Density Estimation by the Penalized Combinatorial Method

Gérard BIAU [a] and Luc DEVROYE [b,*]

[a] Laboratoire de Statistique Théorique et Appliquée,
Université Pierre et Marie Curie – Paris VI,
Boîte 158, 175 rue du Chevaleret, 75013 Paris, France;
[b] School of Computer Science, McGill University,
Montreal, Canada H3A 2K6

Abstract

Let f be an unknown multivariate density belonging to a prespecified parametric class of densities, $\mathcal{F}_k$, where k is unknown, but $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k and each $\mathcal{F}_k$ has finite Vapnik-Chervonenkis dimension. Given an *i.i.d.* sample of size n drawn from f, we show that it is possible to select automatically, and without extra restrictions on f, an estimate $f_{n,\hat{k}}$ with the property that $\mathbf{E}\{\int |f_{n,\hat{k}} - f|\} = \mathrm{O}(1/\sqrt{n})$. Our method is inspired by the combinatorial tools developed in Devroye and Lugosi [16] and it includes a wide range of density models, such as mixture models or exponential families.

Index Terms — Multivariate density estimation, Vapnik-Chervonenkis dimension, mixture densities, penalization.

AMS 2000 Classification: Primary 62G05.

1 Introduction

We consider the general problem of estimating a density f on $\mathbb{R}^d$ that belongs to a prespecified *parametric* class of densities, $\mathcal{F}_k$, where k is unknown, but $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Define

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

In the union above, $\mathcal{F}_k$ denotes, for each fixed $k \geq 1$, a given class of densities, parameterized by one or more parameters, and considered from a topological

*Corresponding author. Email: luc@cs.mcgill.ca .

point of view as a closed metric subspace of the space of all densities on $\mathbb{R}^d$ endowed with the L_1 metric. Note that the requirement that $\mathcal{F}_k$ is closed for the L_1 metric is not restrictive, since any metric subspace of L_1 can be extended into a closed one by the principle of extension by continuity (Dunford and Schwartz [19]). For example, $\mathcal{F}_k$ might be the class of all mixtures of k Gaussians on $\mathbb{R}^d$, see below. Given a random sample $X_1, \ldots, X_n$ drawn from f, this article proposes a general methodology to pick automatically, and without extra restrictions on f, a density estimate $f_{n,\hat{k}}$ in $\mathcal{F}$ with the property that

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

The estimate $f_{n,\hat{k}}$ will be selected by a *penalized combinatorial criterion*, inspired by the combinatorial tools developed in Devroye and Lugosi [16]. Before we present our method, we illustrate the generality of the approach by working out examples for two important parametric classes.

Example 1: Mixture classes. Consider first the classes $\mathcal{F}_k$ of all mixtures of k normal densities over $\mathbb{R}^d$, that is, the classes of all densities of form

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

where $(p_1, \ldots, p_k)$ is a probability vector, $\Sigma_1, \ldots, \Sigma_k$ are positive definite $d \times d$ matrices, and $m_1, \ldots, m_k$ are arbitrary elements of $\mathbb{R}^d$. An enormous body of literature exists regarding the application, computational issues and theoretical aspects of mixture models when the number of components is known, but estimating the unknown number of components remains an area of intense research. The scope of application is vast, as mixture models are routinely employed across the entire diverse application range of statistics, including nearly all of the social and experimental sciences. For early references, see Everitt and Hand [20], Titterington, Smith and Makov [43], McLachlan and Basford [34], and McLachlan and Peel [35]. The commonly used method for estimating the parameters of a mixture is the EM (expectation-maximization) algorithm (see Redner and Walker [37]). While originally designed for fixed mixture classes, such as mixtures of k Gaussians, the problem of the unknown k has received some attention in the Bayesian literature (Diebolt and Robert [17], Richardson and Green [38], Roeder and Wasserman [39], Celeux, Hurn and Robert [9], and Hurn, Justel and Robert [26]). The statistical learning community has also looked in depth at the problem (Bishop [6], Jordan and Jacobs [28], Zeevi and Meir [49], Figueiredo and Jain

[21]). In clustering, or unsupervised learning, one often makes an assumption about the number of clusters and the distribution within each cluster. Estimating the distributions in the clusters and the weights of the clusters then leads to a natural way of clustering. Likelihood ratios have been used for this in most works, from Hartigan [25] to Fukumizu [23]. Dacunha-Castelle and Gassiat [10], [11], [12] on the other hand use the moment method for identification and estimation of the number of components. The most recent attempts at estimating the mixture density parameters and the number of mixture densities jointly are by Priebe [36], James, Priebe and Marchette [27], and Rogers, Marchette and Priebe [40].

Example 2: Increasing exponential families. Each density f in an *exponential family* $\mathcal{F}_k$ may be written in the form

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

where θ belongs to some parameter set Θ, $\psi_1, \ldots, \psi_k : \mathbb{R}^d \to \mathbb{R}$, $\beta : \mathbb{R}^d \to [0, \infty)$, $\alpha, \pi_1, \ldots, \pi_k : \Theta \to \mathbb{R}$ are fixed functions, and c is a normalization constant. Examples of exponential families include classes of Gaussian, gamma, beta, Rayleigh, and Maxwell densities. By allowing k to grow, this model can become very rich and powerful. In fact, by taking the more classical statistical view, and concentrating on identification of the parameters, one is doomed to run into problems of identifiability and unstable or non-converging estimation algorithms. Rather than focusing on the parameters, we will look directly at the performance of the estimate without worrying about the consistency in the space of all unknown parameters.

The paper is organized as follows. In section 2, we present our estimation procedure as well as some useful related tools. The main result, L_1-optimality of $f_{n,\hat{k}}$, is stated in section 3. The proofs are gathered in section 4.

2 The penalized combinatorial method

Using ideas from Yatracos [48], Devroye and Lugosi explore in [16] a new paradigm for the data-based or automatic selection of the free parameters of density estimates in general, so that the expected error is within a given constant multiple of the best possible error. To summarize in the present context, fix $k \geq 1$, and define a density estimate $f_{n,k}$ in $\mathcal{F}_k$ as follows. First introduce the class of sets

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

($\mathcal{A}_k$ is the so-called *Yatracos class* associated with $\mathcal{F}_k$) and the goodness criterion for a density $g \in \mathcal{F}_k$:

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

where $\mu_n(A) = (1/n) \sum_{i=1}^n \mathbf{1}_{[X_i \in A]}$ is the empirical measure associated with the sample $X_1, \ldots, X_n$. For each $k \geq 1$, the *minimum distance estimate* $f_{n,k}$ is defined as any density estimate selected from among those densities $f \in \mathcal{F}_k$ with

$$\Delta_k(f) < \inf_{g \in \mathcal{F}_k} \Delta_k(g) + \frac{1}{n}.$$

Note that the $1/n$ term here is added to ensure the existence of such a density estimate. For each minimum distance estimate $f_{n,k}$, we have (Devroye and Lugosi [16], Theorem 6.4)

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

The uniform convergence of empirical measures as developed by Vapnik and Chervonenkis [46] can now be applied to density estimation via the term $\Delta_k(f)$. To this aim, we let $\mathcal{V}_k$ be the *Vapnik-Chervonenkis dimension* of the class of sets $\mathcal{A}_k$ (Vapnik and Chervonenkis [46]). Recall that $\mathcal{V}_k$ is defined as the largest integer p such that

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

where $\mathcal{S}_{\mathcal{A}_k}(p)$ is the *Vapnik-Chervonenkis shatter coefficient*, defined by

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

If $\mathcal{S}_{\mathcal{A}_k}(p) = 2^p$ for all p, then we say that $V = \infty$. A standard inequality from empirical process theory (Dudley [18]) shows that if $\mathcal{A}_k$ has Vapnik-Chervonenkis dimension bounded by V_k, then

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

where C is a universal positive constant. Using the original result of Dudley [18], the value of the constant C is found to be around 65. However, this value can undoubtedly be sharpened.

A simple consequence of the well-known *bounded difference inequality* (McDiarmid [33]) tells us that

$$\mathbf{P}\Big\{ \big|\Delta_k(f) - \mathbf{E}\big\{\Delta_k(f)\big\}\big| > t\Big\} \leq 2e^{-2nt^2} \tag{2.3}$$

for any $n \geq 1$ and $t > 0$. This shows that for any class $\mathcal{A}_k$, the maximal deviation is sharply concentrated around its mean. Combining (2.2) and (2.3) leads to the following useful inequality:

$$\mathbf{P}\left\{\Delta_k(f) > \frac{t}{\sqrt{n}} + C\sqrt{\frac{V_k}{n}}\right\} \leq 2e^{-2t^2}. \tag{2.4}$$

Inequalities (2.1) and (2.2) then imply

$$\mathbf{E}\Big\{ \int |f_{n,k} - f| \Big\} \leq 3 \inf_{g\in\mathcal{F}_k} \int |f - g| + 4C\sqrt{\frac{V_k}{n}} + \frac{3}{n}\,. \tag{2.5}$$

We introduce the *index of the economical representation of* f as

$$k_0 = \min\{k \geq 1 : f \in \mathcal{F}_k\}\,.$$

Naturally, as it is assumed that $f \in \mathcal{F}$, one has $k_0 < \infty$. Thus k_0 represents the index of the most parsimonious model for f. Since $f \in \mathcal{F}_{k_0}$, inequality (2.5) reduces to

$$\mathbf{E}\Big\{ \int |f_{n,k} - f| \Big\} \leq 4C\sqrt{\frac{V_k}{n}} + \frac{3}{n} \tag{2.6}$$

as soon as $k \geq k_0$. As k typically grows with n–it is a parameter of our choice–, we see that the rate of convergence guaranteed by (2.6) is not quite $\mathrm{O}(1/\sqrt{n})$. Indeed, if $V_k \to \infty$ as $k \to \infty$, then we cannot conclude that a square root n rate is obtained. If we do not let k tend to ∞, then we have no guarantees that $k \geq k_0$, and we cannot even conclude consistency. Since the class $\mathcal{F}$, the infinite union, is too "rich" to be dealt with directly by the original combinatorial method, as its Vapnik-Chervonenkis dimension is typically infinite, we cannot apply the original combinatorial method to it. It is to correct this situation that we present the penalized combinatorial method.

We define a *penalized minimum distance estimate* $f_{n,\hat{k}}$ as any estimate selected among the family of minimum distance estimates $f_{n,k}$, $k \geq 1$, with

$$\hat{k} \in \mathrm{argmin}_{k\geq 1}\left\{ \sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) \right\}, \tag{2.7}$$

where $\text{pen}_n(k)$ is a *penalty function* to be specified later but tending to infinity with k. This, together with the fact that the right-hand term of (2.7) is at most $1 + \text{pen}_n(1)$, shows that we need only do the computations for those k for which $\text{pen}_n(k) \leq 1 + \text{pen}_n(1)$–and there are a finite number of such terms.

The idea of minimizing the sum of an empirical term and a term penalizing the complexity has been investigated in various statistical problems. It was first introduced by Vapnik and Chervonenkis [47] and Vapnik [44] in pattern recognition as *structural risk minimization*. It was applied to regression estimation as *complexity regularization principle* in Barron [1] and was further investigated by Barron, Birgé and Massart [2], Krzyżak and Linder [30], and Kohler [29]. For further references, we refer the reader to Devroye, Györfi and Lugosi [15], Chapter 18, and Györfi, Kohler, Krzyżak and Walk [24], Chapter 12.

We will not be concerned with the actual details of the minimization algorithm. We realize however that more work is needed to make the present method computationally feasible (see the discussion at the end of section 3).

3 Results

Here and below, $\mathcal{B}$ denotes the class of all Borel sets of $\mathbb{R}^d$. Recall that a class $\mathcal{A}$ of subsets of $\mathcal{B}$ is a *π-system* if it is closed under the formation of finite intersections: $A, B \in \mathcal{A}$ implies $A \cap B \in \mathcal{A}$. See Billingsley [5] for details.

Our result is as follows:

Theorem 3.1 *Let $(\mathcal{A}_k)_{k \geq 1}$ be the sequence of Yatracos classes associated with the models $(\mathcal{F}_k)_{k \geq 1}$. For each $k \geq 1$, assume that the Vapnik-Chervonenkis dimension $\mathcal{V}_k$ of $\mathcal{A}_k$ is finite and let V_k be a real number satisfying $\mathcal{V}_k \leq V_k$. Consider some universal constant M and some family of nonnegative weights $(x_k)_{k \geq 1}$ such that*

$$\sum_{k \geq 1} e^{-2x_k^2} \leq M \,. \tag{3.1}$$

Then, provided $\mathcal{A}_1$ contains a π-system that generates $\mathcal{B}$, the penalized minimum distance estimate $f_{n,\hat{k}}$ defined with

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

satisfies, for all n large enough,

$$\mathbf{E}\left\{ \int |f_{n,\hat{k}} - f| \right\} \leq \frac{D}{\sqrt{n}} + \frac{2}{n} + 4Me^{-2n^{2/3}} ,$$

where

$$D = 4\left(x_{k_0} + C\sqrt{V_{k_0}} + \frac{M\sqrt{2\pi}}{2} \right)$$

and

$$k_0 = \min\{k \geq 1 : f \in \mathcal{F}_k\} .$$

The requirement that $\mathcal{A}_1$–and thus each $\mathcal{A}_k$–contains a π-system generating the Borel sets $\mathcal{B}$ is essentially of technical nature and in no way restrictive. Observe for example that it is satisfied for $d = 1$ as soon as the Yatracos class contains the π-system of all intervals, and more generally for $d \geq 1$ the π-system of all d-dimensional rectangles. The penalty function depends on the weights $(x_k)_{k\geq 1}$ satisfying (3.1). A reasonable way of choosing those weights is to make them depend on k only through the majorizing sequence $(V_k)_{k\geq 1}$. One could be for instance interested in the choice $x_k = \sqrt{V_k}$. In this case, condition (3.1) reads

$$\sum_{k\geq 1} e^{-2V_k} \leq M . \tag{3.2}$$

As an illustration, consider the examples of section 1. It can be shown (see Devroye and Lugosi [16], Chapter 8) that $\mathcal{V}_k = \mathrm{O}(k^4)$ for the univariate Gaussian mixtures with k components, and $\mathcal{V}_k \leq k + 1$ for the exponential families. Inequality (3.2) is thus satisfied by a large collections of models.

It is strictly speaking not necessary that $V_k \to \infty$, although such situations are of little general interest. Indeed, if $\sup_{k\geq 1} V_k < \infty$, then the Vapnik-Chervonenkis dimension of $\mathcal{F}$ is finite, and one could just apply the ordinary combinatorial method. The idea of using the additional x_k's in the definition of the penalty is due to Barron, Birgé and Massart [2], who study performance bounds for model selection based on an empirical loss or contrast function with an added penalty term motivated by empirical process theory, and roughly proportional to the number of parameters needed to describe the model divided by the number of observations. See also Castellan [8] and Massart [32].

Summarizing, we have thus shown, assuming that $f \in \mathcal{F}$, how to pick a mixture complexity and a density from the given mixture, and still guarantee

an $O(1/\sqrt{n})$ rate of convergence for the expected error, just as if we had been given the mixture complexity beforehand. The constant D in the bound of Theorem 3.1 can undoubtedly be improved a bit. On the other hand, we realize that an important situation occurs when f is not assumed to be in one of the models $\mathcal{F}_k$. The penalized combinatorial method may be applied in this case, but one needs to carefully assess the expected L_1 error of the resulting density estimate as a function of a quantity like

$$\inf_{k\geq 1}\left\{\inf_{g\in\mathcal{F}_k}\int|f-g|+\sqrt{\frac{V_k}{n}}\right\}.$$

We will deal with this problem elsewhere. The challenge here will be to make off the trade-off between the bias resulting from using a small model and the variance term. In the present paper the bias is assumed away, the variance dominates, and we obtain the fast convergence rate $1/\sqrt{n}$. The papers of Stone [42] and Barron and Sheu [3] on approximation of density functions by sequences of log-splines and exponential families, respectively, should provide good starting points.

The present paper, by virtue of its universality, is intended to describe a basis for future work in this area. Indeed, several questions need to be addressed, including that of the data-based choice of $(x_k)_{k\geq 1}$ (which will necessarily involve an optimization with respect to the second term in the error rates). Equally important is the computationally efficient selection of the minimum distance estimate. It is important to realize that the number of classes $\mathcal{F}_k$ that need to be considered, if $x_k=\sqrt{V_k}$, is conservatively bounded by the largest k for which $(C+1)\sqrt{V_k}\leq(C+1)\sqrt{V_1}+\sqrt{n}$ (as pointed out in the text). In fact, as the argument in (2.7) evolves with k as a global minimum m (at k_0, most likely) plus a term that grows roughly as $\sqrt{V_k}/\sqrt{n}$, we see that one can stop when $\sqrt{V_k}>m\sqrt{n}$, which in view of $m=O(1/\sqrt{n})$ in fact implies that with high probability, only a finite number of classes (independent of n) need ever be considered. So then, the main computational burden is in the computation of the minimum distance estimate $f_{n,k}$. To date, we do not know any method for its precise computation. Discretized methods and randomized methods that provide acceptable and computationally feasible approximation have been used in the simulation study of Devroye [14]. However, those simulations only involve one-dimensional problems, and thus, much more work is needed. In fact, the exploration of the relationship between class complexity, computational complexity and approximation seems very interesting. One may follow the role model of pattern recognition and machine learning, where these connections have been thoroughly studied. In

machine learning, the criterion to minimize is the empirical probability of error over a class of estimates (Devroye, Györfi and Lugosi [15]). No one knows how to do this exactly using limited computation. This has sparked the emergence of methods such as boosting (Freund and Shapire [22], Lugosi and Vayatis [31]) and approximate support vector machines (Schölkopf [41], Vapnik [45]). One should approach the computation of $f_{n,k}$ in the same manner.

The results of this paper are being used to test model complexity (is $f \in \mathcal{F}_k$ or not?), see Biau and Devroye [4]. Our methods also influenced the penalized model selection procedure of Bunea and Wegkamp [7] in nonparametric regression.

4 Proofs

Throughout this section, we let $(\mathcal{A}_k)_{k\geq 1}$ be the sequence of Yatracos classes associated with the models $(\mathcal{F}_k)_{k\geq 1}$. For each $k \geq 1$, we assume that the Vapnik-Chervonenkis dimension $\mathcal{V}_k$ of $\mathcal{A}_k$ is finite and we let V_k be a real number satisfying $\mathcal{V}_k \leq V_k$. Recall that the letter C stands for the universal constant of inequality (2.4).

Before proving Theorem 3.1, we state a technical proposition and two lemmas that are interesting by themselves. Proof of Lemma 4.1 is a straightforward consequence of inequality (2.4).

Proposition 4.1 *Assume that $\mathcal{F}_k$ is closed for the L_1 metric on densities and that $\mathcal{A}_k$ contains a π-system that generates $\mathcal{B}$. Then $\mathcal{F}_k$ is closed for the D_k metric on densities, defined by*

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

Proof First observe that the fact that $\mathcal{A}_k$ contains a π-system generating $\mathcal{B}$ forces D_k to be a metric on densities (see, for example, Billingsley [5]). Note also that, according to Scheffé's identity (Devroye [13], page 2), for two densities f and g,

$$D_k(f,g) \leq \frac{1}{2}\int |f-g| \tag{4.1}$$

and, whenever $f, g \in \mathcal{F}_k$,

$$D_k(f,g) = \frac{1}{2}\int |f-g|\,. \tag{4.2}$$

Now, let $(f_n)_{n\geq 1}$ be a Cauchy sequence in $\mathcal{F}_k$ for the D_k metric. Clearly, according to (4.2), $(f_n)_{n\geq 1}$ is also a Cauchy sequence for the L_1 metric. By assumption, $\mathcal{F}_k$ is closed for the L_1 metric on densities. Since the subspace of densities is closed in the complete space L_1, $\mathcal{F}_k$ is also complete as a subspace of densities. Therefore, one deduces that there exists $f \in \mathcal{F}_k$ such that $\int |f_n - f| \to 0$ as $n \to \infty$. According to (4.1), this implies that $D_k(f_n, f) \to 0$ as $n \to \infty$. Thus the set $\mathcal{F}_k$ is complete, and therefore closed, for the D_k metric. ■

Lemma 4.1 *Consider some arbitrary family of positive numbers $(y_k)_{k\geq 1}$. Then*

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

Lemma 4.2 *Assume that $\mathcal{A}_1$ contains a π-system that generates $\mathcal{B}$. Consider some universal constant M and some family of nonnegative weights $(x_k)_{k\geq 1}$ such that*

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

Let $f_{n,\hat{k}}$ be the penalized minimum distance estimate defined in section 2 and $k_0 = \min\{k \geq 1 : f \in \mathcal{F}_k\}$. Then, with the choice

$$\mathrm{pen}_n(k) = \frac{x_k + C\sqrt{V_k}}{\sqrt{n}} , \tag{4.3}$$

one has, for all n large enough,

$$\mathbf{P}\{\hat{k} < k_0\} \leq 2Me^{-2n^{2/3}} .$$

Proof For each $n \geq 1$, denote by Ω_n the event

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

From Lemma 4.1, we know that

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

If $k_0 = 1$, the proof is clear. Assume that $k_0 > 1$. We have

$$\begin{aligned} &\mathbf{P}\{\hat{k} \geq k_0\} \\ &\quad \geq \mathbf{P}\left\{ \sup_{A\in\mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| + \mathrm{pen}_n(k_0) \right. \\ &\qquad \left. < \min_{1\leq k\leq k_0-1} \left[\sup_{A\in\mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) \right] \right\} . \end{aligned}$$

By the triangle inequality, for $k = 1, \ldots, k_0 - 1$,

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

Using the particular choice (4.3) for the penalty function, we deduce from above that, on Ω_n,

$$\begin{aligned} &\min_{1 \leq k \leq k_0 - 1} \left[\sup_{A \in \mathcal{A}_k} \left| \int_A f_{n,k} - \mu_n(A) \right| + \mathrm{pen}_n(k) \right] \\ &\quad \geq \min_{1 \leq k \leq k_0 - 1} \inf_{g \in \mathcal{F}_k} \sup_{A \in \mathcal{A}_k} \left| \int_A f - \int_A g \right| - \frac{1}{n^{1/6}} \\ &\quad := m - \frac{1}{n^{1/6}}. \end{aligned}$$

By assumption, $\mathcal{A}_1$–and thus each $\mathcal{A}_k$–contains a π-system generating $\mathcal{B}$. Since the $\mathcal{F}_k$'s are closed for the L_1 metric on densities, we know from Proposition 4.1 that they are also closed for the D_k metric. Therefore, the definition of k_0 implies $m > 0$.

Moreover

$$\begin{aligned} \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f_{n,k_0} - \mu_n(A) \right| + \mathrm{pen}_n(k_0) &\leq \sup_{A \in \mathcal{A}_{k_0}} \left| \int_A f - \mu_n(A) \right| + \mathrm{pen}_n(k_0) + \frac{1}{n} \\ &\quad \text{(by definition of } f_{n,k_0}) \\ &\leq 2\,\mathrm{pen}_n(k_0) + \frac{1}{n^{1/6}} + \frac{1}{n}, \\ &\quad \text{(on the set } \Omega_n) \end{aligned}$$

and this bound is (strictly) smaller than $m - 1/n^{1/6}$ for all n large enough. Therefore, for all n large enough,

$$\mathbf{P}\{\hat{k} \geq k_0\} \geq \mathbf{P}\{\Omega_n\} \geq 1 - 2Me^{-2n^{2/3}}.$$

■

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1 Let $t > 0$ be given. Denote by Ω_t the event

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

The very definition of the penalized estimate leads to

$$\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f_{n,\hat{k}}-\mu_n(A)\right|+\mathrm{pen}_n(\hat{k})\le\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0}-\mu_n(A)\right|+\mathrm{pen}_n(k_0)\,.$$

Since, for every $k\ge 1$,

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

we obtain

$$\begin{aligned}\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f_{n,\hat{k}}-\int_A f\right|-\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f-\mu_n(A)\right|+\mathrm{pen}_n(\hat{k})\\ \le\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0}-\mu_n(A)\right|+\mathrm{pen}_n(k_0)\,.\end{aligned}$$

Thus we deduce that

$$\begin{aligned}\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f_{n,\hat{k}}-\int_A f\right|&\le\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f-\mu_n(A)\right|+\mathrm{pen}_n(k_0)\\ &\quad+\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f-\mu_n(A)\right|-\mathrm{pen}_n(\hat{k})+\frac{1}{n}\\ &\quad\text{(by definition of } f_{n,k_0})\\ &\le\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f-\mu_n(A)\right|+\mathrm{pen}_n(k_0)+\frac{t}{\sqrt{n}}+\frac{1}{n}\\ &\quad\text{(on the set }\Omega_t)\\ &\le 2\,\mathrm{pen}_n(k_0)+\frac{2t}{\sqrt{n}}+\frac{1}{n}\,.\qquad(4.4)\\ &\quad\text{(on the set }\Omega_t)\end{aligned}$$

Rephrasing Lemma 4.1, we know that

$$\mathbf{P}\{\Omega_t\}\ge 1-2Me^{-2t^2}\,.\qquad(4.5)$$

Define now

$$Z=\frac{\sqrt{n}}{2}\left[\sup_{A\in\mathcal{A}_{\hat{k}}}\left|\int_A f_{n,\hat{k}}-\int_A f\right|-2\,\mathrm{pen}_n(k_0)-\frac{1}{n}\right].$$

Inequalities (4.4) and (4.5) lead to

$$\mathbf{P}\{Z>t\}\le 2Me^{-2t^2}\,.$$

Therefore, using the fact that $\mathbf{E}\{Z\} \leq \int_0^\infty \mathbf{P}\{Z > t\}\, \mathrm{d}t$ for any integrable random variable Z, we obtain

$$\mathbf{E}\left\{ \sup_{A \in \mathcal{A}_{\hat{k}}} \left| \int_A f_{n,\hat{k}} - \int_A f \right| \right\} \leq 2\, \mathrm{pen}_n(k_0) + \frac{M\sqrt{2\pi}}{\sqrt{n}} + \frac{1}{n}\,.$$

According to Lemma 4.2, for all n large enough, $\mathbf{P}\{\hat{k} < k_0\} \leq 2Me^{-2n^{2/3}}$. Observing finally that, for $\hat{k} \geq k_0$,

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

one deduces that, for all n large enough,

$$\mathbf{E}\left\{ \int |f_{n,\hat{k}} - f| \right\} \leq 4\, \mathrm{pen}_n(k_0) + \frac{2M\sqrt{2\pi}}{\sqrt{n}} + \frac{2}{n} + 4Me^{-2n^{2/3}}\,.$$

■

Acknowledgment. The authors thank an Associate Editor and a referee for many valuable comments and suggestions.

References

[1] Barron, A. (1991). Complexity regularization with application to artificial neural networks, in *Nonparametric Functional Estimation and Related Topics*, ed. Roussas, G., pp. 501–576, NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

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

[3] Barron, A.R. and Sheu, C.H. (1991). Approximation of density functions by sequences of exponential families, *The Annals of Statistics*, Vol. 19, pp. 1347–1369.

[4] Biau, G. and Devroye, L. (2004). A note on density model size testing, *IEEE Transactions on Information Theory*, in press.

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

[6] 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.

[7] Bunea, F. and Wegkamp, M. (2003). A note on penalized minimum distance estimation in nonparametric regression, *The Canadian Journal of Statistics*, Vol. 31, pp. 267–274.

[8] 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.

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

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

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

[12] Dacunha-Castelle, D. and 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.

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

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

[15] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer–Verlag, New York.

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

[17] Diebolt, J. and 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.

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

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

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

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

[22] Freund, Y. and Schapire, R.E. (1999). A short introduction to boosting, *Journal of Japanese Society for Artificial Intelligence*, Vol. 14, pp. 771–780.

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

[24] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer–Verlag, New York.

[25] 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.

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

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

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

[29] Kohler, M. (1998). Nonparametric regression function estimation using interaction least squares splines and complexity regularization, *Metrika*, Vol. 47, pp. 147–163.

[30] Krzyżak, A. and Linder, T. (1998). Radial basis function networks and complexity regularization in function learning, *IEEE Transactions on Neural Networks*, Vol. 9, pp. 247–256.

[31] Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods (with discussion), *The Annals of Statistics*, in press.

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

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

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

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

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

[37] Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, Vol. 26, pp. 195–239.

[38] Richardson, S. and 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.

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

[40] Rogers, G.W., Marchette, D.J. and 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.

[41] Schölkopf, B. (1997). *Support Vector Learning*, Oldenbourg Verlag, Munich.

[42] Stone, C.J. (1990). Large sample inference for log-spline models, *The Annals of Statistics*, Vol. 18, pp. 717–741.

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

[44] Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*, Springer–Verlag, New York.

[45] Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory, 2nd Edition*, Springer–Verlag, New York.

[46] Vapnik, V.N. and 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.

[47] Vapnik, V.N. and Chervonenkis, A.Ya. (1974). *Theory of Pattern Recognition*, Nauka, Moscow (in Russian). German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin (1979).

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

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