\mathcal{M}}$ satisfying the condition (11).*

Let $\theta \in (0,1)$ and $\kappa > 2-\theta$ and we assume that there exists a finite (possibly empty) subset $\overline{\mathcal{M}}$ of $\mathcal{M}$ such that the penalty function pen satisfies

$$pen(m) \geq Q_m \text{ for } m \in \mathcal{M}\backslash\overline{\mathcal{M}},$$

with

$$Q_m = \varepsilon^2 D_m(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m) \text{ for all } m \in \mathcal{M}.$$

Then the corresponding penalized projection estimator $\tilde{s} = \hat{s}_{\hat{m}}$ exists a.s. and satisfies

$$\begin{aligned}(1-\theta)\mathbb{E}_s[\|s-\tilde{s}\|_n^2] \leq\ & \inf_{m\in\mathcal{M}}\Big\{\|s-\overline{s}_m\|_n^2 + pen(m) - \varepsilon^2 D_m\Big\} + \sup_{m\in\overline{\mathcal{M}}}\{Q_m - pen(m)\},\\ & +\varepsilon^2\Sigma\Big[(2-\theta)^2(k+\theta-2)^{-1} + 2\theta^{-1}\Big].\end{aligned} \tag{12}$$

It follows from this theorem that there exists a minimal penalty of a general form :

$$pen_{min}(m) = \varepsilon^2\, pen_{shape}(m), \tag{13}$$

where the function $pen_{shape} : \mathcal{M} \to \mathbb{R}^+$ can be defined as the optimal shape of the penalty according to D_m. This function depends on the number of models. In the next paragraph, we determine the shape of the minimal penalty in the spline regression context.

2.2.1. The minimal penalty for spline regression

Spline regression belongs to the gaussian linear process framework. With respect to the function pen_{shape} a simple choice is to take as a constant ; this choice is reasonable with a small number of models, for example, if we consider only one model per dimension. Our situation is slightly different, since we have at our disposal a large number of models of same dimension. We denote by $\mathcal{M}_D$ the set of all $m \in \mathcal{M}$ such that $D_m = D + d + 1$, i.e.

$$\mathcal{M}_D = \{m \in \mathcal{M} \,|\, D_m = D + d + 1\}.$$

The cardinality of $\mathcal{M}_D$ is given by the binomial coefficient C_N^D where N is defined in Section 2.1. The following results determine the shape of the penalty for our context. By taking $L_m = 2 + log(\frac{N}{|m|})$ for $m \in \mathcal{M}$, we get

$$\begin{aligned}
\sum_{m\in\mathcal{M}} exp(-L_m\, D_m) &= \sum_{D=1}^{N} |\mathcal{M}_{\mathcal{D}}|\, exp[-(D+d+1)(2+log(N)-log(D)] \\
&\leq \sum_{D=1}^{N} exp[-D-2\,(d+1)-(d+1)log(\frac{N}{D})] \\
&\leq \sum_{D=1}^{N} exp[-D-2\,(d+1)] \\
&\leq e^{-2\,(d+1)}\,(e-1)^{-1} < +\infty,
\end{aligned}$$

as, indeed the cardinality of $\mathcal{M}_{\mathcal{D}}$ satisfies

$$\begin{aligned}
log|\mathcal{M}_{\mathcal{D}}| = log(C_N^D) &\leq log\Big(\frac{N\,e}{D}\Big)^D \\
&= D\Big[log(N) + 1 - log(D)\Big].
\end{aligned}$$

By applying the theorem (2.1) with the weights L_m, we obtain the following proposition :

Proposition 1. — *They exist two positive constants K_1 and K_2 such that if the penalty is defined for all partitions $m \in \mathcal{M}$ by*

$$pen(m) = \varepsilon^2 D_m \left(K_1 log\left(\frac{N}{|m|}\right) + K_2\right), \tag{14}$$

then, the risk associated with the penalized estimator $\tilde{s}$ satisfies

(15)

$$\mathbb{E}_s[\|s - \tilde{s}\|_n^2] \leq C(K_1, K_2)\Big[\inf_{m\in\mathcal{M}} \Big\{\|s - \overline{s}_m\|_n^2 + \varepsilon^2 D_m \left[K_1 log\left(\frac{N}{|m|}\right) + K_2 - 1\right]\Big\} + \varepsilon^2\Big].$$

This result relies on the proposition of Lebarbier (2005). However, the optimal values of both constants are unknown. So, the next step is to estimate the constants K_1, K_2 from simulation studies by assuming the variance σ^2 known. We look for K_1 and K_2 minimizing the following risk ratio uniformly for all functions s and sample size n. We define the ratio risk depending on n by

$$R_n(s, m) = \frac{E_s[\|s - \hat{s}_m\|_n^2]}{\inf_{m\in\mathcal{M}} \mathbb{E}_s[l(s, \hat{s}_m)]}. \tag{16}$$

Note that the penalty (14) depends on the sequence of knots m only through its dimension D_m. Thus, the penalization strategy used gives the same penalty to all models of a given dimension. As a consequence, we define the best estimator for a fixed dimension $D + d + 1$ noted $\hat{s}_D$ by

$$\begin{aligned}\hat{s}_D &= \arg \min_{\{t\in\bigcup_{\{m\in\mathcal{M}|D_m=D+d+1\}} \mathcal{S}_m\}} \gamma_n(t) \\ &= \arg \min_{\{m\in\mathcal{M}|\,|m|=D\}} \gamma_n(\hat{s}_m).\end{aligned}$$

We define the associated natural benchmark considered as a reference of quality

$$\mathcal{O}_{(n)}(s, \mathcal{S}) = \inf_{D\in\mathcal{D}} \mathbb{E}_s[\|s - \hat{s}_D\|_n^2] \text{ with } \mathcal{D} = \{1, \ldots, N\}.$$

Thus, the performance of a penalized least-squares estimator is mesured with the ratio :

$$R_n(s, D) = \frac{E_s[\|s - \hat{s}_D\|_n^2]}{\mathcal{O}_{(n)}(s, \mathcal{S})} \tag{17}$$

The following paragraph gives us an estimation of two unknown constants K_1 and K_2.

2.2.2. The estimates of K_1 and K_2

The simulation schema is the same that Lebarbier (2005). We note $\tilde{s}(K_1, K_2)$ and $E_s[\|s - \tilde{s}(K_1, K_2)\|_n^2]$ respectively the penalized estimator by the penalty (14) and the associated quadratic risk. The main issue is to determine a penalty leads to an oracle inequality : we want to obtain an quadratic risk as close as possible of the oracle risk defined by $\mathcal{O}_{(n)}(s, \mathcal{S})$. The simulation procedure consists in the simulation of several spline functions, the constituted set is noted $\mathcal{L}$ for different values of n belong to a set noted $\mathcal{N}$. We fix $\sigma^2 = 1$. As explained by Lebarbier, we search for suitable K_1 and K_2 for all the functions and minimizing

$$R_n(K_1, K_2) = \sup_{s \in \mathcal{L}} R_n(s, K_1, K_2), \tag{18}$$

with

$$R_n(s, K_1, K_2) = \frac{\mathbb{E}_s[\|s - \tilde{s}(K_1, K_2)\|_n^2]}{\mathcal{O}_{(n)}(s, \mathcal{S})}. \tag{19}$$

We compute the ratio (19) for each s and n over 50 simulations, then we calculate the supremum for each n. We have at disposal a collection of values $\{R_n(K_1, K_2), K_1, K_2 > 0, n \in \mathcal{N}\}$. We plot the functions $K_1 \to R_n(K_1, K_2)$ for $K_2 = 3, 5, 8$. The idea is to determine the optimal value of K_1 satisfying $K_1^* = K_1^*(n, K_2^*)$ for any n with $K_1^*(n, K_2^*)$ the minimizer of $R_n(K_1, K_2)$ with respect to K_1. According to Figure 1 the value $K_2 = 8$ is an optimal constant, indeed this value stabilizes $K_1^*(n, K_2^*)$ with respect to n for the value K_1 approximately equal to 6.

In the next section, different methods are proposed to estimate the penalty from the data when the variance is unknown.

3. Data driven penalties

From the previous results, the penalty for the spline regression is given by

$$pen(m) = 2\,\varepsilon^2\, f(D_m) \;\text{ with }\; f(D_m) = D_m\,(3\,log\big(\frac{N}{|m|}\big) + 4). \tag{20}$$

However, in most cases the variance σ^2 is unknown and we cannot use directly the penalty. We present here two methods based on slope heuristics to evaluate the penalty from the data themselves, and a data driven estimation is obtained.

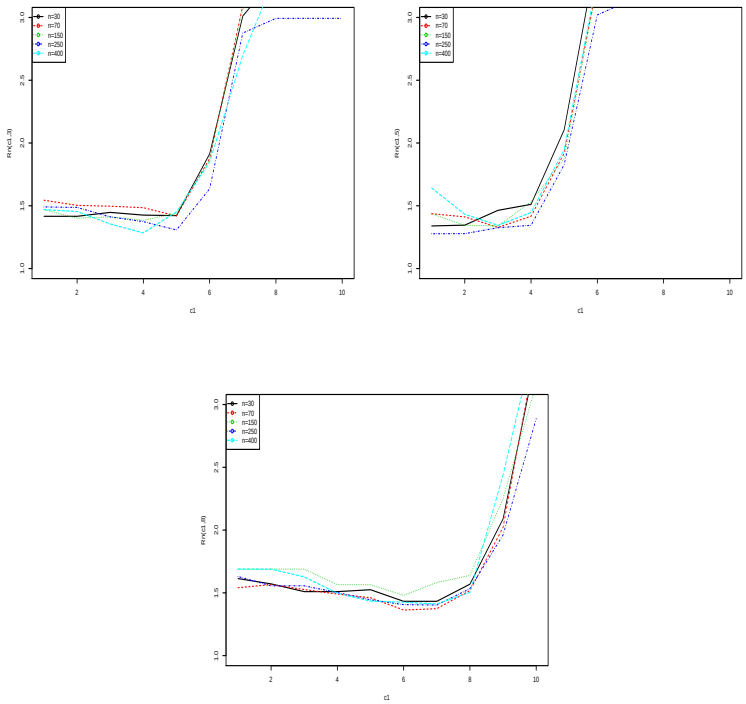

FIGURE 1. Functions $K_1 \to R_n(K_1, 3)$, $K_1 \to R_n(K_1, 5)$ *and* $K_1 \to R_n(K_1, 8)$

3.0.3. First approach

Like the Mallows'C_p or Akaike's AIC, the Birgé and Massart's criterion relies on the idea of unbiased estimation of the quadratic risk :

$$\mathbb{E}_s[l(s, \hat{s}_m)] = l(s, \overline{s}_m) + \mathbb{E}_s[l(\overline{s}_m, \hat{s}_m)].$$

We define for every $m \in \mathcal{M}$:

$$b_m = l(s, \overline{s}_m); V_m = l(\overline{s}_m, \hat{s}_m); \hat{b}_m = \gamma_n(\overline{s}_m) - \gamma_n(s); \hat{V}_m = \gamma_n(\overline{s}_m) - \gamma_n(\hat{s}_m).$$

The model $\hat{m}$ selected by the penalization procedure minimizes the penalized criterion (9). In practice, this comes down to minimizing :

$$\gamma_n(\hat{s}_m) - \gamma_n(s) + pen(m) \quad = \quad \hat{b}_m - \hat{V}_m + pen(m). \tag{21}$$

Arlot and Massart (2008) prove that $\hat{V}_m$ and V_m are concentrated around their expectation in a general framework for $\hat{V}_m$ and in the case of the regressogram for V_m. They also demonstrate that for every $m \in \mathcal{M}$ $\mathbb{E}[V_m] \approx \mathbb{E}[\hat{V}_m]$ in the

particular case of the regressogram. This last result is the main hypothesis of slope heuristics. Thus, we conjecture that these results are applicable at least in the general least squares regression framework on a fixed design and especially for spline regression. Given that in our context $\hat{b}_m$ is an unbiased estimator of the bias term b_m, minimizing (21) is equivalent to minimize

$$b_m - \mathbb{E}[\hat{V}_m] + pen(m).$$

The goal is to make this quantity close to the quadratic risk $\mathbb{E}_s[l(s, \hat{s}_m)]$. An ideal penalty therefore seems to be : $pen(m) = \mathbb{E}[\hat{V}_m] + \mathbb{E}[V_m]$. The different concentration arguments above lead to the penalty :

$$pen(m) \approx 2\hat{V}_m.$$

Furthermore, we can rewrite $\hat{V}_m$ under the following form :

$$\begin{aligned} \hat{V}_m = \gamma_n(\overline{s}_m) - \gamma_n(\hat{s}_m) &= \gamma_n(\overline{s}_m) - \gamma_n(s) + \gamma_n(s) - \gamma_n(\hat{s}_m) \\ &= \hat{b}_m + \gamma_n(s) - \gamma_n(\hat{s}_m). \end{aligned}$$

For large dimensions the bias term $\hat{b}_m$ tends to stabilize itself. Furthermore, by assuming that the penalty is proportional to the function $f(D_m)$, we can estimate the behaviour of $\hat{V}_m$ according to the function $f(D_m)$ via $\gamma_n(\hat{s}_m)$. Let α be the slope of the linear part of $-\gamma_n(\hat{s}_m)$; the final penalty is given by :

$$pen(m) = 2\alpha f(D_m).$$

In practice, we use the robust regression (Huber, 1981) to estimate the slope of the linear part of the function $-\gamma_n(\hat{s}_m)$. In the context of spline regression, the bias term seems to become constant quickly. In the next paragraph, we apply a calibration algorithm defined in the least squares regression framework for general-shape penalties and proposed by Arlot and Massart (2008).

3.0.4. A data driven calibration algorithm

The algorithm proposed by Arlot and Massart (2008) relies mainly on slope heuristics and is inspired by the practical procedure explained by Birgé and Massart (2007). We explain this algorithm in the case of spline regression by letting $pen_K(D) = Kf(D)$. The goal is to determine some unknown constant $K_{min} > 0$ such that $2K_{min}f(D)$ is approximately optimal. Denoting by $crit_K(D) = \gamma_n(\hat{s}_D) + Kf(D)$ the penalized criterion, we define $\hat{D}_K$ the associated dimension such that

$$\hat{D}(K) \in \arg \min_{D \in \{1, \ldots, N\}} \{\gamma_n(\hat{s}_D) + Kf(D)\}.$$

The following algorithm describes the different steps in order to obtain an optimal calibration of the penalty.

Algorithm 1. —

1. *Compute* $\hat{D}(K)$ *for* $K > 0$
2. *Take* $\hat{K}_{min} > 0$ *such that for* $K < \hat{K}_{min}$ *the associated dimension* $\hat{D}(K)$ *is very large and for* $K > \hat{K}_{min}$ *the dimension is much smaller.*
3. *Select the model of associated dimension* $\hat{D}(K) = \hat{D}(2\hat{K}_{min})$.

The second algorithm permits computation of $\hat{K}_{min}$, which is clearly considered as a dimension jump. We use the notations of Arlot and Massart (2008).

Algorithm 2. —

1. *Init :* $K_0 = 0, D_0 = \arg\min_{D\in\{1,\dots,N\}}\{\gamma_n(\hat{s}_D)\}$,
2. *Step* $i, i \geq 1$ *: Let*

$$G(D_{i-1}) = \{d \in \{1, \dots, N\} \text{ s.t. } \gamma_n(\hat{s}_d) > \gamma_n(\hat{s}_{D_{i-1}}) \text{ and } f(d) < f(D_{i-1})\}.$$

If $G(D_{i-1}) = \varnothing$, *then put* $K_i = +\infty$, $i_{max} = i - 1$ *and stop. Otherwise,*

$$K_i = \inf\left\{\frac{\gamma_n(\hat{s}_d) - \gamma_n(\hat{s}_{D_{i-1}})}{f(D_{i-1}) - f(d)} \text{ s.t. } d \in G(D_{i-1})\right\}$$

and

$$D_i = \min_i F_i \text{ with } F_i = \arg\min_{d\in G(D_{i-1})}\left\{\frac{\gamma_n(\hat{s}_d) - \gamma_n(\hat{s}_{D_{i-1}})}{f(D_{i-1}) - f(d)}\right\}$$

Arlot and Massart (2008) proved that $K \to \hat{D}(K)$ is piecewise constant and non-increasing. Thus, the trajectory of $(\hat{D}(K))_{K\geq 0}$ is characterized by the number of jumps noted i_{max}. We have $i_{max} \in \{0, \dots, N-1\}$. The position of the jump i is noted K_i. The set of positions is denoted by an increasing sequence of positive reals $(K_i)_{0\leq i\leq i_{max}+1}$ verifying the following limit conditions : $K_0 = 0$ and $K_{i_{max}+1} = +\infty$. The sequence of selected models is defined by $(D_i)_{0\leq i\leq i_{max}}$ with for all $i \in \{0, \dots, i_{max}\}$and for all $K \in [K_i, K_{i+1}), \hat{D}(K) = D_i$. We assume that the passage to the minimal penalty to be marked by a sudden fall of the dimension, which corresponds to the "dimension jump". Two definitions for $\hat{K}_{min}$ are given by Arlot and Massart (2008). The first definition uses a threshlod value $D_{threshold}$ considered as the largest "reasonably small" dimension ; in short, models with overly large dimensions can not be selected. Choosing $D_{threshold} \propto n/(ln(n))$ or $n/(ln(n))^2$ seems logical. We therefore define :

$$\hat{K}_{min} = inf\{K > 0 \text{ s.t. } \hat{D}(K) \leq D_{threshold}\}.$$

The following alternative also seems logical :

$$\hat{K}_{min} = K_{i_{jump}} \text{ with } i_{jump} = \arg \max_{i \in \{0,\dots,i_{max}-1\}} \{D_{i+1} - D_i\}.$$

When there is one clear jump, both definitions give the same value of $\hat{K}_{min}$, or at least the same selected model. A problematic case occurs for distant values of K, where several jumps are observed. It would be useful to add models with large dimensions to have a clear jump or to perform a mixture of both definitions. In our experiments, we deal with the second definition of K_{min}, and most of the time we observe a clear dimension jump.

We present an application of the method estimating the slope of the linear part of the function $-\gamma_n(\hat{s}_D)$. In the sequel we call this method : method 1, and the method using the calibration algorithm is called the method 2. Three functions are considered to observe the behaviour of the function $-\gamma_n(\hat{s}_D)$ according to $f(D)$. Let $\sigma^2 = 1$, $d = 1$ and $N = 10$, we consider three configurations :

- $n = 450$ and $s = (s_1)$,
- $n = 60$ and $s = (s_2)$,
- $n = 250$ and $s = (s_3)$.

We plot the functions s_1, s_2, s_3 in Figure 1. The graphics of the associated function $-\gamma_n(\hat{s}_D)$ according to $f(D)$ are represented by the Figure 2. For any n, $-\gamma_n(\hat{s}_D)$ reaches quickly linearity with respect to the function $f(D)$, thus we can estimate the slope of the linear part for N small. It is therefore not necessary, with regards to spline regression, to have many models in order to apply the slope heuristic.

4. Applications

In this section, some simulation experiments are performed to compare the performance of the previous methods with the "basic" criteria : the C_p Mallows and the BIC criterion for which asymptotic properties have been established. With regards to the penalties provided by the slope heuristics, we define :

- $pen_{\sigma^2}(D) = 2\,\varepsilon^2\,f(D)$ with $f(D) = (D + d + 1)\,(3\,log(\frac{N}{D}) + 4)$,
- $pen_1(D) = 2\,\alpha\,f(D)$ with α the slope of the linear part of the function $-\gamma_n(\hat{s}_D)$,
- $pen_2(D) = 2\,K_{min}\,f(D)$ with K_{min} the constant obtained by the data driven calibration algorithm.

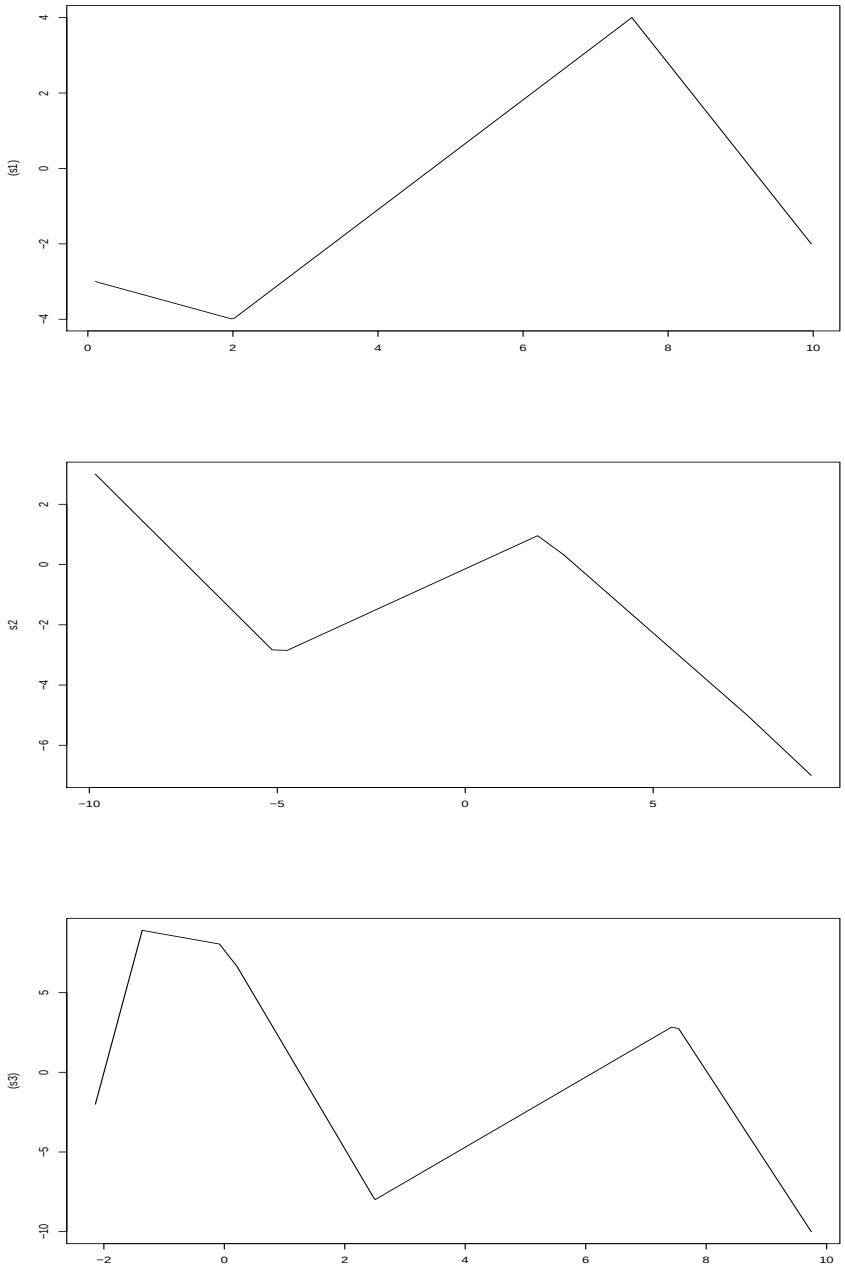

FIGURE 2. Three functions : (s_1), (s_2), (s_3)

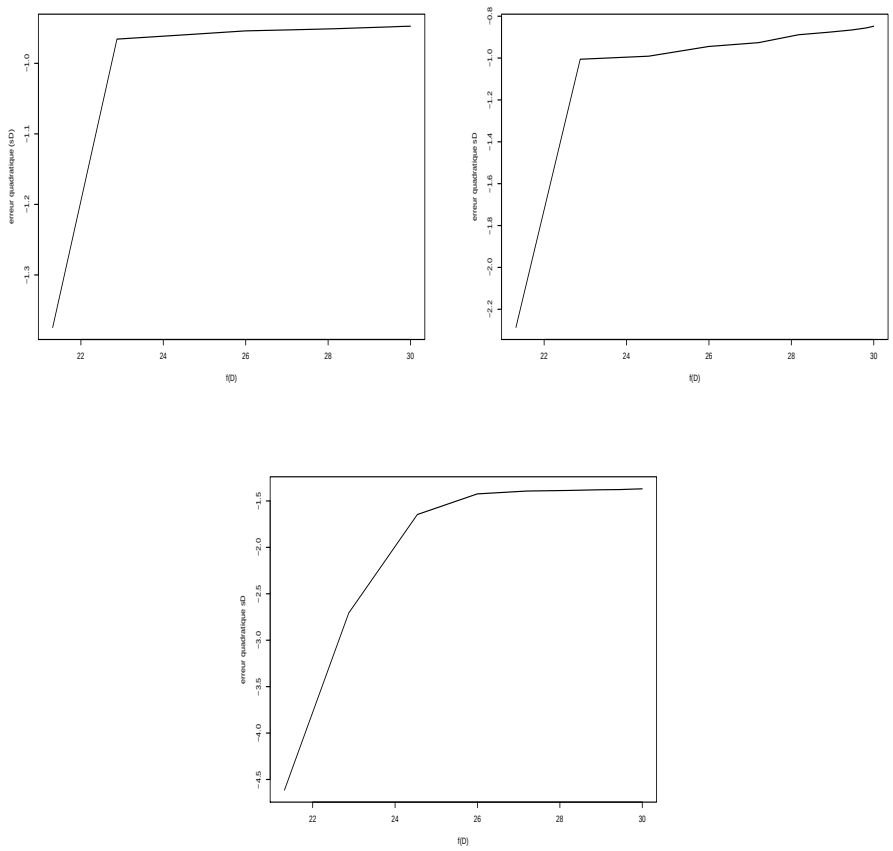

FIGURE 3. Graphics of $(f(D), -\gamma_n(\hat{s}_D))$ for $D = 1, \ldots, 10$ for three functions s_1, s_2, s_3.

The penalties pen_1 and pen_2 correspond to methods 1 and 2. In the sequel, we call the method associated to the first penalty the "theoretic" method and the "theoretic estimated" method with an estimated variance.

4.1. Introduction

Firstly, we consider a function s and two realizations for $n = 60$ and $n = 300$ noted respectively $s(1)$ and $s(2)$. These three functions are plotted in Figure 3. Table 1 gives the dimension selected using the method 1, the method 2, the

theoretic method with the variance known and the estimated variance (noted respectively T and TE).

		Method 1	Method 2	T	TC	$\inf_{D\geq 1} \|s - \hat{s}_D\|_n^2$
$s(1)$	$\hat{D}$	3	2	3	6	3
	$\|s - \hat{s}_D\|_n^2$	0.077	0.094	0.077	0.099	0.077
$s(2)$	$\hat{D}$	2	2	3	4	2
	$\|s - \hat{s}_D\|_n^2$	0.023	0.023	0.0284	0.043	0.023

TABLE 1. Dimensions selected by the differents penalties and the associated loss.

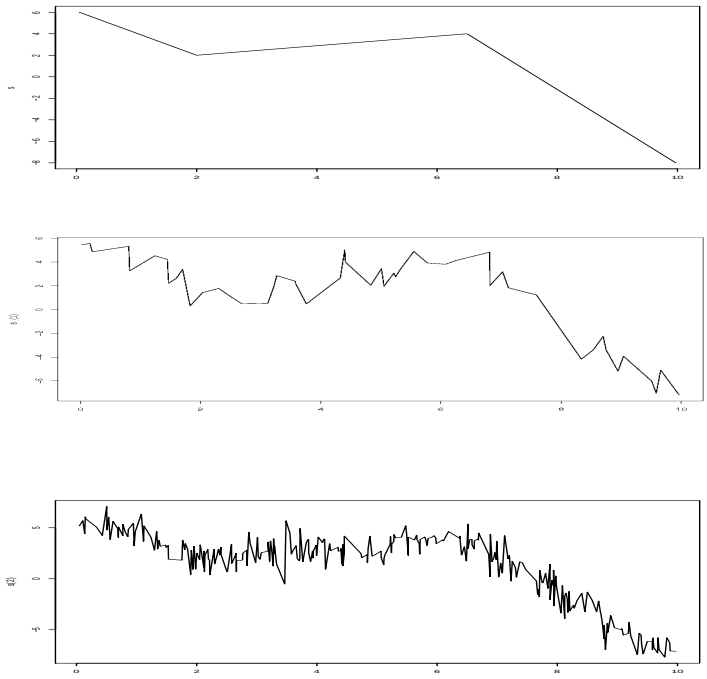

FIGURE 4. The function s and two realizations for $n = 60$ and $n = 300$.

Their associated loss function $\|s - \hat{s}_D\|_n^2$ is given.

In Figure 4, are plotted the estimators selected by each method (if two methods give same results, we plot only one graph). From these results, method 1 works well, indeed it selects the oracle dimension. The theoretic penalty with

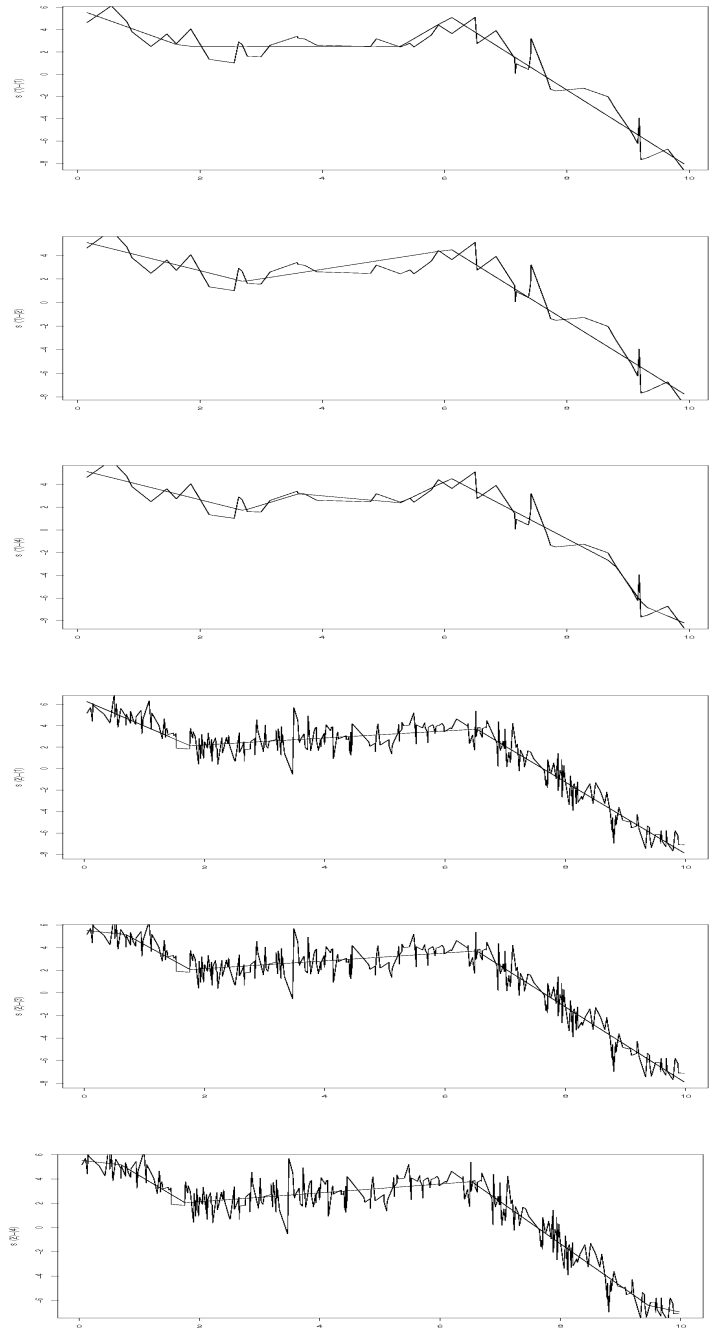

FIGURE 5. The estimators selected by each method for both realizations ((1)=Method 1, (2)=Method 2, (3)=Theorical method, (4)=Theoretical estimated method).

estimated variance seems to select a larger dimension than other methods. Finally, four methods provide a good estimation according to the Figures.

4.2. Simulated datasets

The study considers four spline functions of degre 1 (numbered from $s_1 - s_4$) with the variance $\sigma^2 = 1$ plotted in Figure 5.

At each function is associated a subset of knots $\{m_1, \ldots, m_N\}$ defined by the N-quantiles on the range of data $(x_1, \ldots, x_n)$. N is defined according to data, in our simulations the choice of $N = 10$ provides good results as seen in some examples presented at the end of Section 3.

Both methods for estimating the minimal penalty are applied. Futhermore, the theoretic method is used with known and unknown variance.

The main issue is to compare the performance of methods based on the slope heuristics with the "classical" penalties defined by

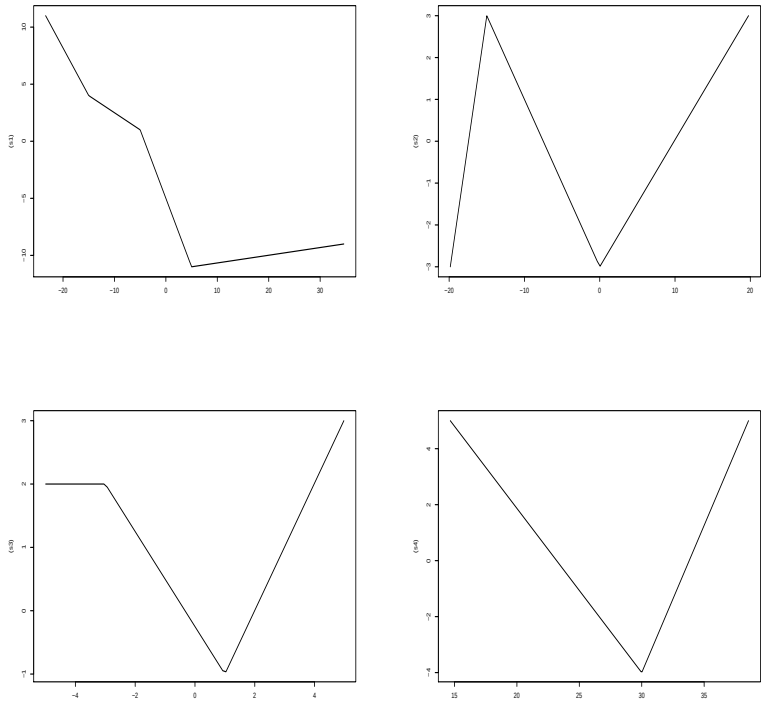

FIGURE 6. The functions $(s_1 - s_4)$.

	$n = 60$					
	Method 1	Method 2	Theoric	Theoric E.	BIC	C_p
s_1	1.03	1.04	1.06	1.17	1.05	1.19
$\%p_{min}$	0.36	0.32	0.35	0.19	0.34	0.18
s_2	1.12	1.18	1.2	1.2	1.14	1.18
$\%p_{min}$	0.37	0.24	0.36	0.25	0.43	0.22
s_3	1.12	1.16	1.04	1.26	1.06	1.35
$\%p_{min}$	0.37	0.30	0.45	0.29	0.44	0.18
s_4	1.79	1.73	1.19	1.97	1.38	2.16
$\%p_{min}$	0.26	0.32	0.76	0.28	0.63	0.16
	$n = 300$					
	Method 1	Method 2	Theoric	Theoric E.	BIC	C_p
s_1	1.02	1.04	1.04	1.04	1.05	1.04
$\%p_{min}$	0.39	0.18	0.39	0.36	0.34	0.23
s_2	1.09	1.08	1.03	1.02	1.18	1.01
$\%p_{min}$	0.21	0.21	0.20	0.22	0.20	0.21
s_3	1.11	1.22	1.23	1.14	1.89	1.20
$\%p_{min}$	0.32	0.27	0.34	0.36	0.05	0.31
s_4	1.85	1.79	1.23	2.04	1.35	2.24
$\%p_{min}$	0.26	0.32	0.76	0.28	0.68	0.16

TABLE 2. For each function s_i and each n, estimation of the risk ratio of the penalized estimator obtained by each criterion and percentage of the number of times that the considered criterion leads to the minimal loss estimator over the 150 simulations.

- Mallows'C_p : $pen_{C_p}(D) = 2\,\sigma^2\,\frac{(D+d+1)}{n}$.

- BIC : $pen_{BIC}(D) = log(n)\,(D + d + 1)$,

With regards to the Mallows'C_p, we independently estimate the variance σ^2.

We have to compare the different criteria for several functions.

We consider $n = 60, 300$ for each function, the risk ratio (17) defined in section (2.2.1) is estimated over 150 simulations and $\%p_{min}$, the percentage of the number of times that the considered criterion leads to the minimal loss estimator over 150 simulations is calculated. Table 2 illustrates the different results.

In Figure 8, the distribution of $\hat{D}$ (the selected dimension) is represented for the different criteria and for each function. The value of the oracle D_{oracle} is noted, as well as the real dimension D_{true}. The column "Meth1" corresponds to the criterion using the slope estimation, the column "Meth2" corresponds to the criterion obtained by the data driven calibration algorithm, the column "Theo" corresponds to the criterion defined by the theoric penalty with the variance known and the column "Theo-E" corresponds to the criterion defined by the theoric penalty with the variance unknown.

Thus, by assuming that the penalty is proportional to the dimension $f(D)$, the slope of the linear part provides an efficient penalty.

From these results, we can say :

- The criterion defined by the method 1 performs better in terms of quadratic risk than Mallows' C_p. With respect to the BIC criterion, the results are mixed : when the true dimension is equal to the oracle's dimension, the BIC criterion seems to be the best.

- The theoretic penalty with known variance can be worser than other methods based on slope heuristics. Indeed, the constants K_1 and K_2 have been chosen to be optimal in most situations and can be suboptimal for specific n and s values. Moreover, the theoretic penalty with unknown variance tends to select a larger dimension, like seen in the previous paragraph.

- However, the BIC criterion seems to select a dimension close to the true dimension, whereas both criteria defined by the slope heuristics tend to select a dimension close to the oracle dimension. This difference can be explained by the different aims of the criteria. The slope heuristics's goal is to select the minimal risk estimator while the BIC criterion provides a consistent procedure.

- Method 2 works worser than the method 1 and, sometimes, worser than the Mallows'C_p, perhaps it is due to the number N.

5. Discussions

We have applied in this paper the so-called "slope heuristics" for spline regression to estimate the correct number of interior knots. Two methods have been used : the data driven calibration algorithm developped by Arlot and Massart (2008) and slope estimation for empirical risk. This last method is efficient when we observe a linear behaviour of the empirical risk. In our experiments, such linear behaviour often appears quickly. Thus, the criterion based on the slope estimation provides good results. However, the BIC criterion seems be the best when the true and oracle dimensions are the same.

With respect to the shape of the optimal penalty, it would be interesting to improve the estimation procedure like Lebarbier (2005), although the proposed estimation here provides good results. Another problem arises from the practical problems, in fact when the linear part in the method 1 is not observed or when there is lots of jumps in the method 2, a calibration method could be considered. One perspective is to adopt calibration procedures to improve the method.

Références

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor,* 1971*)*, 267–281. Akadémiai Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19,** 716–723.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least squares regression. *Journal of Machine Learning Research* 245–279.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc* **3,** 203–268.

Birgé, L. and Massart, P. (2001). A generalized C_p criterion for Gaussian model selection, Technical report No 647 , Publication Université Paris-VI 639–50.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab Theory Related Fields* **138(1-2),** 33–73.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Comput. Stat. Data Anal.* **50,** 967–991

DiMatteo, I., Genovese, C.R. and Kass, R.E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika.* **8,** 1055-1071.

Holmes, C. C. and Mallick, B. K. (2003). Generalized nonlinear modeling with multivariate free-knot regression splines. *Journal of the American Statistical Association.***98(462),** 352–368.

Lebarbier, émilie. (2005). Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.* **85,** 717–736.

Lindstrom, Mary J. (2002). Bayesian estimation of free-knot splines using reversible jumps. *Comput. Stat. Data Anal.* **41,** 255–269.

McQuarrie, A.D.R. and Tsai, C.L.(1998). Regression and Time Series Model Selection. *World Scientic.* Singapore.

Mallows, C. (1973). Some comments on C_p. *Technometric.* **37,** 362–372.

Martin-Magniette, M. L. (2005). Nonparametric estimation of the hazard function by using a model selection method : estimation of cancer deaths in Hiroshima atomic bomb survivors. *Applied statistics.* **54(2),** 317–331.

Maugis, C. and Michel, B. (2008). Slope heuristics for variable selection and clustering via Gaussian mixtures. Rapport de recherche.

Molinari, N., Daurès, J.P. and Durand, J.F. (2001). Regression splines for threshold selection in survival data analysis. *Statistics in Medicine.* **20,** 237-247.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika.* **68,** 45–54.

Wenxin, M. and Zhao, L.H. Free-knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society. Series B, Statistical methodology.* **65,** 901–919

Marie Denis, IURC, 641 avenue du doyen Gaston GIRAUD, 34093, Montpellier, France.
E-mail : marie.denis@inserm.fr

Nicolas Molinari, Hôpital Carremeau, CHU Nîmes, Place du Pr. R. Debré, 30029 Nîmes cedex 9, France.

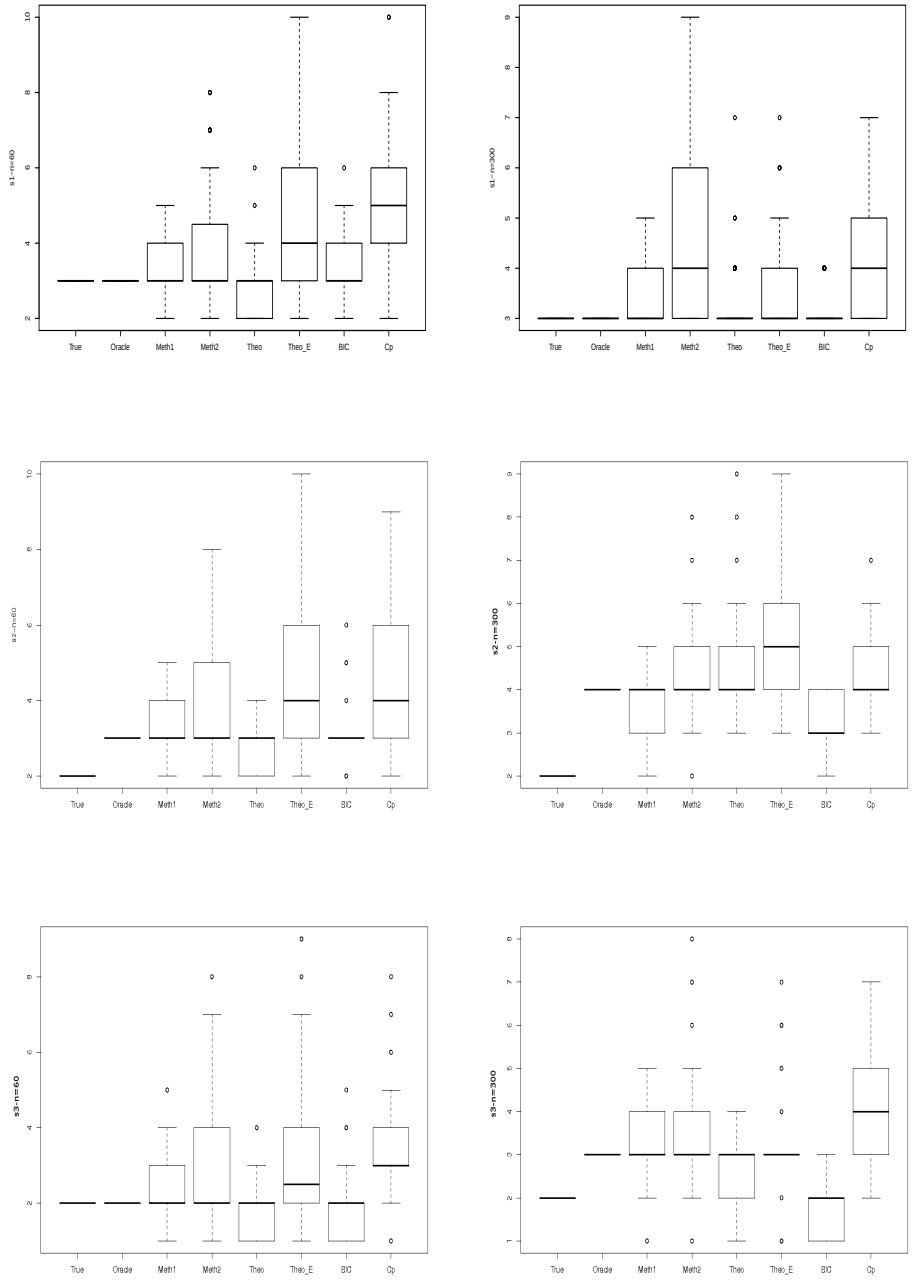

FIGURE 7. Distribution of $\hat{D}$ for the functions $(s_i)_{i=1,\dots,4}$ for each penalized criterion

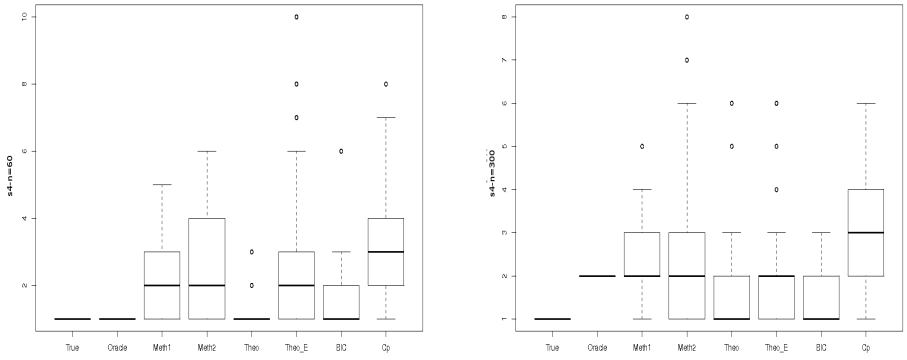

FIGURE 8. Distribution of $\hat{D}$ for the functions $(s_i)_{i=1,\dots,4}$ for each penalized criterion

3.3 Conclusion

Ce troisième chapitre a permis de développer à partir des données un critère des moindres carrés pénalisé afin de déterminer le nombre et la position des nœuds dans le cadre de la régression spline. L'approche non asymptotique de Birgé et Massart (2001, 2007) a été utilisée et a permis d'obtenir une méthode donnant, dans certaines situations, des résultats aussi bons voir meilleurs que le C_p de Mallows et le critère BIC.

Cependant les résultats obtenus ne sont pas extrément ceux que l'on attendait. En effet, on observe une variance importante des estimations, de plus le risque quadratique associé à ces différentes méthodes n'est pas systématiquement optimal.

Cet article constitue la révision d'un premier article dans lequel nous avions d'autres hypothèses plus restrictives. Expliquons-les.

Tout d'abord, la principale différence concerne le choix du nombre de modèles par dimension : dans le premier article il y avait un seul modèle par dimension. En effet, pour une dimension donnée on associait une séquence de nœuds fixée. Dans cette révision plusieurs modèles sont associés à une même dimension. On a donc un ensemble de séquence de nœuds pour une seule dimension.

Nous avons effectué ce changement afin d'avoir plusieurs choix pour la séquence de nœuds. Le choix fait à l'origine conduisait à une pénalité dont la forme était égale à D_m. L'estimation de cette pénalité s'effectuait par l'intermédiaire des deux méthodes expliquées dans l'article précédent et qui reposent sur l'heuristique des pentes. Les résultat obtenus sont bons, en voici quelques exemples.

Le tableau 3.1 donne les résultats concernant le rapport des risques (risque de l'estimateur sélectionné sur le risque de l'oracle) pour quelques fonctions.

On observe, d'une part, une amélioration de l'estimation lorsque n augmente, d'autre part les résultats donnés par les méthodes basées sur l'heuristique des pentes donnent de meilleurs résultats que le BIC ou le C_p. De plus, le C_p de Mallows est particulièrement mauvais lorsque n est petit. Notons également que, comme pour les résultats de la nouvelle version de l'article, lorsque la dimension de l'oracle est égale à la vraie dimension, le BIC tend à être meilleur.

La figure 3.1 représente les boxplot de la fonction s_1 pour $n = 100$ et $n = 500$ ainsi que la représentation de la fonction s_1 et $-\gamma_n(s_1)$ en fonction de D_m. La colonne Slope correspond à la première méthode et la colonne Algo à la deuxième méthode. On observe clairement la partie linéaire de la fonction $-\gamma_n(s_1)$.

Ainsi ces différents résultats présentent une première aproche simple qui fonctionne bien.

La nouvelle approche proposée dans l'article, bien que présentant de bons résultats en ce qui concerne la méthode estimant la pente, pose quelques difficultés et quelques questions.

Tout d'abord la nouvelle forme de la pénalité dépend de deux constantes inconnues, leur estimation s'est effectuée par simulation. Cette méthode peut conduire à des choix non optimaux pour K_1 et K_2. Certaines améliorations sont à envisager afin d'avoir des estimations plus pertinentes.

TAB. 3.1 – Estimation du rapport des risques pour chaque critère sur 500 simulations.

	$n = 100$			
	BIC	Methode 1	Methode 2	C_p
s_1	1.55	1.14	1.14	3.47
s_2	1.40	1.29	1.27	2.31
s_3	1.55	1.17	1.17	1.69
	$n = 500$			
	BIC	Methode 1	Methode 2	C_p
s_1	1.36	1.10	1.09	1.85
s_2	1.46	1.11	1.12	1.85
s_3	1.87	1.15	1.14	1.23

D'autre part il se peut que dans certaines situations l'estimation de la pente de la partie linéaire soit difficile à obtenir. De même, la détection du "bon" saut peut s'avérer problématique. Il serait donc intéressant de développer une méthode de calibration fonctionnant dans toutes les situations possibles. Cette méthode pourrait améliorer la deuxième approche reposant sur l'algorithme d'Arlot et Massart. En effet, cette méthode fonctionne moins bien que celle basée sur l'estimation de la pente.

Ces résultats obtenus avec plusieurs modèles par dimension ne sont pas aussi pertinents que ceux de l'approche considérant un seul modèle par dimension. Dans la première version ces deux approches reposant sur l'heuristiques des pentes menaient aux même résultats. Cette différence est peut-être due au choix de N ou aux choix des différents estimateurs $s_{\hat{D}}$.

Cependant, les différents résultats trouvés ainsi que les différentes applications réalisées dans cet article ont permis d'accéder à la performance des méthodes proposées dans diverses situations. De plus, la forme de la pénalité associée à une telle situation a été déterminée.

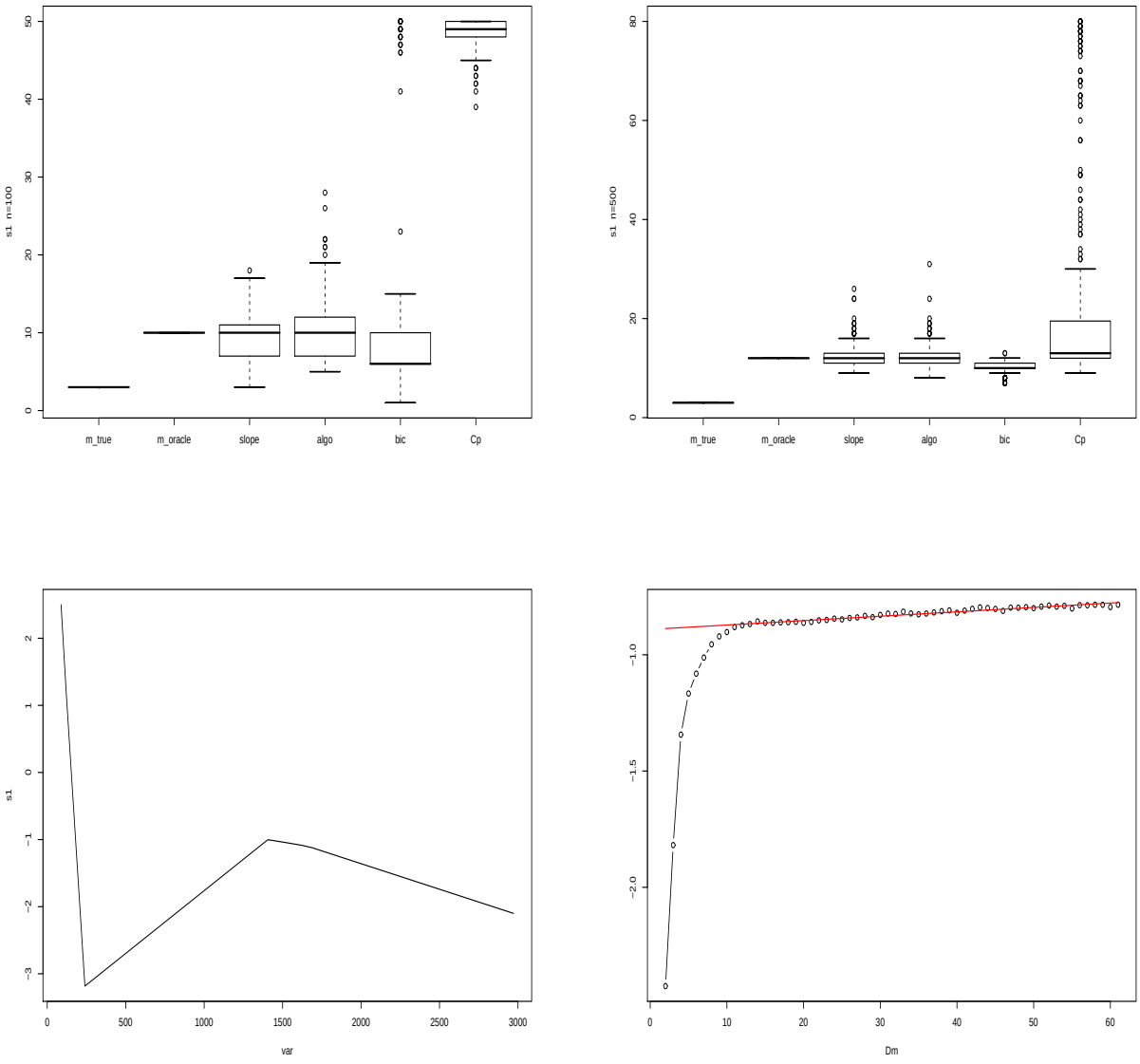

FIG. 3.1 – Les boxplot de la fonction s_1 pour $n = 100$ et $n = 500$ (en haut) et la représentation de la fonction s_1 et de $-\gamma_n(s_1)$ en fonction de D_m (en bas).

Chapitre 4

La modélisation bayésienne non paramétrique

Nous avons besoin en biostatistique de nouveaux outils afin de modéliser des données de grandes dimensions ou de dimensions infinies. L'utilisation de modèles bayésiens semi-paramétriques ou non paramétriques semble appropriée et offre une alternative face aux méthodes non paramétriques "classiques" qui ne sont pas performantes pour ce type de problème. Ceci est notamment dû au problème de la "malédiction" de la dimension. Les modèles non paramétriques bayésiens incorporent des paramètres de dimension infinie au travers de priors centrés sur un modèle de base paramétrique. Ces méthodes fournissent plus de flexibilité notamment pour les modèles avec beaucoup de paramètres et ont une analyse plus robuste que les inférences paramétriques classiques et bayésiennes. De nombreux problèmes d'inférence statistique sont traités par l'inférence bayésienne non paramétrique : l'estimation de densité, la régression, les modèles hiérarchiques, la validation de modèles, ...

C'est à partir des travaux fondateurs de Ferguson ([27], [28]) sur le processus de Dirichlet que le domaine de la statistique bayésienne non paramétrique a vraiment pris de l'ampleur. Ces processus définissent une distribution sur l'ensemble des distributions de probabilité. De nombreux travaux suivirent notamment sur les modèles de mélange de processus de Dirichlet (DPM). Ces modèles forment une classe de modèles bayésiens non paramétriques très importante. Plusieurs approches ont été développées afin de les estimer. Escobar [25] fût l'un des premier à mettre en œuvre une méthode d'inférence basée sur l'échantillonneur de Gibbs. Une représentation du processus de Dirichlet a été mis en évidence par Sethuraman [65] : la représentation *stick breaking*.

Nous nous intéressons dans ce chapitre au problème de la classification à travers l'estimation de la densité. Une première partie sera consacrée à la définition des principaux concepts utilisés en statistique bayésienne non paramétrique, la seconde partie mettra en avant les modèles de mélanges de processus de Dirichlet pour la classification. Enfin, on developpera un modèle non paramétrique hiérarchique pour un problème donné.

4.1 Introduction

Nous présentons dans cette première partie les définitions et les concepts utilisés en statistique bayésienne non paramétrique.

Considérons un ensemble de données $x = (x_1, \ldots, x_n)$ avec $x_i \in \mathcal{X}$ distribuées indépendamment selon une loi de probabilité inconnue F de densité f par rapport à la mesure de Lebesgue :

$$x_i \sim f(x_i), \quad i = 1, \ldots, n. \tag{4.1}$$

Dans la suite de notre exposé nous allons modéliser la densité d'où sont issues les données par un mélange de densités de probabilité. On s'intéresse donc à la classe des densités de probabilité pouvant s'écrire sous la forme du modèle de mélange suivant :

$$f(x) = \int_{\Theta} f(x|\theta)\, dG(\theta), \tag{4.2}$$

où θ est une variable latente, $f(.|\theta)$ une densité mélangée connue et $G(.)$ une distribution de mélange inconnue.

Dans un contexte bayésien l'objectif est d'estimer la loi de probabilité F à partir des données $x = (x_1, \ldots, x_n)$. Nous sommes donc amenés à définir une distribution a priori pour G notée $P(G)$. G est appelée mesure de probabilité aléatoire et appartient à un espace fonctionnel $\mathcal{G}$ de dimension infinie. Le modèle (4.2) peut se réécrire sous la forme hiérarchique suivante :

$$\begin{aligned} x_i|\theta_i &\overset{ind}{\sim} f(.|\theta_i) \\ \theta_i|G &\overset{i.i.d}{\sim} G \\ G &\sim P(G). \end{aligned} \tag{4.3}$$

On a ainsi reformulé un problème d'estimation de mélange de densités (4.2) sous la forme d'un modèle bayésien non paramétrique hiérarchique. Notons que dans le cas paramétrique la loi de probabilité G est caractérisée par un paramètre de dimension finie qui est notre paramètre inconnu. Ce type de modélisation contraint la densité de probabilité à prendre une certaine forme paramétrique ce qui peut limiter l'inférence. Les modèles non paramétriques fournissent un support plus large.

Le prior le plus couramment utilisé pour G est le processus de Dirichlet. Le modèle obtenu est appelé modèle de mélange de processus de Dirichlet, noté DPM. Nous étudierons ce type de modèle dans la section 1.2. En ce qui concerne le processus de Dirichlet, il est expliqué dans la section suivante.

4.1.1 La distribution de Dirichlet

Le processus de Dirichlet est défini à partir de la distribution de Dirichlet. La plupart des propriétés du processus de Dirichlet (conjugaison, formulation en urne de Pólya) sont analogues aux propriétés de la distribution de Dirichlet. Commençons par définir la distribution de Dirichlet.

Soit S_{K-1} le simplexe de $\mathbb{R}^{K-1}$ défini par

$$S_{K-1} = \{p = (p_1, \ldots, p_{K-1}) \in \mathbb{R}^{K-1} \, : \, p_i \geq 0 \text{ pour } i = 1, 2, \ldots, K-1, \sum_{i=1}^{K-1} p_i \leq 1\}.$$

La distribution de Dirichlet est définie de la façon suivante :

Définition 4.1.1. ***La distribution de Dirichlet*** *est une distribution sur le simplexe S_{K-1} caractérisée par la densité de $p = (p_1, \ldots, p_{K-1})$ par rapport à la mesure de Lebesgue dans $\mathbb{R}^{K-1}$ vérifiant :*

$$f(p) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \Big(\prod_{j=1}^{K-1} p_j^{\alpha_j - 1} \Big) \Big(1 - \sum_{j=1}^{K-1} p_j \Big)^{\alpha_K - 1} \mathbb{I}_{\{p \in S_{K-1}\}} \tag{4.4}$$

où $\alpha = (\alpha_1, \ldots, \alpha_K)$ est un jeu de paramètres avec $\alpha_j > 0$, avec Γ la fonction Gamma définie par :

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha - 1} \, e^{-x} \, dx, \quad \alpha > 0.$$

Dans la suite on notera la distribution de Dirichlet de paramètre α par $\mathcal{D}ir(\alpha)$ ou $\mathcal{D}ir(\alpha_1, \ldots, \alpha_K)$. La distribution de Dirichlet est une généralisation de la loi Beta, on a égalité de ces deux lois de probabilité si $K = 2$. Comme pour la distribution Beta, la distribution de Dirichlet admet une représentation utile en terme de variables Gamma.

Proposition 4.1.2. *Si $Y_1, \ldots, Y_K$ sont des variables aléatoires Gamma indépendantes de paramètres α_j et 1 avec $\alpha_j \geq 0$ pour tout j et $\sum_{j=1}^{K} \alpha_j > 0$, alors*

(i) le vecteur

$$\Big(\frac{Y_1}{\sum_{j=1}^{K} Y_j}, \ldots, \frac{Y_K}{\sum_{j=1}^{K} Y_j} \Big)$$

est distribué selon une loi de Dirichlet $\mathcal{D}ir(\alpha_1, \ldots, \alpha_K)$;

(ii)

$$\Big(\frac{Y_1}{\sum_{j=1}^{K} Y_j}, \ldots, \frac{Y_K}{\sum_{j=1}^{K} Y_j} \Big)$$

est indépendant de $\sum_{j=1}^{K} Y_j$,

(iii) Si $p = (p_1, \ldots, p_K) \sim Dir(\alpha_1, \ldots, \alpha_K)$, alors pour toute partition $B_1, \ldots, B_m$ de $\mathcal{X}$, le vecteur

$$(P(B_1), \ldots, P(B_m)) = (\sum_{j \in B_1} p_j, \ldots, \sum_{j \in B_m} p_j) \sim Dir(\alpha'_1, \ldots, \alpha'_K),$$

où $\alpha'_i = \sum_{j \in B_i} \alpha_j$.

En particulier, la distribution marginale de p_j est une loi Beta de paramètres $(\alpha_j, \sum_{j \neq i} \alpha_i)$.

Remarque 4.1.3. *On peut donc interpréter le paramètre α comme une mesure sur $\mathcal{X}$ en posant $\alpha(\{j\}) = \alpha_j$, ainsi $\alpha(\mathcal{X}) = \sum_{j=1}^{K} \alpha_j$.*

Pour obtenir un échantillon selon une distribution de Dirichlet $\mathcal{D}ir(\alpha)$, il suffit de tirer K variables indépendantes $(y_1, \ldots, y_K)$ de loi Gamma $\mathcal{G}(\alpha_j, 1)$. Le vecteur $x = (x_1, \ldots, x_K)$ contruit de la façon suivante

$$x_i = \frac{y_i}{\sum_{j=1}^{k} y_j}, \quad i = 1, \ldots, k.$$

sera donc une réalisation de la loi (4.4).

Un modèle intéressant est le modèle appelé "multinomial-Dirichlet". Commençons par définir la notion de loi multinomiale.

Soient n variables aléatoires $(X_1, \ldots, X_n)$ i.i.d. à valeurs dans $\mathcal{X}$ distribuées selon la loi de probabilité P caractérisée par le vecteur $p = (p_1, \ldots, p_{K-1})$. Le vecteur p est un élèment du simplexe S_{K-1} et vérifie $\mathbb{P}(X_i = j) = p_j$. La vraisemblance associée à l'échantillon $x = (x_1, \ldots, x_n)$ est donnée par

$$l(x|p) = \prod_{j=1}^{K} p_j^{n_j}, \tag{4.5}$$

où n_j correspond au nombre d'observations égales à j, i.e. $n_j = \sum_{i=1}^{n} \mathbb{I}_{\{X_i = j\}}$. Ainsi la statistique $\mathbf{n} = (n_1, \ldots, n_K)$ est dite multinomiale de paramètre p et sa fonction de masse s'écrit :

$$\frac{n!}{n_1! \ldots n_K!} \prod_{j=1}^{K} p_j^{n_j}.$$

Ce qui nous amène à la proposition suivante.

Proposition 4.1.4. *La distribution de Dirichlet est conjuguée à la distribution multinomiale. Soit $\boldsymbol{n}|p \sim Multinomiale(.|p)$ et $p \sim \mathcal{D}ir(\alpha)$ alors la loi a posteriori de $p|\boldsymbol{n}$ est encore une ditribution de Dirichlet de paramètre α' où $\alpha' = \alpha + \boldsymbol{n}$. La loi prédictive a priori de $\boldsymbol{n}$ est une loi Multinomiale-Dirichlet de paramètre α de densité*

$$\frac{n!}{(\sum_{j=1}^{K} \alpha_k)^{[n]}} \prod_{j=1}^{K} \frac{\alpha_j^{[n_j]}}{n_j!}$$

où $\alpha^{[s]} = \alpha\,(\alpha + 1)\,\ldots\,(\alpha + s - 1) = \frac{\Gamma(\alpha+s)}{\Gamma(\alpha)}$

Le concept d'urne de Pólya va nous servir pour la représentation des lois marginales de variables distribuées selon un processus de Dirichlet. Ce qui suit donne une idée du concept de l'urne de Pólya ainsi qu'une relation entre l'urne de Pólya et la distribution de Dirichlet.

Soit une urne avec $\alpha(\mathcal{X})$ boules et α_j boules de couleur j, $j = i, \ldots, K$. On suppose $\alpha_j \geq 0$ et $\sum_{j=1}^{K} \alpha_j > 0$. On tire succesivement n boules, à chaque tirage on remplace la boule tirée par deux boules de la même couleur, c'est- à-dire que l'on replace la boule et on ajoute une boule de la même couleur.

Posons $X_i = j$ si la i ième boule est de couleur j. Ainsi

$$\mathbb{P}(X_1 = j) = \frac{\alpha_j}{\alpha(\mathcal{X})}$$

$$\mathbb{P}(X_2 = j|X_1) = \frac{\alpha_j + \delta_{X_1}(j)}{\alpha(\mathcal{X}) + 1}.$$

Ce qui donne de manière générale :

$$\mathbb{P}(X_n = j|X_1, \ldots, X_{n-1}) = \frac{\alpha_j + \sum_{i=1}^{n-1} \delta_{X_i}(j)}{\alpha(\mathcal{X}) + (n-1)}.$$

La distribution jointe de $X_1, \ldots, X_n$ est donnée par

$$\begin{aligned}\mathbb{P}(X_{=}x_1, \ldots, X_n = x_n) &= \mathbb{P}(X_1 = x_1)\,\mathbb{P}(X_2 = x_2|X_1 = x_1) \ldots \\ &\quad \mathbb{P}(X_n = x_n|X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \\ &= \frac{\alpha_{x_1}}{\alpha(\mathcal{X})} \frac{\alpha_{x_2} + \delta_{x_1}(x_2)}{\alpha(\mathcal{X}) + 1} \cdots \frac{\alpha_{x_n} + \sum_{i=1}^{n-1} \delta_{x_i}(x_n)}{\alpha(\mathcal{X}) + n - 1} \\ &= \frac{\alpha_1^{[n_1]} \ldots \alpha_K^{[n_K]}}{\alpha(\mathcal{X})^{[n]}}\end{aligned}$$

avec la convention que $\alpha_j^{[0]} = 1$. La suite $(X_1, \ldots, X_n)$ est appelée une suite de Pólya de paramètre α.

Notons que l'on retrouve la densité de la loi prédictive a priori, ce qui permet d'obtenir la proposition suivante :

Proposition 4.1.5. *Soit α une mesure non nulle sur $\mathcal{X}$. Soit $X = (X_1, \ldots, X_n)$ defini par*

$$\begin{aligned} X_i|P &\sim P, \text{ i.i.d.} \\ P &\sim \mathcal{D}ir(\alpha)\end{aligned}$$

et soit $X' = (X'_1, \ldots, X'_n)$ une suite de Pólya de paramètre α. Alors X et X' ont la même loi.

On a donc obtenu une relation entre la distribution de Dirichlet et la représentation en urne de Pólya. Cette relation va nous servir dans le cadre des processus de Dirichlet.

4.1.2 Le processus de Dirichlet

Soit $(\mathcal{X}, \mathcal{A})$ un espace mesurable. Notons $\mathcal{P}$ l'ensemble de toutes les lois de probabilité sur $(\mathcal{X}, \mathcal{A})$.

Le processus de Dirichlet est défini comme une mesure de probabilité sur l'espace des mesures de probabilités.

Définition 4.1.6 (Processus de Dirichlet)**.** *Soient $(\mathcal{X}, \mathcal{A})$ un espace mesurable, P_0 une mesure de probabilité sur cet espace et α_0 un nombre réel positif. Une mesure de probabilité G est distribuée selon* ***un processus de Dirichlet*** *de paramètres G_0 et α_0 si pour toutes partitions finies $(B_l)_{l=1,\ldots,r}$ de $\mathcal{X}$, la loi de*

$(G(B_1), \ldots, G(B_r))$ est une loi de Dirichlet $\mathcal{D}(\alpha_0 G_0(B_1), \ldots, \alpha_0 G_0(B_r))$.

On note ceci par

$$G \sim DP(\alpha_0 G_0)$$

Ce processus est donc défini par deux paramètres : α_0 qui correspond à un paramètre de concentration (ou paramètre d'échelle) et G_0 qui est une mesure de probabilité de base.

Proposition 4.1.7. *Si $G \sim DP(\alpha_0 G_0)$ alors $G(B)$ suit une loi Beta de paramètres $\alpha_0 G_0(B)$ et $\alpha_0 (1 - G_0(B))$. Ce qui implique en particulier que*

$$\mathbb{E}[G(B)] = G_0(B)$$

pour tout ensemble mesurable B de $\mathcal{X}$.

De même :

$$\mathbb{V}ar(G(B)) = \frac{G_0(B)\,(1 - G_0(B))}{\alpha_0 + 1},$$

on montre donc que α_0 est un paramètre d'échelle du processus.

Démonstration. On considère la partition (B, B^c), $G(B)$ a une distribution $\mathcal{B}eta(\alpha_0 G_0(A), \alpha_0 (1 - G_0(A)))$. Ainsi la preuve est immédiate. □

Le théorème suivant est très important. Il justifie le choix des processus de Dirichlet dans un modèle bayésien non paramétrique en montrant la simplicité de la mise à jour de la distribution a posteriori.

Théorème 4.1.8. *Si G est a priori distribuée suivant un processus de Dirichlet $DP(\alpha_0 G_0)$ et si $\theta = (\theta_1, \ldots, \theta_n)$ est un échantillon i.i.d. de loi G alors la loi a posteriori de G est un processus de Dirichlet $DP(\alpha_0' G_0')$ tel que :*

$$\alpha_0' = \alpha_0 + n, \;\; G_0' = \frac{\alpha_0}{n + \alpha_0} G_0 + \frac{n}{n + \alpha_0} G_n,$$

où $G_n = (1/n) \sum \delta_{\theta_i}$ est la loi empirique de l'échantillon.

Démonstration. En ce qui concerne la preuve, se référer au livre de Droesbeke, Fine et Saporta [23]. □

Ce théorème permet d'observer deux choses importantes :

(*i*) La distribution de Dirichlet constitue une famille fermée pour l'inférence dans un échantillonnage i.i.d. non paramétrique.

(*ii*) On remarque que si α_0 est petit, la loi a posteriori devient le processus de Dirichlet $DP(n G_n)$. Ce processus ne dépend que de l'échantillon et il est centré sur G_n. Ainsi, dans le cas où $\alpha_0 = 0$, le processus de Dirichlet $DP(\alpha_0 G_0)$ fournirait une distribution a priori non informative. Le paramètre α_0 représente donc le degré de la connaissance a priori et G_0 reflète la connaissance a priori sous la forme de la distribution.

Le processus de Dirichlet possède une propriété importante qu'il faut prendre en compte : une réalisation G d'un processus de Dirichlet est presque sûrement discrète [3]. Sethurman [65] a établi la représentation stick-breaking des réalisations d'un processus de dirichlet. On peut donc écrire :

$$G(.) = \sum_{k=1}^{\infty} w_k \delta_{\xi_k}(.), \tag{4.6}$$

avec les atomes $\xi_k \overset{i.i.d.}{\sim} G_0$, les poids $w_k = v_k \prod_{j=1}^{k-1}(1-v_j)$ et $v_k \overset{i.i.d}{\sim} \mathcal{B}eta(1, \alpha_0)$.

De plus, les suites $(\xi_k)_{k\geq 1}$ et $(v_k)_{k\geq 1}$ sont indépendantes.

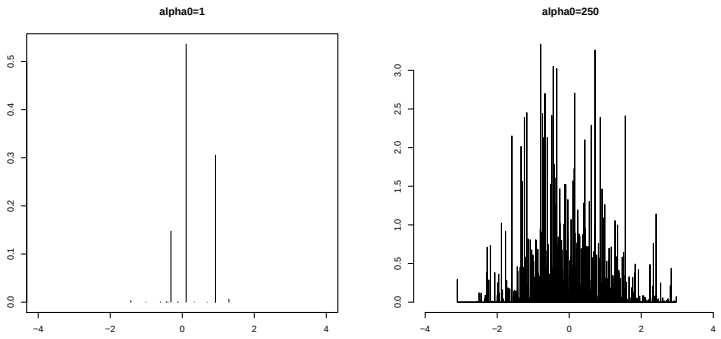

FIG. 4.1 – Représentation de réalisations du processus de Dirichlet $G \sim DP(G_0, \alpha_0)$ de distribution de base $G_0 = \mathcal{N}(0, 1)$ pour $\alpha_0 = 1$ à gauche et $\alpha_0 = 250$ à droite.

La réalisation d'un processus de Dirichlet peut donc être vue comme un mélange infini dénombrable de mesures de Dirac. Les localisations ξ_k de ces fonctions de Dirac sont notées sous le nom de "cluster" en anglais et "classes" en français.

Remarquons que si α_0 tend vers 0, les poids w_k tendent très rapidement vers 0. En effet pour des valeurs de α_0 proche de 0, $v_1 \approx 1$ et seul les premiers atomes ont un poids non négligeable. A contrario, pour de larges valeurs de α_0, la distribution de G ressemble à celle de G_0 et tous les atomes ont un poids non négligeable.

Afin d'illustrer ces quelques propriétés, nous représentons sur la Figure 4.1 des réalisations de processus de Dirichlet $G \sim DP(\alpha_0 G_0)$ avec $G_0 = \mathcal{N}(0, 1)$ et $\alpha_0 = 1$, $\alpha_0 = 250$.

D'autres processus a priori admettent également une représentation stick-breaking, comme par exemple :

– le processus de beta à deux paramètres [42],

– le processus Pitman-Yor, aussi appelé processus de Poisson-Dirichlet à deux paramètres [55].

Cependant comme nous l'avons précisé précedemment, l'utilisation du processus de Dirichlet comme mesure de probabilité aléatoire a priori est priviligiée. En effet la mise à jour de la distribution a posteriori est simple (théorème 4.1.8).

De plus, Blackwell et MacQueen [10] ont démontré que la distribution prédictive, obtenue en marginalisant selon la mesure de probabilité aléatoire G, admet une représentation en urne de Pólya. On peut donc décomposer la loi marginale de $(\theta_1, \ldots, \theta_n)$ de la façon suivante :

$$\theta_i|\theta_1, \ldots, \theta_{i-1} \sim \frac{\alpha_0}{i-1+\alpha_0} G_0 + \frac{1}{i-1+\alpha_0} \sum_{j=1}^{i-1} \delta_{\theta_j}, \tag{4.7}$$

avec δ_θ la mesure de Dirac au point θ.

Ainsi, conditionnellement aux valeurs des variables $\theta_1, \ldots, \theta_{i-1}$ déjà échantillonnées la probabilité que le nouvel échantillon soit identique à un échantillon précédent est $\frac{1}{i-1+\alpha_0}$ et la probabilité que le nouvel échantillon soit distribué indépendamment selon la distribution de base G_0 est $\frac{\alpha_0}{i-1+\alpha_0}$. Plusieurs variables peuvent avoir la même valeur, elles sont donc associées à la même "classe" ξ_k. Le nombre de valeurs distinctes de θ_i notée N est, par conséquent, inférieur ou égal à n.

On observe un effet de classification, si l'on note n_k le nombre de variables θ_i dont la valeur est égale à ξ_k, la probabilité que le nouvel échantillon soit égal à cette valeur sera $\frac{n_k}{\alpha_0+n_k}$. Le paramètre α_0 règle le nombre de valeurs distinctes N. Ainsi pour n grand, Antoniak [3] a montré que $\mathbb{E}[N|\alpha_0, n] \simeq \alpha_0 \, log(1+\frac{n}{\alpha_0})$.

Ce qui montre que lorsque α_0 tend vers 0, la plupart des variables θ_i partagent la même valeur. Inversement quand α_0 tend vers l'infini, les θ_i sont comme des échantillons i.i.d. de la distribution de base G_0. Le paramètre d'échelle α_0 joue un rôle essentiel dans la distribution de θ.

L'illustration de cet effet de classification se fait au travers du processus du restaurant chinois. C'est un processus aléatoire dans lequel nous avons n clients

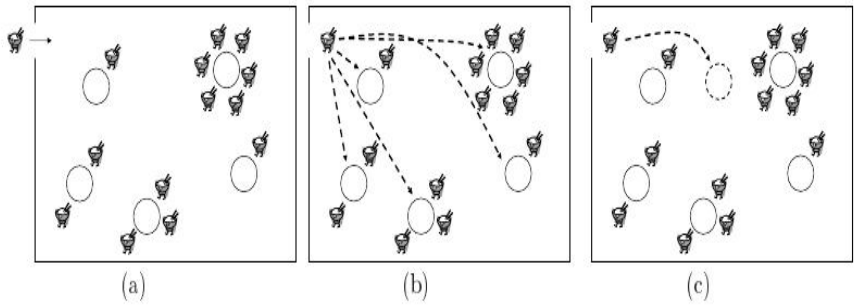

FIG. 4.2 – Le processsus du restaurant chinois.

assis dans un restaurant avec un nombre infini de tables. Le premier s'assoit à la première table, les i clients qui suivent vont s'assoir de façon aléatoire à une table selon la distribution suivante :

$$\begin{aligned} \mathbb{P}(\text{que la table précedemment occupée soit } l|\mathcal{F}_{i-1}) &\propto n_l \\ \mathbb{P}(\text{la prochaine table inoccupée}|\mathcal{F}_{i-1}) &\propto \alpha_0 \end{aligned} \tag{4.8}$$

avec n_l le nombre de clients assis à la table l couramment, et $\mathcal{F}_{i-1}$ dénote l'état du restaurant après que les $i-1$ clients se soient assis. Le schéma 4.2 permet de visualiser ce processus, le client arrive dans le restaurant (a) avec un nombre infini de tables, puis soit il s'assoit à une table déjà occupée (b), soit il s'installe à une table inoccupée (c).

Les données correspondent aux clients et les tables aux classes. Cette approche permet de visualiser de façon concrète le "pouvoir" de classification du processus de Dirichlet et justifie le choix de tels modèles pour traiter de problèmes de classification. La prochaine section s'intéresse aux modèles de mélange de processus de Dirichlet ainsi qu'à la classification.

4.2 Les modèles de mélange de processus de Dirichlet

Les modèles de mélange de processus de Dirichlet ont été introduits par Lo [45], ils exploitent le processus de Dirchlet comme une mesure mélangeante. Ils ont été considérablement développés d'un point de vue pratique par Escobar et West [26], MacEachern et Müller [46]. Comme mentionné dans l'introduction ces modèles forment une classe très importante de modèles bayésiens non paramétriques. Cette section va permettre d'expliquer en détails ces modèles et les algorithmes qui permettent de les estimer. Une application en oncologie réalisée dans le cadre d'une collaboration avec des biologistes sera donnée.

On considére un processus de Dirichlet $DP(\alpha_0\, G_0)$ comme distribution a priori pour G dans le modèle (4.3). On obtient un modèle de mélange de processus de Dirichlet (en anglais Dirichlet process mixture model) noté DPM.

$$\begin{aligned} x_i|\theta_i &\overset{ind}{\sim} f(.|\theta_i) \\ \theta_i|G &\overset{i.i.d}{\sim} G \\ G|\alpha_0,\, G_0 &\sim DP(\alpha_0\, G_0). \end{aligned} \tag{4.9}$$

En utilisant la représentation stick-breaking du processus de Dirichlet (4.6) et l'équation (4.3), on peut réécrire la distribution inconnue de la manière suivante :

$$f(.) = \sum_{k=1}^{\infty} w_k\, f(.\,|\xi_k).$$

Avec $\xi_k \overset{i.i.d.}{\sim} G_0$, $w_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l)$ où $\beta_l \overset{i.i.d.}{\sim} \mathcal{B}e(1,\, \alpha_0)$.

f est donc un mélange infini dénombrable de densités de probabilités $f(.\,|\xi_k)$.

L'inférence sur les modèles de mélange de processus de Dirichlet se fait au travers de différentes approches que l'on décrira dans le paragraphe suivant. Elles sont, pour la plupart, basées sur un échantillonnage de Gibbs.

Dans la suite on s'intéressera au problème de classification. En effet, ces modèles de mélange de processus de Dirichlet constituent une alternative attractive pour la modélisation de la distribution des variables latentes. Les mélanges de DP permettent une certaine incertitude sur le choix des formes paramétriques ainsi que sur le nombre de composantes du mélange.

4.2.1 Méthodes d'approximation de la distribution a posteriori

L'objectif est d'estimer la densité a posteriori complète $p(G, \theta|x)$, pour se faire deux grandes approches basées sur l'échantillonnage de Gibbs existent :

- l'approche marginale,
- l'approche conditionnelle.

L'approche marginale consiste à intégrer analytiquement suivant la mesure de probabilité aléatoire G et à estimer la densité a posteriori marginale $p(\theta|x)$ en utilisant la représentation en urne de Pólya (c.f. (4.7) et le processus du restaurant chinois). Différents algorithmes ont été développés notamment par Escobar et al., [25], [26] et Neal [53]. Il s'agit des méthodes MCMC par échantillonnage de Gibbs qui utilisent la représentation en urne de Pólya du processus de Dirichlet. D'autres auteurs se sont également penchés sur ce type d'approche [46].

L'approche conditionnelle utilise soit l'échantillonnage rétrospectif, ce qui est effectué par Papaspiliopoulos et Roberts [54] avec l'algorithme du "retrospective sampling", soit la modélisation sous la forme stick-breaking afin d'estimer la densité a posteriori complète. Ce sont également des méthodes MCMC qui utilisent l'échantillonneur de Gibbs.

Dans ces différentes approches les hyperparamètres peuvent être considérés comme inconnus, ce qui ajoute un degré de flexibilité à l'algorithme.

Nous arborderons dans cette thèse les méthodes conditionnelles. En effet, bien que les méthodes marginales fournissent de bons résultats, elles présentent de nombreuses limites. Tout d'abord, la mise à jour des paramètres se fait "pas à pas" conduisant à une convergence lente de l'algorithme.

De plus la marginalisation sur G induit des effets indésirables, en effet l'inférence sur la distribution a posteriori de G est seulement basée sur les valeurs a posteriori de θ_i. Enfin, bien que différentes améliorations aient été proposées, le cas non conjugué présente toujours quelques difficultés. On pourra se reporter à différents articles ([25], [26], [46], [53]) pour un exposé complet de l'échantillonnage marginal pour les DPMs.

La méthode conditionnelle que nous utiliserons dans la suite est celle développée par Ishwaran et Zarepour [42], Ishwaran et James ([40], [41]). C'est l'algorithme du Blocked Gibbs Sampling (BGS). La clé de cette méthode est l'utilisation d'une approximation du processus de Dirichlet au moyen d'une troncation de la représentation stick-breaking. La dimension finie de tels priors permet d'exprimer le modèle entier en terme de nombre fini de variables aléatoires.

Ainsi cet algorithme met à jour des "blocs" de paramètres qui sont distribués selon des distributions multivariées simples. Au final on obtient des valeurs échantillonnées directement selon la distribution a posteriori non paramétrique. De nombreux avantages en résultent : aux niveaux calculatoire et inférentiel.

En ce qui concerne la méthode de Papaspiliopoulos et Roberts [54] son principal interêt est d'éviter la troncation du processus de Dirichlet. Cependant le temps de calcul associé à cet algorithme représente un véritable inconvénient.

L'article de Ishwaran et James [40] compare le blocked gibbs sampler et les méthodes marginales améliorées. Les deux algorithmes montrent de bonnes performances en terme de convergence. Cependant le BGS présente un avantage par la simplicité de sa programmation, de plus il fournirait de meilleurs résultats dans le cas de modèles non conjugués.

Le paragraphe suivant explique l'algorithme du blocked gibbs sampler.

4.2.2 Le Blocked Gibbs Sampler (BGS)

Afin d'estimer le modèle (4.9) le BGS considère le modèle bayésien non paramétrique hiérarchique suivant :

$$\begin{aligned} x_i|\theta_i &\overset{ind}{\sim} f(.|\theta_i), \quad i = 1, \ldots, n \\ \theta_i|G &\overset{i.i.d}{\sim} G \\ G &\sim \mathcal{P}_N, \end{aligned} \tag{4.10}$$

où

$$\mathcal{P}_N(.) = \sum_{i=1}^{N} w_k \, \delta_{\xi_k}(.)$$

représente la troncation de la représentation stick breaking du processus de Dirichlet $DP(\alpha_0 G_0)$. D'autres priors de dimension finie peuvent être utilisés, nous nous limiterons ici aux processus de Dirichlet tronqués.

Les poids ξ_k sont des variables aléatoires i.i.d. de distribution G_0 indépendantes des poids $w = (w_1, \ldots, w_N)$. Ces poids sont définis grâce à la représentation stick breaking (4.6), on a que :

$$w_1 = v_1 \;\; et \;\; w_k = (1 - v_1)(1 - v_2) \ldots (1 - v_{k-1})\, v_k \;\; k = 2, \ldots, N, \tag{4.11}$$

où $v_1, \ldots, v_{N-1}$ sont des variables aléatoires i.i.d. de loi Beta de paramètres 1 et α_0. On pose $v_N = 1$ pour s'assurer que la somme des poids est égale à un, i.e. $\sum_{k=1}^{N} w_k = 1$. Dans la suite on notera $w \sim GEM(\alpha_0)$ lorsque les poids sont définis de la façon précédente.

Il est facile de montrer [52] que $\mathcal{P}_N$ converge presque sûrement vers le processus de Dirichlet de paramètres (α_0, G_0). C'est à dire que :

$$\mathcal{P}_N \overset{p.s.}{\to} DP(\alpha_0 G_0).$$

Le BGS va permettre d'obtenir une inférence directe pour $\mathcal{P}_N$ et de construire une méthode MCMC efficace. La clé de son succès est l'utilisation d'un prior de dimension finie permettant, ainsi, d'exprimer notre modèle (4.10) en terme de variables aléatoires dont la mise à jour se fait assez facilement.

Le modèle (4.10) peut se réécrire de la façon suivante :

$$\begin{aligned} x_i|\xi,\, K &\overset{ind}{\sim} f(.|\xi_{K_i}) \;\; i = 1,\, \ldots,\, n \\ K_i|w &\overset{i.i.d}{\sim} \sum_{k=1}^{N} w_k\, \delta_k(.) \\ (w,\, \xi) &\sim GEM(\alpha_0) \times G_0^N(\xi), \end{aligned} \tag{4.12}$$

avec $K = (K_1,\, \ldots,\, K_n)$, $\xi = (\xi_1,\, \ldots,\, \xi_N)$ avec $\xi_k \overset{i.i.d}{\sim} G_0$. Il est important de remarquer que l'on a l'égalité $\theta_i = \xi_{K_i}$. Les variables K_i correspondent à des variables de classification permettant d'identifier la valeur ξ_k associée à chaque θ_i. Ainsi sachant le vecteur de classification $K = (K_1,\, \ldots,\, K_n)$ on est en mesure de décrire la classification des θ_i.

Le choix de N

Le choix de N est important, il doit mener à une approximation précise. Une méthode est de choisir une valeur de N telle que la densité marginale de x soit presque indissociable de sa limite.

Posons :

$$m_N(x) = \int \Big(\prod_{i=1}^{n} \int_{\Theta} f(x_i|\theta_i)\, G(d\theta_i) \Big)\, \mathcal{P}_N(dG),$$

la densité marginale du modèle (4.10).

Notons m_∞ la densité marginale de ce même modèle dans la cas où l'on prend un processus de Dirichlet $DP(\alpha_0\, G_0)$ comme mesure aléatoire pour G.

Théorème 4.2.1. *Nous avons que :*

$$\begin{aligned} \int_{\mathbb{R}^n} |m_N(x) - m_\infty(x)|\, dx &\leq 4\Big[1 - \mathbb{E}\Big\{ \Big(\sum_{k=1}^{N-1} w_k \Big)^n \Big\}\Big] \\ &\approx 4\, n\, exp(-(N-1)/\alpha_0), \end{aligned} \tag{4.13}$$

où w_k sont les poids aléatoires de la représentation stick-breaking (4.11).

Démonstration. Se référer à la preuve développée dans l'article de Ishwaran et James [41]. □

N n'a pas besoin d'être très grand pour avoir une bonne approximation. De plus pour une valeur de N raisonnablement grande, la taille de l'échantillon a peu d'effet sur cette borne. Dans la suite, on choisira N en fonction de la taille de l'échantillon n, du paramètre α_0 et de la borne obtenue au théorème (4.2.1).

Le choix du paramètre d'échelle

Dans la plupart des applications on suppose que le paramètre de concentration du processus de Dirichlet est inconnu. On doit donc l'estimer. Escobar et West [26] propose une loi Gamma comme distribution a priori pour α_0 afin de traduire l'incertitude a priori sur α_0. Les hyperparamètres a et b sont fixés. α_0 conditionne le nombre de classes ayant un poids significatif. Comme vu dans la section précédente lorsque α_0 est petit, la distribution de mélange G va se

concentrer sur quelques classes ayant un poids important. Pour α_0 grand, la distribution de mélange a plusieurs points de support, le modèle non paramétrique devient plus proche de la distribution de base G_0.

Ce problème de sensibilité du prior a motivé de nombreux auteurs à adopter une approche bayésienne empirique où l'on estime par maximum de vraisemblance α_0. L'inférence sur les autres paramètres du modèle se fait conditionnellement à cette estimation. Ce type d'approche nécessite des calculs très importants et peut être numériquement instable dans plusieurs situations. D'où une nouvelle approche développée par Dorazio [21] pour calculer un prior pour α_0 qui peut être utilisé en présence ou en absence d'information a priori sur le niveau de classification des θ_i.

Une approche très intéressante proposée par Lijoi et al [44] utilise un prior non paramétrique plus général que le processus de Dirichlet et obtenu à partir d'un processus Gamma généralisé. Le principal avantage de cette approche est d'ajouter un paramètre pour gérer le comportement de classification des variables latentes.

Contrairement aux processus de Dirichlet qui ne peuvent utiliser qu'un seul paramètre libre α_0 pour gérer la distribution du nombre de composantes, Lijoi et al [44] font dépendre G et le nombre de classes de deux paramètres

- $\beta \in (0, \infty)$ qui joue le même rôle que α_0,
- $\gamma \in (0, 1)$ qui influence le groupement des observations dans les différentes classes (ce paramètre est fixé dans le processus de Dirichlet).

Ce paramètre γ est responsable du nombre distinct de variables latentes. Les auteurs mettent en place une procédure de renforcement qui se va renforcer les classes ayant la plus grande fréquence, plus γ va être grand plus le mécanisme de renforcement va être fort.

Ainsi, ces nouveaux processus, notés $GG(\beta, \gamma)$ peuvent se voir de la façon suivante. Le processus de renforcement va permettre d'allouer les masses à certaines classes plutôt qu'à d'autres en pénalisant les classes de petites tailles et en favorisant celles exhibant des preuves empiriques. Lors de l'implémentation du BGS, nous avons utilisé une loi Gamma comme prior pour α_0.

L'algorithme du Blocked Gibbs Sampler

La réécriture de notre modèle sous la forme (4.12) permet d'utiliser l'échantillonneur de Gibbs pour explorer la distribution a posteriori $\mathcal{P}_N|x$. Pour implémenter cet algorithme, on échantillonne itérativement des valeurs selon les distributions conditionnelles suivantes :

$$(\xi|K, x), \tag{4.14}$$
$$(K|\xi, w, x), \tag{4.15}$$
$$(w|K).$$

Cette méthode permet d'obtenir des valeurs échantillonnées selon la distribution a posteriori complète $(\xi, K, w|x)$. A chaque étape on peut garder un

échantillon (ξ', K', w'). On produit ainsi une mesure de probabilité aléatoire :

$$\mathcal{P}'_N(.) = \sum_{k=1}^{N} w'_k \, \delta_{\xi'_k}(.),$$

qui est un échantillon de la distribution a posteriori $\mathcal{P}_N|\mathbf{x}$. On peut donc utiliser $\mathcal{P}'_N$ pour estimer directement $\mathcal{P}_N|\mathbf{x}$ et ses fonctions.

Passons à la description de l'algorithme.

Soit $\{K_1^*, \ldots, K_m^*\}$ un ensemble qui dénote les m uniques valeurs de K. Pour chaque itération de l'échantillonneur de Gibbs, nous échantillonnons différentes valeurs dans cet ordre :

(*i*) Conditionnelle pour ξ : pour chaque $k \in K - \{K_1^*, \ldots, K_m^*\}$, on simule $\xi_k \overset{i.i.d.}{\sim} G_0$. Puis, on échantillonne $(\xi_{K_j^*}|K, x)$ selon la densité

$$f(\xi_{K_j^*}|K, x) \propto G_0(d\xi_{K_j^*}) \prod_{\{i:K_i=K_j^*\}} f(x_i|\xi_{K_j^*}),$$
$$j = 1, \ldots, m. \tag{4.16}$$

(*ii*) Conditionnelle pour K :

$$(K_i|\xi, w, x) \overset{ind}{\sim} \sum_{k=1}^{N} w_{k,i} \, \delta_k(.), \quad i = 1, \ldots, n,$$

où

$$(w_{1,i}, \ldots, w_{N,i}) \propto \Big(w_1 \, f(x_i|\xi_1), \ldots, w_N \, f(x_i|\xi_N)\Big).$$

(*iii*) Conditionnelle pour w : par la propriété de la ditribution de Dirichlet et de la distribution multinomiale, on a

$$w_1 = v_1^* \;\; et \;\; w_k = (1 - v_1^*)\,(1 - v_2^*) \ldots (1 - v_{k-1}^*)\, v_k^* \;\; k = 2, \ldots, N-1,$$

où

$$v_k^* \overset{ind}{\sim} \mathcal{B}e\Big(1 + M_k, \, \alpha_0 + \sum_{l=k+1}^{N} M_l\Big), \;\; pour \;\; k = 1, \ldots, N-1$$

et M_k indique le nombre de valeurs de K_i égales à k.

Pour compléter la spécification a priori de $\mathcal{P}_N$, nous utilisons le prior suivant pour α_0 :

$$(\alpha_0|\eta_1, \eta_2) \sim \mathcal{G}amma(\eta_1, \eta_2)$$

Nous écrivons $\mathcal{G}amma(\eta_1, \eta_2)$ pour dénoter une distribution gamma avec η_1 comme paramètre de forme et η_2 comme paramètre d'échelle.

Les différents résultats obtenus au niveau de la convergence avec cet algorithme sont équivalents ou supérieurs aux méthodes faisant appel à la représentation par urne de Pólya.

Le processus de Dirichlet est donc utilisé pour spécifier les variables latentes. Différentes situations peuvent être étudiées : le cas univarié simple, le cas multivarié simple puis le cas où la plupart des paramètres sont aléatoires. Dans la suite on considère la densité de probabilité de la loi normale. Le Blocked Gibbs Sampler a été implémenté avec le logiciel R dans ces différentes situations.

4.2.3 Application

Dans cette section, on s'intéresse aux modèles de mélange de lois normales, c'est-à-dire que $F = \mathcal{N}(\mu, \sigma)$. Le modèle (4.3) se réécrit sous la forme :

$$\begin{aligned} x_i|\mu_i &\sim \mathcal{N}(\mu_i, \sigma),\ i = 1, \ldots, n \\ \mu_i|G &\sim G \\ G|\alpha_0, G_0 &\sim DP(\alpha_0\, G_0), \end{aligned}$$

où $\sigma > 0$ est une variance connue. Pour compléter ce modèle de mélange nous introduisons un prior pour α_0 et des hyperprior pour ξ. En utilisant la représentation (4.12),on a :

$$\begin{aligned} (x_i|\xi, K) &\overset{ind}{\sim} \mathcal{N}(\xi_{K_i}, \sigma)\ \ i = 1, \ldots, n \\ (K_i|w) &\overset{i.i.d}{\sim} \sum_{k=1}^{N} w_k\, \delta_k(.) \\ (\xi_k|\mu_0, \sigma_\xi) &\sim \mathcal{N}(\mu_0, \sigma_\xi) \\ (\sigma_\xi^{-1}|\tau_1, \tau_2) &\sim \mathcal{G}amma(\tau_1, \tau_2), \\ (\alpha_0|\nu_1, \nu_2) &\sim \mathcal{G}amma(\nu_1, \nu_2), \end{aligned} \tag{4.17}$$

avec la distribution de w spécifiée par la représentation stick breaking. Une loi inverse Gamma est utilisée comme prior pour σ_ξ. Si l'on veut s'assurer que les priors sont non informatifs, on sélectionne de petites valeurs pour les hyperparamètres τ_1, τ_2, c'est à dire $\tau_1 = \tau_2 = 0.001$.

Le choix de la distribution a priori de α_0 est important, en effet comme nous l'avons déjà dit la valeur de α_0 est directement relié au nombre de valeurs θ_i distinctes. Le choix d'une distribution Gamma [26] est intéressant par sa flexibilité. Le choix d'un paramètre d'échelle ν_2 élevé est pertinent pour la modélisation de mélanges finis, en effet il favorise les répétitions dans les θ_i et peut être utilisé comme un outil pour étudier le nombre de composantes du mélange. Le processus de Dirichlet permet d'échantillonner de façon exacte α_0 lorsqu'on utilise une distribution Gamma comme prior.

La loi a posteriori obtenue pour σ_ξ^{-1} est donnée par :

$$(\sigma_\xi^{-1}|\xi, \mu_0) \sim \mathcal{G}amma\Big(\tau_1 + \frac{N}{2}, \tau_2 + \sum_{k=1}^{N}(\xi_k - \mu_0)^2/2\Big),$$

et celle obtenue pour α_0 :

$$(\alpha_0|w) \sim \mathcal{G}amma\Big(N + \nu_1 - 1,\, \nu_2 - \sum_{k=1}^{N-1} log(1 - v_k^*)\Big).$$

En introduisant σ comme un paramètre, on ajoute un paramètre de dimension finie au modèle hiérarchique (4.17). Un choix judicieux est de prendre une loi inverse Gamma :

$$(\sigma^{-1}|\gamma_1,\, \gamma_2) \sim \mathcal{G}amma(\gamma_1,\, \gamma_2).$$

Pour un prior non informatif, on choisit $\gamma_1 = \gamma_2 = 0.001$. La loi a posteriori se calcule facilement :

$$(\sigma^{-1}|x,\, \xi,\, K) \sim \mathcal{G}amma\Big(\gamma_1 + \frac{n}{2},\, \gamma_2 + \sum_{k=1}^{n}(x_i - \xi_{K_i})^2/2\Big).$$

Nous sommes ici dans le cas univarié, l'approche multivariée diffère peu, il suffit de prendre une loi normale inverse Wishart pour gérer la matrice de variance-covariance.

Jeu de données

L'algorithme BGS a été appliqué dans le cadre d'une collaboration avec des biologistes, un article est en préparation. Les différents résultats présentés sont tels qu'ils apparaissent dans l'article.

Leur problèmatique était la suivante : Comment la protéine Syk se déplace-t-elle jusqu'au centrosome quand elle est activée ?

Explicitons un peu plus le contexte. La protéine Syk est une kynase ayant un rôle vital, en effet elle permet de réguler la réponse immunitaire, de plus elle semble être corrélée aux tumeurs dans les cellules epitheliales du sein. La présence de protéine Syk indique donc l'absence de cellules métastasiques. Lorsqu'elle est activée cette kynase se trouve au niveau du centrosome. La question est donc de savoir comment la protéine Syk se déplace jusqu'au centrosome.

La principale hypothèse consiste à dire que les protéines Syk se déplacent jusqu'au centrosome grâce aux microtubules. Les microtubules serviraient donc de "rails".

Pour vérifier cette hypothèse, on étudie le comportement dynamique de Syk dans les cellules vivantes du cancer du sein, en se focalisant plus spécialement sur les mouvements de la kynase de et vers le centrosome. Différents produits chimiques ont été utilisés afin de montrer l'importance des microtubules :

(*i*) Le Nocodaxol qui permet de casser les microtubules,

(*ii*) Le Taxol qui augmente la fiabilité et la solidité des microtubules.

On associe la protéine RFP (Red Fluorescent Protein) à la protéine Syk afin d'obtenir une protéine fluorescente. Une activation laser est effectuée autour du centrosome. On tape juste dans cette zone afin de détruire la RFP associée à

Syk (la protéine n'est plus fluorescente). On regarde alors le temps de retour de la fluorescence.

Deux types de transports existent :
- Le transport "passif" concerne les Syk présentent dans le cytoplasme et qui se déplacent "librement", la fluorescence va donc revenir rapidement,
- Le transport "actif" concerne les Syk transportées par les microtubules, c'est donc un transport régulé.

On va étudier les temps de demi-retour, noté τ, de la fluorescence des protéines à l'état sauvage et avec les différentes substances mentionnées ci-dessus.

Une autre question a été traitée en parallèle : la protéine Syk activée a-t-elle besoin d'aller au centrome ? Pour répondre à cette question, on a simulé des Syk activées grâce au $y130$. La méthode d'activation laser a été réalisée afin d'observer les temps de demi-retour.

Les analyses statistiques réalisées grâce au BGS ont permis d'estimer la distribution des temps de demi-retour et de déterminer le nombre de classes pour chaque condition expérimentale.

En ce qui concerne les paramètres du BGS, on a pris : $N = 20$, $\alpha_0 \sim \mathcal{G}amma(1, 10)$, $\sigma^{-1} \sim Gamma(0.01, 0.01)$. Le nombre d'itérations est égal à 10000 avec un temps de chauffe de 5000.

A

Treatment	DsRed Syk wt	
	τ_1	τ_2
Nocodazole	7.55 ± .214 s	21.28 ± 5.92 s
Taxol	16.76 ± 4.12 s	31.52 ± 4.18 s

B

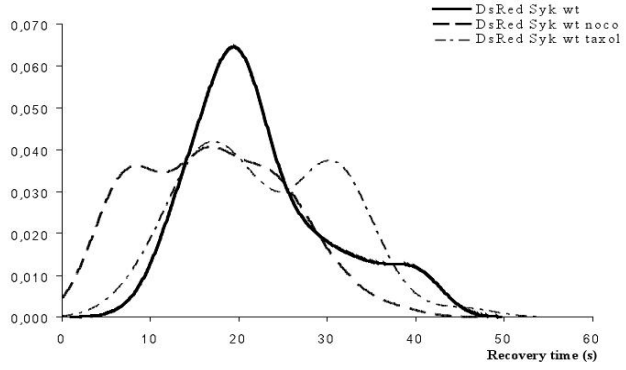

FIG. 4.3 – Estimation de la distribution des temps de demi-retour pour le Nocodaxol, le Taxol et les Syk "sauvages". Les moyennes et les écarts types pour les groupes des différentes populations sont donnés dans le tableau du haut.

Les premiers résultats concernent le Nocodazole, l'analyse statistique du temps de demi-retour τ révèle deux populations principales avec les valeurs

moyennes suivantes 7.55 ± 2.14 et 21.28 ± 5.92 secondes. Ces deux populations sont visibles sur la Figure 4.3.

La différence entre la distribution de la première population traitée au Nocodazole et la population des cellules non traitées est hautement significative ($p = 2 \times 10^{-16}$) tandis que la seconde population ayant des temps de demi retour plus longs n'est pas significativement différente de la population de cellules non traitées ($p = 0.09$). Ainsi la perturbation du cytosquelette des microtubules avec le Nocodazole affecte le transport des protéines Syk au centrosome en accélerant significativement le recrutement d'une population importante de cellules.

En ce qui concerne la stabilisation des microtubules avec le Taxol, deux populations de cellules sont détectées avec, pour la première population un temps de demi-retour de moyenne 16.76 ± 4.12 avec une p-value de 0.2. Pour la deuxième population, on a la moyenne 31.52 ± 4.18 et une p-value hautement significative à 2×10^{-6}. Ces résultats sont représentés sur la Figure 4.3.

Ces deux résultats nous confortent dans l'idée qu'une partie du transport des protéines Syk jusqu'au centrosome dépend des microtubules.

A

Type of experiment	DsRed Syk wt	DsRed Syk Y130E
FRAP	τ=18.54 ± 3.63 sec	τ=11.67 ± 4.59 sec
Photoactivation	τ=0.32 ± 0.14 sec τ=3.55 ± 2.98 sec	τ=0.4 ± 0.13 sec τ=6.4 ± 2.8 sec

B

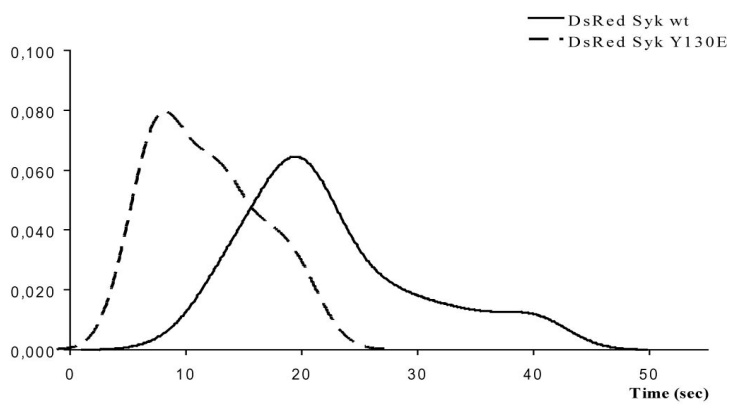

FIG. 4.4 – Estimation de la distribution des temps de demi-retour pour les Syk activées par le $y130$ et les Syk "sauvages". Les moyennes et les écarts types pour les différentes populations sont donnés dans le tableau du haut.

Pour ce qui est de la deuxième question soulevée par nos collègues biologistes, la Figure 4.4 représente les estimations des distributions des temps de demi-retour pour les protéines "sauvages" et celles activées par $y130$. Les formes des deux distributions estimées sont semblables cependant la différence entre les deux moyennes 11.67 ± 4.59 secondes et 18.54 ± 3.63 secondes est hautement significative 2×10^{-8}. Notons qu'une seule population est détectée pour le $y130$ et les protéines "sauvages" contrairement au Nocodazole et au Taxol.

Ces résultats montrent que la protéine Syk a besoin d'aller rapidement au

centrosome lorsqu'elle est activée.

L'approche bayésienne non paramétrique hiérarchique a donc permis de détecter les différents comportements de la protéine Syk en fonction des différentes substances.

4.3 Les processus de Dirichlet hiérarchiques

La section précédente a mis en place les modèles de mélange de processus de Dirichlet, ainsi que l'algorithme du blocked gibbs sampler qui permet d'estimer la loi a posteriori complète.

Cette nouvelle section a été développée suite à un problème soulevé par un de nos collègues médecins. Il avait à disposition un ensemble de données réparties dans plusieurs groupes. Le but était de déterminer les éléments communs aux différents groupes ainsi que leurs répartitions.

On a donc eu l'idée d'associer à chaque groupe de données un processus de Dirichlet et de "lier" ces processus par un processus de Dirichlet commun. Ce principe constitue l'idée générale développée au travers des processus de Dirichlet hiérarchiques notés HDP ([67], [68]). En ce qui concerne l'inférence, l'idée a été de développer l'algorithme du blocked Gibbs sampler dans ce cadre spécifique.

4.3.1 Définition du modèle

Dans un cadre général, les processus de Dirichlet hiérarchiques ont été proposés dans le cas où l'on a plusieurs groupes de données. Le modèle pour chaque groupe intègre une variable discrète de cardinal inconnu, le but est de lier ces variables à travers les groupes. Supposons par exemple que l'on ait J groupes de données, chaque groupe est associé à un modèle de mélange. Le but est

(i) de déterminer pour chaque groupe le nombre de composantes et les paramètres associés,

(ii) de "lier" ces différents groupes.

Une première idée serait d'affecter à chaque groupe de données un processus de Dirichlet de paramètres α_0, G_0, ces groupes étant liés par la mesure de base G_0. Cette première approche peut être schématisée de la figure 4.5.

Les atomes générés par les mesures de probabilité aléatoires G_j seront distincts. Il n'y aura pas de partage d'atomes entre les différents groupes, pas de partage de classes et donc pas d'effets de classification.

On a besoin d'une mesure de probabilité G_0 aléatoire, flexible et discrète. Une idée est de faire l'hypothèse que G_0 suit un processus de Dirichlet de paramètres γ, H :

$$G_0 \sim DP(\gamma\, H),$$

et ainsi

$$G_j|\alpha_0, G_0 \sim DP(\alpha_0\, G_0) \text{ pour tout } j = 1, \ldots, J.$$

Les mesures de probabilité aléatoires G_j ont donc la même mesure de base G_0. De plus, comme cette mesure est atomique les échantillons de G_j seront rééchan-

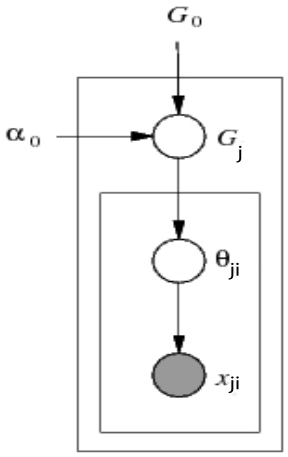

FIG. 4.5 – Première idée : un processus de Dirichlet pour chaque groupe, ces processus sont liés par un processus sous-jacent G_0.

tillonés parmi les atomes de G_0. On a donc rajouté une hiérarchie bayésienne. La figure 4.6 permet de visualiser cette idée.

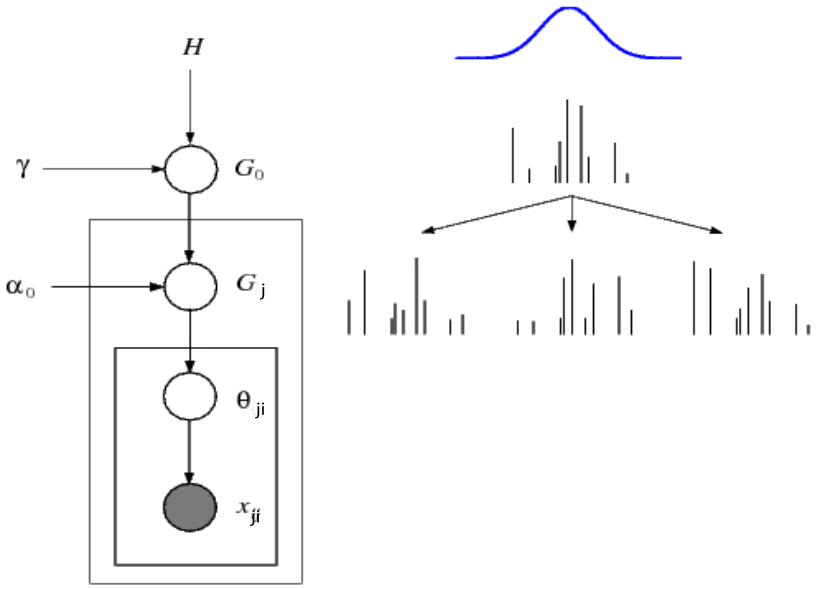

FIG. 4.6 – Les mélanges de processus de Dirichlet hiérarchiques

Les composantes de chaque mélange sont partagées par les différents groupes. La variabilité à travers les groupes est modélisée en permettant différentes proportions de mélange dans chaque groupe.

Cette approche offre un cadre flexible. La mesure de base G_0 est appelée mesure globale. En notant x_{ji} l'observation i dans le groupe j et θ_{ji} le paramètre

associé, on obtient le modèle suivant :

$$\begin{aligned} x_{ji}|\theta_{ji} &\sim f(.|\,\theta_{ji}) \\ \theta_{ji}|G_j &\overset{ind}{\sim} G_j \\ G_j|\alpha_0,\, G_0 &\overset{ind}{\sim} DP(\alpha_0,\, G_0) \\ G_0|\gamma,\, H &\sim DP(\gamma,\, H). \end{aligned} \tag{4.18}$$

Comme pour les modèles de mélange de processus de Dirichlet, on utilise la représentation stick-breaking des processus de Dirichlet (4.6). Ainsi la mesure de base G_0 s'écrit sous la forme :

$$G_0 \sim \sum_{k\geq 1} \beta_k\, \delta_{\phi_k},$$

où $\beta_k = v_k \prod_{l=1}^{k-1}(1-v_l)$ avec $v_k \sim \mathcal{B}e(1,\,\gamma)$ et $\phi_k \overset{ind}{\sim} H$ pour tout $k \geq 1$. Les atomes $\phi = (\phi_k)_{k\geq 1}$ constituent le support de G_0.

Comme $G_j|G_0 \sim DP(\alpha_0,\, G_0)$, on écrit pour tout $j \geq 1$:

$$G_j = \sum_{k\geq 1} \pi_{jk}\, \delta_{\phi'_k},$$

avec $\pi_{jk} = v'_k \prod_{l=1}^{k-1}(1-v'_l)$ où $v'_k \sim \mathcal{B}e(1,\,\alpha_0)$ et $\phi'_k|G_0 \sim G_0$.

Par construction (Teh et al. [67]) les mesures de probabilité aléatoires G_j ont le même support que la mesure de base G_0, on peut réécrire G_j comme suit :

$$G_j = \sum_{k\geq 1} \pi_{jk}\, \delta_{\phi_k}.$$

La représentation stick breaking des G_j est donc une somme repondérée des atomes de G_0. Notons que les mesures aléatoires G_j sachant G_0 sont simulées indépendamment selon G_0, ainsi les poids $\pi_j = (\pi_{jk})_{k\geq 1}$ sachant $\beta = (\beta_k)_{k\geq 1}$ sont indépendants.

Il est important pour la suite de définir la relation entre les poids β et π_j pour tout $j \geq 1$.

Soit $(B_1,\, \ldots,\, B_r)$ une partition mesurable de Θ. Par définition d'un processus de Dirichlet, on a pour chaque j :

$$(G_j(B_1),\, \ldots,\, G_j(B_r)) \sim Dir(\alpha_0\, G_0(B_1),\, \ldots,\, \alpha_0\, G_0(B_r)).$$

On définit $N_l = \{k : \phi_k \in B_l\}$ pour $l = 1,\, \ldots,\, r$ tel que $(N_1,\, \ldots,\, N_r)$ constitue une partition finie d'entiers positifs. L'hypothèse que H est non atomique permet de déduire que les atomes ϕ_k sont tous distincts avec une probabilité égale à 1. Ainsi toute partition d'entiers positifs correspond à une partition de Θ. On a pour chaque j :

$$(\sum_{k\geq 1} \pi_{jk}\delta_{\phi_k}(B_1),\, \ldots,\, \sum_{k\geq 1} \pi_{jk}\delta_{\phi_k}(B_r)) \sim Dir(\alpha_0 \sum_{k\geq 1} \beta_k\delta_{\phi_k}(B_1),\, \ldots,\, \alpha_0 \sum_{k\geq 1} \beta_k\delta_{\phi_k}(B_r)),$$

ce qui est équivalent à

$$(\sum_{k\in N_1}\pi_{jk},\ldots,\sum_{k\in N_r}\pi_{jk})\sim Dir(\alpha_0\sum_{k\in N_1}\beta_k,\ldots,\alpha_0\sum_{k\in N_r}\beta_k). \tag{4.19}$$

Ainsi chaque π_j est independemment distribué selon un processus de Dirichlet $DP(\alpha_0, \beta)$, où l'on interprète β et π_j comme des mesures de probabilités sur les entiers positifs. On a donc : $\pi_j \sim DP(\alpha_0, \beta)$.

Essayons de formuler une relation explicite entre les élèments de β et π_j. En utilisant la définition des β_k obtenue avec la représentation stick-breaking et (4.19), on obtient une mesure de probabilité aléatoire $\pi_j \sim DP(\alpha_0, \beta)$ tel que :

$$\pi_{jk}=\pi'_{jk}\prod_{l=1}^{k-1}(1-\pi'_{jl})\ \text{ où }\ \pi'_{jk}\sim\mathcal{B}eta(\alpha_0\beta_k,\alpha_0(1-\sum_{l=1}^{k}\beta_l)). \tag{4.20}$$

On a une relation entre les coefficients β et π_j pour tout $j \geq 1$. Une description complète de la représentation stick-breaking du processus de Dirichlet hiérarchique est ainsi donnée.

On décrit une analogie avec le processus du restaurant chinois dans le cadre des processus de Dirichlet hiérarchique que l'on appelera la franchise des restaurants chinois. La métaphore est la suivante : on a une franchise de restaurants avec un menu commun à l'ensemble de ces restaurants. A chaque table un poisson est désigné par le premier client qui s'assoit, et ce même poisson est partagé par tous les clients s'asseyant à cette table. L'ensemble des poissons étant commun à tous les restaurants, un même poisson peut être servi à plusieurs tables dans plusieurs restaurants.

Ainsi les restaurants correspondent aux différents groupes ($j \in J$) et les clients sont les paramètres θ_{ji}. Les variables $\phi_1, \ldots, \phi_K$ représentent l'ensemble des poissons disponibles dans tous les restaurants. Elles sont indépendantes et identiquement distribuées selon H. L'article de Teh et Jordan [68] décrit bien ce processus, ainsi que les différentes distributions marginales qui en découlent.

Dans la section suivante nous décrivons une méthode d'inférence pour les modèles de mélange de processus de Dirichlet hiérarchique. La méthode du blocked gibbs sampling est utilisée.

4.3.2 L'inférence

Teh et al. [67] proposent trois types de méthodes basée sur une approche MCMC pour estimer les modèles de processus de Dirichlet hiérarchiques. Ces méthodes reposent sur l'échantillonneur de Gibbs et la franchise du restaurant chinois avec plus ou moins de variation.

En ce qui nous concerne nous nous sommes plutôt intéressés à l'utilisation du blocked gibbs sampler pour les processus de Dirichlet hiérarchiques. L'algorithme du BGS a été présenté dans la section (1.2.3) dans le cadre de modèles de mélange de processus de Dirichlet.

Comme pour les DPMs, on utilise la troncation de la représentation stick-breaking des processus de Dirichlet G_j pour $j = 1, \ldots, J$ et G_0. Le modèle

(4.18) se réécrit sous la forme suivante :

$$\begin{aligned} x_{ji}|\theta_{ji} &\sim f(.|\,\theta_{ji}) \\ \theta_{ji}|G_j &\overset{ind}{\sim} G_j \\ G_j|\alpha_0,\, G_0 &\overset{ind}{\sim} \sum_{k=1}^{N} \pi_{jk}\,\delta_{\phi_k} \\ G_0|\gamma,\, H &\sim \sum_{k=1}^{N} \beta_k\,\delta_{\phi_k} \end{aligned} \tag{4.21}$$

On introduit une variable indicatrice associée à l'observation i dans le groupe j notée K_{ji} et telle que $\theta_{ji} = \phi_{K_{ji}}$. Ainsi sachant K_{ji}, la variable x_{ji} suit la loi de probabilité F de densité $f(.|\phi_{K_{ji}})$.

Une représentation équivalente du modèle de mélange définie par (4.21) est donnée par :

$$\begin{aligned} x_{ji}|K_{ji},\,\phi &\sim f(.|\,\phi_{K_{ji}}) \\ K_{ji}|\pi_j &\overset{ind}{\sim} \sum_{k=1}^{N} \pi_{jk}\,\delta_k(.) \\ \phi_k|H &\overset{ind}{\sim} H \\ \pi_j|\alpha_0,\,\beta &\overset{ind}{\sim} DP(\alpha_0,\,\beta) \\ \beta|\gamma &\sim GEM(\gamma). \end{aligned} \tag{4.22}$$

On prend les densités de probabilité $F(.|\,\phi_l) = \mathcal{N}(\phi_l,\,\sigma)$ et $H = \mathcal{N}(\mu_0,\,\sigma_0)$. Soit $K = \{K_{ji},\, i = 1,\,\ldots,n_j,\; j = 1,\,\ldots,\,J\}$, on note K^* l'ensemble des valeurs distinctes de K.

L'algorithme du blocked gibbs sampler se déroule de la façon suivante :

(i) Loi conditionnelle pour ξ : pour chaque $k \in K - K^*$, on simule $\xi_k \overset{i.i.d.}{\sim} H$. Puis, on échantillonne $(\phi_{K^*_{ji}}|K,\,x)$ selon la densité

$$f(\phi_{K^*_{ji}}|K,\,x) \propto H(d\phi_{K^*_{ji}}) \prod_{\{K_{ji}=K^*_{ji}\}} f(x_{ji}|\phi_{K^*_{ji}}), \tag{4.23}$$

(ii) Pour $j = 1,\,\ldots,\,J$, on fait :

– Loi conditionnelle pour $K_j = (K_{j1},\,\ldots,\,K_{jn_j})$:

$$(K_{ji}|\phi,\,\pi_j,\,\beta,\,x) \overset{ind}{\sim} \sum_{k=1}^{N} \pi^i_{jk}\,\delta_k(.),\quad i = 1,\,\ldots,\,n,$$

où

$$(\pi^i_{j1},\,\ldots,\,\pi^i_{jN}) \propto (\pi_{j,\,1}\,f(x_{ji}|\phi_1),\,\ldots,\,\pi_{j,\,N}\,f(x_{ji}|\phi_N))$$

– Loi conditionnelle pour π_j :

$$\pi_{jk} = v_k^* \prod_{l=1}^{k-1}(1 - v_l^*)$$

$$v_k^* \sim Beta\big(\alpha_0\,\beta_k + m_{jk};\, \alpha_0\,(1 - \sum_{l=1}^{k}\beta_l) + \sum_{l=k+1}^{N} m_{jl}\big)$$

où m_{jk} correspond au nombre d'observations égales à k dans le groupe j.

(*iii*) Loi conditionnelle pour β :

$$\beta_k = v_k^{**} \prod_{l=1}^{k-1}(1 - v_l^{**})$$

$$v_l^{**} \sim Beta(1 + m_k,\, \gamma + \sum_{l=k+1}^{N} m_l),$$

où m_l correspond au nombre d'observations égales à k.

On obtient ainsi l'estimation de la loi a posteriori complète (ϕ, π, $\beta|x$).

4.3.3 Application

Une application réalisée par Teh, Jordan, Beal, Blei [67] traitait de la modélisation de la relation parmi un ensemble de documents. Un document étant vu comme un "sac de mots". De plus, la modélisation des mots dans un document comme un modèle de mélange où une composante correspond à un topic. L'objectif était de modéliser un corpus de documents dans le but de permettre aux topics d'être partagés dans les documents d'un corpus.

Comme précisé au début de cette section nous avons été amenés à considérer cette d'approche pour résoudre une problèmatique soulevée par les médecins dans le cadre d'un projet ANR-Institut de Recherche en Santé Publique.

En effet, nous avions à disposition un ensemble de données décrivant la distance moyenne parcourue par l'ensemble des patients dans un hôpital j pour une une pathologie i. La question était de savoir si certains hôpitaux étaient plus attractifs que d'autres. Nous disposons de 58 hôpitaux et de 759 pathologies.

Le BGS décrit dans le paragraphe ci-dessus a été appliqué avec les paramètres : $N = 20$, $\gamma \sim \mathcal{G}amma(7.5,\, 20)$, μ_0 est égal à la moyenne des observations, σ_0 est égal à l'estimation de la variance sur les observations. Le paramètre α_0 est fixé à 1. A chaque itération on met à jour le paramètre de concentration γ. Le nombre d'itération est égal à 10000 avec un temps de chauffe de 5000.

Les différents résultats obtenus permettent de mettre en évidence trois types d'hôpitaux : les hôpitaux de "références", de "recours" et de "proximités". Le graphique 4.7 permet de visualiser le nombre de groupes distincts parmi tous les hôpitaux considérés. Ce graphique représente la distribution a posteriori du nombre de valeurs θ_{ji} distinctes.

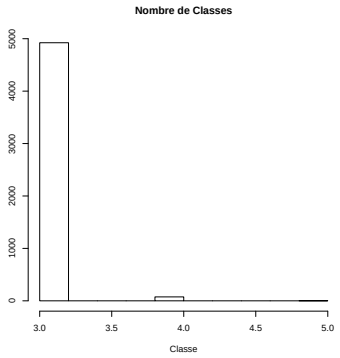

FIG. 4.7 – Nombre de classes

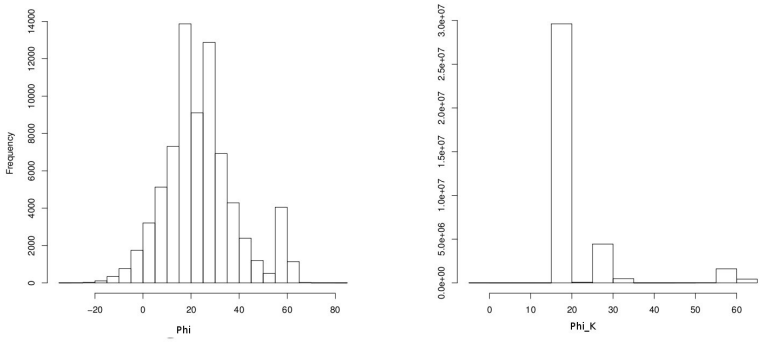

FIG. 4.8 – Distributions a posteriori de ϕ (à gauche) et de ϕ_K (à droite).

D'autre part quatres hôpitaux considérés comme des hôpitaux de référence ressortent au travers de cette expérience : l'hôpital local de Lodève, l'hôpital de la lutte contre le cancer, la clinique du souffle de Vallonie et le CHU de Montpellier.

Ces observations sont en accord avec les données déjà connues des médecins. Les deux autres graphiques (Figure 4.8) permettent d'observer les distributions a posteriori de ϕ, on observe clairement trois modes. Cet aspect est vérifié avec le graphique qui représente la distribution a posteriori de ϕ_K.

Ces résultats concernent la région Languedoc Roussillon. Un point intéressant à étudier serait de comparer les différentes régions hospitalières françaises.

4.4 Conclusion

Ce dernier chapitre a permis d'introduire la statistique bayésienne non paramétrique ainsi que les différentes applications possibles.

Dans un premier temps, nous avons implémenté l'algorithme du blocked Gibbs sampler dans un contexte biologique. L'utilisation de cet algorithme a donné des résultats intéressants pour l'interprétation de plusieurs phénomènes biologiques de la protéine Syk. Ces résultats ont d'ailleurs contribué à l'élabo-

ration d'un article en collaboration et en préparation avec des biologistes.

Dans un second temps, une généralisation de cet algorithme a été développée dans le contexte de processus de Dirichlet hiérarchiques. Cet algorithme a été mis en œuvre pour répondre à un problème soulevé par les médecins dans le cadre d'un projet ANR-Institut de Recherche en Santé Publique. Son but était de différencier les hôpitaux en fonction de la distance parcourue par les patients. Cette approche a été concluante et a permis de mettre en avant certains hôpitaux de la région Languedoc-Roussillon en les considérant comme attractifs. Les hôpitaux attractifs correspondant aux établissements où les patients parcourent de grandes distances.

Cette approche bayésienne non paramétrique est d'autant plus intéressante qu'elle permet de prendre en compte l'information partagée entre les différents groupes pour déterminer les caractéristiques de l'ensemble des groupes. Cette approche constitue donc une méthode de modélisation non paramétrique intéressante pour le traitement de plusieurs groupes de données liées.

Ces deux applications donnent des exemples d'utilisation des méthodes bayésiennes non paramétriques dans divers contextes.

Une perspective intéressante est à envisager : l'intégration de la méthode de Lojoi et al. [44] dans le cadre des processus de Dirichlet hiérarchiques. En effet, nous nous sommes intéressés à cette approche qui semble donner de bons résultats.

Conclusion

L'objectif de ce travail de thèse a été d'étudier plusieurs méthodes de modélisation bayésienne avec des applications en recherche clinique.

Ainsi la plupart des chapitres se sont construits autour d'hypothèses, ou problématiques soulevées par nos collégues médecins, biologistes, épidémiologistes. Une première étape naturelle a été de commencer par introduire la statistique bayésienne. De nombreux concepts fondamentaux sont à connaître, à maîtriser. Ainsi les notions de distributions *a priori*, *a posteriori* ont été introduites. Dans une suite logique, une partie a été consacrée à la théorie de la décision en analyse bayésienne.

Ces rappels étaient nécessaires pour introduire le travail effectué lors d'un stage doctoral à l'Université de Sherbrooke et en collaboration avec Éric Marchand. Ce travail a permis d'étudier les intervalles de confiance bayésiens pour un paramètre contraint. Ces problèmes ont été étudiés par Marchand et Strawderman [49] et ont permis d'obtenir des bornes inférieures pour les probabilités de recouvrement fréquentistes associées à l'intervalle de confiance HPD. Le contexte dans lequel nous nous sommes plaçés était légèrement différent par la forme de la contrainte sur le paramètre.

Lors des deux mois passés à Sherbrooke nous avons réussi à établir des propriétés théoriques pour la distribution *a posteriori* du paramètre étudié. En ce qui concerne le comportement de la probabilité de recouvrement fréquentiste pour l'intervalle ad hoc, des approximations ont été obtenues. Des perspectives pour ce travail sont à envisager. En effet bien que certains de nos résultats reposent sur des approximations, une tendance positive en ressort.

Ainsi l'un des premiers objectifs est donc d'obtenir des résultats théoriques sur le comportement de la probabilité de recouvrement de l'intervalle ad hoc. Ces premiers résultats permettront d'obtenir les intervalles de croissance et de décroissance et éventuellement une borne inférieure. De plus, l'intervalle d'interêt étant l'intervalle HPD, un autre objectif est de déterminer analytiquement cet intervalle et d'en déduire sa probablité de recouvrement. Ainsi ces travaux préléminaires ont permis "d'explorer" une forme de contrainte qui semble intéressante.

La dernière partie de ce premier chapitre a été dédiée à deux méthodes d'approximation : les méthodes de Monte Carlo par chaînes de Markov et le filtrage particulaire. Bien que motivées par des problèmes différents, ces deux méthodes sont basées sur le même principe de simulation de Monte Carlo. Les méthodes particulaires sont utilisées lorsque l'on souhaite traiter les données en ligne, ce sont des méthodes séquentielles. A contrario, les méthodes MCMC

traitent les données *a posteriori*.

Un travail realisé lors de mon stage de Master a donc été brievement présenté. Ce travail a permis de comparer ces deux types de méthodes dans le cas particulier d'un modèle d'évolution de la biomasse. Les résultats obtenus avec les méthodes MCMC et les méthodes particulaires sont similaires et suggèrent de s'intéresser au contrôle du nombre de particules dans l'approximation particulaire.

Le deuxième chapitre a permis de traiter plusieurs questions posées en recherche clinique. Le modèle de Cox et le modèle logistique ont été mis en œuvre. Ces deux modèles permettent d'établir diverses questions ou questionnements notamment au sujet de la modélisation linéaire ou log linéaire qui peut s'avérer trop restrictive et pas en adéquation avec ce que l'on cherche à modéliser.

En parallèle, un autre objectif fut déterminé : avoir une interprétation simple des résultats dans le but de déterminer des valeurs seuils.

Ainsi l'idée a été de se tourner vers la représentation B-spline, ce qui explique la structure du deuxième chapitre qui commence par un rappel sur ces deux modèles et sur les splines.

La représentation par les B-splines est donc une réponse aux différents questionnements en permettant un ajustement par des polynômes par morceaux. Cependant, un problème venant des fonctions splines concerne le choix du nombre et de la position des nœuds. La plupart des méthodes utilisent un critère de sélection de modèles pour déterminer le nombre de nœuds.

L'objectif, ici, a été de déterminer ces deux paramètres dans un même algorithme. Une approche bayésienne a été mise en place : le Reversible Jump Markov Chain Monte Carlo (RJMCMC) ou méthode Monte Carlo par chaînes de Markov à sauts réversibles. Cette méthode est une méthode MCMC qui diffère un peu des méthodes classiques en permettant des simulations selon des distributions cibles sur des espaces de dimension variable.

Ainsi, cette méthode MCMC associée à la représentation B-spline dans le modèle de Cox et le modèle logistique a fait l'objet des deux articles présentés dans ce chapitre. Le premier a été accepté dans *Communication in Statistics, Theory and Methods*. Les résultats obtenus après implémentation de l'algorithme avec le logiciel *R*, sur différents jeux de données, ont permis une interprétation simple et sont en adéquation avec les études précédentes. Ce deuxième chapitre a permis de mettre en place une méthode MCMC dans le cadre de modèles couramment utilisés en recherche clinique.

Le troisième chapitre se place dans un cadre plus général de la régression spline. L'objectif était d'appliquer une méthode de sélection de modèle proposée par Birgé et Massart ([8], [9]) pour la régression spline. Bien que ce chapitre sorte un peu du cadre bayésien, il permet tout de même une approche intéressante pour la régression spline et le choix du nombre et de la position des nœuds en développant à partir des données un critère des moindres carrés pénalisé.

Ce critère permet donc d'obtenir une méthode donnant, dans certaines situations, des résultats aussi bons voir meilleurs que le C_p de Mallows et le critère BIC. Ces recherches ont permis d'écrire un article qui est en révision dans le journal de la Sfds. Dans cet article, nous développons la forme de la pénalité minimale pour le cas de la régression spline. Cette pénalité dépend de deux

constantes que nous avons estimées par simulations. De plus, différentes simulations ont été réalisées et ont permis d'accéder à la performance de la méthode proposée. Les résultats montrent clairement que l'heuristique des pentes donne de bons résultats en terme de risque quadratique.

Certaines améliorations doivent être envisagées. D'une part, l'étude d'une procédure de simulation plus pertinente pour l'estimation des constantes K_1, K_2 doit être développée. D'autre part, il se peut que dans certaines situations, l'estimation de la pente de la partie linéaire soit difficile à estimer, de même la détection du "bon" saut peut s'avérer problématique. Il serait donc intéressant de développer une méthode de calibration fonctionnant dans toutes les situations possibles.

Enfin, le quatrième chapitre a permis de présenter les applications possibles des méthodes statistiques bayésiennes non paramétriques. Une première application en oncologie des modèles de mélange de processus de Dirichlet a été réalisée. Les résultats obtenus par cette méthode ont permis d'interpréter le comportement de la protéine Syk qui est une protéine vitale. Ces travaux ont d'ailleurs donné lieu à l'élaboration d'un article en collaboration avec les biologistes. L'algorithme qui a été mis en œuvre avec le logiciel R est le blocked Gibbs sampling. Ce même algorithme a été développé et utilisé dans un autre contexte.

En effet, dans le cadre d'un projet ANR-Institut de Recherche en Santé Publique, nous avons obtenu un ensemble de données représentant la distance moyenne parcourue par les patients pour une pathologie et un hôpital donné. Le problème soulevé était le suivant : y a-t-il des hôpitaux plus attractifs que d'autres ? quels sont ces hôpitaux ? Pour répondre à ces questions nous avons utilisé les mélanges de processus de Dirichlet hiérarchiques. Cette approche permet le traitement de plusieurs groupes de données liés et utilise l'information apportée par chaque groupe. Nous avons donc adapté et implémenté le blocked Gibbs sampler dans ce contexte.

Les résultats obtenus sur les données du médecin ont permis de mettre en avant certains hôpitaux, ces résultats sont cohérents avec les hypothèses formulées et donnent un exemple d'application concret des méthodes bayésiennes non paramétriques.

Une perspective qui semble intéressante et à laquelle nous avons commencé à nous intéresser est l'adaptation de la méthode de Lojoi et al. [44] dans le cadre des processus de Dirichlet hiérarchiques. Cette méthode propose une nouvelle approche pour le paramètre de précision α_0, avec notamment la mise en place d'une procédure de renforcement pour donner un poids supplémentaire aux classes ayant la plus forte féquence.

Un des autres domaines pour lequel les méthodes de modélisation bayésienne semblent appropriées est l'environnement. Ainsi mon nouvel emploi au CIRAD va me permettre de mettre en œuvre les différentes méthodes développées dans le cadre des Eucalyptus et des Palmiers à huile.

Bibliographie

[1] H Akaike. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (Tsahkadsor,* 1971*)*, pages 267–281, 1973.

[2] H. Akaike. A new look at thes staatistical model identification. *IEEE Trans. Automatic Control*, AC-19(System identification and time-series analysis) :716–723, 1974.

[3] C.E. Antoniak. Mixtures of Dirichlet processes with applications to bayesian nonparametric problem. *Annals of Statistics*, 2 :1152–1174, 1974.

[4] S. Arlot and P. Massart. Data-driven calibration of penalties for least-quares regression. *Journal of Machine Learning Research*, 10 :245–279, 2009.

[5] B. C. Arnold and R. J. Beaver. *Elliptical models subject to hidden truncation and selective sampling.* 2004.

[6] J. Berger. *Statistical Decision Theory and Bayesian Analysis.* 1985.

[7] J.M. Bernardo. Reference posterior distributions for bayesian inference (with discussion). *Journal Royal Statist. Soc. (Ser. B)*, 41(113–147), 1979.

[8] M. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc*, 3 :203–268, 2001.

[9] M. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability Theory Related Fields*, 138(1-2) :33–73, 2007.

[10] David Blackwell and James B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2) :353–355, 1973.

[11] Kevin Bosa. Estimation bayésienne par intervalle sur un espace paramétrique restreint. Master's thesis, Faculté des sciences, Université de Sherbrooke, 2007.

[12] A.V. Boyd. Inequalities for mill's ratio. *Rep. Stat. Appl. Res.*, (6) :44–46, 1959.

[13] Stephen P. Brooks and Gareth O. Roberts. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8(4) :319–335, 1998.

[14] L.D. Brown. *Foundations of Exponential Families*, volume 6 of *IMS Lecture Notes — Monograph Series.* IMS, Hayward, 1986.

[15] G. Casella and R. Berger. *Statistical Inference.* Wadsworth, Belmont, CA, second edition, 2001.

[16] M.K. Cowles and B.P. Carlin. Markov chain monte carlo convergence diagnostics : A comparative review. *Journal of the American Statistical Association*, (91) :883–904, 1996.

[17] D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, (34) :187–220, 1972.

[18] A. DasGuptas. *Asymptotic Theory of Statistics and Probability*. Springer, 2008.

[19] C. de Boor. *A Practical Guide to Splines*. New-York : Springer-Verlag, 1978.

[20] Marie Denis. Analyse bayésienne de modèles d'évolution de ressources naturelles. Master's thesis, Université de Montpellier 1, 2007.

[21] R.M. Dorazio. On selecting a prior of the precision parameter of Dirichlet process mixture models. *Journal of Statistical Planning and Inference*, (139) :3384–3390, 2009.

[22] Arnaud Doucet, Nando de Freitas, and Neil J. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer–Verlag, New York, 2001.

[23] Fine J. et Saporta G. Droesbeke, J-J. *Méthodes Bayésiennes en statistique*. Editions Technip, 2002.

[24] Lebarbier E. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85 :717–736, 2005.

[25] M. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89 :268–277, 1994.

[26] M. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90 :577–588, 1995.

[27] T.S Ferguson. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1 :209–230, 1973.

[28] T.S. Ferguson. Prior distributions in spaces of probability measures. 2 :615–629, 1974.

[29] D. Fourdrinier. *Statistique Inférentielle*. 2002.

[30] D. Gamerman. *Markov Chain Monte Carlo : Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. Chapman & Hall / CRC Press, London, 1997.

[31] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410) :398–409, jun 1990.

[32] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7 :457–511, 1992.

[33] Richardson S. Gilks, W.R. and D.J. Spiegelhalter. *Markov Chain Monte Carlo in practice*. Chapman and Hall, 1996.

[34] Simon J. Godsill, Olivier Cappé, and Eric Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. In *Proceedings of the IEEE s*, volume 95, pages 899–924, 2007.

[35] I.J. Good. *Some history of the hierarchical Bayesian methodology.* In Bayesian Statistics II, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (Eds.), North-Holland, Amsterdam, 1980.

[36] I.J. Good. *Good Thinking : The Foundations of Probability and Its Applications.* University of Minnesota Press, Minneapolis, 1983.

[37] Neil J. Gordon, David J. Salmond, and Adrian F. M. Smith. Novel approach to nonlinear/non–Gaussian Bayesian state estimation. *IEE Proceedings, Part F*, 140, 1993.

[38] C. Hill, C. Com-Nougué, A. Kramar, T. Moreau, J. O'quigley, R. Senoussi, and C. Chastang. *Analyse statistique des données de survie.* 1990.

[39] K.L. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76 :297–307, 1989.

[40] H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(161–173), 2001.

[41] H. Ishwaran and L.F. James. Approximate dirichlet process computing in finite normal mixtures : Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3) :508–532, 2002.

[42] H. Ishwaran and M. Zarepour. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87 :371–390, 2000.

[43] H. Jeffreys. *Theory of Probability (third edition).* 1961.

[44] Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. Controlling the reinforcement in bayesian non-parametric mixture models. *Journal Of The Royal Statistical Society Series B*, 69(4) :715–740, 2007.

[45] A.Y. Lo. On a class of bayesian nonparametric estimates : I, density estimates. *Ann. Statist.*, 12 :351–357, 1984.

[46] Steven N. MacEachern and Peter Müller. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2) :223–238, 1998.

[47] C. Mallows. Some comments on cp. *Technometrics*, 37 :362–372, 1973.

[48] É. Marchand and W.E. Strawderman. Estimation in restricted parameter spaces : A review. *Festschrift for Herman Rubin, Institute of Mathematical Statistics Lecture Notes-Monograph Serie*, pages 21–44, 2004.

[49] É. Marchand and W.E. Strawderman. On the behaviour of Bayesian credible intervals for some restricted parameter space problems. *IMS Lecture Notes - Monograph Series*, pages 112–126, 2006.

[50] É. Marchand, W.E. Strawderman, K. Bosa, and A. Lmoudden. On the frequentist coverage of bayesian credible intervals for lower bounded means. *Electronic Journal of Statistics, 2*, pages 1028–1042, 2008.

[51] K.L. Mengersen, C.P. Robert, and C. Guihenneuc-Jouyaux. *MCMC convergence diagnostics : A review (with discussion).* In Bayesian Statistics 6, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.), 1999.

[52] Pietro Muliere and Luca Tardella. Approximating distributions of random functionals of ferguson-dirichlet priors. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 26(2) :283–297, 1998.

[53] R.M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2) :249–265, 2000.

[54] O. Papaspiliopoulos and G. O. Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical mode. *Scandinavian Journal of Statistics*, 30 :241–251, 2000.

[55] Jim Pitman and Marc Yor. The two-parameter Poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25(2) :855–900, 1997.

[56] Michael K. Pitt and Neil Shephard. Auxiliary variable based particle filters. In Arnaud Doucet, Nando de Freitas, and Neil J. Gordon, editors, *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science, chapter 13, pages 273–293. Springer–Verlag, New York, 2001.

[57] C.P. Robert. *The Bayesian Choice*. 1994.

[58] C.P. Robert. *Convergence assessment for Markov Chain Monte Carlo algorithms*. 1995.

[59] C.P. Robert. *Méthodes de Monte Carlo par Chaînes de Markov*. Economica, Paris, 1996.

[60] C.P. Robert. *Discretization and MCMC convergence assessment*. Springer-Verlag, New York, 1998.

[61] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Berlin, 1999.

[62] Gareth O. Roberts and Jeffrey S. Rosenthal. Markov chain Monte Carlo. In S. Asmussen, editor, *Encyclopedia of the Actuarial Sciences*, 2003.

[63] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1, 2004.

[64] H. Ruben. A convergent asymptotic expansion for Mill's ratio and the normal probability integral in terms of rational functions. *Mathematische Annalen*, 4 :355–364, 1963.

[65] J. Sethuraman. A constructive definition of Dirichlet priors. *Statist. Sinica*, 2 :639–650, 1994.

[66] D. Sorensen, S. Andersen, D. Gianola, and I.R. Korsgraard. Bayesian inference in threshold models using gibbs sampling. *Genetics, Selection, Evolution*, 27 :229–249, 1995.

[67] Jordan M.I. Beal M.J. Teh, Y.W. and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006.

[68] Y.W. Teh and M.I. Jordan. Hierarchical bayesian nonparametric models with application. *The Annals of Statistics*, 30(3) :631–682, 2008.