Test (2008) 17: 531–545
DOI 10.1007/s11749-006-0042-6

ORIGINAL PAPER

Strongly consistent model selection for densities

Gérard Biau · Benoît Cadre · Luc Devroye · László Györfi

Received: 24 March 2006 / Accepted: 21 November 2006 /
Published online: 27 February 2007

Abstract Let f be an unknown multivariate density belonging to a set of densities $\mathcal{F}_{k^*}$ of finite associated Vapnik–Chervonenkis dimension, where the complexity k^* is unknown, and $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Given an i.i.d. sample of size n drawn from f, this article presents a density estimate $\hat{f}_{K_n}$ yielding almost sure convergence of the estimated complexity K_n to the true but unknown k^* and with the property $\mathbf{E}\{\int |\hat{f}_{K_n} - f|\} = \mathrm{O}(1/\sqrt{n})$. The methodology is inspired by the combinatorial tools developed in Devroye and Lugosi (Combinatorial methods in density estimation. Springer, New York, 2001) and it includes a wide range of density models, such as mixture models and exponential families.

Keywords Histogram-based estimate · Mixture densities · Multivariate density estimation · Strong consistency · Vapnik–Chervonenkis dimension

Mathematics Subject Classification (2000) 62G10

G. Biau (✉) · B. Cadre
Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Equipe de Probabilités et Statistique, Université Montpellier II, CC 051, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
e-mail: biau@math.univ-montp2.fr

L. Devroye
School of Computer Science, McGill University, Montreal, Canada

L. Györfi
Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary

1 Introduction

Let $(\mathcal{F}_k)_{k\geq 1}$ be a sequence of nested parametric models of density functions on $\mathbb{R}^d$. Define

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

In the above union $\mathcal{F}_k$ denotes, for each fixed $k \geq 1$, a given class of densities parametrized by one or more parameters and such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Typically, $\mathcal{F}_k$ may be the class of all mixtures of k Gaussian densities on $\mathbb{R}^d$, but many other nested models are possible, see below. In the present paper we consider the general problem of estimating a density f which belongs to $\mathcal{F}$. Formally, we let the complexity associated with f be defined as

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

Clearly, as it is supposed that $f \in \mathcal{F}$, we have $k^* < \infty$. Thus, k^* just represents the index of the most parsimonious model for f. Given an i.i.d. random sample $X_1, \ldots, X_n$ drawn from f, this article presents a density estimate $\hat{f}_{K_n}$ yielding almost sure convergence of the estimated complexity K_n to the true but unknown k^* and with the property that $\mathbf{E}\{\int |\hat{f}_{K_n} - f|\} = \mathrm{O}(1/\sqrt{n})$. Thus, we show how to pick a model complexity and a density from the given model and still guarantee an $\mathrm{O}(1/\sqrt{n})$ rate of convergence for the expected error, just as if we were given the model complexity beforehand. The strongly consistent estimate K_n is obtained by minimizing the L_1 error between candidate models and the standard histogram estimate. Based on this estimate, the model parameters are selected using the general combinatorial tools developed in Devroye and Lugosi (2001). Our methodology is close in spirit to Biau and Devroye (2004, 2005), who use a penalized combinatorial criterion to select a density from a nested sequence of models. However, these authors do not consider the estimation of the complexity parameter k^* and they employ a penalty depending on a sequence of arbitrary weights which renders the method difficult to implement. In contrast, the selection procedure presented in the present paper does not rely on any arbitrary penalty choice, and it seems, therefore, more tractable.

The paper is organized as follows. In Sect. 2 we present our new histogram-based estimate K_n for the complexity k^* and show its strong consistency. In Sect. 3 we develop our density estimation procedure and state its L_1-optimality. For the sake of clarity, the proofs are gathered in Sect. 4.

2 Complexity estimation

Without loss of generality, we assume that the sample of independent random vectors distributed according to the probability measure μ with density f is of even size $2n$. Let μ_{2n} be its empirical measure, i.e., $\mu_{2n}(A) = (1/(2n))\sum_{i=1}^{2n} \mathbf{1}_{\{X_i \in A\}}$. Split the sample into two subsamples: $X_1, \ldots, X_n$ and $\{X'_1, \ldots, X'_n\} = \{X_{n+1}, \ldots, X_{2n}\}$, and denote by μ_n and μ'_n the respective empirical measures. Let $\mathcal{P}_n = \{A_{n,j} : j \geq 1\}$ be a cubic partition of $\mathbb{R}^d$ with volume h_n^d. Introduce the statistic

$$d_{n,k} = \inf_{g \in \mathcal{F}_k} \sum_{A \in \mathcal{P}_n} \left| \int_A g - \mu_{2n}(A) \right|.$$

Roughly speaking, $d_{n,k}$ should be small if $f \in \mathcal{F}_k$. Consequently, let the threshold be

$$T_n = \sum_{A \in \mathcal{P}_n} \left| \mu_n(A) - \mu'_n(A) \right|.$$

Note that the statistic T_n has been introduced in Biau and Györfi (2005), while its kernel version was studied in Cao and Lugosi (2005). We say that $d_{n,k}$ is small if it is smaller than T_n, so our estimate of k^* is

$$K_n = \min\{k \geq 1 : d_{n,k} \leq T_n\},$$

with the convention $\min\{\emptyset\} = \infty$. Clearly, the statistic $d_{n,k}$ is nonincreasing in k, so that this definition makes sense.

From a topological point of view we will suppose throughout the paper that each $\mathcal{F}_k$ is a closed metric subspace of the space of all densities on $\mathbb{R}^d$ endowed with the weak convergence topology. In other words, for any sequence (g_n) in $\mathcal{F}_k$, satisfying

$$\lim_{n \to \infty} \int g_n(x) \varphi(x) \, \mathrm{d}x = \int g(x) \varphi(x) \, \mathrm{d}x,$$

for every bounded, continuous real function φ, one has $g \in \mathcal{F}_k$. This requirement is not restrictive. For example, one may check that this prerequisite holds for the class $\mathcal{F}_k$ of all mixtures of k Gaussian densities on $\mathbb{R}^d$. In this case each density f_k in $\mathcal{F}_k$ may be written as

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

where $k < \infty$ is the mixture complexity, $p_i \geq 0$, $i = 1, \ldots, k$, are the mixture weights satisfying $\sum_{i=1}^k p_i = 1$, $m_1, \ldots, m_k$ are arbitrary elements of $\mathbb{R}^d$ and $\Sigma_1, \ldots, \Sigma_k$ are positive definite $d \times d$ matrices.

Methodologies for consistent estimation of a mixture distribution are well known. 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. Recent attempts at estimating the mixture density parameters and the number of mixture densities jointly are by Priebe (1994), James et al. (2001), and Rogers et al. (1995). For an updated list of references we refer the reader to Biau and Devroye (2005) and the discussion therein.

As another interesting collection of models, we might also consider the class of increasing exponential families. Each density f_k in an exponential family $\mathcal{F}_k$ may be written in the form

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

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.

Theorem 2.1 *Assume that, for each $k \geq 1$, $\mathcal{F}_k$ is closed with respect to the weak convergence topology. Choose $h_n = n^{-\delta}$ with $0 < \delta < 1/d$. Then there exists a positive constant κ, depending on f, such that*

$$\mathbf{P}\{K_n \neq k^*\} \leq \exp(-\kappa\, n^{d\delta}),$$

and consequently, almost surely,

$$K_n = k^*$$

for all n large enough.

3 Fast density estimate

In the same way as in Biau and Devroye (2004, 2005), based on the complexity estimation presented in the previous section, we consider a minimum distance type estimate. Using ideas from Yatracos (1985), Devroye and Lugosi (2001) 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 $\hat{f}_k$ in $\mathcal{F}_k$ as follows. First introduce the class of sets

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

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

$$\Delta_k(g) = \sup_{A \in \mathcal{A}_k} \left| \int_A g - \mu_{2n}(A) \right|.$$

Then the minimum distance estimate $\hat{f}_k$ is defined as any density estimate selected from those densities $f_k \in \mathcal{F}_k$ with

$$\Delta_k(f_k) < \inf_{g \in \mathcal{F}_k} \Delta_k(g) + \frac{1}{2n}.$$

Note that the $1/(2n)$ term here is added to ensure the existence of such a density estimate. From now on we let V_k be the Vapnik–Chervonenkis dimension of the class of sets $\mathcal{A}_k$ (Vapnik and Chervonenkis 1971).

Corollary 3.1 *Assume that V_{k^*} is finite. Then, under the conditions of Theorem 2.1, the minimum distance estimate $\hat{f}_{K_n}$ satisfies*

$$\mathbf{E}\left\{\int |\hat{f}_{K_n} - f|\right\} = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right). \tag{3.1}$$

Proof For the minimum distance estimate $\hat{f}_k$, we have (Devroye and Lugosi 2001, Theorem 6.4, p. 56)

$$\int |\hat{f}_k - f| \le 3 \inf_{g\in\mathcal{F}_k} \int |g - f| + 4\Delta_k(f) + \frac{3}{2n}. \tag{3.2}$$

Since the inequality (3.2) holds for every $k \ge 1$, it holds, in particular, for K_n. The corresponding minimum distance estimate, $\hat{f}_{K_n}$, is a natural candidate for the estimation of f. Clearly,

$$\int |\hat{f}_{K_n} - f| \le \int |\hat{f}_{K^*} - f| \,\mathbf{1}_{\{K_n=k^*\}} + 2\,\mathbf{1}_{\{K_n\neq k^*\}}.$$

By (3.2), we have, for all $n \ge 1$,

$$\int |\hat{f}_{K^*} - f| \le 4\Delta_{K^*}(f) + \frac{3}{2n}.$$

Using Theorem 2.1, we deduce that

$$\mathbf{E}\left\{\int |\hat{f}_{K_n} - f|\right\} \le 4\mathbf{E}\{\Delta_{k^*}(f)\} + \frac{3}{2n} + 2\exp(-\kappa\, n^{d\delta}), \tag{3.3}$$

where κ is the positive constant of Theorem 2.1. The uniform convergence of empirical measures, as developed by Vapnik and Chervonenkis (1971), can now be applied to density estimation via the term $\Delta_{k^*}(f)$. A standard inequality from the empirical process theory (Dudley 1978) shows that if $\mathcal{A}_{k^*}$ has Vapnik–Chervonenkis dimension V_{k^*} then

$$\mathbf{E}\{\Delta_{k^*}(f)\} \le C\sqrt{\frac{V_{k^*}}{n}}, \tag{3.4}$$

where C is a positive universal constant. Inequalities (3.3) and (3.4) then imply

$$\mathbf{E}\left\{\int |\hat{f}_{K_n} - f|\right\} \le 4C\sqrt{\frac{V_{k^*}}{n}} + \frac{3}{2n} + 2\exp(-\kappa\, n^{d\delta}).$$

□

As an illustration, consider the examples of Sect. 2. It can be shown (see Devroye and Lugosi 2001, Chap. 8) that $V_k = \mathrm{O}(k^4)$ for the univariate Gaussian mixtures with k components, and $V_k \le k + 1$ for the exponential families. Equality (3.1) is thus satisfied by a large collection of models.

It is, strictly speaking, not necessary that $V_k \to \infty$, although such situations are of little general interest. Indeed, if $\sup_{k\ge 1} V_k < \infty$ then the Vapnik–Chervonenkis dimension of the infinite union $\mathcal{F}$ is finite, and one could just apply the ordinary combinatorial method. However, in most situations of interest, the Vapnik–Chervonenkis dimension of $\mathcal{F}$ is infinite, and we cannot use the original combinatorial method. It is to correct this situation that we have presented the two step procedure.

Observe also that the finiteness of the associated Vapnik–Chervonenkis dimension is only used at k^*. We thus allow this dimension to be infinite starting at any $k > k^*$. Considering a silly example, we could take $\mathcal{F}_{k^*+1}$ to be the class of all densities. Then the $d_{n,k}$ is zero at $k^* + 1$, so the selected k is never more than $k^* + 1$. But the method is clever enough to stop at k^*. If you stop at $k^* + 1$ then the minimum distance estimate includes any density. So the penalty for overshooting is high.

We note that our model selection procedure is simpler than that of Biau and Devroye (2004), since the projection over a cubic partition is easier than the projection over a Yatracos class. However, practically speaking, several questions need to be addressed, including that of the effective computation of the minimum distance estimate $\hat{f}_k$. 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 (1997). 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 model of pattern recognition and machine learning, where these connections have been thoroughly studied.

Summarizing, we have shown, assuming that $f \in \mathcal{F}$, how to 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 error, just as if we had been given the mixture complexity beforehand. On the other hand, we realize that an important situation occurs when our infinite union is dense in the set of all densities, all Vapnik–Chervonenkis dimensions are finite, and the density is not in any $\mathcal{F}_k$. Note that T_n depends upon n and h_n in the standard way. Now $d_{n,k}$ tends to zero with k, so we end up, anyway, with a finite K_n. As $d_{n,k}$ is below T_n by our selection criterion, it seems that we can have an error bound of the order of the bound for T_n. This would mean that we can, in the worst case, have a bound as for the classical histogram (because that is what T_n would give). We believe, however, that such results are beyond the scope of the present paper and we will deal with this problem elsewhere.

Remark 3.1 Recall that when the set $\{k \geq 1 : d_{n,k} \leq T_n\}$ is empty, we take $K_n = \infty$ by convention. In this case any choice for $\hat{f}_{K_n}$ in $\mathcal{F}$ will do.

Remark 3.2 For the actual density estimate we could not avoid the projection over the Yatracos class. Thus, we could not decrease the computational complexity. It is an open problem whether the projection over a cubic partition is simply a proper density estimate. Note also that the histogram is only used to select k^*. The actual density estimate depends upon the Yatracos classes—the histogram is thrown away! This point is essential, as a histogram estimate cannot converge at the rate $\mathrm{O}(1/\sqrt{n})$. So we have the apparent paradox that a density estimate with a bad rate can be used to obtain—in two steps—a great rate! In a sense, the histogram is used to choose the complexity (or lack of smoothness) of another class of functions. The choice of the histogram's width, h_n, thus has a secondary effect. To stay within the limits of the theorem, one could easily replace h_n by one of the many data-dependent and automated bandwidths. Some of these are discussed or surveyed in Devroye and Györfi (1985) and Devroye and Lugosi (2001).

Interestingly, in the second step, and under some conditions, just the projection of another, non-consistent (asymptotically biased) histogram can result in a density estimate with good rate, too. In respect to this denote by $\xi_{n,r}$ a histogram estimate constructed using the sample $X_1, \ldots, X_{2n}$ and a cubic partition $\{B_{r,j} : j \geq 1\}$ of $\mathbb{R}^d$ with volume r^d, and let $\pi_r(g)$ be the expectation of this histogram for the density g, that is,

$$\pi_r(g)(x) = \frac{1}{r^d} \int_{B_r(x)} g,$$

where $B_r(x)$ is the cell of the partition containing x. Let the estimate $\tilde{f}_{n,r}$ be defined as

$$\tilde{f}_{n,r} = \arg\min_{g \in \mathcal{F}_{K_n}} \int \left|\pi_r(g) - \xi_{n,r}\right|.$$

Suppose, moreover, that the bin width r is such that

$$C_{f,r} = \sup_{g \in \mathcal{F}_{k^*}} \frac{\int |g - f|}{\int |\pi_r(g) - \pi_r(f)|} < \infty. \tag{3.5}$$

If $K_n = k^*$ then

$$\begin{aligned}\int \left|\pi_r(\tilde{f}_{n,r}) - \pi_r(f)\right| &\leq \int \left|\pi_r(\tilde{f}_{n,r}) - \xi_{n,r}\right| + \int \left|\pi_r(f) - \xi_{n,r}\right| \\ &= \min_{g \in \mathcal{F}_{k^*}} \int \left|\pi_r(g) - \xi_{n,r}\right| + \int \left|\pi_r(f) - \xi_{n,r}\right| \\ &\leq \int \left|\pi_r(f) - \xi_{n,r}\right| + \int \left|\pi_r(f) - \xi_{n,r}\right|.\end{aligned}$$

Therefore, by the condition (3.5),

$$\begin{aligned}\int |\tilde{f}_{n,r} - f| &\leq C_{f,r} \int \left|\pi_r(\tilde{f}_{n,r}) - \pi_r(f)\right| \\ &\leq 2C_{f,r} \int \left|\pi_r(f) - \xi_{n,r}\right|.\end{aligned}$$

Using Theorem 2.1, it easily follows, under the additional condition that f has compact support (see Devroye and Györfi 1985, Theorem 6, p. 99), that

$$\mathbf{E}\left\{\int |\tilde{f}_{n,r} - f|\right\} \leq 2C_{f,r}\mathbf{E}\left\{\int \left|\pi_r(f) - \xi_{n,r}\right|\right\} \leq C_{f,r} \frac{\mathrm{C}}{\sqrt{nr^d}},$$

where C is a positive constant. Therefore, for such fixed r, we can have rate $\mathrm{O}(1/\sqrt{n})$. Unfortunately, (3.5) is hard to verify, and requires that the operator π_r is invertible on $\mathcal{F}_{k^*}$ and the inverse is continuous. We cannot handle our standard examples. However, for the class of normal densities, this condition holds for any fixed r. In other words, there is no need to choose small r, i.e., the projection of non-consistent (asymptotically biased) histogram results in a good density estimate.

4 Proofs

From now on all limits are taken as $n \to \infty$.

4.1 Proof of Theorem 2.1

We split the proof into two steps in which bounds for $\mathbf{P}\{K_n > k^*\}$ (STEP 1) and $\mathbf{P}\{K_n < k^*\}$ (STEP 2) are obtained.

STEP 1. Note that, since f belongs to $\mathcal{F}_{k^*}$, one has

$$d_{n,k^*} \le \sum_{A \in \mathcal{P}_n} \left| \int_A f - \mu_{2n}(A) \right| = \sum_{A \in \mathcal{P}_n} \left| \mu(A) - \mu_{2n}(A) \right|.$$

Consequently,

$$\begin{aligned} \{K_n > k^*\} &\subset \{d_{n,k^*} > T_n\} \\ &\subset \left\{ \sum_{A \in \mathcal{P}_n} \left| \mu(A) - \mu_{2n}(A) \right| > \sum_{A \in \mathcal{P}_n} \left| \mu_n(A) - \mu_n'(A) \right| \right\}. \end{aligned}$$

For $A \in \mathcal{P}_n$, set

$$I_n(A) = \left| \mu(A) - \mu_{2n}(A) \right| - \left| \mu_n(A) - \mu_n'(A) \right|,$$

and

$$\varepsilon_n = - \sum_{A \in \mathcal{P}_n} \mathbf{E}\{I_n(A)\}.$$

In the next paragraph we prove that, for all n large enough,

$$\varepsilon_n \ge \frac{\kappa_1}{\sqrt{n h_n^d}}, \tag{4.1}$$

for some positive constant κ_1 depending on f. We can apply McDiarmid's inequality (Devroye et al. 1996; McDiarmid 1989, Theorem 9.2, p. 136). We have $2n$ random variables and the fluctuation of $\sum_{A \in \mathcal{P}_n} I_n(A)$ is at most $3/n$. Therefore,

$$\begin{aligned} \mathbf{P}\{K_n > k^*\} &\le \mathbf{P}\left\{ \sum_{A \in \mathcal{P}_n} I_n(A) > 0 \right\} \\ &= \mathbf{P}\left\{ \sum_{A \in \mathcal{P}_n} \left[I_n(A) - \mathbf{E}\{I_n(A)\} \right] > \varepsilon_n \right\} \end{aligned}$$

$$\leq \exp\left(-\frac{2\varepsilon_n^2}{2n(3/n)^2}\right) \leq \exp\left(-\frac{\kappa_1^2}{9}h_n^{-d}\right)$$
$$= \exp\left(-\frac{\kappa_1^2}{9}n^{d\delta}\right).$$

STEP 2. Denote by f_n (resp., f_{2n}) the standard histogram estimate drawn from the data $X_1, \ldots, X_n$ (resp., $X_1, \ldots, X_{2n}$) and the partition $\mathcal{P}_n$. That is, for $x \in \mathbb{R}^d$,

$$f_n(x) = \frac{\mu_n(A_n(x))}{h_n^d} \quad \text{and} \quad f_{2n}(x) = \frac{\mu_{2n}(A_n(x))}{h_n^d},$$

where $A_n(x)$ is the cell of the partition $\mathcal{P}_n$ containing x. For any density g, let $\pi_n(g)$ be defined by

$$\pi_n(g)(x) = \frac{1}{h_n^d}\int_{A_n(x)} g.$$

Clearly, for any g in $\mathcal{D}$, by the triangle inequality,

$$\sum_{A\in\mathcal{P}_n}\left|\int_A g - \mu_{2n}(A)\right| = \int \left|\pi_n(g) - f_{2n}\right|$$
$$\geq \int \left|\pi_n(g) - \pi_n(f)\right| - \int \left|\pi_n(f) - f_{2n}\right|.$$

Therefore,

$$\{K_n < k^*\} \subset \{d_{n,k^*-1} \leq T_n\}$$
$$\subset \left\{\inf_{g\in\mathcal{F}_{k^*-1}} \int \left|\pi_n(g) - \pi_n(f)\right| \leq \int \left|\pi_n(f) - f_{2n}\right| + T_n\right\}.$$

We shall prove that, under the closure condition on the models $\mathcal{F}_k$, there exists a $\Delta > 0$ satisfying

$$\liminf_{n\to\infty} \inf_{g\in\mathcal{F}_{k^*-1}} \int \left|\pi_n(g) - \pi_n(f)\right| = 3\Delta. \tag{4.2}$$

Then, because of

$$\mathbf{P}\{T_n \geq \Delta\} \leq 2\mathbf{P}\left\{\sum_{A\in\mathcal{P}_n}\left|\mu_n(A) - \int_A f\right| \geq \Delta/2\right\}$$
$$= 2\mathbf{P}\left\{\int \left|f_n - \pi_n(f)\right| \geq \Delta/2\right\},$$

we deduce that there exists a positive constant κ_2, depending on f, such that, for all n large enough,

$$\mathbf{P}\{K_n < k^*\} \leq 2\mathbf{P}\left\{\int \left|f_n - \pi_n(f)\right| \geq \Delta/2\right\} + \mathbf{P}\left\{\int \left|\pi_n(f) - f_{2n}\right| \geq \Delta\right\}$$
$$\leq 7\exp(-\kappa_2 n).$$

The last exponential inequality arises from Devroye and Györfi (1985, Lemma 4, p. 22). Since $nh_n^d \to \infty$, it follows that, for all n large enough,

$$\mathbf{P}\{K_n < k^*\} \le 7\exp(-\kappa_2 h_n^{-d}) = 7\exp(-\kappa_2 n^{d\delta}).$$

In order to prove (4.2) we use an indirect argument. Assume that

$$\liminf_{n\to\infty} \inf_{g\in\mathcal{F}_{k^*-1}} \int \left|\pi_n(g) - \pi_n(f)\right| = 0,$$

and, possibly extracting a subsequence, that

$$\lim_{n\to\infty} \inf_{g\in\mathcal{F}_{k^*-1}} \int \left|\pi_n(g) - \pi_n(f)\right| = 0.$$

By the Abou-Jaoude theorem, we obtain

$$\lim_{n\to\infty} \inf_{g\in\mathcal{F}_{k^*-1}} \int \left|\pi_n(g) - f\right| = 0.$$

Thus, if this is true, there exists a sequence (g_n) in $\mathcal{F}_{k^*-1}$ such that

$$\lim_{n\to\infty} \int \left|\pi_n(g_n) - f\right| = 0. \tag{4.3}$$

Observe now that for any bounded Lipschitz function φ on $\mathbb{R}^d$, the following chain of equalities is valid:

$$\begin{aligned}\int \pi_n(g_n)\varphi &= \sum_{A\in\mathcal{P}_n} \int_A \frac{1}{h_n^d}\left(\int_A g_n\right)\varphi = \sum_{A\in\mathcal{P}_n} \int_A g_n\left(\frac{1}{h_n^d}\int_A \varphi\right) \\ &= \sum_{A\in\mathcal{P}_n} \int_A g_n\pi_n(\varphi) = \int g_n\pi_n(\varphi).\end{aligned}$$

Moreover, as φ is Lipschitz, there exists a constant $c > 0$ such that, for all $n \ge 1$,

$$\sup_{\mathbb{R}^d}\left|\pi_n(\varphi) - \varphi\right| \le c \sup_{A\in\mathcal{P}_n} \operatorname{diam} A,$$

and the right-hand term above goes to 0 because the partition $\mathcal{P}_n$ is cubic. Using the fact that g_n is a density function and collecting the above results, we obtain

$$\lim_{n\to\infty}\left(\int \pi_n(g_n)\varphi - \int g_n\varphi\right) = 0.$$

Therefore, using (4.3) and the fact that φ is bounded, we are led to

$$\lim_{n\to\infty} \int g_n\varphi = \int f\varphi.$$

Since this equality holds for any bounded Lipschitz function, it also holds for any bounded and continuous function (see Dudley 2002, Theorem 11.3.3, p. 395). Invoking the closure of the class $\mathcal{F}_{k^*-1}$, we finally deduce that $f \in \mathcal{F}_{k^*-1}$, which is a contradiction. This proves equality (4.2).

4.2 Proof of (4.1)

By Jensen's inequality,

$$\begin{aligned}\mathbf{E}\{|\mu_n(A)-\mu_n'(A)|\} &= \mathbf{E}\{\mathbf{E}\{|\mu_n(A)-\mu_n'(A)| \mid X_1,\dots,X_n\}\}\\ &\geq \mathbf{E}\{|\mu_n(A)-\mathbf{E}\{\mu_n'(A)\mid X_1,\dots,X_n\}|\}\\ &= \mathbf{E}\{|\mu_n(A)-\mu(A)|\}.\end{aligned}$$

Therefore,

$$\begin{aligned}\varepsilon_n &= \sum_{A\in\mathcal{P}_n}\big[\mathbf{E}\{|\mu_n(A)-\mu_n'(A)|\}-\mathbf{E}\{|\mu_{2n}(A)-\mu(A)|\}\big]\\ &\geq \sum_{A\in\mathcal{P}_n}\big[\mathbf{E}\{|\mu_n(A)-\mu(A)|\}-\mathbf{E}\{|\mu_{2n}(A)-\mu(A)|\}\big] \triangleq \varepsilon_n',\end{aligned}$$

so instead of (4.1) it suffices to show that, for some positive constant c_f and all n large enough,

$$\varepsilon_n' \geq \frac{c_f}{\sqrt{nh_n^d}}. \tag{4.4}$$

Let $B(n,p)$ denote a binomial random variable.

Lemma 4.1 *There exists a positive universal constant γ such that*

$$\left|\mathbf{E}\{|B(n,p)-np|\}-\sqrt{\frac{2np(1-p)}{\pi}}\right| \leq \gamma.$$

Proof Let Φ denote the standard normal distribution function, let F denote the distribution function of $B(n,p)$, let $m=np$ and $\sigma^2=np(1-p)$. We will use the facts:

$$\mathbf{E}\{|B(n,p)-np|\} = \int_m^\infty \big(1-F(x)\big)\,\mathrm{d}x + \int_{-\infty}^m F(x)\,\mathrm{d}x$$

and

$$\big|\mathbf{P}\{B(n,p)\leq m+x\sigma\}-\Phi(x)\big| \leq \frac{Cn\mathbf{E}\{|B(1,p)-p|^3\}}{\sigma^3(1+|x|^3)} \tag{4.5}$$

by a local version of the Berry–Esseen inequality, where C is a positive universal constant (cf. Nagaev 1965). Working out (4.5), we note that

$$\mathbf{E}\{|B(1,p)-p|^3\} = p(1-p)^3+(1-p)p^3 = p(1-p)\big((1-p)^2+p^2\big) \leq p(1-p).$$

Thus, the bound becomes

$$\frac{Cnp(1-p)}{(np(1-p))^{3/2}(1+|x|^3)} = \frac{C}{\sqrt{np(1-p)}(1+|x|^3)}. \tag{4.6}$$

Therefore,

$$\begin{aligned}&\left|\int_m^\infty \big(1-F(x)\big)\,\mathrm{d}x - \int_m^\infty \left(1-\Phi\left(\frac{x-m}{\sigma}\right)\right)\mathrm{d}x\right| \\ &\quad\le \int_m^\infty \frac{C}{\sqrt{np(1-p)}(1+|(x-m)/\sigma|^3)}\,\mathrm{d}x,\end{aligned}$$

by (4.6). After replacing $(x-m)/\sigma$ by y, we obtain

$$\left|\int_m^\infty \big(1-F(x)\big)\,\mathrm{d}x - \sigma\int_0^\infty \big(1-\Phi(y)\big)\,\mathrm{d}y\right| \le \int_0^\infty \frac{C}{1+|y|^3}\,\mathrm{d}y \triangleq \gamma/2.$$

Thus,

$$\left|\int_m^\infty \big(1-F(x)\big)\,\mathrm{d}x - \frac{\sigma}{\sqrt{2\pi}}\right| \le \gamma/2.$$

Similarly,

$$\left|\int_{-\infty}^m F(x)\,\mathrm{d}x - \frac{\sigma}{\sqrt{2\pi}}\right| \le \gamma/2.$$

Finally,

$$\left|\mathbf{E}\big\{\big|B(n,p)-np\big|\big\} - \frac{2\sigma}{\sqrt{2\pi}}\right| \le \gamma. \qquad \square$$

Lemma 4.2

$$\mathbf{E}\big\{\big|B(n,p)-np\big|\big\} - \frac{1}{2}\mathbf{E}\big\{\big|B(2n,p)-2np\big|\big\} \ge 0.$$

Proof If $B(n,p)$ and $B'(n,p)$ are i.i.d. then

$$B(2n,p) = B(n,p) + B'(n,p).$$

Thus,

$$\begin{aligned}\mathbf{E}\big\{\big|B(2n,p)-2np\big|\big\} &\le \mathbf{E}\big\{\big|B(n,p)-np\big|\big\} + \mathbf{E}\big\{\big|B'(n,p)-np\big|\big\} \\ &= 2\mathbf{E}\big\{\big|B(n,p)-np\big|\big\}.\end{aligned} \qquad \square$$

Lemma 4.3

$$\left|\mathbf{E}\big\{\big|B(n,p)-np\big|\big\} - \frac{1}{2}\mathbf{E}\big\{\big|B(2n,p)-2np\big|\big\} - \alpha\sqrt{np(1-p)}\right| \le \frac{3\gamma}{2},$$

where

$$\alpha = \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{\frac{2}{\pi}}.$$

Proof By Lemma 4.1, using ζ, ζ' for numbers in $[-\gamma, \gamma]$,

$$\begin{aligned}
&\mathbf{E}\{|B(n,p) - np|\} - \frac{1}{2}\mathbf{E}\{|B(2n,p) - 2np|\} \\
&\quad = \sqrt{\frac{2np(1-p)}{\pi}} - \frac{1}{2}\sqrt{\frac{4np(1-p)}{\pi}} + \zeta + \zeta'/2 \\
&\quad = \left(1 - \frac{1}{\sqrt{2}}\right)\sqrt{\frac{2np(1-p)}{\pi}} + \zeta + \zeta'/2.
\end{aligned}$$

□

We are now in position to prove (4.1). To this aim we show that, under the conditions of Theorem 2.1,

$$\liminf_{n\to\infty} \sqrt{nh_n^d}\,\varepsilon_n' \ge \alpha \int \sqrt{f}.$$

Note that, when the right-hand side is infinite, inequality (4.4) is trivially true.

Proof Take $1/2 > \varepsilon > 0$ arbitrarily. Because of the absolute continuity of μ and $h_n \to 0$,

$$\sup_{A\in\mathcal{P}_n} \mu(A) < \varepsilon,$$

for all n large enough. Thus, by Lemmas 4.2 and 4.3, we obtain

$$\begin{aligned}
\varepsilon_n' &= \sum_{A\in\mathcal{P}_n} \left[\mathbf{E}\{|\mu_n(A) - \mu(A)|\} - \mathbf{E}\{|\mu_{2n}(A) - \mu(A)|\}\right] \\
&= \frac{1}{n}\sum_{A\in\mathcal{P}_n} \left[\mathbf{E}\{|B(n,\mu(A)) - n\mu(A)|\} - \frac{1}{2}\mathbf{E}\{|B(2n,\mu(A)) - 2n\mu(A)|\}\right] \\
&\ge \frac{1}{n}\sum_{A\in\mathcal{P}_n} \left[\alpha\sqrt{n\mu(A)(1-\mu(A))} - \frac{3\gamma}{2}\right]^+ \\
&\ge \frac{1}{n}\sum_{A\in\mathcal{P}_n} \left[\alpha\sqrt{1-\varepsilon}\sqrt{n\mu(A)} - \frac{3\gamma}{2}\right]^+ \\
&\ge \frac{1}{n}\sum_{A\in\mathcal{P}_n,\,\sqrt{n\mu(A)}\ge\beta} \alpha(1-\varepsilon)^{3/2}\sqrt{n\mu(A)} = \frac{\alpha(1-\varepsilon)^{3/2}}{\sqrt{n}}\sum_{A\in\mathcal{P}_n,\,\sqrt{n\mu(A)}\ge\beta} \sqrt{\mu(A)},
\end{aligned}$$

where

$$\beta = \frac{3\gamma}{2\alpha\varepsilon\sqrt{1-\varepsilon}}.$$

Thus, by Jensen's inequality,

$$\begin{aligned}
\sqrt{nh_n^d}\varepsilon_n' &\geq \alpha(1-\varepsilon)^{3/2} \sum_{A\in\mathcal{P}_n,\, \sqrt{n\mu(A)}\geq\beta} \sqrt{h_n^d\mu(A)} \\
&= \alpha(1-\varepsilon)^{3/2} \sum_{A\in\mathcal{P}_n,\, \sqrt{n\mu(A)}\geq\beta} \sqrt{h_n^d \int_A f(x)\,\mathrm{d}x} \\
&\geq \alpha(1-\varepsilon)^{3/2} \sum_{A\in\mathcal{P}_n,\, \sqrt{n\mu(A)}\geq\beta} \int_A \sqrt{f(x)}\,\mathrm{d}x \\
&= \alpha(1-\varepsilon)^{3/2} \int \sqrt{f(x)}\,\mathbf{1}_{\{\sqrt{n\mu(A_n(x))}\geq\beta\}}\,\mathrm{d}x.
\end{aligned}$$

By Fatou's lemma,

$$\begin{aligned}
\liminf_{n\to\infty}\sqrt{nh_n^d}\varepsilon_n' &\geq \alpha(1-\varepsilon)^{3/2}\int \sqrt{f(x)}\liminf_{n\to\infty}\mathbf{1}_{\{\sqrt{n\mu(A_n(x))}\geq\beta\}}\,\mathrm{d}x \\
&\geq \alpha(1-\varepsilon)^{3/2}\int_{\{x:f(x)>0\}} \sqrt{f(x)}\liminf_{n\to\infty}\mathbf{1}_{\{\sqrt{n\mu(A_n(x))}\geq\beta\}}\,\mathrm{d}x.
\end{aligned}$$

By the Lebesgue density theorem, for almost all x, $\mu(A_n(x))/h_n^d \approx f(x)$; therefore, the condition $nh_n^d \to \infty$ implies that, for almost all x satisfying $f(x) > 0$ and for any β,

$$\liminf_{n\to\infty}\mathbf{1}_{\{\sqrt{n\mu(A_n(x))}\geq\beta\}} = \liminf_{n\to\infty}\mathbf{1}_{\left\{\sqrt{nh_n^d\frac{\mu(A_n(x))}{h_n^d}}\geq\beta\right\}} = 1.$$

Thus,

$$\liminf_{n\to\infty}\sqrt{nh_n^d}\varepsilon_n' \geq \alpha(1-\varepsilon)^{3/2}\int\sqrt{f(x)}\,\mathrm{d}x.$$

As $\varepsilon > 0$ is arbitrary, (4.4) follows. □

Acknowledgements The authors greatly thank an Associate Editor and a referee for a careful reading of the paper.

References

Biau G, Devroye L (2004) A note on density model size testing. IEEE Trans Inf Theory 50:576–581

Biau G, Devroye L (2005) Density estimation by the penalized combinatorial method. J Multivar Anal 94:196–208

Biau G, Györfi L (2005) On the asymptotic properties of a nonparametric L_1-test of homogeneity. IEEE Trans Inf Theory 51:3965–3973

Cao R, Lugosi G (2005) Goodness-of-fit tests based on the kernel density estimator. Scand J Stat 32:599–616

Devroye L (1997) Universal smoothing factor selection in density estimation: theory and practice (with discussion). Test 6:223–320

Devroye L, Györfi L (1985) Nonparametric density estimation: the L_1 view. Wiley, New York

Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York
Devroye L, Lugosi G (2001) Combinatorial methods in density estimation. Springer, New York
Dudley RM (1978) Central limit theorems for empirical measures. Ann Probab 6:899–929
Dudley RM (2002) Real analysis and probability. Cambridge University Press, Cambridge
James LF, Priebe CE, Marchette DJ (2001) Consistent estimation of mixture complexity. Ann Stat 29:1281–1296
McDiarmid C (1989) On the method of bounded differences. In: Surveys in combinatorics 1989, pp 148–188. Cambridge University Press, Cambridge
Nagaev SV (1965) Some limit theorems for large deviations. Teor Verojatn Primen 10:231–254 (Russian)
Priebe CE (1994) Adaptive mixtures. J Am Stat Assoc 89:796–806
Rogers GW, Marchette DJ, Priebe CE (1995) A procedure for model complexity selection in semiparametric model mixture density estimation. In: Proceedings of the 10th international conference on mathematics and computer modelling, 1995
Vapnik VN, Chervonenkis AYa (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab Appl 16:264–280
Yatracos YG (1985) Rates of convergence of minimum distance estimators and Kolmogorov's entropy. Ann Stat 13:768–774