A Note on Density Model Size Testing

Gérard Biau and Luc Devroye

***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 a random sample $X_1, \ldots, X_n$ drawn from f, an integer $k_0 \geq 1$ and a real number $\alpha \in (0,1)$, we introduce a new, simple, explicit α-level consistent testing procedure of the null hypothesis $\{\mathbf{H}_0 : k^* = k_0\}$ versus the alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. Our method is inspired by the combinatorial tools developed in Devroye and Lugosi [1] and it includes a wide range of density models, such as mixture models, neural networks or exponential families.**

***Index Terms*—Hypothesis testing, mixture densities, nonparametric estimation, penalization, Vapnik-Chervonenkis dimension.**

I. Introduction

LET f be an unknown density on $\mathbb{R}^d$ belonging to a prespecified 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–often parameterized by one or more parameters–and considered from a topological 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 [2]). For example, $\mathcal{F}_k$ might be the class of all mixtures of k Gaussians on $\mathbb{R}^d$. Since $f \in \mathcal{F}$, it is natural to define the *index of the economical representation of* f (James, Priebe, and Marchette [3]) as

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

Naturally, as it is assumed that $f \in \mathcal{F}$, one has $k^* < \infty$. Roughly speaking, k^* represents the index of the most parsimonious model for f. Now, given a random sample $X_1, \ldots, X_n$ drawn from f, an integer $k_0 \geq 1$ and a real number $\alpha \in (0,1)$, the purpose of this paper is to introduce a new, simple, explicit α-level consistent testing procedure of the null hypothesis $\{\mathbf{H}_0 : k^* = k_0\}$ versus the alternative $\{\mathbf{H}_1 : k^* \neq k_0\}$. The problem of testing hypotheses on k^* has received some attention in the literature, essentially for mixture classes, such as mixtures of Gaussians. Due to a lack of identifiability, the limiting distribution for the likelihood ratio test statistic was until recently unavailable, and so bootstrap testing methodologies have been developed (see, *e.g.*, McLachlan [4]). This lack of identifiability leads to the degeneracy of the Fisher information of the model, so that the classical chi-square theory does not apply. Dacunha-Castelle and Gassiat [5], [6] proposed a theory of reparametrization to solve such problems, that they called "locally conic parametrization". Roughly speaking, the idea is to approach the null hypothesis using directional submodels in which the Fisher information is normalized to be uniformly equal to one.

In the present manuscript, we propose a new testing methodology, which is close in spirit to Biau and Devroye [7], where we show how to pick automatically, and without extra restrictions on f, a density estimate $f_{n,\hat{k}}$ in $\mathcal{F}$ with the property that

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

Our approach is inspired by the combinatorial tools developed in Devroye and Lugosi [1] and recently extended to the problem of robust hypothesis testing in Devroye, Györfi, and Lugosi [8]. As we explain it in [7], our methodology is not proper to mixtures, and it includes a wide range of models, such as neural networks or exponential families. We refer the reader to [7] for a detailed discussion, examples and references. 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 II, we present our testing procedure as well as some useful related tools. The main results, level and consistency of the testing methodology, are stated in Section III. Proofs are gathered in Section IV.

II. The testing procedure

A. Presentation

In [1], Devroye and Lugosi explore 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

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

Manuscript received January 20, 2002; revised November 18, 2002.

G. Biau is with the Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie – Paris VI, 175 rue du Chevaleret, 75013 Paris, France (e-mail: biau@ccr.jussieu.fr).

L. Devroye is with the School of Computer Science, McGill University, Montreal, Canada H3A 2K6 (e-mail: luc@cs.mcgill.ca).

($\mathcal{A}_k$ 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.

Having disposed of this preliminary step, we let $\mathcal{V}_k$ be the *Vapnik-Chervonenkis dimension* of the class of sets $\mathcal{A}_k$ (Vapnik and Chervonenkis [9]). 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 $V = \infty$. Using combinatorial arguments, one can show (see Devroye and Lugosi [1], Chapter 4, and the references therein) that if $\mathcal{A}_k$ has Vapnik-Chervonenkis dimension $\mathcal{V}_k$, then

$$\mathbf{E}\big\{\Delta_k(f)\big\} \leq C\sqrt{\frac{\mathcal{V}_k}{n}} , \tag{1}$$

where C is a universal positive constant. Now, denoting by $\alpha \in (0,1)$ the required testing level, the proposed testing procedure is as follows: accept the null hypothesis $\mathbf{H}_0$ if

$$k_0 \in \mathrm{argmin}_{k\geq 1}\left\{ \sup_{A\in\mathcal{A}_k}\left| \int_A f_{n,k} - \mu_n(A)\right| + \mathrm{pen}_n(k)\right\} . \tag{2}$$

Here, the term $\mathrm{pen}_n(k)$ denotes a *penalty function* defined by

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

where the letter C stands for the universal constant of inequality (1) and $(x_k)_{k\geq k_0+1}$ is a sequence of nonnegative real numbers satisfying the condition

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

Note that $\mathrm{pen}_n(k)$ tends to infinity with k. This, together with the fact that the right-hand term of (2) is at most $1+\mathrm{pen}_n(1)$, shows that we need only do the computations for those k for which $\mathrm{pen}_n(k) \leq 1 + \mathrm{pen}_n(1)$. Therefore, a good choice for the sequence $(x_k)_{k\geq k_0+1}$ is one for which this sequence has a fast rate of divergence. For example, the choice

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

is preferable to the choice

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

Note however that there is a price of letting x_k grow rapidly, see Remark 2 in Section III.

B. Some useful results

For each minimum distance estimate $f_{n,k}$, we have (Devroye and Lugosi [1], Theorem 6.4)

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

Since $f \in \mathcal{F}_{k^*}$, this inequality reduces to

$$\int |f_{n,k} - f| \leq 4\Delta_k(f) + \frac{3}{n}$$

as soon as $k \geq k^*$. Note that if not all members of $\mathcal{F}_k$ are densities (just take for $\mathcal{F}_k$ the class of all neural networks with k hidden nodes or the class of series estimates with k basis functions) then, as shown in Exercise 6.2 of Devroye and Lugosi [1], (3) can be replaced by

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

The only impact of this is that $3/n$ has to be replaced by $5/n$. The effect is so slight that the reader can easily carry through the necessary adjustments in the sections that follow. Thus, we will assume in the sequel that all members in all classes $\mathcal{F}_k$ are densities.

With respect to the term $\Delta_k(f)$ in (3), a simple consequence of the well-known *bounded difference inequality* (McDiarmid [10]) tells us that

$$\mathbf{P}\Big\{ \big|\Delta_k(f) - \mathbf{E}\{\Delta_k(f)\}\big| > t\Big\} \leq 2e^{-2nt^2} \tag{4}$$

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 (1) and (4) leads to the following useful inequality:

$$\mathbf{P}\left\{\Delta_k(f) > \frac{t}{\sqrt{n}} + C\sqrt{\frac{\mathcal{V}_k}{n}}\right\} \leq 2e^{-2t^2}. \tag{5}$$

Finally, we will also use the following result, due to Talagrand [11], which states that there exists a universal positive constant D such that for all $n \geq 1$ and $\varepsilon > 0$,

$$\mathbf{P}\big\{\Delta_k(f) > \varepsilon\big\} \leq \frac{D}{\sqrt{n}\varepsilon}\left(\frac{Dn\varepsilon^2}{\mathcal{V}_k}\right)^{\mathcal{V}_k} e^{-2n\varepsilon^2} , \tag{6}$$

whenever $\mathcal{V}_k$ is finite.

III. 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 [12] for details. We remind the reader that k^*, the index of the economical representation of f, is defined by $k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}$. Our first result insures that the testing procedure defined in Section II is of level α.

Theorem 1: Let $(\mathcal{A}_k)_{k\geq 1}$ be the sequence of Yatracos classes associated with the models $(\mathcal{F}_k)_{k\geq 1}$. Let $k_0 \geq 1$ be an integer and α a real number in $(0,1)$. Assume that the Vapnik-Chervonenkis dimension $\mathcal{V}_{k_0}$ of $\mathcal{A}_{k_0}$ is finite. Then, provided $\mathcal{A}_1$ contains a π-system that generates $\mathcal{B}$, the testing procedure of $\{\mathbf{H}_0 : k^* = k_0\}$ *vs* $\{\mathbf{H}_1 : k^* \neq k_0\}$ as defined in (2) satisfies, for all n large enough,

$$\mathbf{P}_0\big(\text{rejecting } \mathbf{H}_0\big) \leq \alpha\,,$$

where $\mathbf{P}_0$ stands for the probability under the null hypothesis.

Remark 1: Theorem 1 holds *for all n large enough*, depending on the unknown f. From the proof it is easy to get estimates for the value of n necessary for the theorem to hold in terms of how well f can be approximated by densities in $\mathcal{F}_{k_0}$.

The next theorem deals with the consistency of the proposed test.

Theorem 2: Under the notations and assumptions of Theorem 1, with the additional condition that $\mathcal{V}_{k^*} < \infty$, there exists a positive constant K (depending on f) such that, for all $n \geq 1$,

$$\mathbf{P}_1\big(\text{rejecting } \mathbf{H}_0\big) \geq 1 - Kn^{-1/2+\mathcal{V}_{k^*}}e^{-2n^{1/3}},$$

where $\mathbf{P}_1$ stands for the probability under the alternative hypothesis.

Remark 2: It is seen from the proof that the value of the constant K increases with the value of the particular weight x_{k^*} whenever $k^* > k_0$. Since k^* is unknown, a good choice for the sequence $(x_k)_{k\geq k_0+1}$ is thus one for which this sequence has a slow rate of divergence. This, together with the discussion at the end of Subsection II-A, shows that the sequence $(x_k)_{k\geq k_0+1}$ should increase *reasonably fast* towards infinity.

As an illustration, just consider the classes $\mathcal{F}_k$ of all mixtures of k normal densities over $\mathbb{R}$, that is, the classes of all densities of form

$$f(x) = \sum_{i=1}^{k} \frac{p_i}{\sqrt{2\pi\sigma_i^2}}\, e^{-\frac{1}{2}(x-m_i)^2/\sigma_i^2}\,, \qquad (7)$$

where $(p_1,\ldots,p_k)$ is a probability vector, $\sigma_1,\ldots,\sigma_k$ are positive real numbers, and $m_1,\ldots,m_k$ are arbitrary elements of $\mathbb{R}$. 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 and testing 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 [13], Titterington, Smith, and Makov [14], McLachlan and Basford [15], and McLachlan and Peel [16]. The commonly used method for estimating the parameters of a mixture is the EM (expectation-maximization) algorithm (see Redner and Walker [17]). 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 [18], Richardson and Green [19], Roeder and Wasserman [20], Celeux, Hurn, and Robert [21], and Hurn, Justel, and Robert [22]). The statistical learning community has also looked in depth at the problem (Bishop [23], Jordan and Jacobs [24], Zeevi and Meir [25], Figueiredo and Jain [26]). 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 [27] to Fukumizu [28]. Dacunha-Castelle and Gassiat [5], [29], [6] 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 [30], James, Priebe, and Marchette [3], and Rogers, Marchette, and Priebe [31].

We draw attention on the fact that the conditions required in Theorem 1 and 2 are in no way restrictive and are in particular satisfied by a large choice of models. 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. In particular, it is satisfied by example (7). Moreover, it can be shown (see Devroye and Lugosi [1], Chapter 8) that $\mathcal{V}_k = \mathrm{O}(k^4)$, what ensures here that $\mathcal{V}_k < \infty$ for all $k \geq 1$. For more details and examples, we refer the reader to Biau and Devroye [7], where we use a close penalized combinatorial criterion to automatically pick a mixture complexity and a density from the given mixture, and still guarantee an $\mathrm{O}(1/\sqrt{n})$ rate of convergence for the expected L_1 error, just as if we had been given the mixture complexity beforehand. Observe, however, that the role played by the penalty in the present testing problem is slightly different from the role played by the penalty in the density selection problem studied in [7]. Roughly speaking, the penalty function allows here to control the level of the testing procedure, whereas in [7], it guarantees a good rate of convergence by limiting the number of selected mixture components.

Note that the idea of using the additional x_k's in the definition of the penalty is due to Barron, Birgé, and Massart [32], 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 [33] and Massart [34]. Rissanen [35], [36] and Rissanen, Speed, and Yu [37] proposed model selection based on minimal stochastic complexity or description length. The focus there was principally on rate of convergence. It is not directly obvious how to modify these methods for hypothesis testing.

Finally, we draw attention on the fact that we can surely use the test for testing $k \leq k_0$ versus $k > k_0$. Just apply the test k_0 times for $k = \ell$ versus $k \neq \ell$, $\ell = 1,\ldots,k_0$. If it is successful in one of these k_0 tests, then we say that $k \le k_0$ is successful. The details can easily be worked out by the reader.

IV. Proofs

A. Proof of Theorem 1

Before proving Theorem 1, we state a technical proposition.

Proposition 1: Let $k \ge 1$ and $\mathcal{A}_k$ be the Yatracos class associated with $\mathcal{F}_k$. 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 [12]). Note also that, according to Scheffé's identity (Devroye [38], page 2), for two densities f and g,

$$D_k(f,g) \le \frac{1}{2}\int |f-g| \tag{8}$$

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

$$D_k(f,g) = \frac{1}{2}\int |f-g|\,. \tag{9}$$

Now, let $(f_n)_{n\ge1}$ be a Cauchy sequence in $\mathcal{F}_k$ for the D_k metric. Clearly, according to (9), $(f_n)_{n\ge1}$ 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 (8), 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. ■

Proof of Theorem 1 The following chain of inequalities is valid.

$$\begin{aligned}
&\mathbf{P}_0\{\text{rejecting } \mathbf{H}_0\}\\
&\quad= \mathbf{P}_0\left\{\min_{k\neq k_0}\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k)\right\}\right.\\
&\qquad\qquad \left.< \sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0} - \mu_n(A)\right|\right\}\\
&\quad\le \sum_{\substack{k\ge1\\k\neq k_0}} \mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k)\right.\\
&\qquad\qquad \left.< \sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0} - \mu_n(A)\right|\right\}\\
&\quad\le \sum_{\substack{k\ge1\\k\neq k_0}} \mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k) - \frac{1}{n}\right.\\
&\qquad\qquad \left.< \Delta_{k_0}(f)\right\},\\
&\qquad\text{(since, under } \mathbf{H}_0,\ f\in\mathcal{F}_{k_0} \text{ and by definition of } f_{n,k_0})
\end{aligned}$$

where we recall that

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

Therefore, we obtain that

$$\begin{aligned}
&\mathbf{P}_0\{\text{rejecting } \mathbf{H}_0\}\\
&\quad\le \sum_{\substack{k\ge1\\k\neq k_0}} \mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k) - \frac{1}{n}\right.\\
&\qquad\qquad \left.< \Delta_{k_0}(f)\right\}\\
&\quad= \sum_{k=1}^{k_0-1} \mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k) - \frac{1}{n}\right.\\
&\qquad\qquad \left.< \Delta_{k_0}(f)\right\}\\
&\qquad + \sum_{k\ge k_0+1} \mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k) - \frac{1}{n}\right.\\
&\qquad\qquad \left.< \Delta_{k_0}(f)\right\}.
\end{aligned}\tag{10}$$

We shall first examine the first of the two terms in the above expression (10) (which makes sense only if $k_0 > 1$). For $k = 1,\ldots,k_0-1$, write

$$\begin{aligned}
&\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right|\\
&\quad\ge \sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \int_A f\right| - \sup_{A\in\mathcal{A}_k}\left|\int_A f - \mu_n(A)\right|\\
&\quad\text{(by the triangle inequality)}\\
&\quad\ge \min_{1\le k\le k_0-1}\ \inf_{g\in\mathcal{F}_k}\ \sup_{A\in\mathcal{A}_k}\left|\int_A f - \int_A g\right| - \Delta_{k_0}(f)\\
&\quad\text{(since } \mathcal{A}_k \subset \mathcal{A}_{k_0})\\
&:= m - \Delta_{k_0}(f)\,.
\end{aligned}$$

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 1 that they are also closed for the D_k metric. Therefore, the definition of k_0 (recall that $k^\star = k_0$ under $\mathbf{H}_0$) implies $m > 0$. Thus we are led to

$$\begin{aligned}
&\sum_{k=1}^{k_0-1} \mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k} - \mu_n(A)\right| + \text{pen}_n(k) - \frac{1}{n}\right.\\
&\qquad\qquad \left.< \Delta_{k_0}(f)\right\}\\
&\quad\le \sum_{k=1}^{k_0-1} \mathbf{P}_0\left\{2\Delta_{k_0}(f) > m + \text{pen}_n(k) - \frac{1}{n}\right\}\\
&\quad\le (k_0-1)\,\mathbf{P}_0\left\{\Delta_{k_0}(f) > \frac{m}{4}\right\}\quad \text{for all } n \text{ large enough}\\
&\quad\le \alpha/2 \quad \text{for all } n \text{ large enough,}
\end{aligned}$$

where, in the last inequality, we used the finiteness of $\mathcal{V}_{k_0}$ together with inequality (5).

Let us now turn to the analysis of the second term in expression (10). We have

$$\begin{aligned}
&\sum_{k\geq k_0+1}\mathbf{P}_0\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k}-\mu_n(A)\right|+\mathrm{pen}_n(k)-\frac{1}{n}\right.\\
&\qquad\qquad\left.<\Delta_{k_0}(f)\right\}\\
&\leq\sum_{k\geq k_0+1}\mathbf{P}_0\left\{\mathrm{pen}_n(k)-\frac{1}{n}<\Delta_{k_0}(f)\right\}\\
&=\sum_{k\geq k_0+1}\mathbf{P}_0\left\{\Delta_{k_0}(f)>\frac{x_k+C\sqrt{\mathcal{V}_{k_0}}}{\sqrt{n}}\right\}\\
&\quad\text{(by definition of the penalty function for } k\geq k_0+1)\\
&\leq 2\sum_{k\geq k_0+1}e^{-2x_k^2}\quad\big(\text{by inequality (5)}\big)\\
&\leq\alpha/2\quad(\text{by definition of the } x_k\text{'s}).
\end{aligned}$$

Putting all pieces together leads to the desired result.

B. Proof of Theorem 2

To prove Theorem 2, we show that $\mathbf{P}_1\{\text{accepting }\mathbf{H}_0\}$ goes to 0 as n grows. We have

$$\begin{aligned}
&\mathbf{P}_1\{\text{accepting }\mathbf{H}_0\}\\
&=\mathbf{P}_1\left\{\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0}-\mu_n(A)\right|\right.\\
&\qquad\left.=\min_{k\geq1}\left\{\sup_{A\in\mathcal{A}_k}\left|\int_A f_{n,k}-\mu_n(A)\right|+\mathrm{pen}_n(k)\right\}\right\}\\
&\leq\mathbf{P}_1\left\{\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0}-\mu_n(A)\right|\right.\\
&\qquad\left.\leq\sup_{A\in\mathcal{A}_{k^*}}\left|\int_A f_{n,k^*}-\mu_n(A)\right|+\mathrm{pen}_n(k^*)\right\}\\
&\leq\mathbf{P}_1\left\{\sup_{A\in\mathcal{A}_{k_0}}\left|\int_A f_{n,k_0}-\mu_n(A)\right|-\mathrm{pen}_n(k^*)-\frac{1}{n}\right.\\
&\qquad\left.<\Delta_{k^*}(f)\right\},
\end{aligned}$$

where, in the last inequality, we use the definition of f_{n,k^*} and the fact that $f\in\mathcal{F}_{k^*}$.

Again, we distinguish the case $k_0<k^*$ from the case $k_0>k^*$ (recall that $k^*\neq k_0$ under $\mathbf{H}_1$). In the first situation (which makes sense only if $k^*>1$), we have, acting as in the proof of Theorem 1,

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

where m is a positive constant. Therefore,

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

for all n large enough (depending on f), and this last term is bounded above by $K_1n^{-1/2+\mathcal{V}_{k^*}}e^{-nm^2/8}$, according to Talagrand's inequality (6), where K_1 is a positive constant depending on k^*. Finally, if k_0 meets the condition $k_0>k^*$ (which makes sense only if $k_0>1$), then

$$\begin{aligned}
\mathbf{P}_1\{\text{accepting }\mathbf{H}_0\}&\leq\mathbf{P}_1\left\{-\mathrm{pen}_n(k^*)-\frac{1}{n}<\Delta_{k^*}(f)\right\}\\
&\leq\mathbf{P}_1\left\{\Delta_{k^*}(f)>\frac{1}{n^{1/3}}\right\},
\end{aligned}$$

by definition of the penalty function for $k=1,\ldots,k_0-1$. We deduce again from Talagrand's inequality that the last term is bounded above by $K_2n^{-1/6+\mathcal{V}_{k^*}/3}e^{-2n^{1/3}}$ for all $n\geq1$, where K_2 is a positive constant depending on k^*.

ACKNOWLEDGMENT

The authors thank both referees for their comments and suggestions.

REFERENCES

[1] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2001.

[2] N. Dunford and J. T. Schwartz, *Linear Operators Part I*. New York: Wiley, 1963.

[3] L. F. James, C. E. Priebe, and D. J. Marchette, "Consistent estimation of mixture complexity," *Ann. Statist.*, vol. 29, pp. 1281–1296, 2001.

[4] G. J. McLachlan, "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *J. Appl. Stat.*, vol. 36, pp. 318–324, 1987.

[5] D. Dacunha-Castelle and E. Gassiat, "Testing in locally conic models, and application to mixture models," *ESAIM Probab. Stat.*, vol. 1, pp. 285–317, 1997.

[6] ——, "Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes," *Ann. Statist.*, vol. 27, pp. 1178–1209, 1999.

[7] G. Biau and L. Devroye, *Density estimation by the penalized combinatorial method*, McGill University, Technical Report, 2002.

[8] L. Devroye, L. Györfi, and G. Lugosi, "A note on robust hypothesis testing," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2111–2114, 2002.

[9] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.

[10] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics 1989*. Cambridge: Cambridge University Press, 1989.

[11] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Probab.*, vol. 22, pp. 28–76, 1994.

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

[13] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*. London: Chapman and Hall, 1981.

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

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

[16] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: John Wiley, 2000.

[17] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, pp. 195–239, 1984.

[18] J. Diebolt and C. P. Robert, "Estimation of finite mixture distributions through Bayesian sampling," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 56, pp. 363–375, 1994.

[19] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 59, pp. 731–792, 1997.

[20] K. Roeder and L. Wasserman, "Practical Bayesian density estimation using mixtures of normals," *J. Amer. Statist. Assoc.*, vol. 92, pp. 894–902, 1997.

[21] G. Celeux, M. Hurn, and C. P. Robert, "Computational and inferential difficulties with mixture posterior distributions," *J. Amer. Statist. Assoc.*, vol. 95, pp. 957–970, 2000.

[22] M. Hurn, A. Justel, and C. P. Robert, "Estimating mixtures of regressions," *J. Comput. Graph. Statist.*, vol. 12, pp. 1–25, 2003.

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

[24] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

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

[26] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 381–396, 2002.

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

[28] K. Fukumizu, "Likelihood ratio of unidentifiable models and multilayer neural networks," *Ann. Statist.*, vol. 31, pp. 833–851, 2003.

[29] D. Dacunha-Castelle and E. Gassiat, "The estimation of the order of a mixture model," *Bernoulli*, vol. 3, pp. 279–299, 1997.

[30] C. E. Priebe, "Adaptive mixtures," *J. Amer. Statist. Assoc.*, vol. 89, pp. 796–806, 1994.

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

[32] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, pp. 301–413, 1999.

[33] G. Castellan, "Sélection d'histogrammes à l'aide d'un critère de type Akaike," *C. R. Acad. Sci. Paris Sér. I Math*, vol. 330, pp. 729–732, 2000.

[34] P. Massart, "Some applications of concentration inequalities to statistics," *Ann. Fac. Sci. Toulouse Math.*, vol. 9, pp. 245–303, 2000.

[35] J. Rissanen, "On Bayesian analysis of mixtures with an unknown number of components," *Automatica*, vol. 14, pp. 465–471, 1978.

[36] ——, "Stochastic complexity (with discussion)," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 49, pp. 223–239, 253–265, 1987.

[37] J. Rissanen, T. Speed, and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 38, pp. 315–323, 1992.

[38] L. Devroye, *A Course in Density Estimation*. Boston: Birkhäuser, 1987.