Estimation de la Densité et Tests par la Méthode Combinatoire Pénalisée

Gérard BIAU

Laboratoire de Statistique Théorique et Appliquée
Université Pierre et Marie Curie – Paris VI
Boîte 158, 175 rue du Chevaleret
75013 Paris, France

biau@ccr.jussieu.fr

Résumé

Soit $(\mathcal{F}_k)_{k\geq 1}$ une famille emboîtée de classes paramétriques de densités possédant une dimension de Vapnik-Chervonenkis finie. Soit f une densité de probabilité appartenant à $\mathcal{F}_{k^*}$, où k^*, inconnu, est le plus petit entier tel que $f \in \mathcal{F}_k$. Etant donné un échantillon i.i.d. de taille n de variables aléatoires admettant f comme densité commune, nous présentons dans la première partie de ce travail une méthode combinatoire générale permettant de sélectionner automatiquement, et sans restrictions supplémentaires sur f, un estimateur $f_{n,\hat{k}}$ dans $\cup_{k\geq 1}\mathcal{F}_k$ satisfaisant la propriété $\mathbf{E}\{\int |f_{n,\hat{k}}-f|\} = \mathbf{O}(1/\sqrt{n})$. L'estimateur $f_{n,\hat{k}}$ est choisi à partir d'un critère combinatoire pénalisé inspiré des travaux de Devroye et Lugosi [13]. Etant donné un entier $k_0 \geq 1$ et un nombre réel $\alpha \in]0,1[$, nous explicitons dans la seconde partie de l'article une procédure permettant de tester l'hypothèse nulle $\{\mathbf{H}_0 : k^* = k_0\}$ contre l'alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. Nous montrons que ce nouveau test est de niveau asymptotique α et de puissance asymptotique 1.

Mots-clés **— Estimation d'une densité multivariée, dimension de Vapnik-Chervonenkis, mélanges de densités, estimation non paramétrique, test, pénalisation.**

Classification AMS **: 62G05 et 62G10.**

Abstract

Let $(\mathcal{F}_k)_{k\geq 1}$ be a nested family of parametric classes of densities with finite Vapnik-Chervonenkis dimension. Let f be a probability density belonging to $\mathcal{F}_{k^*}$, where k^* is the unknown smallest integer such that $f \in \mathcal{F}_k$. Given an i.i.d. sample of size n drawn from f, we present in the first part of the paper a general combinatorial method to select automatically, and without extra restrictions on f, an estimate $f_{n,\hat{k}}$ in $\cup_{k\geq 1}\mathcal{F}_k$ with the property that $\mathbf{E}\{\int |f_{n,\hat{k}} - f|\} = \mathbf{O}(1/\sqrt{n})$. The estimate $f_{n,\hat{k}}$ will be selected via a penalized combinatorial criterion inspired by the works of Devroye and Lugosi [13]. Given an integer $k_0 \geq 1$ and a real number $\alpha \in]0,1[$, we develop in the second part of the paper a testing procedure of the null hypothesis $\{\mathbf{H}_0 : k^* = k_0\}$ versus the alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. We show that this new test is asymptotically of level α and has asymptotic power 1.

Keywords and phrases **— Multivariate density estimation, Vapnik-Chervonenkis dimension, density mixtures, nonparametric estimation, test, penalization.**

AMS Classification **: 62G05 and 62G10.**

1 Introduction

Dans cet article, nous nous intéressons au problème de l'estimation d'une densité de probabilité inconnue f à partir d'un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de même loi à densité f. Nous supposons que f est définie sur $\mathbb{R}^d$ et qu'elle appartient à une certaine classe *paramétrique* de densités $\mathcal{F}_k$, où l'entier k est inconnu mais $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ pour chaque $k \geq 1$. La classe $\mathcal{F}_k$ peut, par exemple, représenter l'ensemble des mélanges de k densités gaussiennes multivariées (voir plus bas pour des exemples détaillés). Soit alors

$$\mathcal{F} = \bigcup_{k\geq 1} \mathcal{F}_k\,.$$

Dans la première partie de ce travail, nous présentons un algorithme combinatoire très général permettant de *sélectionner automatiquement* (à partir des seules données $X_1, \ldots, X_n$), et sans restrictions supplémentaires sur f, un estimateur particulier (ou *modèle*) $f_{n,\hat{k}}$ dans $\mathcal{F}$–c'est-à-dire un entier $\hat{k}$ et un jeu de paramètres dans $\mathcal{F}_{\hat{k}}$–satisfaisant la propriété

$$\mathbf{E}\left\{\int |f_{n,\hat{k}} - f|\right\} = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right).$$

Notre algorithme de sélection s'appuie sur des techniques combinatoires pour l'estimation de la densité développées dans l'ouvrage récent de Devroye et Lugosi [13].

Définissons maintenant l'*indice de la représentation économique de f* comme

$$k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}.$$

Cet indice, qui a toujours un sens, représente d'un point de vue intuitif la dimension du modèle le plus "parcimonieux" permettant de décrire f. Etant donné un entier $k_0 \geq 1$, il est alors naturel de s'intéresser à la mise en œuvre d'une procédure de test de l'hypothèse nulle $\{\mathbf{H}_0 : k^* = k_0\}$ contre l'alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. Cette problématique fait l'objet de la deuxième partie de l'article, où nous montrons, en utilisant des outils combinatoires similaires, comment construire une telle procédure de test qui soit de surcroît convergente.

Mais avant d'expliciter les différents résultats, nous illustrons la généralité de notre approche au travers de deux exemples significatifs.

Classes de mélanges. Dans ce premier exemple important, chaque classe $\mathcal{F}_k$ est composée des mélanges de k densités gaussiennes multivariées, c'est-à-dire des densités de la forme

$$f(x) = \sum_{i=1}^{k} \frac{p_i}{\sqrt{(2\pi)^d \,\mathrm{dét}(\Sigma_i)}}\, e^{-\frac{1}{2}(x-m_i)^T \Sigma_i^{-1}(x-m_i)},$$

où $(p_1, \ldots, p_k)$ est un vecteur de probabilité, $\Sigma_1, \ldots, \Sigma_k$ des matrices $d \times d$ définies positives et $m_1, \ldots, m_k$ des vecteurs arbitraires de $\mathbb{R}^d$. Beaucoup a été dit et écrit dans les cinquante dernières années sur les divers aspects (théoriques ou appliqués) de l'estimation statistique dans les modèles de mélanges (gaussiens ou pas) dans le cadre, forcément restrictif, où le nombre de composantes k est un entier supposé connu. Pour des références récentes sur cette question, nous renvoyons le lecteur à Everitt et Hand [17], Titterington, Smith et Makov [33], McLachlan et Basford [27], ou encore McLachlan et Peel [28]. Lorsque le paramètre k devient variable, la problématique de l'estimation se complique singulièrement et fait, à ce titre, l'objet d'une recherche encore très active. Etant donné que les modèles de mélanges interviennent dans de nombreuses branches de la statistique, le champ des applications potentielles est très vaste. On comprend dès lors que des communautés scientifiques parfois éloignées se soient intéressées au problème. C'est par exemple le cas de la communauté bayésienne, notamment au travers des travaux de Diebolt et Robert [14], Richardson et Green [30], Roeder et Wasserman [31], Celeux, Hurn et Robert [6] ou encore Hurn, Justel et Robert [21]. On trouve également des solutions au problème du k inconnu dans la littérature relative à l'apprentissage statistique (Bishop [4], Jordan et Jacobs [23], Zeevi et Meir [36], Figueiredo et Jain [18]), en clustering (Hartigan [20], Fukumizu [19]) ou encore en statistique plus traditionnelle (Priebe [29], Dacunha-Castelle et Gassiat [7], [8], [9], James, Priebe et Marchette [22], et Rogers, Marchette et Priebe [32]).

Familles exponentielles croissantes. Chaque densité f appartenant à une *famille exponentielle* $\mathcal{F}_k$ admet une représentation de la forme

$$f(x) = c\alpha(\theta)\beta(x)e^{\sum_{i=1}^{k}\pi_i(\theta)\psi_i(x)},$$

où θ appartient à un ensemble de paramètres Θ, $\psi_1, \ldots, \psi_k : \mathbb{R}^d \to \mathbb{R}$, $\beta : \mathbb{R}^d \to [0, +\infty[$, $\alpha, \pi_1, \ldots, \pi_k : \Theta \to \mathbb{R}$ sont des fonctions fixes, et c est une constante de normalisation. Les classes de densités gaussiennes, gamma, bêta, Rayleigh et Maxwell sont des exemples à la fois riches et classiques de familles exponentielles.

La suite du présent article se divise en trois parties. Dans la première partie (Section 2) nous présentons en détail notre méthode de sélection, dite *méthode combinatoire pénalisée*. La seconde partie (Section 3) est consacrée à l'étude du test de l'hypothèse $\{\mathbf{H}_0 : k^* = k_0\}$ contre $\{\mathbf{H}_1 : k^* \neq k_0\}$. Nous avons regroupé l'ensemble des preuves dans la dernière partie (Section 4).

2 La méthode combinatoire pénalisée

2.1 Présentation

Notre algorithme de sélection s'appuie sur les résultats combinatoires de Devroye et Lugosi [13]. Il s'articule en deux étapes.

Première étape. Le principe de cette première étape consiste à sélectionner, pour chaque $k \geq 1$, une densité particulière $f_{n,k}$ dans $\mathcal{F}_k$ à partir de la méthode combinatoire développée par Devroye et Lugosi dans [13]. Dans le contexte qui est le nôtre, il s'agit, à partir de la classe d'ensembles

$$\mathcal{A}_k = \big\{\{x : g_1(x) \geq g_2(x)\} : g_1, g_2 \in \mathcal{F}_k\big\}$$

(appelée *classe de Yatracos* associée à $\mathcal{F}_k$, en référence aux travaux de Yatracos [35]), de minimiser le critère suivant, valable pour tout g dans $\mathcal{F}_k$:

$$\Delta_k(g) = \sup_{A \in \mathcal{A}_k} \left| \int_A g - \mu_n(A) \right|,$$

où $\mu_n(A) = (1/n) \sum_{i=1}^n \mathbf{1}_{[X_i \in A]}$ est la *mesure empirique* associée à l'échantillon $X_1, \ldots, X_n$. Pour chaque $k \geq 1$, l'*estimateur de la distance minimum* $f_{n,k}$ est alors défini comme n'importe quelle densité g^* dans $\mathcal{F}_k$ vérifiant

$$\Delta_k(g^*) < \inf_{g \in \mathcal{F}_k} \Delta_k(g) + \frac{1}{n},$$

le terme $1/n$ visant simplement à assurer l'existence d'une telle densité.

Rappelons à ce stade que la *distance en variation totale* pour deux mesures de probabilité μ_1 et μ_2 est définie par

$$\mathrm{T}(\mu_1, \mu_2) = \sup_{B \in \mathcal{B}} \left|\mu_1(B) - \mu_2(B)\right|,$$

où $\mathcal{B}$ représente la tribu borélienne de $\mathbb{R}^d$. D'après un résultat de Scheffé (Devroye [10], page 2), lorsque μ_1 (*resp.* μ_2) admet une densité g_1 (*resp.* g_2) par rapport à la mesure de Lebesgue, on a

$$\mathrm{T}(\mu_1, \mu_2) = \frac{1}{2} \int |g_1 - g_2|,$$

la dernière intégrale étant la distance L_1 entre g_1 et g_2, la distance naturelle sur l'ensemble des densités. On en déduit donc que le supremum tend vers 0 selon un mode stochastique lorsque la distance L_1 entre g_1 et g_2 tend vers 0 selon le même mode. Ce critère L_1 est calculable sans aucune hypothèse additionnelle sur la densité et possède une signification claire en termes de différences de probabilités : si nous savons que $\int |g_1 - g_2| \leq 0.004$, alors nous savons aussi que les différences entre les probabilités d'ensemble sont au plus de 0.002. Le théorème de Scheffé entraîne également que la distance L_1 reste invariante par transformation bijective de l'espace $\mathbb{R}^d$. Cette propriété peut être exploitée lorsque l'on visualise sur écran l'erreur L_1 qui est commise dans la queue de distribution pour une densité ou un estimateur à support infini. Pour ce faire, il suffit de transformer la partie intéressante de l'axe réel d'une façon continue monotone en un intervalle borné. Alors que les formes des densités changent sous des transformations non linéaires, les distances L_1 entre elles restent invariantes. Les contributions dues aux queues peuvent être par conséquent rendues visibles. Enfin, l'erreur L_1 commise entre g_1 et g_2 peut être aisément visualisée : elle correspond à l'aire comprise entre les courbes représentatives des deux fonctions. Les raisons évoquées ci-dessus font du critère L_1 un des critères privilégiés des statisticiens.

Il est facile de voir que $\mathrm{T}(\mu_1, \mu_2) \leq 1$. Lorsque μ_1 possède une densité et $\mu_2 = \mu_n$ (la mesure empirique), on a même $\mathrm{T}(\mu_1, \mu_n) = 1$! En particulier, toute tentative de sélectionner une densité g en minimisant un critère du genre

$$\sup_{B \in \mathcal{B}} \left| \int_A g - \mu_n(A) \right|$$

est sans espoir, puisque alors la quantité à optimiser est constamment égale à 1. Bien entendu, cela ne se produit plus lorsque l'on remplace le supremum sur $\mathcal{B}$ par le supremum sur une classe plus petite, comme par exemple une classe de Yatracos. L'estimateur de la distance minimum est donc construit en minimisant un critère empirique *plus petit* que la variation totale.

Au terme de cette première étape de l'algorithme, nous disposons donc d'une famille d'estimateurs $(f_{n,k})_{k \geq 1}$, où chaque $f_{n,k}$, élément de $\mathcal{F}_k$, vérifie l'inégalité suivante (Devroye et Lugosi [13], Théorème 6.4) :

$$\int |f_{n,k} - f| \leq 3 \inf_{g \in \mathcal{F}_k} \int |f - g| + 4\Delta_k(f) + \frac{3}{n}. \tag{2.1}$$

Le terme $\inf_{g \in \mathcal{F}_k} \int |f - g|$ représente la plus petite erreur qui puisse être commise lorsque l'on approche f par un élément de $\mathcal{F}_k$. Evidemment, la valeur de ce terme d'erreur optimale, qui dépend de la cible f, nous est inconnue. Heuristiquement, l'inégalité (2.1) signifie donc que l'erreur L_1 commise par l'estimateur $f_{n,k}$ ne dépasse pas trois fois l'erreur minimum sur la classe plus un terme résiduel, $\Delta_k(f)$, qu'il va falloir s'attacher à contrôler. Le contrôle de ce terme peut être effectuée via un détour par la théorie de Vapnik et Chervonenkis [34] sur la convergence uniforme de la mesure empirique. A cet effet, désignons par $\mathcal{V}_k$ la *dimension de Vapnik-Chervonenkis* de la classe d'ensembles $\mathcal{A}_k$. Rappelons que $\mathcal{V}_k$ est définie comme le plus grand entier p tel que

$$\mathcal{S}_{\mathcal{A}_k}(p) = 2^p,$$

où $\mathcal{S}_{\mathcal{A}_k}(p)$ est le *coefficient de pulvérisation*, défini par

$$\mathcal{S}_{\mathcal{A}_k}(p) = \max_{x_1,\ldots,x_p \in \mathbb{R}^d} \mathrm{Card}\big\{\{x_1,\ldots,x_p\} \cap A : A \in \mathcal{A}_k\big\}\,.$$

Dit autrement, le coefficient $\mathcal{S}_{\mathcal{A}_k}(p)$ n'est autre que le nombre maximum de sous-ensembles de p points pouvant être obtenus à l'aide de recouvrements par des éléments de $\mathcal{A}_k$. La dimension de Vapnik-Chervonenkis (infinie par convention lorsque $\mathcal{S}_{\mathcal{A}_k}(p) = 2^p$ pour tout $p \geq 1$) mesure donc, en un certain sens, la richesse combinatoire discrète d'une classe donnée. Des arguments de nature combinatoire (voir Dudley [15]) montrent que si $\mathcal{A}_k$ a une dimension de Vapnik-Chervonenkis bornée par V_k, alors

$$\mathbf{E}\big\{\Delta_k(f)\big\} \leq C\sqrt{\frac{V_k}{n}}\,, \tag{2.2}$$

où C est une constante universelle (*i.e.* ne dépendant pas des paramètres du problème) strictement positive. Cette dernière majoration, combinée avec l'inégalité (2.1), nous conduit au résultat fondamental suivant :

$$\mathbf{E}\left\{\int |f_{n,k} - f|\right\} \leq 3 \inf_{g \in \mathcal{F}_k} \int |f-g| + 4C\sqrt{\frac{V_k}{n}} + \frac{3}{n}\,. \tag{2.3}$$

Cette inégalité est centrale dans les travaux de Devroye et Lugosi [13]. Elle est remarquable, et ceci pour au moins trois raisons. D'abord, il s'agit d'un résultat *non asymptotique* : nul besoin d'un passage à la limite pour avoir de l'information. Ce point de vue est d'ailleurs typique d'une problématique de type sélection de modèles : étant donné n observations et une famille de modèles candidats, comment choisir, à partir des seules données, un modèle qui, à distance finie, ne s'éloigne pas trop de l'optimum ? Ensuite, toutes les constantes sont *connues* et *explicitement calculables*. La seule difficulté, qui réside dans le calcul de V_k, est désormais de nature combinatoire. Enfin, l'inégalité (2.3) ne nécessite pour ainsi dire *aucune hypothèse de régularité* sur la densité cible f. Il nous semble que ce fait est suffisamment rare pour mériter d'être souligné.

Il nous reste maintenant à choisir un estimateur particulier parmi les $(f_{n,k})_{k\geq 1}$, ce qui revient donc à sélectionner le paramètre k. Rappelons pour ce faire que l'indice de la représentation économique de f a été défini dans l'introduction comme la dimension du modèle le plus "parcimonieux" permettant de décrire f, *i.e.*

$$k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}\,.$$

Puisque, par construction, $f \in \mathcal{F}_{k^*}$, l'inégalité (2.3) se réduit à

$$\mathbf{E}\left\{ \int |f_{n,k} - f| \right\} \leq 4C\sqrt{\frac{V_k}{n}} + \frac{3}{n} \tag{2.4}$$

dès que $k \geq k^*$. Attendu que k^* est inconnu, une première idée consiste à laisser k croître avec n, de sorte que V_k/n tende vers 0. Mais, dans un tel cas de figure, la vitesse de convergence garantie par (2.4) est alors un $\mathrm{O}(\sqrt{V_k/n})$, loin (hormis les cas triviaux) de notre objectif $\mathrm{O}(1/\sqrt{n})$. Notons également qu'il n'est pas raisonnable d'appliquer directement la méthode combinatoire à la classe $\mathcal{F}$ (en considérant donc cette fois-ci k comme un paramètre *initial* à sélectionner) car, dans un tel cas de figure, la dimension de Vapnik-Chervonenkis des classes de Yatracos associées au problème devient beaucoup trop grande, typiquement infinie. C'est pour corriger cet ensemble de défauts que nous introduisons une étape supplémentaire dans l'algorithme.

Seconde étape. Dans cette seconde étape, nous sélectionnons parmi la famille $(f_{n,k})_{k\geq 1}$ un estimateur particulier, que nous notons $f_{n,\hat{k}}$ et auquel nous donnons le nom d'*estimateur de la distance minimum pénalisée.* La sélection s'effectue sur le paramètre k, à partir du critère

$$\hat{k} \in \mathrm{argmin}_{k\geq 1}\left\{ \sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) \right\},$$

où $\mathrm{pen}_n(k)$ est une *fonction de pénalité* (ou *pénalité*), connue mais à spécifier, à laquelle nous imposons de tendre vers l'infini avec k à n fixé, ce qui assure bien de l'existence d'un minimum. En outre, comme

$$\min_{k\geq 1}\left\{ \sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) \right\} \leq 1 + \mathrm{pen}_n(1),$$

on a $\mathrm{pen}_n(\hat{k}) \leq 1 + \mathrm{pen}_n(1)$. Il en résulte que les calculs pourront être limités aux seuls entiers k vérifiant $\mathrm{pen}_n(k) \leq 1 + \mathrm{pen}_n(1)$ (qui sont en nombre fini) ce qui, d'un point de vue algorithmique, simplifie considérablement la tâche.

Remarque 1. Lorsque les éléments de $\mathcal{F}_k$ ne sont pas tous des densités de probabilité, l'Exercice 6.2 de Devroye et Lugosi [13] montre que l'inégalité (2.1) peut être remplacée par

$$\int |f_{n,k} - f| \leq 5 \inf_{g\in\mathcal{F}_k} \int |f - g| + 4\Delta_k(f) + \frac{5}{n}.$$

Il est donc très facile, moyennant une légère perte dans les constantes, d'adapter tous nos résultats à ce contexte plus général.

Terminons ce paragraphe en signalant qu'il reste encore beaucoup à faire pour rendre les algorithmes de type Devroye-Lugosi utilisables et performants en pratique. Le lecteur curieux trouvera une première approche dans Devroye [11].

2.2 Résultats

Dans tout ce qui suit, la notation $\mathcal{B}$ représente la tribu borélienne de $\mathbb{R}^d$. Rappelons qu'une sous-famille $\mathcal{A}$ de $\mathcal{B}$ est un π*-système* si cette famille est stable par intersection finie : A, $B \in \mathcal{A}$ implique $A \cap B \in \mathcal{A}$. Voir Billingsley [3] pour plus de détails. Notre résultat essentiel est alors le suivant.

Théorème 2.1 *Soit $(\mathcal{A}_k)_{k\geq 1}$ la suite des classes de Yatracos associées aux modèles $(\mathcal{F}_k)_{k\geq 1}$. Pour chaque $k \geq 1$, on suppose que la dimension de Vapnik-Chervonenkis $\mathcal{V}_k$ de $\mathcal{A}_k$ est finie et on désigne par V_k un nombre réel satisfaisant $\mathcal{V}_k \leq V_k$. Soient $(x_k)_{k\geq 1}$ une famille de poids réels positifs ou nuls vérifiant la relation*

$$\sum_{k\geq 1} e^{-2x_k^2} < \infty\,. \tag{2.5}$$

Alors, à condition que $\mathcal{A}_1$ contienne un π-système qui engendre $\mathcal{B}$, l'estimateur de la distance minimum pénalisée $f_{n,\hat{k}}$ défini avec

$$\mathrm{pen}_n(k) = \frac{x_k + C\sqrt{V_k}}{\sqrt{n}}$$

satisfait

$$\mathbf{E}\Big\{\int |f_{n,\hat{k}} - f|\Big\} = \mathrm{O}\Big(\frac{1}{\sqrt{n}}\Big)\,.$$

L'hypothèse selon laquelle $\mathcal{A}_1$ (et donc chacune des classes $\mathcal{A}_k$) contient un π-système engendrant la tribu borélienne $\mathcal{B}$ est essentiellement technique et demeure vraie dans tous les exemples classiques. Elle n'est en aucun cas restrictive. Cette hypothèse est par exemple satisfaite en dimension un ($d = 1$) dès lors que la classe de Yatracos contient le π-système des intervalles et, plus généralement pour $d \geq 1$, le π-système des hyperrectangles d-dimensionnels. Il est d'ailleurs intéressant de mentionner que, toujours sous cette hypothèse, le critère $\sup_{A\in\mathcal{A}_k} |\int_A g_1 - \int_A g_2|$ devient une distance sur l'ensemble des densités (*cf.* la Proposition 4.1).

La fonction de pénalité, quant à elle, dépend de la famille de modèles $(\mathcal{F}_k)_{k\geq 1}$ au travers du choix des poids $(x_k)_{k\geq 1}$, soumis à la condition (2.5). L'idée la plus naturelle consiste à choisir $x_k = \sqrt{V_k}$. Dans ce cas, la condition (2.5) s'écrit

$$\sum_{k\geq 1} e^{-2V_k} < \infty\,.$$

On vérifie aisément que cette relation est satisfaite par les modèles classiques. On a, par exemple, $\mathcal{V}_k = \mathrm{O}(k^4)$ pour les mélanges de gaussiennes univariées et $\mathcal{V}_k \leq k+1$ pour les familles exponentielles. Pour plus de précisions, nous renvoyons le lecteur à Devroye et Lugosi [13], Chapitre 8, ainsi qu'à Anthony et Bartlett [1]. Soulignons enfin que l'idée de la pondération des modèles est due à Barron, Birgé et Massart [2] (voir également Castellan [5] et Massart [24] et noter à ce sujet que la quantité $\Delta_k(g)$ à optimiser n'est pas stricto sensu un contraste au sens statistique usuel).

3 Test de la taille d'un modèle

3.1 Procédure

Etant donné un entier $k_0 \geq 1$ et un nombre réel $\alpha \in]0,1[$, nous présentons dans cette section une procédure de test explicite relative à l'hypothèse nulle $\{\mathbf{H}_0 : k^* = k_0\}$ contre l'alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$, de niveau asymptotique α et de puissance asymptotique 1. Cette problématique de test, connexe au problème de la sélection d'un meilleur modèle développé dans la Section 2, a fait l'objet de plusieurs études. Dans le cas des mélanges, et en raison de la mauvaise identifiabilité des modèles, la loi limite du rapport de vraisemblance n'a été disponible que récemment (Dacunha-Castelle et Gassiat [7], [9]) et ce sont donc essentiellement des procédures de type bootstrap qui ont été majoritairement développées (McLachlan [26]). Pour de plus amples références sur ces questions, nous renvoyons le lecteur aux travaux précédemment cités.

Soit donc $\alpha \in]0,1[$ le niveau du test requis. La procédure que nous proposons est la suivante : on *accepte* l'hypothèse $\mathbf{H}_0$ si

$$k_0 \in \mathrm{argmin}_{k\geq 1}\left\{ \sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A)\right| + \mathrm{pen}_n(k)\right\}. \tag{3.1}$$

Ici, le terme $\mathrm{pen}_n(k)$ représente la *fonction de pénalité* définie par

$$\mathrm{pen}_n(k) = \begin{cases} -n^{-1/3} - 1/n & \text{pour } k = 1,\ldots,k_0-1 \\ 0 & \text{pour } k = k_0 \\ \left(x_k + C\sqrt{\mathcal{V}_{k_0}}\right)/\sqrt{n} + 1/n & \text{pour } k \geq k_0+1\,, \end{cases}$$

où la lettre C désigne la constante universelle de l'inégalité (2.2) et $(x_k)_{k\geq k_0+1}$ une suite de nombres réels positifs ou nuls vérifiant la relation

$$2\sum_{k\geq k_0+1} e^{-2x_k^2} \leq \alpha/2\,.$$

Remarquons que $\mathrm{pen}_n(k)$ tend vers l'infini avec k à n fixé, ce qui assure bien de l'existence d'un minimum dans (3.1). Afin de minimiser les calculs, on aura plutôt intérêt à choisir une suite $(x_k)_{k\geq k_0+1}$ tendant rapidement vers l'infini. Par exemple, le choix

$$2e^{-2x_k^2} = \frac{\alpha}{2^{k+1}}$$

sera préférable au choix

$$2e^{-2x_k^2} = \frac{\alpha}{2k(k+1)}\,.$$

Il y a néanmoins un prix à payer à laisser croître les x_k trop rapidement vers l'infini, *cf.* la Remarque 2 plus bas. Notons finalement le lien étroit qui unit la présente procédure de test et la procédure de sélection développée dans la Section 2. Hormis la définition de la pénalité, distincte dans les deux cas, le protocole de test peut s'énoncer ainsi : on garde $\mathbf{H}_0$ si k_0 fait partie des entiers "sélectionnables" par la méthode combinatoire pénalisée (voir également à ce sujet Devroye, Györfi et Lugosi [12]).

3.2 Résultats

Notre premier résultat assure que la procédure de test définie au paragraphe précédent a bien, asymptotiquement, le niveau requis.

Théorème 3.1 *Soit $(\mathcal{A}_k)_{k\geq 1}$ la suite des classes de Yatracos associées aux modèles $(\mathcal{F}_k)_{k\geq 1}$. Soit $k_0 \geq 1$ un entier et α un nombre réel dans $]0,1[$. On suppose que la dimension de Vapnik-Chervonenkis $\mathcal{V}_{k_0}$ de $\mathcal{A}_{k_0}$ est finie. Alors, à condition que $\mathcal{A}_1$ contienne un π-système qui engendre $\mathcal{B}$, la procédure de test de l'hypothèse $\{\mathbf{H}_0 : k^* = k_0\}$ contre l'hypothèse $\{\mathbf{H}_1 : k^* \neq k_0\}$ définie en (3.1) est de niveau asymptotique α.*

Le théorème suivant établit alors la convergence du test.

Théorème 3.2 *Avec les notations et hypothèses du Théorème 3.1, et sous la condition supplémentaire $\mathcal{V}_{k^*} < +\infty$, la procédure de test de l'hypothèse $\{\mathbf{H}_0 : k^* = k_0\}$ contre l'hypothèse $\{\mathbf{H}_1 : k^* \neq k_0\}$ définie en (3.1) est de puissance asymptotique* 1.

Remarque 2. La preuve du Théorème 3.2 montre que les constantes intervenant dans la vitesse de convergence du test diminuent avec la valeur du poids particulier x_{k^*} dès lors que $k^* > k_0$. Comme k^* est inconnu, il vaut donc mieux travailler avec une suite $(x_k)_{k \geq k_0+1}$ qui tende lentement vers l'infini. En accord avec la remarque de la fin du Paragraphe 3.1, il semble judicieux de choisir in fine une suite de poids $(x_k)_{k \geq k_0+1}$ qui tende *raisonnablement vite* vers l'infini.

Le lecteur vérifiera aisément que les différents exemples et commentaires présentés à l'issue de l'énoncé du Théorème 2.1 se transposent sans difficultés particulières aux Théorèmes 3.1 et 3.2. Soulignons néanmoins que le rôle joué par la pénalité dans la présente procédure de test diffère du rôle joué par la pénalité dans le problème de sélection de la Section 2. Intuitivement, la fonction de pénalité permet ici de contrôler le niveau de la procédure de test, alors qu'elle visait plutôt à limiter le nombre de composantes du modèle à choisir dans le problème de sélection.

4 Preuves

4.1 Preuve du Théorème 2.1

Dans tout ce paragraphe, $(\mathcal{A}_k)_{k \geq 1}$ représente la suite des classes de Yatracos associées aux modèles $(\mathcal{F}_k)_{k \geq 1}$. Pour chaque $k \geq 1$, on suppose que la dimension de Vapnik-Chervonenkis $\mathcal{V}_k$ de $\mathcal{A}_k$ est finie et on désigne par V_k un nombre réel satisfaisant $\mathcal{V}_k \leq V_k$. On rappelle que la lettre C désigne la constante universelle apparaissant dans l'inégalité (2.2).

Une conséquence facile de l'inégalité de McDiarmid (McDiarmid [25]) nous apprend que

$$\mathbf{P}\Big\{ \big|\Delta_k(f) - \mathbf{E}\{\Delta_k(f)\}\big| > t\Big\} \leq 2e^{-2nt^2} \tag{4.1}$$

pour tout $n \geq 1$ et $t > 0$. La combinaison de (2.2) et (4.1) fournit alors l'inégalité suivante, très utile par la suite :

$$\mathbf{P}\left\{\Delta_k(f) > \frac{t}{\sqrt{n}} + C\sqrt{\frac{V_k}{n}}\right\} \leq 2e^{-2t^2}. \tag{4.2}$$

Dans un souci de simplification, nous supposerons que chaque classe $\mathcal{F}_k$ est un sous-espace métrique *fermé* de l'ensemble des densités sur $\mathbb{R}^d$ muni de la distance L_1. Il ne s'agit en aucun cas d'une hypothèse restrictive, puisque, en vertu du principe d'extension par continuité, tout sous-espace métrique de L_1 peut être prolongé en un espace fermé (Dunford et Schwartz [16]).

Avant de prouver le Théorème 2.1, nous établissons une proposition et deux lemmes préliminaires. La preuve du Lemme 4.1 est une conséquence immédiate de l'inégalité (4.2).

Proposition 4.1 *Soit $k \geq 1$ et $\mathcal{A}_k$ la classe de Yatracos associée à $\mathcal{F}_k$. Supposons que $\mathcal{A}_k$ contienne un π-système qui engendre $\mathcal{B}$. Alors $\mathcal{F}_k$ est fermé en tant que sous-espace métrique de l'ensemble des densités muni de la distance D_k, définie par*

$$D_k(g_1, g_2) = \sup_{A \in \mathcal{A}_k} \left| \int_A g_1 - \int_A g_2 \right|.$$

Preuve On remarque en premier lieu que l'hypothèse *$\mathcal{A}_k$ contient un π-système engendrant $\mathcal{B}$* implique que D_k est bien une distance sur les densités (pour davantage de détails, on pourra se reporter à Billingsley [3]). On note également, d'après l'identité de Scheffé (Devroye [10], page 2), que pour deux densités g_1 et g_2,

$$D_k(g_1, g_2) \leq \frac{1}{2} \int |g_1 - g_2| \tag{4.3}$$

et, dès lors que $g_1, g_2 \in \mathcal{F}_k$,

$$D_k(g_1, g_2) = \frac{1}{2} \int |g_1 - g_2| . \tag{4.4}$$

Soit maintenant $(g_n)_{n \geq 1}$ une suite de Cauchy dans $\mathcal{F}_k$ pour la distance D_k. Clairement, d'après (4.4), $(g_n)_{n \geq 1}$ est aussi une suite de Cauchy pour la distance L_1. Par hypothèse (*cf.* le début de cette section), $\mathcal{F}_k$ est fermé en tant que sous-espace métrique de l'ensemble des densités muni de la distance L_1. Etant donné que l'ensemble des densités est lui-même fermé dans l'espace complet L_1, $\mathcal{F}_k$ est aussi complet pour la distance L_1. On en déduit qu'il existe $g \in \mathcal{F}_k$ telle que $\int |g_n - g| \to 0$ quand $n \to +\infty$. D'après (4.3), cela implique donc $D_k(g_n, g) \to 0$ quand $n \to +\infty$. En conclusion, l'ensemble $\mathcal{F}_k$ est complet, donc fermé, pour la distance D_k. ∎

Lemme 4.1 *Soit $(y_k)_{k \geq 1}$ une famille de nombres réels strictement positifs. Alors*

$$\mathbf{P}\left\{ \bigcup_{k \geq 1} \left[\sup_{A \in \mathcal{A}_k} \left| \int_A f - \mu_n(A) \right| > \frac{y_k + C\sqrt{V_k}}{\sqrt{n}} \right] \right\} \leq 2 \sum_{k \geq 1} e^{-2y_k^2} .$$

Lemme 4.2 *Supposons que $\mathcal{A}_1$ contienne un π-système qui engendre $\mathcal{B}$. Soient $(x_k)_{k\geq 1}$ une famille de poids réels positifs ou nuls et une constante universelle M vérifiant la relation*

$$\sum_{k\geq 1} e^{-2x_k^2} \leq M\,.$$

Soit $f_{n,\hat{k}}$ l'estimateur de la distance minimum pénalisée et $k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}$. Alors, avec le choix

$$\operatorname{pen}_n(k) = \frac{x_k + C\sqrt{V_k}}{\sqrt{n}}\,, \tag{4.5}$$

on a, pour n assez grand,

$$\mathbf{P}\left\{\hat{k} < k^*\right\} \leq 2Me^{-2n^{2/3}}.$$

Preuve Introduisons, pour $n \geq 1$, l'événement Ω_n défini par

$$\bigcap_{k\geq 1}\left[\sup_{A\in\mathcal{A}_k}\left|\int_A f - \mu_n(A)\right| \leq \frac{x_k + C\sqrt{V_k}}{\sqrt{n}} + \frac{1}{n^{1/6}}\right].$$

Une application directe du Lemme 4.1 montre que

$$\mathbf{P}\left\{\Omega_n\right\} \geq 1 - 2Me^{-2n^{2/3}}\,.$$

Si $k^* = 1$, la preuve est terminée. Supposons donc $k^* > 1$. On a

$$\begin{aligned}
&\mathbf{P}\left\{\hat{k} \geq k^*\right\}\\
&\quad\geq \mathbf{P}\Bigg\{\sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f_{n,k^*} - \mu_n(A)\right| + \operatorname{pen}_n(k^*)\\
&\qquad\qquad < \min_{1\leq k\leq k^*-1}\left[\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \operatorname{pen}_n(k)\right]\Bigg\}.
\end{aligned}$$

Par l'inégalité triangulaire, pour $k = 1, \ldots, k^* - 1$,

$$\begin{aligned}
\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \operatorname{pen}_n(k) \geq{}& \sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \int_A f\right|\\
&- \sup_{A\in\mathcal{A}_k}\left|\int_A f - \mu_n(A)\right| + \operatorname{pen}_n(k).
\end{aligned}$$

En utilisant le choix particulier (4.5) pour la fonction de pénalité, on déduit de ce qui précède que, sur Ω_n,

$$\begin{aligned}
&\min_{1\le k\le k^*-1}\left[\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k}-\mu_n(A)\right|+\mathrm{pen}_n(k)\right]\\
&\ge \min_{1\le k\le k^*-1}\inf_{g\in\mathcal{F}_k}\sup_{A\in\mathcal{A}_k}\left|\int_A f-\int_A g\right|-\frac{1}{n^{1/6}}\\
&:= m-\frac{1}{n^{1/6}}\,.
\end{aligned}$$

Par hypothèse, $\mathcal{A}_1$–et donc chaque $\mathcal{A}_k$–contient un π-système qui engendre $\mathcal{B}$. Nous en déduisons alors (Proposition 4.1) que les classes $\mathcal{F}_k$ sont fermées pour la distance D_k. La définition de k^* implique donc $m>0$.

Par ailleurs

$$\begin{aligned}
\sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f_{n,k^*}-\mu_n(A)\right|+\mathrm{pen}_n(k^*) &\le \sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f-\mu_n(A)\right|+\mathrm{pen}_n(k^*)+\frac{1}{n}\\
&\qquad\text{(par définition de } f_{n,k^*})\\
&\le 2\,\mathrm{pen}_n(k^*)+\frac{1}{n^{1/6}}+\frac{1}{n}\,,\\
&\qquad\text{(sur l'ensemble } \Omega_n)
\end{aligned}$$

et cette borne est (strictement) plus petite que $m-1/n^{1/6}$ pour n assez grand. Il s'ensuit donc, pour n assez grand,

$$\mathbf{P}\{\hat{k}\ge k^*\}\ge\mathbf{P}\{\Omega_n\}\ge 1-2Me^{-2n^{2/3}}.$$

■

Nous sommes maintenant en mesure de prouver le Théorème 2.1.

Preuve du Théorème 2.1 Soit, pour $t>0$, l'événement Ω_t défini par

$$\bigcap_{k\ge 1}\left[\sup_{A\in\mathcal{A}_k}\left|\int_A f-\mu_n(A)\right|\le\frac{x_k+C\sqrt{V_k}}{\sqrt{n}}+\frac{t}{\sqrt{n}}\right].$$

La définition même de l'estimateur pénalisé implique

$$\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f_{n,\hat{k}}-\mu_n(A)\right|+\mathrm{pen}_n(\hat{k})\le\sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f_{n,k^*}-\mu_n(A)\right|+\mathrm{pen}_n(k^*)\,.$$

Comme, pour tout $k \geq 1$,

$$\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \int_A f\right| - \sup_{A\in\mathcal{A}_k}\left|\int_A f - \mu_n(A)\right| \leq \sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right|,$$

il vient

$$\begin{aligned}\sup_{A\in\mathcal{A}_{\hat k}}\left|\int_A f_{n,\hat k} - \int_A f\right| - \sup_{A\in\mathcal{A}_{\hat k}}\left|\int_A f - \mu_n(A)\right| + \mathrm{pen}_n(\hat k)\\ \leq \sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f_{n,k^*} - \mu_n(A)\right| + \mathrm{pen}_n(k^*)\,.\end{aligned}$$

Nous en déduisons que

$$\begin{aligned}\sup_{A\in\mathcal{A}_{\hat k}}\left|\int_A f_{n,\hat k} - \int_A f\right| &\leq \sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f - \mu_n(A)\right| + \mathrm{pen}_n(k^*)\\ &\quad + \sup_{A\in\mathcal{A}_{\hat k}}\left|\int_A f - \mu_n(A)\right| - \mathrm{pen}_n(\hat k) + \frac{1}{n}\\ &\quad \text{(par définition de } f_{n,k^*})\\ &\leq \sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f - \mu_n(A)\right| + \mathrm{pen}_n(k^*) + \frac{t}{\sqrt{n}} + \frac{1}{n}\\ &\quad \text{(sur l'ensemble } \Omega_t)\\ &\leq 2\,\mathrm{pen}_n(k^*) + \frac{2t}{\sqrt{n}} + \frac{1}{n}\,. \qquad (4.6)\\ &\quad \text{(sur l'ensemble } \Omega_t)\end{aligned}$$

Le Lemme 4.1 nous apprend que

$$\mathbf{P}\{\Omega_t\} \geq 1 - 2Me^{-2t^2}\,.$$

En intégrant alors l'inégalité (4.6) par rapport à t, on obtient

$$\mathbf{E}\left\{\sup_{A\in\mathcal{A}_{\hat k}}\left|\int_A f_{n,\hat k} - \int_A f\right|\right\} \leq 2\,\mathrm{pen}_n(k^*) + \frac{M\sqrt{2\pi}}{\sqrt{n}} + \frac{1}{n}\,. \qquad (4.7)$$

Or, d'après le Lemme 4.2, pour n assez grand, $\mathbf{P}\{\hat k < k^*\} \leq 2Me^{-2n^{2/3}}$. En remarquant de plus que, pour $\hat k \geq k^*$,

$$\sup_{A\in\mathcal{A}_{\hat k}}\left|\int_A f_{n,\hat k} - \int_A f\right| = \frac{1}{2}\int |f_{n,\hat k} - f|\,,$$

on en déduit que, pour n assez grand,

$$\mathbf{E}\left\{\int |f_{n,\hat k} - f|\right\} \leq 4\,\mathrm{pen}_n(k^*) + \frac{2M\sqrt{2\pi}}{\sqrt{n}} + \frac{2}{n} + 4Me^{-2n^{2/3}}\,.$$

■

Remarque 3. Comme l'a souligné un des relecteurs de l'article, il est intéressant de noter que l'inégalité (4.7) fournit un résultat non asymptotique qui, certes, ne porte pas sur la distance en variation totale, mais a l'avantage d'être valable même si les modèles ne sont pas emboîtés. L'utilisation du Lemme 4.2 donne effectivement un résultat asymptotique pour la variation totale et uniquement pour des modèles emboîtés.

5 Preuve des Théorèmes 3.1 et 3.2

Dans cette section, $(\mathcal{A}_k)_{k\geq 1}$ désigne la suite des classes de Yatracos associées aux modèles $(\mathcal{F}_k)_{k\geq 1}$ et $\mathcal{V}_k$ la dimension de Vapnik-Chervonenkis de $\mathcal{A}_k$.

5.1 Preuve du Théorème 3.1

En désignant par $\mathbf{P}_0$ la probabilité sous l'hypothèse nulle, nous pouvons écrire :

$$\begin{aligned}
&\mathbf{P}_0\{\text{rejeter } \mathbf{H}_0\}\\
&\quad= \mathbf{P}_0\left\{ \min_{k\geq 1}\left\{ \sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \mathrm{pen}_n(k)\right\} < \sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0} - \mu_n(A)\right|\right\}\\
&\quad\leq \sum_{\substack{k\geq 1\\ k\neq k_0}} \mathbf{P}_0\left\{ \sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \mathrm{pen}_n(k) < \sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0} - \mu_n(A)\right|\right\}\\
&\quad\leq \sum_{\substack{k\geq 1\\ k\neq k_0}} \mathbf{P}_0\left\{ \sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f)\right\},\\
&\qquad\text{(car, sous } \mathbf{H}_0,\ f\in\mathcal{F}_{k_0} \text{ et par définition de } f_{n,k_0})
\end{aligned}$$

où l'on rappelle que

$$\Delta_{k_0}(f) = \sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f - \mu_n(A)\right|.$$

Ainsi

$$
\begin{aligned}
&\mathbf{P}_0\{\text{rejeter } \mathbf{H}_0\} \\
&\quad \leq \sum_{\substack{k\geq 1\\ k\neq k_0}} \mathbf{P}_0\bigg\{ \sup_{A\in\mathcal{A}_k} \bigg| \int_A f_{n,k} - \mu_n(A)\bigg| + \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f)\bigg\} \\
&\quad = \sum_{k=1}^{k_0-1} \mathbf{P}_0\bigg\{ \sup_{A\in\mathcal{A}_k} \bigg| \int_A f_{n,k} - \mu_n(A)\bigg| + \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f)\bigg\} \\
&\qquad + \sum_{k\geq k_0+1} \mathbf{P}_0\bigg\{ \sup_{A\in\mathcal{A}_k} \bigg| \int_A f_{n,k} - \mu_n(A)\bigg| + \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f)\bigg\} .
\end{aligned}
\tag{5.1}
$$

Examinons tout d'abord le premier des deux termes de l'expression (5.1) (qui n'a évidemment de sens que si $k_0 > 1$). Pour $k = 1, \dots, k_0 - 1$, il vient

$$
\begin{aligned}
\sup_{A\in\mathcal{A}_k} \bigg| \int_A f_{n,k} - \mu_n(A)\bigg| &\geq \sup_{A\in\mathcal{A}_k} \bigg| \int_A f_{n,k} - \int_A f\bigg| - \sup_{A\in\mathcal{A}_k} \bigg| \int_A f - \mu_n(A)\bigg| \\
&\quad \text{(par l'inégalité triangulaire)} \\
&\geq \min_{1\leq k\leq k_0-1} \inf_{g\in\mathcal{F}_k} \sup_{A\in\mathcal{A}_k} \bigg| \int_A f - \int_A g\bigg| - \Delta_{k_0}(f) \\
&\quad (\text{car } \mathcal{A}_k \subset \mathcal{A}_{k_0}) \\
&:= m - \Delta_{k_0}(f) .
\end{aligned}
$$

Par hypothèse, $\mathcal{A}_1$–et donc chaque $\mathcal{A}_k$–contient un π-système qui engendre $\mathcal{B}$. Nous en déduisons (Proposition 4.1) que les classes $\mathcal{F}_k$ sont fermées pour la distance D_k. La définition de k^* (rappelons que $k^* = k_0$ sous $\mathbf{H}_0$) implique donc $m > 0$. On obtient alors

$$
\begin{aligned}
&\sum_{k=1}^{k_0-1} \mathbf{P}_0\bigg\{ \sup_{A\in\mathcal{A}_k} \bigg| \int_A f_{n,k} - \mu_n(A)\bigg| + \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f)\bigg\} \\
&\quad \leq \sum_{k=1}^{k_0-1} \mathbf{P}_0\bigg\{ 2\Delta_{k_0}(f) > m + \mathrm{pen}_n(k) - \frac{1}{n}\bigg\} \\
&\quad \leq (k_0 - 1)\, \mathbf{P}_0\bigg\{ \Delta_{k_0}(f) > \frac{m}{4}\bigg\} \quad \text{pour } n \text{ assez grand} \\
&\quad \leq \alpha/2 \quad \text{pour } n \text{ assez grand,}
\end{aligned}
$$

où, dans la dernière inégalité, nous avons utilisé la finitude de $\mathcal{V}_{k_0}$ et l'inégalité (4.2).

Passons maintenant à l'analyse du second terme de l'expression (5.1). Nous avons

$$
\begin{aligned}
&\sum_{k\geq k_0+1} \mathbf{P}_0\left\{ \sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f) \right\} \\
&\quad \leq \sum_{k\geq k_0+1} \mathbf{P}_0\left\{ \mathrm{pen}_n(k) - \frac{1}{n} < \Delta_{k_0}(f) \right\} \\
&\quad = \sum_{k\geq k_0+1} \mathbf{P}_0\left\{ \Delta_{k_0}(f) > \frac{x_k + C\sqrt{\mathcal{V}_{k_0}}}{\sqrt{n}} \right\} \\
&\qquad \text{(par définition de la fonction de pénalité pour } k \geq k_0+1\text{)} \\
&\quad \leq 2 \sum_{k\geq k_0+1} e^{-2x_k^2} \quad \big(\text{par l'inégalité (4.2)}\big) \\
&\quad \leq \alpha/2 \quad \text{(par définition des } x_k\text{)}.
\end{aligned}
$$

Le résultat souhaité s'ensuit. ■

5.2 Preuve du Théorème 3.2

Désignons par $\mathbf{P}_1$ la probabilité sous l'hypothèse alternative. Afin de prouver le Théorème 3.2, nous devons montrer que $\mathbf{P}_1\{\text{accepter } \mathbf{H}_0\}$ tend vers 0 lorsque n croît. On écrit pour cela

$$
\begin{aligned}
&\mathbf{P}_1\{\text{accepter } \mathbf{H}_0\} \\
&\quad = \mathbf{P}_1\left\{ \sup_{A\in\mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| = \min_{k\geq 1} \left\{ \sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) \right\} \right\} \\
&\quad \leq \mathbf{P}_1\left\{ \sup_{A\in\mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| \leq \sup_{A\in\mathcal{A}_{k^*}} \left| \int_A f_{n,k^*} - \mu_n(A) \right| + \mathrm{pen}_n(k^*) \right\} \\
&\quad \leq \mathbf{P}_1\left\{ \sup_{A\in\mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| - \mathrm{pen}_n(k^*) - \frac{1}{n} < \Delta_{k^*}(f) \right\},
\end{aligned}
$$

où, pour la dernière inégalité, nous utilisons la définition de f_{n,k^*} et le fait que $f \in \mathcal{F}_{k^*}$.

Il convient ici de distinguer le cas $k_0 < k^*$ du cas $k_0 > k^*$ (rappelons que $k^* \neq k_0$ sous $\mathbf{H}_1$). Dans la première situation (qui n'a évidemment de sens que si $k^* > 1$), en raisonnant comme dans la preuve du Théorème 2.1, on trouve

$$\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0}-\mu_n(A)\right| \geq m-\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f-\mu_n(A)\right|$$
$$\geq m-\Delta_{k^*}(f)\,,$$

où m est une constante strictement positive. Ainsi, dans ce cas,

$$\mathbf{P}_1\{\text{accepter } \mathbf{H}_0\} \leq \mathbf{P}_1\left\{\Delta_{k^*}(f) > \frac{m}{4}\right\}$$

pour n assez grand. D'après l'inégalité (4.2), ce dernier terme tend bien vers 0. Finalement, si k_0 vérifie la condition $k_0 > k^*$ (qui n'a de sens que si $k_0 > 1$), alors

$$\mathbf{P}_1\{\text{accepter } \mathbf{H}_0\} \leq \mathbf{P}_1\left\{-\operatorname{pen}_n(k^*)-\frac{1}{n} < \Delta_{k^*}(f)\right\}$$
$$\leq \mathbf{P}_1\left\{\Delta_{k^*}(f) > \frac{1}{n^{1/3}}\right\},$$

par définition de la fonction de pénalité pour $k = 1,\ldots,k_0-1$. On déduit alors de l'inégalité (4.2) le résultat souhaité. ■

Remerciements. Cet article effectue une synthèse de plusieurs résultats obtenus en collaboration avec Luc Devroye (Université McGill de Montréal), que j'associe au prix qui m'a été décerné par la Société Française de Statistique. Qu'il me soit également permis de remercier Henri Caussinus, Editeur en Chef du Journal de la SFdS, ainsi que deux relecteurs anonymes : leurs commentaires, leurs critiques et leurs suggestions constructives m'ont permis d'améliorer substantiellement la qualité de ce travail.

Références

[1] Anthony, M. et Bartlett, P.L. (1999). *Neural Network Learning : Theoretical Foundations*, Cambridge University Press, Cambridge.

[2] Barron, A., Birgé, L. et Massart, P. (1999). Risk bounds for model selection via penalization, *Probability Theory and Related Fields*, **Vol. 113**, pp. 301–413.

[3] Billingsley, P. (1995). *Probability and Measure, 3rd Edition*, Wiley, New York.

[4] Bishop, C.L. (1994). Mixture density networks, *Neural Computing Research Group Report NCRG/94/004*, Department of Computer Science and Applied Mathematics, Aston University, Birmingham.

[5] Castellan, G. (2000). Sélection d'histogrammes à l'aide d'un critère de type Akaike, *Comptes Rendus de l'Académie des Sciences de Paris*, **Vol. 330**, pp. 729–732.

[6] Celeux, G., Hurn, M. et Robert, C.P. (2000). Computational and inferential difficulties with mixture posterior distributions, *Journal of the American Statistical Association*, **Vol. 95**, pp. 957–970.

[7] Dacunha-Castelle, D. et Gassiat, E. (1997). Testing in locally conic models, and application to mixture models, *ESAIM : Probability and Statistics*, **Vol. 1**, pp. 285–317.

[8] Dacunha-Castelle, D. et Gassiat, E. (1997). The estimation of the order of a mixture model, *Bernoulli*, **Vol. 3**, pp. 279–299.

[9] Dacunha-Castelle, D. et Gassiat, E. (1999). Testing the order of a model using locally conic parametrization : population mixtures and stationary ARMA processes, *The Annals of Statistics*, **Vol. 27**, pp. 1178–1209.

[10] Devroye, L. (1987). *A Course in Density Estimation*, Birkhäuser, Boston.

[11] Devroye, L. (1997). Universal smoothing factor selection in density estimation : theory and practice, *Test*, **Vol. 6**, pp. 223–320.

[12] Devroye, L., Györfi, L. et Lugosi, G. (2002). A note on robust hypothesis testing, *IEEE Transactions on Information Theory*, **Vol. 48**, pp. 2111–2114.

[13] Devroye, L. et Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*, Springer–Verlag, New York.

[14] Diebolt, J. et Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling, *Journal of the Royal Statistical Society, Series B*, **Vol. 56**, pp. 363–375.

[15] Dudley, R.M. (1978). Central limit theorems for empirical measures, *The Annals of Probability*, **Vol. 6**, pp. 899–929.

[16] Dunford, N. et Schwartz, J.T. (1963). *Linear Operators Part I*, Wiley, New York.

[17] Everitt, B.S. et Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman and Hall, London.

[18] Figueiredo, M.A.T. et Jain, A.K. (2002). Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **Vol. 24**, pp. 381–396.

[19] Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks, *The Annals of Statistics*, **Vol. 31**, pp. 833–851.

[20] Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume II*, pp. 807–810.

[21] Hurn, M., Justel, A. et Robert, C.P. (2003). Estimating mixtures of regressions, *Journal of Computational and Graphical Statistics*, **Vol. 12**, pp. 1–25.

[22] James, L.F., Priebe, C.E. et Marchette, D.J. (2001). Consistent estimation of mixture complexity, *The Annals of Statistics*, **Vol. 29**, pp. 1281–1296.

[23] Jordan, M.I. et Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm, *Neural Computation*, **Vol. 6**, pp. 181–214.

[24] Massart, P. (2000). Some applications of concentration inequalities to statistics, *Annales de la Faculté des Sciences de Toulouse*, **Vol. 9**, pp. 245–303.

[25] McDiarmid, C. (1989). On the method of bounded differences, in *Surveys in Combinatorics 1989*, pp. 148–188, Cambridge University Press, Cambridge.

[26] McLachlan, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Journal of Applied Statistics*, **Vol. 36**, pp. 318–324.

[27] McLachlan, G.J. et Basford, K.E. (1988). *Mixture Models : Inference and Applications to Clustering*, Marcel Dekker, New York.

[28] McLachlan, G.J. et Peel, D. (2000). *Finite Mixture Models*, John Wiley, New York.

[29] Priebe, C.E. (1994). Adaptive mixtures, *Journal of the American Statistical Association*, **Vol. 89**, pp. 796–806.

[30] Richardson, S. et Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society, Series B*, **Vol. 59**, pp. 731–792.

[31] Roeder, K. et Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association*, **Vol. 92**, pp. 894–902.

[32] Rogers, G.W., Marchette, D.J. et Priebe, C.E. (2002). A procedure for model complexity selection in semiparametric mixture model density estimation, *Technical Report, Naval Surface Warfare Center, Dahlgren Division, Virginia.*

[33] Titterington, D.M., Smith, A.F.M. et Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester.

[34] Vapnik, V.N. et Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and its Applications*, **Vol. 16**, pp. 264–280.

[35] Yatracos, Y.G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy, *The Annals of Statistics*, **Vol. 13**, pp. 768–774.

[36] Zeevi, A. et Meir, R. (1997). Density estimation through convex combinations of densities ; approximation and estimation bounds, *Neural Networks*, **Vol. 10**, pp. 90–109.