JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

AAD VAN DER VAART

Vitesses de convergence de mesures a posteriori

Journal de la société française de statistique, tome 145, n° 1 (2004), p. 7-30

<http://www.numdam.org/item?id=JSFS_2004__145_1_7_0>

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

VITESSES DE CONVERGENCE DE MESURES A POSTERIORI

CONFÉRENCE LUCIEN LE CAM 2002

Aad van der VAART *

RÉSUMÉ

Cet article est une version écrite de la « Conférence Lucien Le Cam » prononcée aux « Journées de Statistique » en 2002 à Bruxelles. Nous considérons des procédures bayésiennes pour des modèles semi- ou non-paramétriques et essayons de caractériser la vitesse de convergence de la mesure a posteriori en utilisant des métriques globales. Nous présentons ces résultats dans un cadre abstrait emprunté à Le Cam, mais donnons aussi des exemples concrets liant les résultats aux développements récents des méthodes numériques du type MCMC.

Mots clés : Entropie de Kolmogorov, dimension de Le Cam, distance de Hellinger, mélange gaussien, statistique bayésienne non-paramétrique, processus de Dirichlet.

Classification AMS : 62G07, 62G15, 62G20, 62F25

ABSTRACT

This paper is a written version of the "Lucien Le Cam Lecture" delivered at the "Journées de Statistique" in 2002 in Brussels. We consider Bayes procedures for semi- and nonparametric models and characterize the speed of convergence of the posterior measure relative to global metrics. We present these results in an abstract setting due to Le Cam, but also give examples that relate the results to recent developments in numerical simulation (MCMC) of posterior distributions.

Keywords : Entropie de Kolmogorov, dimension de Le Cam, distance de Hellinger, mélange gaussien, statistique bayésienne non-paramétrique, processus de Dirichlet.

Classification AMS : 62G07, 62G15, 62G20, 62F25

1. Introduction

Un théorème célèbre en Théorie de la Décision Statistique énonce que l'ensemble des procédures bayésiennes est complet. C'est-à-dire que, pour toute procédure statistique, il existe une procédure bayésienne dont les performances sont au moins aussi bonnes. Ce théorème des classes complètes revient à Wald dans les années 1940-50 et a aussi trouvé son expression dans le travail de Lucien Le Cam, dans un cadre abstrait et plus général (voir par exemple Le Cam (1964)). On peut se demander si ce théorème s'étend aux modèles statistiques non-paramétriques et semi-paramétriques, qui ont été ajoutés au

* Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, Pays-Bas.
E-mail : aad@cs.vu.nl

« toolbox » statistique dans les quarante années qui ont suivi la fondation de la Théorie de la Décision Statistique. Dans ces modèles, le paramètre inconnu (par exemple une fonction de densité ou une fonction de répartition totalement non spécifiée) prend ses valeurs dans un ensemble de dimension infinie plutôt que dans un espace Euclidien.

Une première réponse à cette question est affirmative. En effet, la théorie de Wald et Le Cam a été écrite dans un langage et utilise une notation qui ne font guère mention de la dimension du paramètre. Celui-ci peut prendre ses valeurs dans un espace abstrait et, en particulier dans la théorie de Le Cam, la fonction de perte est entièrement générale.

Néanmoins, en inspectant la théorie de plus près, on se rend compte que le théorème des classes complètes est typiquement évoqué d'une façon assez imprécise. On a suivi cette mauvaise habitude ci-dessus, le fait important étant que la fermeture de la classe des procédures bayésiennes, plutôt que l'ensemble des procédures bayésiennes lui-même, est une classe complète. C'est-à-dire que toute procédure statistique est dominée par la limite d'une suite de procédures bayésiennes. Prenons par exemple la moyenne empirique de n observations d'une loi gaussienne $N(\theta, 1)$, estimateur « attitré » de θ. Cet estimateur est la limite des estimateurs bayésiens associés à un a priori $N(0, \tau)$ lorsque $\tau \to \infty$, mais n'est pas lui-même un estimateur bayésien. Cette nécessité d'une opération de fermeture ne semble pas avoir de conséquences trop importantes dans le cas des modèles de dimension finie classiques, mais pourrait bien mitiger l'enthousiasme pour l'utilisation du principe de Bayes dans les modèles semi- et non-paramétriques. Le fait est que la topologie sous laquelle l'opération de fermeture doit se faire est « faible », et est d'autant plus « faible » que l'ensemble des paramètres est « grand ». Il serait possible que la classe des procédures bayésiennes soit complète, mais dans un sens trop faible pour être d'aucun intérêt pratique.

L'idée de problèmes potentiels avec les procédures bayésiennes pour des modèles de dimension infinie s'est installée avec plus de rigueur à la suite des travaux de Freedman (1963) et de Diaconis et Freedman (1986). Une première propriété remarquable des procédures bayésiennes est que la mesure a posteriori « rétrécit » vers la vraie valeur du paramètre si le nombre d'observations tend vers l'infini. Dans ce cas la mesure a priori est corrigée par les données et son influence disparaît quand ces dernières deviennent dominantes, permettant ainsi de connaître asymptotiquement la vraie valeur du paramètre. Un autre théorème célèbre, démontré par Doob (1948), montre, dans un cadre général, que, pour toute mesure a priori, ce comportement se vérifie pour presque toute vraie valeur du paramètre ; ceci constitue un argument de poids en faveur des méthodes bayésiennes. Diaconis et Freedman ont démontré que cet « presque toute vraie valeur » peut exclure, cependant, de « grands » ensembles de valeurs du paramètre. Les exemples dans le cas des modèles non paramétriques sont particulièrement embarrassants. De plus Diaconis et Freedman ont démontré que ces problèmes sont typiques de la plupart des mesures a priori, et ne peuvent être évités que pour un sous-ensemble

de mesures a priori assez « mince ». Même s'il existe peut-être de « bonnes » mesures à priori, les « mauvaises » mesures sont certainement abondantes.

Arrivés à ce point, il nous faut expliquer que, dans cet article, nous nous plaçons dans un cadre purement fréquentiste. Même si nous étudions des procédures bayésiennes, nous supposons que les données ont été produites par un mécanisme aléatoire gouverné par une loi de probabilité fixée, donnée par un paramètre que nous notons θ. Nous considérons le cadre bayésien comme un outil permettant de construire des procédures statistiques (estimateurs, tests, etc.), sans adopter pour autant la « philosophie bayésienne ». En effet il y a peu à dire de la qualité des procédures bayésiennes si on accepte le paradigme bayésien « philosophique ».

Le problème qui nous concerne est de caractériser les mesures a priori qui produisent des procédures statistiques fréquentistes convergentes, et de caractériser la vitesse de convergence des mesures a posteriori vers la vraie loi des données en fonction du modèle statistique et de la mesure a priori. Ces questions sont très appropriées dans le cadre d'une « Conférence Lucien Le Cam ». En vue du rôle central joué par les procédures bayésiennes en Théorie de la écision Statistique, Le Cam en effet s'y est intéressé pendant toute sa carrière. Le résultat fondamental sur la convergence des mesures a posteriori a été obtenu par Schwartz (1965) dans la proximité de Le Cam. Des résultats sur les vitesses de convergence sont contenus dans les travaux de Le Cam (1956, 1964). Voir Le Cam (1986) pour une présentation finale.

Ces questions ont été reprises récemment avec une attention accrue pour des modèles de dimension infinie. Voir par exemple le livre de Ghosh et Ramamoorthy (2003), qui s'adresse à la construction de mesures a priori concrètes et à la question de leur convergence. Voir aussi les articles de Ghosh, Ghosal et van der Vaart (2000), Ghosal et van der Vaart (2002, 2003). Dans ces articles, les innovations par rapport aux résultats de Le Cam portent sur l'obtention de bornes supérieures plus fines sur la vitesse de convergence (essentielles pour les modèles de dimension infinie) ; un grand nombre de mesures a priori concrètes y sont également étudiées.

Un intérêt renouvelé pour les procédures bayésiennes et pour les mesures a priori concrètes s'est manifesté en liaison avec le développement, dans les années 1990, des méthodes dites MCMC (« Markov Chain Monte Carlo »). Ces méthodes (voir par exemple les livre de Robert (2001), ou Liu (2001)) ont libéré la machinerie bayésienne des difficultés de mise en pratique numérique, qui ont longtemps terni sa popularité. Le calcul d'une mesure a posteriori exige en effet, par application de la règle de Bayes, la détermination de l'intégrale de la fonction de vraisemblance par rapport au paramètre et à la mesure a priori. Avant les progrès récents de l'informatique, le statisticien appliqué était condamné à utiliser une mesure a priori permettant le calcul analytique de cette intégrale. De ce fait, le choix de la mesure a priori était le plus souvent limité aux mesures du type « a priori conjugué ». L'ordinateur a d'abord permis une plus grande liberté dans le choix des mesures a priori, grâce à la possibilité d'évaluer les mesures a posteriori par des méthodes d'intégration numérique. Cependant, pour des paramètres de

dimension élevée cette intégration numérique reste difficile à entreprendre. Il a fallu la conjonction des progrès informatiques et des nouvelles méthodes de simulation de variables aléatoires pour consacrer l'entrée des méthodes bayésienne dans la pratique. À l'heure actuelle, on pourrait affirmer que les estimateurs bayésiens sont devenus plus populaires que l'estimateur du maximum de vraisemblance, en particulier pour les modèles compliqués, où le calcul numérique du maximum de vraisemblance est souvent considéré plus difficile que le calcul par simulation de la mesure a posteriori.

Dans les dix dernières années, une abondante activité de recherche a été consacrée à la mise en pratique du calcul MCMC dans la détermination des mesures a posteriori (un signe clair que ceci n'est pas tout à fait aussi simple que le suggère la théorie générale). Ces recherches s'étendent aussi aux modèles et mesures a priori de dimension infinie. Presque tout l'effort s'est concentré sur la question du calcul de la mesure a posteriori associée à un ensemble de données et une mesure a priori. Un algorithme MCMC produit une chaîne de Markov dont la loi stationnaire est la loi a posteriori recherchée. Comme il est impossible de commencer la chaîne à l' équilibre, il faut se satisfaire d'un algorithme qui converge rapidement vers l'état stationnaire. L'algorithme est dit convergent si la loi marginale de la chaîne de Markov converge vers la loi a posteriori. Dans ce cas on sait qu'après avoir répété les itérations de l'algorithme assez longtemps, l'on produit des variables ayant (à peu près) la loi a posteriori. Cette notion de convergence n'a aucune relation avec la convergence considérée dans cet article, qui est la convergence de la mesure a posteriori quand les données se multiplient.

Une étude de ces propriétés asymptotiques fréquentistes est particulièrement importante pour les modèles semi- et non paramétriques, parce qu'il est difficile d'avoir une bonne intuition pour une mesure a priori définie sur un ensemble de paramètres d'une grande complexité. Par exemple, on a une bonne idée d'une mesure a priori $N(0, \tau)$ pour un paramètre réel, mais il est presque impossible d'avoir une intuition comparable pour la mesure a priori dite « processus de Dirichlet », qui est définie sur l'ensemble de toutes les mesures de probabilité sur un espace mesurable donné.

Dans cet article nous suivons la méthode de Le Cam en commençant par une description abstraite (Section 2). Ensuite nous considérons les implications dans le cadre des données indépendantes et équidistribuées, et donnons quelques exemples concrets. Nous finissons avec une discussion du choix de modèles et l'adaptation dans le cadre bayésien, qui, à notre avis, fournissent une motivation convaincante du paradigme bayésien.

1.1. Notation

Le symbole $\lesssim$ signifie « plus petit à une constante multiplicative près ». Pour deux fonctions non-négatives f et g et une mesure dominante fixée μ, les notations $h(f,g)$, $K(f,g)$ et $V(f,g)$ sont réservées à la distance de Hellinger, la divergence de Kullback-Leibler, et « la variance de la divergence » entre f et g, définies par

$$h^2(f,g) = \int (f^{1/2} - g^{1/2})^2 \, d\mu, \tag{1.1}$$

$$K(f,g) = \int \left(\log \frac{f}{g}\right) f \, d\mu, \tag{1.2}$$

$$V(f,g) = \int \left|\log \frac{f}{g} - K(f,g)\right|^2 f \, d\mu. \tag{1.3}$$

2. Résultat Principal

Soit $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)} : \theta \in \Theta)$ une suite d'expériences statistiques ($n = 1, 2, \ldots$) où le paramètre θ appartient à un ensemble Θ quelconque. C'est-à-dire que $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)})$ est un espace mesurable et que, pour tout $\theta \in \Theta$, $P_\theta^{(n)}$ est une mesure de probabilité définie sur $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)})$. Dans le paradigme bayésien on considère le paramètre θ comme une variable aléatoire distribuée selon une loi Π_n, dite *mesure a priori* sur Θ, et on suppose qu'on a des données $X^{(n)}$ possédant une loi conditionnelle $P_\theta^{(n)}$ étant donné θ. Ceci fait apparaître une loi conditionnelle $\Pi_n(\cdot | X^{(n)})$ de θ étant donné $X^{(n)}$, qui s'exprime selon la règle de Bayes. Si $P_\theta^{(n)}$ est déterminée par une fonction de densité $p_\theta^{(n)}$, cette *mesure a posteriori* est donnée par

$$\Pi_n(B \mid X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)}) \, d\Pi_n(\theta)}{\int p_\theta^{(n)}(X^{(n)}) \, d\Pi_n(\theta)}, \qquad B \in \mathcal{B}. \tag{2.1}$$

De toute évidence, il est nécessaire d'équiper l'ensemble Θ d'une structure de mesurabilité et de poser des conditions assurant que cette mesure est bien définie. On suppose que $(\theta, \mathcal{B})$ est un espace mesurable et que la fonction $(x, \theta) \mapsto p_\theta^{(n)}(x)$ est mesurable par rapport à la tribu-produit $\mathcal{A}^{(n)} \otimes \mathcal{B}$.

Dans cet article, la mesure a posteriori est définie par (2.1), mais nous n'adoptons pas le cadre bayésien. Nous supposons que l'observation $X^{(n)}$ est distribuée selon une loi $P_{\theta_0}^{(n)}$ pour une valeur fixée θ_0 du paramètre, dit « vrai paramètre », et étudions les propriétés de la loi aléatoire $B \mapsto \Pi(B \mid X^{(n)})$, une fonction de $X^{(n)}$, sous cette loi $P_{\theta_0}^{(n)}$.

Si les mesures a priori n'excluent pas le paramètre θ_0, et si les expériences statistiques deviennent de plus en plus informatives, on peut s'attendre à ce que les mesures a posteriori se concentrent autour de θ_0 quand $n \to \infty$. Dans ce cas les mesures a posteriori sont dites convergentes. La convergence est

une propriété très désirable, mais n'est pas en elle-même une garantie de bons résultats pratiques. La *vitesse de convergence* vers θ_0 est plus significative. On peut la mesurer par le rayon de la plus petite boule centrée en θ_0 qui est capable de capturer (presque) toute la probabilité a posteriori.

Notre résultat principal est une borne supérieure sur cette vitesse de convergence, exprimée en fonction de la complexité du modèle statistique et de la concentration autour de θ_0 de la mesure a priori. Le théorème s'applique partout où existent des tests avec erreurs décroissantes à vitesse exponentielle lorsque les hypothèses s'éloignent. Ces tests ont été introduits par Le Cam, et ont été utilisés par Le Cam (1973, 1975, 1986) et Birgé (1983a, 1983b) pour construire des estimateurs optimaux. Dans cet article ils sont des outils abstraits qui caractérisent les types d'expériences statistiques, et surtout les métriques utilisées pour mesurer la vitesse de convergence, pour lesquelles nous savons établir des résultats. Nous n'avons pas besoin de leur forme concrète et on pourrait sauter le paragraphe suivant.

Pour chaque n on se donne une semi-métrique d_n sur Θ telle qu'il existe des constantes universelles positives ξ et K telles que, pour chaque $\varepsilon > 0$: pour tout $\theta_1 \in \Theta$ tel que $d_n(\theta_1, \theta_0) > \varepsilon$ il existe un test ϕ_n satisfaisant

$$\begin{aligned} P_{\theta_0}^{(n)} \phi_n &\leqslant e^{-Kn\varepsilon^2}, \\ \sup_{\theta \in \Theta : d_n(\theta, \theta_1) < \varepsilon\xi} P_\theta^{(n)}(1 - \phi_n) &\leqslant e^{-Kn\varepsilon^2}. \end{aligned} \tag{2.2}$$

(Un test est une fonction mesurable $\phi : \mathcal{X}^{(n)}) \to [0, 1]$, par exemple la fonction indicatrice d'une région de rejet.) Cette condition permet de séparer le vrai paramètre θ_0 d'un autre paramètre θ_1 par un test aux deux erreurs bornées par $\exp(-Knd_n^2(\theta_0, \theta_1))$. De plus on est capable de tester θ_0 contre une petite boule $\{\theta : d_n(\theta, \theta_1) \leqslant \varepsilon\xi\}$ autour de θ_1. (Voir Figure 1.) Les valeurs de ξ et K sont sans importance. Des tests satisfaisant (2.2) avec des distances naturelles existent dans plusieurs cadres : observations indépendantes, chaînes de Markov, processus de diffusion, séries chronologiques gaussiennes, etc. On le verra pour les modèles d'observations indépendantes et équidistribuées dans la Section 3.

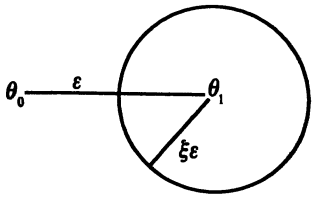

FIG 1. — **Hypothèse nulle et contre-hypothèse utilisées pour les tests.**

Un résultat fondamental de Le Cam (1973, 1975, 1986) et Birgé (1983a, 1983b, 2003) montre que la vitesse de convergence (dans un sense minimax) des meilleurs estimateurs de θ par rapport à la distance d_n est donnée par l'entropie locale, que nous appellerons ici *dimension de Le Cam* de l'ensemble

Θ. Cette dimension est une fonction qui donne pour chaque (petite) valeur de $\varepsilon > 0$ le logarithme du nombre minimal de boules de rayon $\varepsilon\xi$ nécessaire pour couvrir une boule de rayon ε autour de θ_0. Par exemple, dans un espace Euclidien de dimension d une « boule carrée » de rayon ε se laisse couvrir par 2^d boules de rayon $\varepsilon/2$ (voir Figure 2 pour le cas $d = 2$). En conséquence, la dimension de Le Cam est égale à $\log 2^d \sim d$, la « dimension ordinaire » de l'espace, pour chaque valeur de ε. Pour des espaces plus grands la dimension de Le Cam dépend de $\varepsilon > 0$ et typiquement croît vers l'infini si $\varepsilon \downarrow 0$.

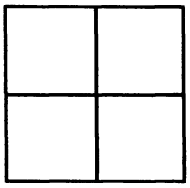

FIG 2. — Couverture d'une boule de rayon ε en dimension 2 avec 2^2 boules de rayon $\varepsilon/2$.

Pour comprendre la dimension de Le Cam pour des espaces plus compliqués, il est utile d'introduire l'*entropie de Kolmogorov*. Si on note $N(\varepsilon, \Theta, d)$ la cardinalité minimale d'une couverture de Θ par des d-boules de rayon ε, l'entropie de Kolmogorov est la fonction $\varepsilon \mapsto \log N(\varepsilon, \Theta, d)$ (voir Figure 3). Cette fonction est décroissante en ε, et on a pour presque chaque espace Θ que $\log N(\varepsilon, \Theta, d) \uparrow \infty$ si $\varepsilon \downarrow 0$. On peut mesurer la complexité de l'espace Θ par la vitesse de divergence de l'entropie de Kolmogorov lorsque $\varepsilon \downarrow 0$.

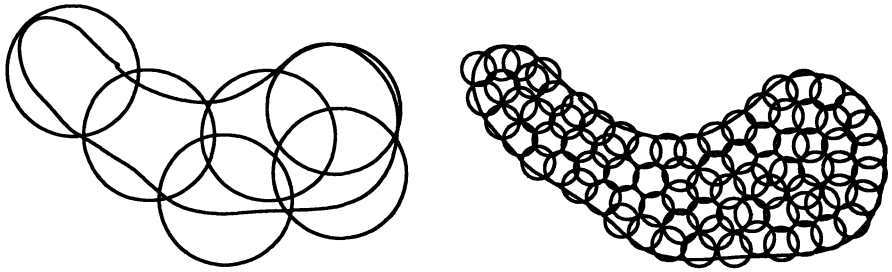

FIG 3. — Couverture d'un ensemble Θ par des boules grandes ou petites.

Exemple 2.1. — On peut couvrir une « boule carrée » de rayon unité dans l'espace Euclidien de dimension d au moyen de boules carrées de taille ε en divisant chacun de ses côtés en $1/\varepsilon$ morceaux de taille ε. (Voir Figure 2.) Comme cette procédure requiert un nombre de boules de l'ordre de $(1/\varepsilon)^d$, l'entropie d'un carré dans l'espace Euclidien de dimension d est de l'ordre $d\log(1/\varepsilon)$. Voir Figure 4 (courbe inférieure) pour l'entropie en fonction de ε pour le cas $d = 1$.

Exemple 2.2. — Considérons l'espace des fonctions $f : [0,1] \to [-1,1]$ qui sont k fois dérivables, avec des dérivées bornées uniformément par 1 en

valeur absolue. Ceci est un espace mesurable pour la métrique uniforme $\|f\|_\infty = \sup_{0 \leqslant x \leqslant 1} |f(x)|$. Kolmogorov et Tikhomirov (1961) ont démontré que l'entropie par rapport à la métrique uniforme est de l'ordre $(1/\varepsilon)^{1/k}$. Voir Figure 4 (courbe supérieure) pour le cas particulier où $k = 1$, c'est-à-dire les fonctions lipschitziennes.

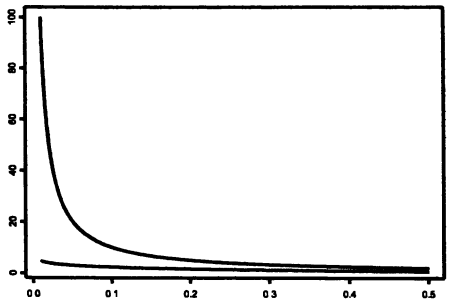

FIG 4. — Entropie d'un carré dans un espace euclidien (courbe inférieure) et entropie de la boule unité dans l'espace des fonctions lipschitziennes de $(0,1)$ dans $[-1,1]$ (courbe supérieure), en fonction de ε (axe horizontal).

On peut trouver beaucoup d'autres exemples dans la littérature récente des processus empiriques (voir van der Vaart et Wellner (1996) ou van der Vaart (1998).)

La dimension de Le Cam peut s'exprimer comme une *entropie de Kolmogorov locale*, définie par

$$\log N(\varepsilon\xi, \{\theta \in \Theta : d_n(\theta, \theta_0) \leqslant \varepsilon\}, d_n).$$

Ceci montre imédiatement que la dimension de Le Cam est bornée par l'entropie de Kolmogorov $\log N(\varepsilon\xi, \Theta, d_n)$ de l'espace Θ entier. Pour beaucoup de modèles de dimension infinie, on ne perd rien en remplaçant la dimension de Le Cam par l'entropie de Kolmogorov, qui est plus intuitive.

Birgé (1983a, 1983b) et Le Cam (1973, 1975, 1986) ont montré qu'il existe des estimateurs $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ tels que $d_n(\hat{\theta}_n, \theta_0) = O_P(\varepsilon_n)$ sous $P_{\theta_0}^{(n)}$, pour $\varepsilon_n \downarrow 0$ satisfaisant

$$\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon\xi, \{\theta \in \Theta : d_n(\theta, \theta_0) \leqslant \varepsilon\}, d_n) \leqslant n\varepsilon_n^2. \tag{2.3}$$

Si (2.3) est satisfaite pour tout $\theta_0 \in \Theta$, alors la vitesse de convergence est uniforme en $\theta_0 \in \Theta$. Sous des conditions générales, cette vitesse est la meilleure possible pour le modèle spécifié par Θ et les lois $P_\theta^{(n)}$.

Ici nous n'avons pas besoin de la force complète du travail de Le Cam et Birgé, mais il nous suffira d'adapter leurs arguments pour établir l'existence de certains tests, qui sont l'ingrédient clé dans la démonstration de notre résultat principal. Noter qu'une suite $\varepsilon_n \downarrow 0$ telle que

$$\log N(\varepsilon_n\xi, \Theta, d_n) \leqslant n\varepsilon_n^2$$

satisfait a fortiori à (2.3). Cette supposition simplifiée suffit pour la plupart des exemples.

Le comportement des mesures a posteriori dépend clairement des mesures a priori. Si celles-ci mettent trop peu de probabilité près de θ_0, la vitesse de convergence se dégrade. Dans notre résultat cet élément est pris en compte au moyen d'une borne sur le quotient des probabilités a priori de grands et de petits voisinages de θ_0. On utilise des voisinages du type

$$B_n(\theta_0, \varepsilon) = \Big\{\theta \in \Theta : K(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \leqslant n\varepsilon^2, V(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \leqslant n\varepsilon^2\Big\}. \tag{2.4}$$

(Les définitions de la divergence de Kullback-Leibler K et de la variance correspondante V sont données en (1.2)-(1.3).)

THÉORÈME 2.1. — *Soit d_n une sémi-métrique sur Θ permettant des tests satisfaisant à (2.2). Supposons que pour une suite de nombres $\varepsilon_n \to 0$ tels que $\liminf n\varepsilon_n^2 > 0$, chaque (grand) entier naturel $j \in$, et des sous-ensembles $\Theta_n \subset \Theta$, les conditions suivantes sont satisfaites :*

$$\sup_{\varepsilon > \varepsilon_n} \log N\Big(\frac{1}{2}\varepsilon\xi, \{\theta \in \Theta_n : d_n(\theta, \theta_0) \leqslant \varepsilon\}, d_n\Big) \leqslant n\varepsilon_n^2,$$

$$\frac{\Pi_n\big(\theta \in \Theta_n : d_n(\theta, \theta_0) \leqslant j\varepsilon_n\big)}{\Pi_n\big(B_n(\theta_0, \varepsilon_n)\big)} \leqslant e^{Kn\varepsilon_n^2 j^2/16}.$$

Alors pour toute suite $M_n \to \infty$ on a

$$P_{\theta_0}^{(n)} \Pi_n(\theta \in \Theta_n : d_n(\theta, \theta_0) \geqslant M_n\varepsilon_n \mid X^{(n)}) \to 0. \tag{2.5}$$

Ici $P_{\theta_0}^{(n)}\phi(X^{(n)})$ désigne l'espérance de la variable $\phi(X^{(n)})$ calculée dans la mesure $P_{\theta_0}^{(n)}$. L'équation (2.5) dit que pour n grand presque toute la masse a posteriori (dans Θ_n) est contenue dans la boule $\{\theta \in \Theta : d_n(\theta, \theta_0) < M_n\varepsilon_n\}$. Comme M_n peut diverger vers l'infini à une vitesse arbitrairement lente, ces boules ont un rayon de l'ordre de $O(\varepsilon_n)$.

La borne (2.4) est presque identique à celle qui figure dans l'équation (2.3) qui décrit la vitesse de convergence optimale pour les expériences statistiques $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)} : \theta \in \Theta_n)$. Cette restriction sur la vitesse de convergence n'a rien à voir avec les méthodes bayésiennes, mais est l'expression de la difficulté d'estimation du paramètre du modèle statistique considéré.

La borne (2.4) a pour interprétation que la mesure a priori doit mettre assez de poids près du paramètre θ_0. Le coté gauche de l'équation est le quotient de la mesure a priori dans un voisinage θ_0 de diamètre $j\varepsilon_n$ par rapport à la mesure a priori d'une boule de de diamètre ε_n et ce quotient doit être borné par une certaine fonction de $j^2 n\varepsilon_n^2$. Nous n'avons pas d'explication intuitive pour la forme de la borne supérieure. Noter que la borne est satisfaite si

$$\Pi_n\big(B_n(\theta_0, \varepsilon_n)\big) \geqslant e^{-n\varepsilon_n^2}. \tag{2.6}$$

Il suffit de remarquer que le numérateur de (2.4), étant une probabilité, est borné par 1.

Le théorème utilise des sous-ensembles $\Theta_n \subset \Theta$ pour assouplir les hypothèses, en particulier la condition d'entropie (2.4). En revanche, l'énoncé du théorème ne concerne que la restriction de la mesure a posteriori aux ensembles $\Theta_n \subset \Theta$, et nous aimerions compléter le théorème par l'énoncé additionnel

$$P_{\theta_0}^{(n)} \Pi_n(\Theta \setminus \Theta_n \mid X^{(n)}) \to 0. \tag{2.7}$$

Ceci est clairement vrai si les ensembles Θ_n sont choisis identiques à Θ pour chaque n, mais dans ce cas la condition d'entropie sera plus restrictive. Dans quelques exemples, il est possible de montrer que la mesure a posteriori des ensembles $\Theta \setminus \Theta_n$ est négligeable par des arguments directs. Comme alternative, le lemme suivant donne une condition simple en fonction de la masse a priori des ensembles $\Theta \setminus \Theta_n$. Le lemme montre essentiellement qu'il n'est pas nécessaire de contrôler le complexité de parties du modèle qui reçoivent très peu de masse a priori.

LEMME 2.2. — *Une condition suffisante pour que (2.7) soit satisfaite est*

$$\frac{\Pi_n(\Theta \setminus \Theta_n)}{\Pi_n\big(B_n(\theta_0, \varepsilon_n)\big)} = o\big(e^{-2n\varepsilon_n^2}\big).$$

3. Observations Indépendantes et Équidistribuées

Tel que Le Cam l'a démontré, le cadre général de la Section 2 s'applique à un grand nombre de structures statistiques. Dans cet article nous nous limitons aux expériences où l'observation est un vecteur $X^{(n)} = (X_1, X_2, \ldots, X_n)$ de variables indépendantes et équidistribuées X_i. Dans ce cas les mesures $P_\theta^{(n)}$ de la Section sont des mesures-produits P_θ^n sur un espace mesurable produit $(\mathcal{X}^n, \mathcal{A}^n)$. On suppose que la loi P_θ a une densité p_θ par rapport à une mesure σ-finie μ sur l'espace mesurable $(\mathcal{X}, \mathcal{A})$.

Dans cette situation la métrique naturelle est la distance de Hellinger h, qui a pour carré

$$h^2(\theta, \theta') = \int \big(\sqrt{p_\theta} - \sqrt{p_{\theta'}}\big)^2 \, d\mu. \tag{3.1}$$

L'existence de tests satisfaisant aux condition (2.2) avec $d_n = h$ est un résultat fondamental de Le Cam (voir Le Cam (1986, pages 475–477)).

La divergence de Kullback-Leibler entre deux mesures produits est la somme des divergences entre les coordonnées individuelles. En conséquence, l'ensemble $B_n(\theta_0, \varepsilon)$ de (2.4) se réduit à

$$B(\theta_0, \varepsilon) = \Big\{\theta \in \Theta : K(\theta_0, \theta) \leqslant \varepsilon^2, \; V(\theta_0, \theta) \leqslant \varepsilon^2\Big\},$$

où $K(\theta_0, \theta) = K(P_{\theta_0}, P_\theta)$ est la divergence de Kullback-Leibler pour une coordonnée, et $V(\theta_0, \theta) = V(P_{\theta_0}, P_\theta)$.

On peut maintenant simplifier le résultat général.

THÉORÈME 3.1. — *Soient $P_\theta^{(n)} = P_\theta^n$ des mesures produits et h la distance Hellinger (3.1). Supposons que pour une suite $\varepsilon_n \to 0$ telle que $\liminf n\varepsilon_n^2 > 0$, tout entier $j \in$ suffisamment grand, et des sous-ensembles $\Theta_n \subset \Theta$,*

$$\sup_{\varepsilon > \varepsilon_n} \log N\big(\varepsilon/36, \{\theta \in \Theta_n : d(\theta, \theta_0) < \varepsilon\}, h\big) \leqslant n\varepsilon_n^2, \tag{3.2}$$

$$\frac{\Pi_n(\Theta \setminus \Theta_n)}{\Pi_n\big(B(\theta_0, \varepsilon_n)\big)} = o\big(e^{-2n\varepsilon_n^2}\big), \tag{3.3}$$

$$\frac{\Pi_n\big(\theta \in \Theta_n : h(\theta, \theta_0) \leqslant j\varepsilon_n\big)}{\Pi_n\big(B(\theta_0, \varepsilon_n)\big)} \leqslant e^{n\varepsilon_n^2 j^2/16}. \tag{3.4}$$

Alors, $P_{\theta_0}^n \Pi_n\big(\theta : h(\theta, \theta_0) \geqslant M_n\varepsilon_n \mid X_1, \ldots, X_n\big) \to 0$ pour chaque $M_n \to \infty$.

Nous avons déja vu que la dimension de Le Cam est majorée par l'entropie de Kolmogorov de l'ensemble des paramètres tout entier. De plus la condition (3.4) peut être remplacée par (2.6), et la condition (3.3) est trivialement satisfaite si $\Theta = \Theta_n$. Nous concluons que les hypothèses du théorème sont satisfaites, et que donc sa conclusion est vérifiée, si

$$\log N\big(\varepsilon_n, \Theta, h\big) \leqslant n\varepsilon_n^2, \tag{3.5}$$

$$\Pi_n\big(B(\theta_0, \varepsilon_n)\big) \geqslant e^{-n\varepsilon_n^2}. \tag{3.6}$$

Ces conditions simplifiées, qui typiquement ne perdent rien en précision, admettent une interprétation de la nature des «bonnes» mesures a priori. Nous avons déja vu que selon les résultats de Le Cam et Birgé la solution minimale ε_n de (3.2) est la vitesse optimale pour l'estimation du paramètre dans l'expérience statistique donnée. Dans sa forme simplifiée la condition (3.5) dit qu'on a besoin de $\exp(n\varepsilon_n^2)$ boules de rayon proportionnel à ε_n pour couvrir l'ensemble des paramètres Θ. Bien que les boules formant une couverture doivent vraisemblablement avoir des intersections non-vides, comme le montre la Figure 5, on peut s'attendre à ce que le nombre de boules disjointes de rayon égal ou inférieur qu'on peut placer dans Θ, comme dans la Figure 3, soit du même ordre $\exp(n\varepsilon_n^2)$, ou peut-être de l'ordre $\exp(cn\varepsilon_n^2)$ pour une constante c. (En effet, le nombre maximal de points $\theta_1, \ldots, \theta_N$ tels que $d(\theta_i, \theta_j) \geqslant \varepsilon$ est plus grand que $N(\varepsilon, \Theta, d)$ et plus petit que $N(\varepsilon/2, \Theta, d)$, et les boules de rayon $\varepsilon/2$ centrées en ces points sont disjointes.) Alors on peut s'attendre à ce qu'on puisse inscrire un nombre de l'ordre $N = \exp(n\varepsilon_n^2)$ de boules disjointes de rayons ε_n en Θ. Si la mesure a priori était «uniforme», alors chacune de ces boules recevrait une partie $1/N = \exp(-n\varepsilon_n^2)$ de la mesure a priori. Dans ce cas la condition (3.6) est satisfaite.

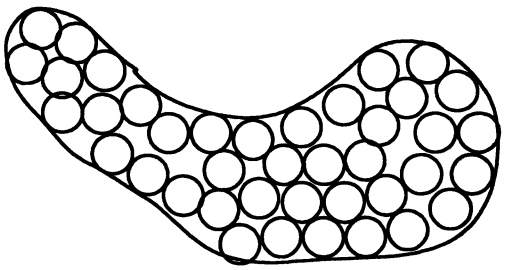

FIG 5. — Boules disjointes inscrites dans un ensemble Θ.

Une interprétation intuitive de la condition que doit satisfaire une bonne mesure a priori est alors que cette loi distribue la masse a priori « uniformément » sur Θ. Si la mesure a priori met trop de probabilité sur une partie de Θ, une autre partie recevra trop peu de masse a priori, et la mesure a posteriori n'atteindra pas la vitesse de convergence optimale. Ici il faut remarquer que des mesures de probabilité qui sont « uniformes » dans un sens exact n'existent que très rarement, même pour les ensembles de paramètres classiques. Dans ce contexte le concept « uniforme » veut dire : chaque voisinage de diamètre de l'ordre ε_n porte une partie proportionelle de la mesure a priori.

Remarquons aussi que nous avons identifié les boules de Hellinger, utilisées en (3.5), avec les voisinages $B(\theta, \varepsilon)$, utilisés en (3.6). Ceci se justifie, par le fait que le carré de la distance de Hellinger, la divergence de Kullback-Leibler et la « distance » V sont souvent comparables.

4. Mélanges de Dirichlet de Lois Gaussiennes

Il y a beaucoup de situations où les résultats des sections précédentes s'appliquent. Dans cet article nous discutons seulement une situation en détail : l'estimation d'une densité en utilisant des mélanges de lois gaussiennes et une mesure a priori de type Dirichlet.

Cette approche est l'approche bayésienne liée à l'estimation non paramétrique de la densité par la méthode du noyau, introduite il y a presque 50 ans par Rosenblatt (1956). Si on choisit un noyau gaussien, l'estimateur à noyau sera un mélange de densités gaussiennes. Une approche bayésienne consiste à mettre une mesure a priori sur les mélanges gaussiens et à calculer la mesure a posteriori par des méthodes MCMC. La méthode a été introduite par Ferguson (1983) et Lo (1984), et les méthodes MCMC ont été développées par West (1992) et Escobar et West (1995).

Notons $\phi_\sigma : x \mapsto \phi_\sigma(x) = (2\sigma^2\pi)^{-1/2} \exp(-x^2/2\sigma^2)$ la densité gaussienne centrée de variance σ^2, et posons

$$p_{F,\sigma}(x) = \int \phi_\sigma(x - z)\, dF(z).$$

On peut mettre une mesure a priori sur les mélanges gaussiens $p_{F,\sigma}$ en mettant une mesure a priori sur les couples (F, σ).

La fenêtre σ est un paramètre réel positif et il est facile de définir une mesure a priori sur $(0, \infty)$. Nous considérons deux possibilités :

(i) $\sigma \sim \pi$ pour une mesure de probabilité π fixée sur un intervalle $[b_1, b_2] \subset (0, \infty)$;

(ii) $\sigma/\sigma_n \sim \pi$ pour la suite $\sigma_n = n^{-1/5}$ et une mesure de probabilité π fixée sur un intervalle $[b_1, b_2] \subset (0, \infty)$.

La vitesse $n^{-1/5}$ en (ii) est la fenêtre classique pour estimer une densité qui est deux fois dérivable en utilisant le noyau gaussien. Nous verrons qu'avec cette fenêtre la procédure bayésienne converge à la même vitesse que les procédures classiques, à un facteur logarithmique près. Le choix (i) d'une mesure a priori fixée est raisonnable quand on cherche à ajuster un mélange de lois gaussiennes fixées. Dans ce cadre il serait plus pratique d'utiliser des mélanges du type $\int \int \phi_\sigma(x - z)\, dF(z)\, dG(\sigma)$, qui permettent aux composantes gaussiennes d'avoir des variances différentes. Avec des mesures a priori processus de Dirichlet sur la variance on obtient des résultats semblables.

Même si le choix de la mesure a priori pour σ est très important, dans le cadre de cet article la mesure a priori pour le paramètre F est plus intéressante. On cherche à mettre une mesure a priori dont le support englobe toutes les mesures de probabilité F sur la droite réelle. Pour le cas (i) la motivation est qu'on ne sait a priori ni les poids des composantes du mélange ni leur localisation sur la droite. Pour le cas (ii) la motivation est que chaque densité arbitraire continue peut être approchée par un mélange p_{σ_n, F_n} pour des paramètres (σ_n, F_n) appropriés. Si la mesure a priori distribue son poids « partout », on est sûr qu'une densité continue arbitraire se laisse approcher par une suite de densités contenues dans le support de la mesure a priori.

Une mesure a priori classique sur l'ensemble de toutes les mesures sur $\mathbb{R}$ est la *mesure processus de Dirichlet*, qu'on peut caractériser de plusieurs façons. Une mesure processus de Dirichlet différente est associée à chaque mesure ordinaire α sur $\mathbb{R}$, qui fonctionne comme paramètre de la famille de mesures processus de Dirichlet. Une mesure de probabilité aléatoire F sur $\mathbb{R}$ est distribuée selon la mesure processus de Dirichlet de paramètre α si pour chaque partition finie $(A_1, \ldots, A_k)$ de $\mathbb{R}$, le vecteur $\big(F(A_1), \ldots, F(A_k)\big)$ est distribué selon la loi de Dirichlet classique avec paramètres $\big(\alpha(A_1), \ldots, \alpha(A_k)\big)$. C'est-à-dire que le vecteur $\big(F(A_1), \ldots, F(A_k)\big)$ a une densité proportionnelle à la fonction

$$(u_1, \ldots, u_k) \mapsto u_1^{\alpha(A_1)-1} u_2^{\alpha(A_2)-1} \cdots u_k^{\alpha(A_k)-1}$$

sur le k-simplexe

$$S_k = \big\{(u_1, \ldots, u_k) : u_i \geqslant 0, u_i = 1\big\}.$$

L'existence d'une mesure aléatoire possédant cette propriété n'est pas évidente, mais a été établie par Ferguson (1973).

Ferguson a aussi montré que la fonction de répartition de F, un processus aléatoire à trajectoires croissantes, s'écrit $x \mapsto \Gamma(\alpha(x))/\Gamma(\alpha(\infty))$ pour un processus Gamma.

Une troisième caractérisation, plus constructive, s'obtient par raffinements successifs. On se donne une suite de partitions $\mathbb{R} = A_{r,1} \cup A_{r,2} \cup \cdots \cup A_{r,2^r}$, où $A_{0,1} = \mathbb{R}$, et les ensembles $A_{r+1,i}$ sont obtenus en divisant chacun des ensembles $A_{r,i}$ en deux sous-ensembles. Naturellement on pose $F(A_{0,1}) = 1$. Étant donné le vecteur de probabilités $\big(F(A_{r,1}), \ldots, F(A_{r,2^r})\big)$ on forme le vecteur $\big(F(A_{r+1,1}), \ldots, F(A_{r+1,2^{r+1}})\big)$ en divisant les masses $F(A_{r,i})$ en deux masses $U_{r,i}F(A_{r,i})$ et $(1-U_{r,i})F(A_{r,i})$ pour des variables aléatoires Beta $U_{r,i}$ indépendantes, de paramètres $\alpha(A_{r+1,i'})/\alpha(A_{r,i})$. En continuant indéfiniment, on construit la mesure F sur une collection dense de sous-ensembles de $\mathbb{R}$.

Nous présentons deux théorèmes séparés pour les cas (i) et (ii). Dans le premier théorème nous supposons que les observations $X_1, \ldots, X_n$ sont issues d'une densité p_0 qui elle-même est un mélange de lois gaussiennes. Dans ce cas on a une vitesse de convergence très rapide. Même si l'ensemble de mélanges gaussiens est de dimension infinie, la vitesse de convergence est presque $1/\sqrt{n}$, la vitesse caractéristique dans les modèles classiques de dimension finie.

THÉORÈME 4.1. — *Supposons que $p_0 = p_{\sigma_0,F_0}$ pour une mesure F_0 avec $F_0[-k_0,k_0] = 1$ pour $k_0 > 0$ donné. En plus supposons que la mesure a priori π pour σ a une densité qui est continue et positive sur un intervalle qui contient σ_0, et que la mesure de base α du processus de Dirichlet a une densité qui est positive sur un intervalle qui contient $[-k_0,k_0]$ et satisfait à $\alpha(|z| > t) \lesssim e^{-b|t|^2}$, pour chaque $t > 0$ et une constante $b > 0$. Alors la vitesse de convergence par rapport à la distance de Hellinger est au moins $(\log n)^{3/2}/\sqrt{n}$.*

La vitesse de convergence dans le deuxième théorème est plus modeste, mais le théorème s'applique à une collection beaucoup plus grande de densités p_0. Si la densité des observations est deux fois dérivable, la mesure a posteriori converge à la vitesse approximative de $n^{-2/5}$. Comme distance nous utilisons la distance de Hellinger tronquée

$$h_k^2(p,q) = \int_{-k}^{k} (p^{1/2} - q^{1/2})^2 \, d\lambda.$$

THÉORÈME 4.2. — *Supposons que p_0 vérifie $P_0[-a,a]^c \leqslant e^{-ca^\gamma}$ pour des nombres positifs c et γ, et est deux fois dérivable avec $\int (p_0''/p_0)^2 \, p_0 \, d\lambda < \infty$ et $\int (p_0'/p_0)^4 \, p_0 \, d\lambda < \infty$. Si la mesure de base α a une densité continue et positive α' satisfaisant à $\alpha'(t) \gtrsim e^{-dt^\gamma}$ pour chaque t et une constante positive d, alors la vitesse de convergence par rapport à la distance h_k est au moins $\varepsilon_n = n^{-2/5}(\log n)^{1+\gamma/2}$.*

4.1. Lemmes Clés

Nous ne présentons pas ici les preuves complètes de ces théorèmes, mais nous les expliquons à l'aide de quelques lemmes. Le lemme le plus important montre qu'on peut « bien » approcher un mélange gaussien arbitraire par des mélanges finis comportant très peu de composantes.

LEMME 4.3. — *Soit* $0 < \varepsilon < 1/4$ *et* $0 < \sigma \lesssim a$. *Pour chaque mesure de probabilité* F *sur un intervalle* $[-a, a]$ *il existe une mesure de probabilité discrète sur l'intervalle* $[-a - \sigma, a + \sigma]$ *avec moins de* $N \lesssim (a/\sigma)\log(1/\varepsilon)$ *atomes, telle que*

$$\|p_{F,\sigma} - p_{F',\sigma}\|_\infty \lesssim \frac{\varepsilon}{\sigma}.$$

Une première conséquence de ce lemme est qu'au niveau de précision ε on peut voir le modèle

$$\mathcal{P}_{\alpha,\tau} = \{p_{F,\sigma} : F[-a, a] = 1, b_1\tau \leqslant \sigma \leqslant b_2\tau\}$$

comme un modèle décrit par $N \leqslant (a/\sigma)\log(1/\varepsilon)$ paramètres. Comme l'entropie d'un modèle de dimension N est de l'ordre $N\log(1/\varepsilon)$, le lemme suggère que l'entropie de $\mathcal{P}_{\alpha,\tau}$ est bornée par $(a/\sigma)(\log(1/\varepsilon))^2$. Le lemme ci-dessous est une justification de cet argument heuristique.

LEMME 4.4. — *Soient* $b_1 < b_2$, $\tau \leqslant 1/4$ *et* $a \geqslant e$ *des nombres positifs arbitraires. Alors pour chaque* $0 < \varepsilon < 1/4$,

$$\log N(\varepsilon, \mathcal{P}_{a,\tau}, h) \lesssim \frac{a}{\tau}\Big(\log\frac{1}{\varepsilon\tau}\Big)\Big(\log\frac{a}{\varepsilon\tau}\Big).$$

Nous pouvons conclure qu'une borne sur la vitesse de convergence optimale ε_n comme en (3.5) pour le modèle des mélanges est la solution de l'inégalité

$$\frac{1}{\sigma_n}\big(\log(1/\varepsilon_n)\big)^2 \leqslant n\varepsilon_n^2.$$

Cette solution est $\varepsilon_n \sim (\log n)/\sqrt{n\sigma_n}$.

Le Lemme 4.3, qui montre qu'un mélange arbitraire peut être approché par un mélange fini, est aussi le point de départ pour obtenir la borne inférieure sur la masse a priori exigée par (3.6). Parce qu'une boule autour d'un mélange arbitraire contient une boule autour d'un certain mélange discret, il suffit de borner la masse a priori près d'un mélange fini. Le lemme suivant montre que cette masse peut être contrôlée par la distance l_1 d'un vecteur du type $\big(F(A_1), \ldots, F(A_N)\big)$ du vecteur des atomes $(p_1, \ldots, p_N)$ du mélange fini, ou p_i est l'atome continue en A_i. Comme ce vecteur a une loi de Dirichlet classique sur le N-simplexe, on peut utiliser des arguments élémentaires. Ceci est pratique, parce que les calculs directs avec des processus de Dirichlet ne sont pas faciles.

LEMME 4.5. — *Soit* $\mathbb{R} = \cup_{j=0}^{N} A_j$ *une partition arbitraire de la droite réelle et soit* $F' = \sum_{j=1}^{N} p_j \delta_{z_j}$ *une mesure de probabilité avec* $z_j \in A_j$ *pour chaque* $j = 1, \ldots, N$. *Alors, pour toute mesure de probabilité* F *sur* $\mathbb{R}$, *on a*

$$\|p_{F,\sigma} - p_{F',\sigma}\|_1 \lesssim \frac{1}{\sigma} \max_{1 \leqslant j \leqslant N} \lambda(A_j) + \sum_{j=1}^{N} |F(A_j) - p_j|.$$

4.2. Méthode MCMC

Dans les sections précédentes nous avons vu que la mesure a priori processus de Dirichlet donne des bonnes vitesses de convergence en conjonction avec des mélanges gaussiens. La mesure Dirichlet est aussi très pratique pour le calcul numérique de la mesure a posteriori par des méthodes MCMC. Dans cette section nous présentons le « Gibbs sampler », un algorithme MCMC qui dans ce cas possède une simplicité et une efficacité remarquables. Pour simplifier la présentation nous nous limitons au cas où le paramètre σ est une constante. Nous suivons Escobar et West (1995).

Nous reprenons la définition de la mesure a priori, par une description en trois étapes, en introduisant des variables auxiliaires non-observées $Z_1, \ldots, Z_n$:

(i) F a pour loi un processus de Dirichlet à mesure de base α ;

(ii) étant donné F, les variables $Z_1, \ldots, Z_n$ sont indépendantes et distribuées selon F ;

(iii) étant donnés F et $Z_1, \ldots, Z_n$, les variables $X_1, \ldots, X_n$ sont indépendantes et X_i est de loi gaussienne $N(Z_i, \sigma^2)$.

Il découle de cette description que les observations $X_1, \ldots, X_n$ sont conditionnellement indépendantes de la mesure aléatoire F étant données les variables $Z_1, \ldots, Z_n$.

On s'intéresse à trouver la loi a posteriori de la fonction de densité $p_{F,\sigma}(x) = \int \phi_\sigma(x - z)\, dF(z)$, pour chaque x fixé, ou, plus généralement, la mesure a posteriori de fonctions linéaires $\int h\, dF$ de la mesure F, pour des fonctions $h : \mathbb{R} \to \mathbb{R}$ données. L'indépendance conditionnelle de F et $X_1, \ldots, X_n$ étant donnés $Z_1, \ldots, Z_n$ montre que

$$\mathrm{E}\Big(\int h\, dF \mid X_1, \ldots, X_n\Big) = \mathrm{E}\Big(\mathrm{E}\Big(\int h\, dF \mid Z_1, \ldots, Z_n\Big) \mid X_1, \ldots, X_n\Big).$$

L'espérance conditionnelle $\mathrm{E}\big(\int h\, dF \mid Z_1, \ldots, Z_n\big)$ concerne seulement F et les variables $Z_1, \ldots, Z_n$, et n'a rien à voir avec la structure des mélanges, qui est seulement introduite dans l'étape (iii). Ferguson (1973) a montré déjà dans son article fondamental sur les processus de Dirichlet que la loi a posteriori de F étant donnés $Z_1, \ldots, Z_n$ est aussi une loi processus de Dirichlet, mais le paramètre change de α à $\alpha + n\mathbb{F}_n^Z$, pour $\mathbb{F}_n^Z = n^{-1} \sum_{i=1}^{n} \delta_{Z_i}$, la mesure empirique de $Z_1, \ldots, Z_n$. Comme l'espérance d'une mesure processus

de Dirichlet de paramètre α est sa mesure de base renormalisée $\alpha/\alpha(\mathbb{R})$, on a

$$\mathrm{E}\Big(\int h\,dF \mid Z_1,\dots,Z_n\Big) = \frac{\int h\,d(\alpha + n\,d\mathbb{F}_n^Z)}{\alpha(\mathbb{R})+n}.$$

En substituant cette identité dans l'équation précédente nous trouvons que

$$\mathrm{E}\Big(\int h\,dF \mid X_1,\dots,X_n\Big) = \frac{\int h\,d\alpha}{\alpha(\mathbb{R})+n} + \frac{n}{\alpha(\mathbb{R})+n}\mathrm{E}\big(h(Z_1) \mid X_1,\dots,X_n\big).$$

Nous concluons que pour calculer la loi a posteriori de $\int h\,dF$, il suffit de calculer la loi a posteriori de Z_1. Ceci ce fait par simulation. Si on pouvait simuler un grand nombre L de vecteurs $(Z_1^{(l)},\dots,Z_n^{(l)})$ selon la loi conditionnelle de $(Z_1,\dots,Z_n)$ étant donné le vecteur $(X_1,\dots,X_n)$, on pourrait estimer l'expression précédente par

$$\frac{\int h\,d\alpha}{\alpha(\mathbb{R})+n} + \frac{n}{\alpha(\mathbb{R})+n}\frac{1}{L}\frac{1}{n}\sum_{l=1}^{L} h(Z_i^{(l)}).$$

Dans ce but on peut utiliser le « Gibbs sampler » étant donné le lemme suivant.

LEMME 4.6. — *La fonction de densité de la loi conditionnelle de Z_1 étant donnés $Z_2,\dots,Z_n,X_1,\dots,X_n$ est proportionnelle à*

$$z \mapsto w_1\frac{\phi_\sigma(X_1-z)\,d\alpha(z)}{\int \phi_\sigma(x-\zeta)\,d\alpha(\zeta)} + (1-w_1)\frac{\phi_\sigma(X_i-z)\,d\mathbb{F}_{n,i}^Z(z)}{\int \phi_\sigma(x-\zeta)\,d\mathbb{F}_{n,2}^Z(\zeta)},$$

où $\mathbb{F}_{n,2}^Z$ désigne la loi empirique de $Z_2,\dots,Z_n$ et

$$w_1 = \frac{\int \phi_\sigma(X_1-\zeta)\,d\alpha(\zeta)}{\int \phi_\sigma(X_1-\zeta)\,d(\alpha+(n-1)\mathbb{F}_{n,2}^Z)(\zeta)}.$$

Preuve. — Nous écrirons les équations pour le cas où la loi F est discrète sur un ensemble fini de points ζ_j. Cette situation peut être étendue au cas général par une opération de passage à la limite.

La variable F et le vecteur $(X_2,\dots,X_n)$ sont conditionnellement indépendantes étant donné le vecteur $(Z_2,\dots,Z_n)$, et le vecteur (Z_1,X_1) est conditionnellement indépendant du vecteur $(X_1,\dots,X_n,Z_n,\dots,Z_n)$ étant donné F. En conséquence on a

$$\begin{aligned}
&\mathrm{P}(Z_1=\zeta_j, X_1\in B \mid X_2,\dots,X_n,Z_2,\dots,Z_n)\\
&\quad= \int \mathrm{P}(Z_1=\zeta_j, X_1\in B \mid F=f, X_2,\dots,X_n,Z_2,\dots,Z_n)\\
&\qquad\qquad dP^{F\mid Z_2,\dots,Z_n,X_2,\dots,X_n}(f)\\
&\quad= \int f\{\zeta_j\}\Phi_\sigma(B-\zeta_j)\,dP^{F\mid Z_2,\dots,Z_n}(f) \qquad (4.1)\\
&\quad= \Phi_\sigma(B-\zeta_j)\mathrm{E}(F\{\zeta_j\} \mid Z_2,\dots,Z_n).
\end{aligned}$$

Parce que la loi conditionnelle de F étant donnés $Z_2, \ldots, Z_n$ est Dirichlet à mesure de base $\alpha + (n-1)\mathbb{F}_{n,2}^Z$, l'espérance conditionnelle du membre de droite de l'équation est égale à $(\alpha\{\zeta_j\} + (n-1)\mathbb{F}_{n,2}^Z\{\zeta_j\})/(\alpha(\mathbb{R}) + n - 1)$. Nous concluons que

$$\mathrm{P}(Z_1 = \zeta_j \mid X_1, \ldots, X_n, Z_2, \ldots, Z_n) = \frac{\phi_\sigma(X_1 - \zeta_j)(\alpha\{\zeta_j\} + (n-1)\mathbb{F}_{n,2}^X\{\zeta_j\})}{\sum_{l=1}^{k} \phi_\sigma(X_1 - \zeta_j)(\alpha\{\zeta_j\} + (n-1)\mathbb{F}_{n,2}^X\{\zeta_j\})}.$$

Ceci est l'énoncé du lemme. □

Le « Gibbs sampler » est un algorithme MCMC qui produit des réalisations d'un vecteur aléatoire $(Z_1, \ldots, Z_n)$ en simulant successivement des lois marginales conditionnelles. Commençant avec des valeurs initiales $z_1^{(0)}, \ldots, z_n^{(0)}$ on tire une nouvelle valeur $z_1^{(1)}$ pour Z_1 de la loi conditionnelle de Z_1 étant donné $Z_2 = z_2^{(0)}, \ldots, Z_n = z_n^{(0)}$, puis on tire une valeur $z_2^{(1)}$ de la loi conditionnelle de Z_2 étant donné $Z_1 = z_1^{(1)}, Z_3 = z_3^{(0)}, \ldots, Z_n = z_n^{(0)}$, et on répète indéfiniment, toujours conditionnant sur les dernières valeurs des « autres » variables.

Cette procédure produit une chaîne de Markov $((Z_1^{(l)}, \ldots, Z_n^{(l)}) : l = 0, 1, 2, \ldots)$ qui a pour loi stationnaire la loi de $(Z_1, \ldots, Z_n)$ qui définit les lois marginales conditionnelles. L'effet d'initialisation à des valeurs $z_1^{(0)}, \ldots, z_n^{(0)}$ disparaît lorsque $l \to \infty$.

Dans la situation présente nous simulons des réalisations du vecteur $(Z_1, \ldots, Z_n)$ étant données les variables $X_1, \ldots, X_n$, et nous utilisons les lois marginales conditionnelles des variables Z_i étant données les autres variables Z_j et les observations $X_1, \ldots, X_n$. Ces lois ont la forme décrite par le lemme précédent.

Cette forme est remarquable parce qu'elle nous indique comment engendrer la nouvelle variable, soit à partir de la loi à densité proportionnelle à $z \mapsto \phi_\sigma(X_1 - z)\, d\alpha(z)$, soit à partir de la loi empirique pondérée $z \mapsto \phi_\sigma(X_1 - z)\, d\mathbb{F}_{n,1}^Z(z)$. La deuxième possibilité revient à tirer la nouvelle variable Z_i des valeurs courantes des autres variables Z_j, de la même façon que le « bootstrap empirique » de Efron, sauf que les poids des Z_j ne sont pas égaux entre eux.

La Figure 6 donne une idée de la sortie de l'algorithme.

5. Sélection de Modèles et Adaptation

Dans la section précédente nous avons introduit deux cadres très différents pour estimer une densité. Avec une fenêtre fixée on a un très bon estimateur si la vraie densité des observations est un mélange de lois gaussiennes.

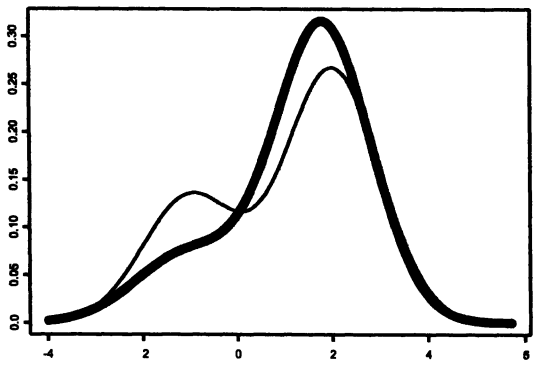

FIG 6. — Densité a posteriori (ligne grasse) basée sur un échantillon de 100 observations d'un mélange de deux lois gaussiennes (ligne simple).

Dans l'autre cas on a une vitesse de convergence modérée, mais l'estimateur converge dès que la vraie densité est deux fois dérivable. On peut se demander s'il est possible de combiner les deux cadres dans une seule procédure d'estimation. Le paradigme bayésien offre une méthode très naturelle pour une telle combinaison, parce qu'on peut combiner deux mesures a priori, dont les supports respectifs sont les deux modèles considérés, en une nouvelle mesure a priori en prenant une combinaison linéaire des deux. Dans cette section nous examinons si une telle combinaison naturelle retient le bon comportement des deux procédures bayésiennes dans les deux cadres. Une question importante est de trouver les poids des deux mesures dans leur combinaison.

Plus généralement nous considérons une collection générale d'expériences statistiques $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, P_{\theta,\alpha}^{(n)} : \theta \in \Theta_\alpha)$, indexées par un indice α contenu dans un ensemble arbitraire A (fini ou dénombrable). Pour chaque $\alpha \in A$ on a une mesure a priori $\Pi_{n,\alpha}$ sur l'ensemble de paramètres Θ_α, choisie pour bien fonctionner avec le modèle donné par les mesures $P_{\theta,\alpha}^{(n)}$. On combine ces mesures avec des poids $\{\lambda_{n,\alpha} : \alpha \in A\}$ sur l'ensemble des indices A tels que $\sum_\alpha \lambda_{n,\alpha} = 1$, et l'on obtient la mesure a priori

$$\Pi_n = \sum_{\alpha \in A} \lambda_{n,\alpha} \Pi_{n,\alpha} \tag{5.1}$$

sur l'union directe $\Theta = \cup_\alpha \Theta_\alpha$ des ensembles de paramètres Θ_α.

La mesure a posteriori correspondante est la mesure aléatoire

$$\Pi_n(B \mid X^{(n)}) = \frac{\sum_{\alpha \in A} \lambda_{n,\alpha} \int_{\theta \in \Theta_\alpha : \theta \in B} p_{\theta,\alpha}^{(n)}(X^{(n)})\, d\Pi_{n,\alpha}(\theta)}{\sum_{\alpha \in A} \lambda_{n,\alpha} \int_{\Theta_\alpha} p_{\theta,\alpha}(X^{(n)})\, d\Pi_{n,\alpha}(\theta)}, \quad B \in \mathcal{B}. \tag{5.2}$$

Celle-ci est exactement (2.1) pour la densité a priori (5.1).

On pourrait inscrire l'ensemble des modèles $(P_{\theta,\alpha}^{(n)} : \theta \in \Theta_\alpha)$ dans le cadre de la Section 2 en considérant la mesure a priori Π_n comme une mesure sur l'union $\Theta = \cup_\alpha \Theta_\alpha$ des ensembles de paramètres. Ceci aurait comme résultat

une caractérisation de la vitesse de convergence de la mesure a posteriori par l'entropie de l'union Θ et la mesure a priori globale Π_n. Naturellement, l'entropie de Θ sera déterminée par le plus complexe des ensembles de paramètres Θ_α. En conséquence, la vitesse de convergence garantie par le Théorème 2.1 serait déterminée par le plus difficile des modèles dans la liste, résultant en la plus lente des vitesses individuelles. Par exemple, dans les deux cadres d'estimation d'une densité considérés en Section 4 on trouve toujours la vitesse $n^{-2/5}(\log n)^{1+/2}$ et jamais la vitesse «super efficace» $n^{-1/2}(\log n)^{3/2}$.

Nous cherchons un résultat plus significatif. Ceci est une énoncé d'*adaptation* : si le vrai paramètre θ_0 est continu dans l'ensemble Θ_β pour un $\beta \in A$ donné, on espère récupérer au moins la vitesse de convergence qu'on peut obtenir en utilisant le modèle indexé par β. Bien sûr, comme β est inconnu, ce programme d'adaptation est très ambitieux : on cherche à obtenir lorsque β est inconnu la même qualité d'estimation que celle qui peut être atteinte lorsque β est connu. Si ce but est trop ambitieux, on peut encore chercher à l'atteindre à un facteur de pénalisation près.

L'on sait des méthodes non-bayésiennes développées dans les années 1990 qu'un programme d'adaptation de ce type est réalisable. Par exemple, pour le problème de l'estimation non paramétrique d'une densité on a trouvé dans les années 1970 que si la vraie densité p d'un échantillon $X_1, \ldots, X_n$ de variables réelles possède α dérivées, on peut estimer p à une vitesse de convergence $\varepsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$, par rapport à des distances naturelles. Des méthodes diverses (noyau, séries tronquées, ondelettes, etc.) permettent d'atteindre cette vitesse $\varepsilon_{n,\alpha}$. Au début ces méthodes dépendaient de la valeur de α et en conséquence exigeaient la connaissance a priori de la régularité de la densité inconnue. Par exemple, il était nécessaire de choisir une fenêtre appropriée lors de l'utilisation d'un estimateur à noyau. Dans les années 1990, plusiers chercheurs ont réussi à libérer ces procédures de cette contrainte peu pratique, et ont développé des procédures qui choisissent une bonne valeur du paramètre de régularité en utilisant seulement les données. Ce sont des résultats théoriques profonds, qui sont encore surprenants à bien des égards. Peut-être est-il plus surprenant encore que ces nouvelles méthodes se comportent bien dans des situations pratiques, et viennent finalement ajouter des procédures non paramétriques opérationnelles aux procédures classiques utilisées quotidiennement en statistique appliquée.

Dans cette section nous examinons la possibilité d'adaptation par des méthodes bayésiennes. Nous nous limitons à une situation simple, mais illustrative : nous présentons une version adaptative du résultat général de la Section 2 dans le cas où le nombre de modèles est fini et en utilisant des poids $\lambda_{n,\alpha}$ spécifiques.

Supposons que pour chaque expérience statistique on ait une vitesse de convergence idéale $\varepsilon_{n,\alpha} \to 0$, par rapport à une distance d_n sur Θ donné. Nous supposons que ces vitesses sont liées à l'entropie des modèles au sens simplifié de la Section 2. Nous supposons, pour chaque α,

$$\log N(\varepsilon_{n,\alpha}, \mathcal{P}_{n,\alpha}, e_n) \leqslant n\varepsilon_{n,\alpha}^2. \tag{5.3}$$

Les poids $\lambda_{n,\alpha}$ sont choisis proportionnels à une puissance de $\exp(-n\varepsilon_{n,\alpha}^2)$, *i.e.* pour une constante C et des nombres donnés positifs λ_α,

$$\lambda_{n,\alpha} = \frac{\lambda_\alpha e^{-Cn\varepsilon_{n,\alpha}^2}}{\sum_\alpha \lambda_\alpha e^{-Cn\varepsilon_{n,\alpha}^2}}. \tag{5.4}$$

Ces poids donnent plus d'importance aux modèles qui sont associés aux vitesses $\varepsilon_{n,\alpha}$ petites, c'est à dire les «petits modèles».

Supposons que l'observation $X^{(n)}$ est distribuée selon $P_{\theta_0,\beta}^{(n)}$ pour un $\beta \in A$ donné. Comme dans la Section 2 on suppose qu'il est possible de tester cette loi par rapport à des contre-hypothèses constituées des boules qui peuvent être contenues dans tous les modèles $(P_{\theta,\alpha}^{(n)} : \theta \in \Theta_\alpha)$. Alors supposons que pour chaque n on dispose d'une semi-métrique d_n sur $\Theta = \cup_\alpha \Theta_\alpha$ telle qu'il existe des constantes universelles positives ξ et K telles que pour chaque $\varepsilon > 0$: pour tout $\theta_1 \in \Theta$ avec $d_n(\theta_1, \theta_0) > \varepsilon$ il existe un test ϕ_n tel que

$$\begin{aligned} P_{\theta_0,\beta}^{(n)}\phi_n &\leqslant e^{-Kn\varepsilon^2}, \\ \sup_{\theta\in\Theta : d_n(\theta,\theta_1)<\varepsilon\xi} P_\theta^{(n)}(1-\phi_n) &\leqslant e^{-Kn\varepsilon^2}. \end{aligned} \tag{5.5}$$

L'on définit $B_{n,\alpha}(\theta_0, \varepsilon)$ comme un voisinage autour de θ_0 du même type que dans la Section 2 :

$$B_{n,\alpha}(\theta_0,\varepsilon) = \left\{\theta \in \Theta_\alpha : K(p_{\theta_0,\beta}^{(n)}, p_{\theta,\alpha}^{(n)}) \leqslant n\varepsilon^2, V(p_{\theta_0,\beta}^{(n)}, p_{\theta,\alpha}^{(n)}) \leqslant n\varepsilon^2\right\}.$$

La condition que la mesure a priori met suffisamment de poids dans le voisinage de θ_0, sous forme simplifiée, est

$$\Pi_{n,\beta}\big(B_{n,\beta}(\theta_0, \varepsilon_{n,\beta})\big) \geqslant e^{-Fn\varepsilon_{n,\beta}^2}. \tag{5.6}$$

THÉORÈME 5.1. — *Soit A un ensemble fini équipé d'un ordre linéaire tel que $\varepsilon_{n,\alpha} \ll \varepsilon_{n,\beta}$ pour chaque $\alpha > \beta \in A$ et $n\varepsilon_{n,\beta}^2 \to \infty$. Supposons que (5.3), (5.5) et (5.6) sont satisfaites. Alors la mesure a posteriori (5.2) avec les poids $\lambda_{n,\alpha}$ donnés en (5.4) satisfait à*

$$P_{\theta_0,\beta}^{(n)}\Pi_n\big(\theta : d_n(\theta_0,\theta) \geqslant M\varepsilon_{n,\beta} \mid X^{(n)}\big) \to 0$$

pour tout M suffisamment grand.

Ce théorème se laisse généraliser dans plusieurs sens. Par exemple, la liste des modèles peut être dénombrable plutôt que finie, si l'on ajoute des conditions un peu techniques. Ensuite, en utilisant cette généralisation et des mesures a priori appropriées, on peut récupérer les résultats d'estimation adaptive de densités décrits ci-dessus. Alors l'adaptation par des méthodes bayésiennes semble une option viable.

Encore plus intéressant d'un point de vue philosophique est l'étude du rôle des poids $\lambda_{n,\alpha}$, qui dans le théorème ont la forme spéciale (5.4), mais peuvent

certainement être variés. Par exemple (5.4) exclut des poids égaux dans le cas d'un nombre fini de modèles de vitesses $\varepsilon_{n,\alpha}$ très différentes.

À l'heure actuelle, cette situation est encore mal comprise. Nous avons déja étudié des situations où des poids uniformes sur une liste de modèles donne une adaptation à un facteur $\log n$ près, et nous connaissons d'autres situations où des poids plus lourds sur des modèles plus grands (l'inverse des poids (5.4) donnent une adaptation exacte. Il semble qu'il y a une certaine liberté dans le choix d'une mesure a priori sur un modèle $(P_{\theta,\alpha}^{(n)} : \theta \in \Theta_\alpha)$ résultant en une vitesse de convergence optimale pour ce modèle, une liberté que l'on perd quand on étudie un ensemble de modèles. Il semble que pour comprendre l'adaptation il faille considérer les poids et les mesures a priori pour les modèles pris dans leur ensemble.

Comme exemple concret nous considérons l'adaptation aux deux cadres d'estimation des mélanges gaussiens de la Section 4. Le théorème précédent donne l'adaptation avec les poids $\lambda_{n,s}$ et $\lambda_{n,ss}$ pour les mélanges «lisses» et «super lisses» du type

$$\lambda_{n,s} \propto \exp\bigl(-Cn^{1/5}(\log n)^k\bigr),$$
$$\lambda_{n,ss} \propto \exp\bigl(-C(\log n)^k\bigr).$$

La différence en grandeur entre $\lambda_{n,s}$ et $\lambda_{n,ss}$ est énorme, le petit modèle de mélanges de lois gaussiennes de variance bornée recevant exponentiellement plus de masse a priori. D'une étude plus générale nous savons que l'on a l'adaptation dès que, pour des constantes positives c et C,

$$\exp\bigl(c(\log n)^k\bigr) < \frac{\lambda_{n,ss}}{\lambda_{n,s}} < \exp\bigl(Cn^{1/5}(\log n)^k\bigr).$$

Ceci donne beaucoup plus de flexibilité dans le choix des poids, mais exclut toujours les poids égaux : $\lambda_{n,s} = \lambda_{n,ss}$. Nous pensons, mais nous n'en avons toujours pas de preuve, que l'adaptation subsiste si

$$\exp\bigl(-c(\log n)^k\bigr) < \frac{\lambda_{n,ss}}{\lambda_{n,s}} < \exp\bigl(Cn^{1/5}(\log n)^k\bigr).$$

6. Remerciements

Cet article est une revue de résultats obtenus avec mes co-auteurs Subhashis Ghosal et Jyri Lember. Voir Ghosal, Ghosh et van der Vaart (2000), Ghosal et van der Vaart (2001), Ghosal et van der Vaart (2003), Ghosal, Lember et van der Vaart (2003). Je remercie Maryse Loranger et Marc Hallin pour la correction d'un grand nombre d'erreurs grammaticales, et je remercie Marc Hallin pour son aide et sa patience.

Références

[1] BIRGÉ L. (1983a). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahr. Verw. Gebiete* **65**, 181-238.

[2] BIRGÉ L. (1983b). Robust testing for independent non-identically distributed variables and Markov chains. In *Specifying Statistical Models. From Parametric to Non-Parametric. Using Bayesian or Non-Bayesian Approaches* (J. P. Florens et al. eds.,) *Lecture Notes in Statistics* **16** Springer-Verlag, New York, 134-162.

[3] DIACONIS P. and FREEDMAN D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1-67.

[4] ESCOBAR M. and WEST M. (1995). On Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577-588.

[5] ESCOBAR M. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268-277.

[6] FERGUSON T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. of Statistics* **1** 209-230.

[7] FERGUSON T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (Rizvi M., Rustagi, J. and Siegmund, D., Eds.) 287-302.

[8] FREEDMAN D. (1963). On the asymptotic distribution of Bayes estimates in the discrete case I. *Ann. Math. Statist.* **34** 1386-1403.

[9] GHOSAL S., GHOSH J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500-531.

[10] GHOSAL S., LEMBER J. and VAN DER VAART A. W. (2003). On Bayesian adaptation. *Acta Applicandae Mathematica* **79** 165-175.

[11] GHOSAL S. and VAN DER VAART A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233-1263.

[12] GHOSAL S. and VAN DER VAART A. W. (2003). Posterior convergence rates of Dirichlet mixtures of normal for smooth densities. Preprint.

[13] GHOSAL S. and VAN DER VAART A. W. (2003). Convergence rates for posterior distributions for noniid observations. Preprint.

[14] GHOSH J. K. and RAMAMOORTHI R. V. (2003). *Bayesian Nonparametrics.* Springer-Verlag, New York.

[15] KOLMOGOROV A. N. and TIHOMIROV V. M. (1961). ε-entropy and ε-capacity of sets in function spaces. *Amer. Math. Soc. Transl. Ser.* 2, **17** 277-364. (Translated from Russian : *Uspekhi Mat. Nauk* **14** 3-86, (1959).)

[16] LE CAM L. M. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics* **1** 277-330.

[17] LE CAM L. M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **22** 38-53.

[18] LE CAM L. M. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic Processes and Related Properties.* Academic Press, New York, 13-54.

[19] LE CAM L. M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

[20] LIU J.S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag, New York.

[21] ROBERT C.P. (2001). *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation.* Springer-Verlag, New York.

[22] ROSENBLATT M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832-837.

[23] SCHWARTZ L. (1965). On Bayes procedures. *Z. Wahr. Verw. Gebiete* **4** 10-26.

[24] VAN DER VAART A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

[25] VAN DER VAART A. W. and WELLNER J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

[26] WEST M. (1992). Modeling with Mixtures. In *Bayesian Statistics* **4** (J.M. Bernardo *et al.*, Eds.) 503-524.