Parameter Selection in Modified Histogram Estimates

A. BERLINET, G. BIAU and L. ROUVIÈRE *

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

Abstract

A multivariate modified histogram density estimate depending on a reference density g and a partition P has recently been proved to have good consistency properties according to several information theoretic criteria. Given an i.i.d. sample, we show how to select automatically both g and P so that the expected L_1 error of the corresponding selected estimate is within a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions. Our method is inspired by the combinatorial tools developed in Devroye and Lugosi [1] and it includes a wide range of reference density and partition models. Results of simulations are presented.

Index Terms **— Modified histogram estimate, nonparametric estimation, partition, Vapnik-Chervonenkis dimension.**

***AMS 2000 Classification*: 62G05.**

1 Introduction

General ϕ-divergences (Liese and Vajda [2]) are widely used in many fields of statistics (data compression, telecommunication networks, classification, pattern recognition, neural networks...), particularly in decision processes based on density estimates and functionals of them. Many authors have put forward their attractive properties as criteria of accuracy. However, considering convergence of estimates of a density in the sense of ϕ-divergences causes some trouble. With standard histograms the situation is even hopeless as

*Corresponding author. Email: rouviere@ensam.inra.fr .

empty cells, occurring with high probability, make most of divergences infinite. The *modified histograms* introduced by Barron [3] and Barron, Györfi and van der Meulen [4] circumvent this problem. They are defined as follows.

Suppose that we observe independent $\mathbb{R}^d$-valued random variables $X_1, \ldots, X_n$ with common unknown density f.

- Denote by g a known density on $\mathbb{R}^d$ and by ν_g the associated probability measure;
- Define a sequence of integers $\{\ell_n\}_{n\geq 1}$ such that $2 \leq \ell_n$ and let $h_n = 1/\ell_n$;
- Introduce a sequence of partitions $P = \{A_{n1}, \ldots, A_{n\ell_n}\}$ such that $\nu_g(A_{ni}) = h_n$ for $i = 1, \ldots, \ell_n$;
- Finally consider, for $a_n = 1/(nh_n + 1)$, the following density estimate f_n:

$$f_n(x) = \left[(1 - a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x) = \frac{n\mu_n\big(A_n(x)\big) + 1}{nh_n + 1}\, g(x)\,, \quad (1)$$

  where μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$, *i.e.*, $\mu_n(A) = (1/n)\sum_{i=1}^n \mathbf{1}_{[X_i \in A]}$, and $A_n(x)$ equals A_{ni} if $x \in A_{ni}$.

The estimate (1) is a mixture of a histogram-type density estimate and the known density g. It can also be regarded as a piecewise transformation of g itself: roughly speaking, this modified histogram results from the comparison of the quantiles of g – the *reference density* – with the empirical quantiles (see Figure 1 for an example).

For further results on modified histograms, we refer the reader to Barron, Györfi and van der Meulen [4] who prove consistency in the sense of information divergence, Berlinet, Györfi and van der Meulen [5] who prove a central limit theorem for Kullback-Leibler divergence, Györfi, Liese, Vajda and van der Meulen [6], and Berlinet, Vajda and van der Meulen [7] who extend the information divergence consistency properties respectively to the χ^2-divergence and to more general ϕ-divergences.

Once the observations are given two parameters have to be chosen to build the modified histogram, namely a reference density g and a partition P. Recent univariate results obtained by Berlinet and Brunel (see [8], [9]) show

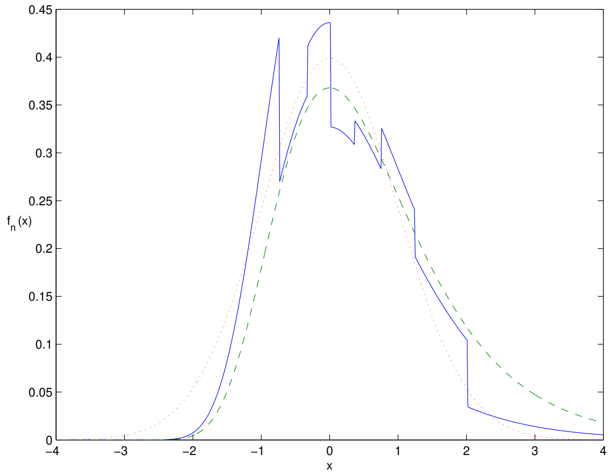

Figure 1: Modified histogram estimate (continuous line) of $\mathbf{n = 100}$ Gaussian (dotted line) data ($\ell_\mathbf{n} = \mathbf{8}$). The reference density is Gumbel (dashed line).

that the Kullback-Leibler cross-validation technique works well to select the partition from the data and that it is asymptotically optimal. As far as we know, no work has been devoted so far to select g and P simultaneously. This article proposes to fill this gap, using a general multivariate data-based combinatorial methodology presented in Devroye and Lugosi [1]. More precisely, we will show how to select both g and P – within given classes – so that the expected L_1 error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions. The paper is organized as follows. In Section 2, we present the multivariate selection procedure and give the main results. Examples are worked out in Section 3 and simulations are presented in Section 4. Proofs are gathered in Section 5.

2 Automatic parameter selection

2.1 The combinatorial method

Using ideas from Yatracos [10], Devroye and Lugosi [1] explore a new paradigm for the data-based or automatic selection of the free parameters of density estimates in general so that the expected L_1 error is within a given constant multiple of the best possible error. To summarize in the present context, assume we are given a class of density estimates parameterized by $\theta \in \Theta$ such that $f_{n,\theta}$ denotes the density estimate with parameter θ. Let $m < n$ be an integer which splits the data $X_1, \ldots, X_n$ into

- a set $X_1, \ldots, X_{n-m}$ used for the construction of the density estimate;
- a validation set $X_{n-m+1}, \ldots, X_n$.

Introduce the class of random sets

$$\mathcal{A}_\Theta = \Big\{ \big\{ x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x) \big\} : (\theta, \theta') \in \Theta^2 \Big\}$$

($\mathcal{A}_\Theta$ is the so-called *Yatracos class* associated with Θ) and define

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|,$$

where $\mu_m(A) = (1/m) \sum_{i=n-m+1}^{n} \mathbf{1}_{[X_i \in A]}$ is the empirical measure associated with the sample $X_{n-m+1}, \ldots, X_n$. Then the *minimum distance estimate* f_n is defined as any density estimate selected among the candidates $f_{n-m,\theta}$ with

$$\Delta_\theta < \inf_{\theta^* \in \Theta} \Delta_{\theta^*} + \frac{1}{n}\,.$$

Note that the $1/n$ term is added to ensure the existence of such a density estimate. According to Devroye and Lugosi [1], Chapter 10, whenever $f_{n-m,\theta}$ integrates to one, the selected f_n satisfies the following inequality:

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \le 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n-m,\theta} - f| \right\} + 8\, \mathbf{E}\left\{ \sqrt{\frac{\log 2\, \mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}} \right\} + \frac{3}{n}\,. \tag{2}$$

Here, $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ is the *Vapnik-Chervonenkis shatter coefficient* of the class of sets $\mathcal{A}_\Theta$ (Vapnik and Chervonenkis [11]), defined by

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1, \ldots, x_m \in \mathbb{R}^d} \mathrm{Card}\big\{ \{x_1, \ldots, x_m\} \cap A : A \in \mathcal{A}_\Theta \big\}.$$

This general methodology provides an automatic procedure to construct a density estimate f_n whose L_1 error is (almost) as small as that of the best estimate among the $f_{n,\theta}$, $\theta \in \Theta$. We emphasize that inequality (2) is nonasymptotic, that is, the bound is valid for all n. The rest of the analysis is then purely combinatorial and merely consists in obtaining upper bounds for the value of $\mathbf{S}_{\mathcal{A}_\Theta}(m)$.

As pointed out by a referee, a challenging question is whether the combinatorial L_1 selection procedure of Devroye and Lugosi [1] can be extended to L_p norms ($1 < p \le \infty$) or to more general ϕ-divergences, such as Kullback-Leibler information or Hellinger distance. According to the authors' experience, the extension to L_p criteria seems feasible, at the price of some technical requirements extending Scheffé's identity [12]. On the other hand, the divergence case presents a more delicate problem. Here, one needs to carefully assess the divergence between two measures as a supremum of functionals over a suitable class of functions. Dual representations of divergences should provide a good starting point, see for example Keziou [13].

2.2 Selecting a modified histogram

In this paragraph, we will be concerned with the selection of a density g and a partition P in the modified histogram estimate, using the general combinatorial tools presented above. Let us first describe the mathematical model. We let $\mathcal{G}$ be a given class of candidate reference densities on $\mathbb{R}^d$, and we denote by ν_g the probability measure associated with each $g \in \mathcal{G}$. Consider $\mathcal{P}$ a family of candidate partitions of $\mathbb{R}^d$ such that each $P \in \mathcal{P}$ has at most r cells ($r \geq 2$, possibly function of n, and to be made precise later on). To each density $g \in \mathcal{G}$ and each partition $P = \{A_1, \ldots, A_\ell\} \in \mathcal{P}$ such that $\nu_g(A_i) = 1/\ell$, $i = 1, \ldots, \ell$, assign the corresponding modified histogram $f_{n,\theta}$ defined as in (1), with $\theta = (g, P)$. We use the minimum distance estimate to select θ from

$$\Theta = \big\{(g, P) : g \in \mathcal{G}, P = \{A_1, \ldots, A_\ell\} \in \mathcal{P},\ \ell \leq r,\ \nu_g(A_i) = 1/\ell\big\}, \qquad (3)$$

the set of all possible pairs of reference densities and partitions. Denote by f_n the resulting minimum distance estimate. Now, to apply (2), we need to obtain upper bounds for the mth shatter coefficient $\mathbf{S}_{\mathcal{A}_\Theta}(m)$ of the Yatracos class associated with Θ. The following theorem is a key combinatorial result towards this direction. Denote by $\mathbf{S}_{\mathcal{D}}(j)$ the jth shatter coefficient of the class of sets

$$\mathcal{D} = \Big\{\big\{(x, z) \in \mathbb{R}^d \times \mathbb{R}_+^* : \alpha z g(x) - g'(x) > 0\big\} : \alpha \in \mathbb{R}_+^*, (g, g') \in \mathcal{G}^2\Big\},$$

and, with a slight abuse of notation, denote by $\mathbf{S}_{\mathcal{P}}(j)$ the jth shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}$.

Theorem 2.1 *If $\mathcal{A}_\Theta$ is the Yatracos class defined by* (3), *then*

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathbf{S}_{\mathcal{D}}(m)\big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{4r}.$$

Consequently

$$\mathbf{E}\Big\{\int |f_n - f|\Big\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\Big\{\int |f_{n-m,\theta} - f|\Big\} + 8\sqrt{\frac{\log 2 + \log \mathbf{S}_{\mathcal{D}}(m) + 4r \log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m}} + \frac{3}{n}. \qquad (4)$$

Since in most cases of interest, bounds for $\mathbf{S}_{\mathcal{D}}(m)$ and $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ are polynomial in m and n (detailed examples are presented in Section 3), one can choose m and r as functions of n such that the terms on the right hand side of (4) are balanced. More precisely:

Corollary 2.1 *Assume that the shatter coefficients $\mathbf{S}_{\mathcal{D}}(m)$ and $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ are polynomial in their arguments. Then the choices*

$$m = \frac{n}{\log n} \qquad \textit{and} \qquad r = n^a, \quad a > 0,$$

lead to

$$\mathbf{E}\Big\{\int |f_n - f|\Big\} \le 3 \inf_{\theta\in\Theta} \mathbf{E}\Big\{\int |f_{n-m,\theta} - f|\Big\} + \mathrm{O}\left(\frac{\log n}{n^{(1-a)/2}}\right).$$

The optimal L_1 error of the univariate modified histogram is known to go to zero, under standard smoothness assumptions, at the rate $n^{-1/3}$, provided $r \sim n^{1/3}$. Therefore, the bound above essentially says that for polynomial shatter coefficients $\mathbf{S}_{\mathcal{D}}(m)$ and $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ and $a = 1/3$, we have asymptotically a performance that is guaranteed to be, up to a logarithm term, within a factor of three of the optimal performance. Roughly, the logarithm term appears as the price to be paid for using unrestricted classes of reference densities.

In order to use Theorem 2.1, we have to make sure that $\inf_{\theta\in\Theta} \mathbf{E} \int |f_{n-m,\theta} - f|$ is not much larger than $\inf_{\theta\in\Theta} \mathbf{E} \int |f_{n,\theta} - f|$, that is, holding out m observations does not cause much trouble. Whereas this result holds for parameter selection by the combinatorial method for most classical nonparametric density estimates (such as histograms, kernel estimates or wavelet estimates, see Devroye and Lugosi [1], Chapter 10), things turn out to be more complicated for the modified histogram estimate under study. Our result is as follows.

Theorem 2.2 *Denote by μ the common distribution of the X_i's, and suppose that there exists a positive real number α such that $\forall \theta \in \Theta$ $\big(\theta = (P, g), P = \{A_1, \ldots, A_\ell\}\big)$*

$$\alpha \le \mu(A_i)\,, \qquad i = 1, \ldots, \ell. \tag{5}$$

Then, for all $m \le n/2$, we have

$$\begin{aligned}&\mathbf{E}\left\{\int |f_n - f|\right\} \le \\ &\quad 3\left(1+\frac{2m}{n-m}+8\sqrt{\frac{m}{n}}+\frac{\sqrt{8}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\right)\inf_{\theta\in\Theta}\mathbf{E}\left\{\int |f_{n,\theta}-f|\right\}\\ &\quad +8\sqrt{\frac{\log 2+\log \mathbf{S}_{\mathcal{D}}(m)+4r\log \mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)}{m}}+\frac{3}{n}.\end{aligned}$$

Corollary 2.2 *Assume that the conditions of Theorem 2.2 are satisfied, and that the shatter coefficients $\mathbf{S}_{\mathcal{D}}(m)$ and $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ are polynomial in their arguments. Then the choices*

$$m=\frac{n}{\log n}\qquad \text{and}\qquad r=n^a,\quad 0<a\le 1/2,$$

lead to

$$\mathbf{E}\left\{\int |f_n-f|\right\}\le 3\left(1+\mathrm{O}\Big(\frac{1}{\sqrt{\log n}}\Big)\right)\inf_{\theta\in\Theta}\mathbf{E}\left\{\int |f_{n,\theta}-f|\right\}+\mathrm{O}\left(\frac{\log n}{n^{(1-a)/2}}\right).$$

Roughly speaking, condition (5) means that the set of candidate reference densities $\mathcal{G}$ is not too far from the target f. It is in particular satisfied when $\mathcal{G}$ is finite or when $\mathcal{G}$ is the class of Gaussian densities with bounded mean and variance parameters, and $\nu_g \ll \mu$ for all $g\in\mathcal{G}$. Let us now discuss some examples.

3 Examples

In this section, we provide various useful bounds for the shatter coefficients $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$ and $\mathbf{S}_{\mathcal{D}}(m)$. We first recall that the *Vapnik-Chervonenkis dimension* V (Vapnik and Chervonenkis [11]) of a class $\mathcal{H}$ of sets is defined as the largest integer p such that

$$\mathbf{S}_{\mathcal{H}}(p)=2^p.$$

If $\mathbf{S}_{\mathcal{H}}(p)=2^p$ for all p, then we say that $V=\infty$. A classical consequence of Sauer's lemma [14] shows that if $\mathcal{H}$ has Vapnik-Chervonenkis dimension $V<\infty$, then

$$\mathbf{S}_{\mathcal{H}}(j)\le (j+1)^V. \tag{6}$$

Let us first derive $\mathbf{S}_{\mathcal{P}}(j)$ for several classes of partitions $\mathcal{P}$ – recall that $\mathbf{S}_{\mathcal{P}}(j)$ means the jth shatter coefficient of the class of sets which are cells of any partition in $\mathcal{P}$. We first consider the univariate case $d=1$.

3.1 Univariate modified histograms

As a simple but important example, consider $d = 1$, and let $\mathcal{P}$ be the class containing all partitions of the real line into at most r intervals. Denoting by G the distribution function associated with any reference density g, the intervals A_i for $P = \{A_1, \ldots, A_\ell\} \in \mathcal{P}$ are defined as follows:

$$\begin{aligned} A_i &= \Big(G^{-1}\Big(\frac{i-1}{\ell}\Big), G^{-1}\Big(\frac{i}{\ell}\Big)\Big]\,,\ i = 1, \ldots, \ell - 1\,, \\ A_\ell &= \Big(G^{-1}\Big(1 - \frac{1}{\ell}\Big), G^{-1}(1)\Big)\,, \end{aligned}$$

where G^{-1} denotes the quantile function defined on $[0, 1]$ by $G^{-1}(u) = \inf\{x \in \mathbb{R} : G(x) \geq u\}$. Within this framework, $\mathbf{S}_{\mathcal{P}}(j)$ is at most the jth shatter coefficient of the class of all intervals, which equals $j(j + 1)/2 + 1$. Note that Berlinet and Brunel [8], [9] study a univariate cross-validation-based method to select ℓ (but not g and ℓ simultaneously).

Let us now focus attention on the shatter coefficient $\mathbf{S}_{\mathcal{D}}(m)$ for two useful classes of univariate reference densities $\mathcal{G}$. Recall that

$$\mathcal{D} = \Big\{\big\{(x, z) \in \mathbb{R}^d \times \mathbb{R}_+^* : \alpha z g(x) - g'(x) > 0\big\} : \alpha \in \mathbb{R}_+^*, (g, g') \in \mathcal{G}^2\Big\}\,.$$

Exponential family. A family $\mathcal{G}$ of densities on $\mathbb{R}$ is called an *exponential family* if each density in $\mathcal{G}$ may be written in the form

$$g_\xi(x) = c\gamma(\xi)\beta(x)e^{\sum_{i=1}^k \pi_i(\xi)\psi_i(x)}\,, \tag{7}$$

where ξ belongs to some parameter set Ξ, $\psi_1, \ldots, \psi_k : \mathbb{R} \to \mathbb{R}$, $\beta : \mathbb{R} \to [0, \infty)$, $\gamma > 0, \pi_1, \ldots, \pi_k : \Xi \to \mathbb{R}$ are fixed functions, and c is a positive normalization constant. Examples of exponential families include classes of Gaussian, gamma, beta, Rayleigh, and Maxwell densities. Note that for $\alpha > 0$, $\alpha z g_\xi(x) > g_{\xi'}(x)$ if and only if

$$\log z + \sum_{i=1}^k \big(\pi_i(\xi) - \pi_i(\xi')\big)\psi_i(x) + \log\frac{\alpha\gamma(\xi)}{\gamma(\xi')} > 0\,. \tag{8}$$

By a mapping that makes each of the functions of x and z a new variable, it is easy to see that inequality (8) is just a homogeneous linear inequality $a_1\lambda_1 + \ldots + a_{k+2}\lambda_{k+2} > 0$, with the coefficients a_i depending upon the pair (ξ, ξ') only. The Vapnik-Chervonenkis dimension for a collection of linear halfspaces in $\mathbb{R}^{k+2}$ is not more than $k+2$ (Devroye and Lugosi [1], Corollary 4.2). As a consequence, by (6),

$$\mathbf{S}_{\mathcal{D}}(m) \leq (m + 1)^{k+2}\,.$$

Series estimates. Let $\psi_1, \ldots, \psi_k$ be fixed nonnegative basis functions from $\mathbb{R}^d$ to $\mathbb{R}$ such that $\int \psi_i = t_i$ for $1 \leq i \leq k$. We define the class $\mathcal{G}$ as the collection of all linear combinations

$$g_\xi(x) = \sum_{i=1}^{k} a_i \psi_i(x)$$

with coefficient $\xi = (a_1, \ldots, a_k)$ satisfying $\sum_{i=1}^k a_i t_i = 1$. Clearly, for $\alpha > 0$, $\alpha z g_\xi(x) > g_{\xi'}(x)$ if and only if

$$\sum_{i=1}^{k} \alpha a_i z \psi_i(x) - \sum_{i=1}^{k} a'_i \psi_i(x) > 0\,.$$

Making again each of the functions $\psi_i(x)$ and $z\psi_i(x)$ a new variable, we are led to a homogeneous linear inequality $b_1\lambda_1 + \ldots + b_{2k}\lambda_{2k} > 0$, with coefficients b_i depending upon the pair (ξ, ξ') only. Therefore

$$\mathbf{S}_{\mathcal{D}}(m) \leq (m+1)^{2k}\,.$$

3.2 Multivariate modified histograms

The aim of this paragraph is to study multivariate modified histograms defined via a multinormal reference density. This leads us to consider the class

$$\mathcal{G} = \left\{ g_{m,\Sigma}(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}}\, e^{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)} \right\},$$

where m is an arbitrary element of $\mathbb{R}^d$ and Σ is a symmetric positive definite $d \times d$ matrix. For a given reference density $g_{m,\Sigma} \in \mathcal{G}$ and a given integer $\ell \geq 2$, we let the partition P be as follows.

- Set $\ell = \ell_1 \ldots \ell_d$, with $\ell_1, \ldots, \ell_d$ positive integers, and let $h_j = 1/\ell_j$ for $j = 1, \ldots, d$;
- For $j = 1, \ldots, d$ and $i_j = 1, \ldots, \ell_j - 1$, compute the quantiles of order $i_j h_j$ of a univariate standard normal $\mathcal{N}(0,1)$; denote by q_{j,i_j} these quantiles, with the convention $q_{j,0} = -\infty$ and $q_{j,\ell_j} = +\infty$;
- Consider the grid defined by the above family $\{q_{j,i_j}\}$; this grid leads to a partition of $\mathbb{R}^d$ into ℓ hyperrectangles, say $\tilde{A}_{i_1,\ldots,i_d}$, $1 \leq j \leq d, 1 \leq i_j \leq \ell_j$;

- Fix $T_{m,\Sigma}$ the affine transformation

$$T_{m,\Sigma}(x) = \Sigma^{1/2}x + m\,,$$

and let $\{A_{i_1,\ldots,i_d}\}$ be the image-partition of $\{\tilde{A}_{i_1,\ldots,i_d}\}$ by $T_{m,\Sigma}$ (see Figure 2 that depicts a bivariate example).

Finally take

$$P = \left\{A_{i_1,\ldots,i_d}\right\}_{\substack{1\le j\le d \\ 1\le i_j\le \ell_j}}\,.$$

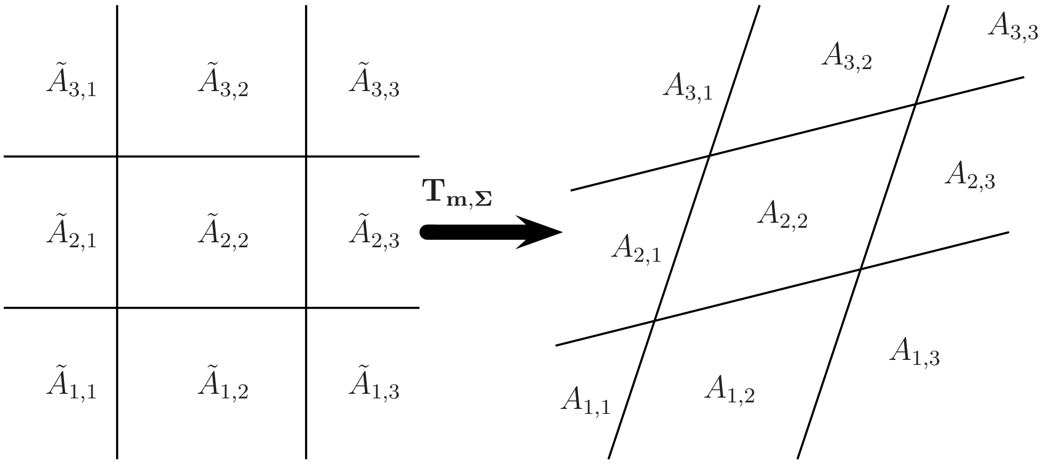

Figure 2: Transformation of a partition in $\mathbb{R}^2$.

Denote by $\nu_{m,\Sigma}$ the probability measure associated with the reference $g_{m,\Sigma}$. It is easily seen that, for any cell $A_{i_1,\ldots,i_d}$ of the partition P,

$$\nu_{m,\Sigma}(A_{i_1,\ldots,i_d}) = 1/\ell\,.$$

Note however that the decomposition $\ell = \ell_1 \ldots \ell_d$ is not necessarily unique. Thus, given $g_{m,\Sigma} \in \mathcal{G}$ and $\ell \ge 2$, we have just constructed a partition of $\mathbb{R}^d$ into ℓ sets of $\nu_{m,\Sigma}$-measure $1/\ell$. Clearly, each set in any such partition is an intersection of at most $2d$ hyperplanes (it is a polytope with at most $2d$ faces). Therefore

$$\mathbf{S}_{\mathcal{P}}(j) \le (j+1)^{2d(d+1)}$$

(see for example Devroye, Györfi and Lugosi [15]).

Let us now consider the shatter coefficient $\mathbf{S}_{\mathcal{D}}(m)$. Here, $\mathcal{G}$ is the class of multinormal densities, hence it is a multivariate exponential family. More precisely, setting $\xi = (m, \Sigma)$, each g_ξ in $\mathcal{G}$ may be written in the form

$$g_\xi(x) = c\gamma(\xi)\beta(x)e^{\sum_{i=1}^{k} \pi_i(\xi)\psi_i(x)},$$

with the notation of (7) – just replace $\mathbb{R}$ with $\mathbb{R}^d$ – and with $k = d(d+3)/2$. We conclude that

$$\mathbf{S}_{\mathcal{D}}(m) \leq (m+1)^{d(d+3)/2+2}.$$

Note that the bounds on the shatter coefficients in the examples presented above are polynomial in their arguments, so that Corollary 2.1 and Corollary 2.2 apply. One can argue that the bound $r = n^a$ is somewhat restrictive. However, extensive simulations (see Berlinet and Biau [16]) reveal that the number of cells ℓ should be very small with respect to n. Therefore, in practice, the bound $r = n^a$ does not harm too much. Moreover, it is consistent with the results of Barron, Györfi and van der Meulen [4], who proved that a univariate Kullback-Leibler-based choice of ℓ is of order $n^{1/3}$.

4 Simulations

In this section, we illustrate the theory with univariate simulation results enlightening the efficiency of the combinatorial method. The density to be estimated, a Beta $(2, 2)$, is shown in Figure 3 .

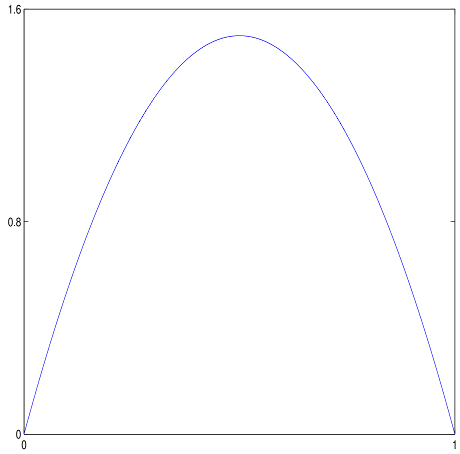

Figure 3: Density Beta $(2, 2)$ to be estimated.

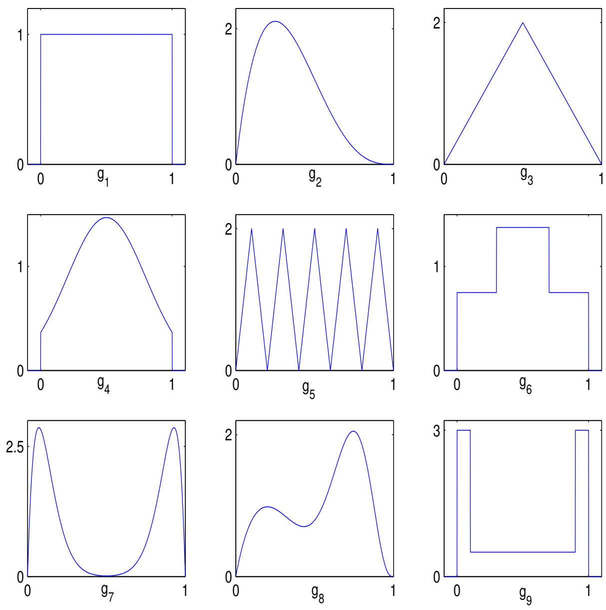

Figure 4: Collection of reference densities.

We consider a class $\mathcal{G}$ of references including 9 densities, presented in Figure 4. Given a reference g in the collection $\mathcal{G}$ and an integer ℓ, the associated partition is constructed via the quantiles of the density g, as explained in Paragraph 3.1. Thus, in this context, the method will automatically select a parameter θ from the set

$$\Theta = \left\{(g, \ell) : g \in \mathcal{G}, 2 \leq \ell \leq r\right\}.$$

The resulting minimum distance estimate is denoted f_n.

As suggested by a referee, we also shed light on the advantages of selecting both the partition and the reference density in contrast to the case where only the partition is selected. To this aim, for each *fixed* reference density $g \in \mathcal{G}$, we run the combinatorial method to select the sole number of cells ℓ from the set $\Theta_g = \{\ell : 2 \leq \ell \leq r\}$, and we denote by $f_{n,g}$ the elected estimate.

To assess the quality of the selected estimates, we compare the L_1 performances of the elected f_n and $f_{n,g}$ with the best estimates $f_{n,\theta^\star}$ and $f_{n,\theta_g^\star}$ in the corresponding classes, that is

$$\theta^\star \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \int |f_{n,\theta} - f| \right\},$$

and, for a fixed g,

$$\theta_g^\star \in \operatorname*{argmin}_{\theta \in \Theta_g} \left\{ \int |f_{n,\theta} - f| \right\}.$$

Table 1 and Table 2 summarize the results. For each of the references g, we display in Table 1 the L_1 error of $f_{n,g}$ and $f_{n,\theta_g^\star}$, and we present in Table 2 the error of the estimates f_n and $f_{n,\theta^\star}$. We also show the number $\hat{\ell}_n$ of selected classes. All results are averaged over 50 repetitions.

	$n=200, m=50, r=16$			$n=1000, m=150, r=30$		
g	$\int \lvert f_{n,g} - f\rvert$	$\int \lvert f_{n,\theta_g^\star} - f\rvert$	$\hat{\ell}_n$	$\int \lvert f_{n,g} - f\rvert$	$\int \lvert f_{n,\theta_g^\star} - f\rvert$	$\hat{\ell}_n$
g_1	0.2060	0.1536	9.68	0.1205	0.0958	17.20
g_2	0.3254	0.2961	12.92	0.2379	0.2228	24.24
g_3	0.1677	0.1103	7.28	0.1043	0.0695	15.12
g_4	0.1767	0.1036	8.28	0.1119	0.0849	14.08
g_5	0.4327	0.4000	14.28	0.3358	0.3176	24.72
g_6	0.2340	0.1891	10.84	0.1419	0.1141	18.08
g_7	0.8241	0.8135	15.64	0.6714	0.6633	29.44
g_8	0.2241	0.1743	9.04	0.1424	0.1144	17.04
g_9	0.2399	0.1728	10.92	0.1370	0.1089	19.12

Table 1: Combinatorial method results for the selection of P.

$n=200, m=50, r=16$			$n=1000, m=150, r=30$		
$\int \lvert f_n - f\rvert$	$\int \lvert f_{n,\theta^\star} - f\rvert$	$\hat{\ell}_n$	$\int \lvert f_n - f\rvert$	$\int \lvert f_{n,\theta^\star} - f\rvert$	$\hat{\ell}_n$
0.2249	0.0995	8.28	0.1469	0.0694	16.32

Table 2: Combinatorial method results for the selection of the pair (g, P).

The L_1 error ratios selected / optimal never exceed 2.26, and all of these results enlighten the good performances of the combinatorial method in general. They also clearly show the advantages of selecting both the partition and the reference density in contrast to the case where only the partition is selected. As a matter of fact, the L_1 performances of f_n over the $f_{n,g}$'s are significantly better for 5 reference models out of 9, and roughly similar for 2. Unsurprisingly, the best performances of $f_{n,g}$ are obtained for the densities g_3 (triangle) and g_4 (truncated Gaussian $\mathcal{N}(0.5, 1)$), which resemble the most the density Beta $(2,2)$. In practice, when one has no or few a priori information on the target density, the selection approach presented in the present paper is preferable.

5 Proofs

5.1 Proof of Theorem 2.1

We just have to prove that

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathbf{S}_{\mathcal{D}}(m)\big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{4r},$$

and the second part of the theorem will directly follow from inequality (2).

Let $y_1, \ldots, y_m$ be m distinct vectors in $\mathbb{R}^d$. For each $\theta = (g, P) \in \Theta$, $P = \{A_1, \ldots, A_\ell\}$, consider the $m \times r$ matrix z_θ such that the element in its tth row and jth column is

$$z_\theta^{(t,j)} = \begin{cases} \mathbf{1}_{[y_t \in A_j]} \sum_{i=1}^{n-m} \mathbf{1}_{[X_i \in A_j]} & \text{for } t \leq m,\ j \leq \ell\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly,

$$\mathbf{1}_{[y_t \in A_j]}\mathbf{1}_{[X_i \in A_j]} = 1 \quad \text{if and only if} \quad (y_t, X_i) \in A_j \times A_j\,.$$

Since there are $m(n-m)$ different pairs (y_t, X_i), the number of different values the jth column $(z_\theta^{(1,j)}, \ldots, z_\theta^{(m,j)})$ of the matrix z_θ can take as we vary $\theta \in \Theta$ is at most the shatter coefficient $\mathbf{S}_{\mathcal{C}}\big(m(n-m)\big)$ of the class of sets $\mathcal{C}$ of the form $A \times A$, where A is any set in any possible partition in $\mathcal{P}$. This shatter coefficient is clearly bounded by the square of the shatter coefficient $\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)$. Hence the jth column of the matrix z_θ can take at most $\big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^2$ values. But since the matrix z_θ has r columns, it can take at most

$$\big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{2r}$$

values. Thus if we set

$$\mathcal{W} = \big\{(z_\theta, z_{\theta'}) : (\theta, \theta') \in \Theta^2\big\},$$

we have

$$\operatorname{Card} \mathcal{W} \leq \big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{4r}.$$

For fixed $(w, w') \in \mathcal{W}$, let $U_{(w,w')}$ denote the collection of all (θ, θ') such that $(z_\theta, z_{\theta'}) = (w, w')$. For $(\theta, \theta') \in U_{(w,w')}$ $\big(\theta = (g, P),\ \theta' = (g', P'),$ $P = \{A_1, \ldots, A_\ell\}$, $P' = \{A'_1, \ldots, A'_{\ell'}\}\big)$ and $t \le m$, we have

$$y_t \in A_{\theta,\theta'} = \big\{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\big\}$$

if and only if

$$\frac{\sum_{j=1}^{\ell} z_\theta^{(t,j)} + 1}{(n-m)h + 1}\, g(y_t) > \frac{\sum_{j=1}^{\ell'} z_{\theta'}^{(t,j)} + 1}{(n-m)h' + 1}\, g'(y_t)\,,$$

where $h = 1/\ell$ and $h' = 1/\ell'$. Within the set $U_{(w,w')}$, $z_\theta^{(t,j)}$ and $z_{\theta'}^{(t,j)}$ are fixed for all t and j. Therefore, with the notation

$$z_t = \frac{\sum_{j=1}^{\ell} z_\theta^{(t,j)} + 1}{\sum_{j=1}^{\ell'} z_{\theta'}^{(t,j)} + 1} \qquad \text{for } 1 \le t \le m\,,$$

we obtain that $y_t \in A_{\theta,\theta'}$ if and only if

$$\frac{(n-m)h' + 1}{(n-m)h + 1}\, z_t\, g(y_t) - g'(y_t) > 0\,.$$

It follows that

$$\begin{aligned}
&\mathrm{Card}\big\{\{\mathbf{1}_{[y_1 \in A_{\theta,\theta'}]}, \ldots, \mathbf{1}_{[y_m \in A_{\theta,\theta'}]}\} : (\theta, \theta') \in U_{(w,w')}\big\} \\
&\quad \le \mathrm{Card}\big\{\{\mathbf{1}_{[\alpha z_1 g(y_1) - g'(y_1) > 0]}, \ldots, \mathbf{1}_{[\alpha z_m g(y_m) - g'(y_m) > 0]}\} : \alpha \in \mathbb{R}_+^*, (g, g') \in \mathcal{G}^2\big\} \\
&\quad \le \mathbf{S}_{\mathcal{D}}(m)\,.
\end{aligned}$$

Putting all pieces together, we obtain

$$\begin{aligned}
\mathrm{Card}\big\{\{y_1, \ldots, y_m\} \cap A_{\theta,\theta'} : (\theta, \theta') \in \Theta^2\big\} &\le \mathbf{S}_{\mathcal{D}}(m)\, \mathrm{Card}\, \mathcal{W} \\
&\le \mathbf{S}_{\mathcal{D}}(m) \big[\mathbf{S}_{\mathcal{P}}\big(m(n-m)\big)\big]^{4r}\,.
\end{aligned}$$

The proof of Theorem 2.1 is finished. ■

5.2 Proof of Theorem 2.2

The proof of Theorem 2.2 is a consequence of Theorem 2.1 and the following lemma.

Lemma 5.1 *Denote by μ the common distribution of the X_i's, and suppose that there exists a positive real number α such that $\forall\theta \in \Theta$ $\big(\theta = (P, g), P = \{A_1, \ldots, A_\ell\}\big)$*

$$\alpha \leq \mu(A_i)\,, \qquad i = 1, \ldots, \ell.$$

Introduce

$$J_{n,\theta} = \int |f_{n,\theta} - f|\,.$$

If m is a positive integer such that $2m \leq n$, then

$$\frac{\inf_{\theta\in\Theta} \mathbf{E}\big\{J_{n-m,\theta}\big\}}{\inf_{\theta\in\Theta} \mathbf{E}\big\{J_{n,\theta}\big\}} \leq 1 + \frac{2m}{n-m} + 8\sqrt{\frac{m}{n}} + \frac{\sqrt{8}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\,.$$

Proof of Lemma 5.1 Note first that the modified histogram is *not* an additive estimate in the sense of Devroye and Lugosi [1] so that their Theorem 10.2 does not apply. Nevertheless we can start with the inequality that they prove:

$$\inf_{\theta\in\Theta} \mathbf{E}\{J_{n-m,\theta}\} \leq \inf_{\theta\in\Theta} \mathbf{E}\{J_{n,\theta}\}\left(1 + 2\sup_{\theta\in\Theta} \frac{\mathbf{E}\big\{\int |f_{n-m,\theta} - f_{n,\theta}|\,\mathrm{d}x\big\}}{\mathbf{E}\big\{\int |f_{n,\theta} - \mathbf{E}f_{n,\theta}|\,\mathrm{d}x\big\}}\right).$$

Fix x and $\theta = (g, P)$ for now and define $K_\theta(x, X_i) = \mathbf{1}_{[X_i \in A(x)]}$. Recall that $A(x)$ denotes the cell of the partition P (which has ℓ cells) in which x falls. Observe that

$$f_{n,\theta}(x) = \frac{1}{nh+1}\Big(1 + \sum_{i=1}^{n} K_\theta(x, X_i)\Big)\, g(x)\,,$$

where $h = 1/\ell$. Introduce

$$Y_i = K_\theta(x, X_i) - \mathbf{E}\big\{K_\theta(x, X_i)\big\}\,,$$

and denote the partial sums of Y_i's by $S_j = Y_1 + \ldots + Y_j$. Observe the following:

$$\begin{aligned}
&(nh+1)|f_{n-m,\theta}(x) - f_{n,\theta}(x)| \\
&\quad = \left|\frac{nh+1}{(n-m)h+1}\Big(1 + \sum_{i=1}^{n-m} K_\theta(x, X_i)\Big) - \Big(1 + \sum_{i=1}^{n} K_\theta(x, X_i)\Big)\right| g(x) \\
&\quad = \left|\frac{mh}{(n-m)h+1}\Big(1 + \sum_{i=1}^{n-m} K_\theta(x, X_i)\Big) - \sum_{i=n-m+1}^{n} K_\theta(x, X_i)\right| g(x) \\
&\quad = \Bigg|\frac{mh}{(n-m)h+1}(Y_1 + \ldots + Y_{n-m}) - (Y_{n-m+1} + \ldots + Y_n) \\
&\qquad + \frac{m}{(n-m)h+1}\Big(h - \mathbf{E}\big\{K_\theta(x, X_1)\big\}\Big)\Bigg|\, g(x)\,,
\end{aligned}$$

so that

$$\mathbf{E}\big\{(nh+1)|f_{n-m,\theta}(x)-f_{n,\theta}(x)|\big\} \le \Big[\frac{m}{n-m}\mathbf{E}\big\{|S_{n-m}|\big\}+\mathbf{E}\big\{|S_m|\big\} + \frac{m}{(n-m)h+1}\big|h-\mathbf{E}\big\{K_\theta(x,X_1)\big\}\big|\Big]\, g(x)\,.$$

Also,

$$(nh+1)|f_{n,\theta}(x)-\mathbf{E}f_{n,\theta}(x)| = |S_n|\, g(x)\,,$$

which implies

$$\mathbf{E}\big\{(nh+1)|f_{n,\theta}(x)-\mathbf{E}f_{n,\theta}(x)|\big\} = \mathbf{E}\big\{|S_n|\big\}\, g(x)\,.$$

If $2m \le n$, a straightforward consequence of Lemma 10.1 and Lemma 10.3 in Devroye and Lugosi (2001) leads to

$$\frac{\mathbf{E}\big\{|f_{n-m,\theta}-f_{n,\theta}|\big\}}{\mathbf{E}\big\{|f_{n,\theta}-\mathbf{E}f_{n,\theta}|\big\}} \le \frac{m}{n-m}+4\sqrt{\frac{m}{n}}+\frac{m}{(n-m)h+1}\,\frac{\sqrt{8}\,\big|h-\mathbf{E}\big\{K_\theta(x,X_1)\big\}\big|}{\sqrt{n}\,\mathbf{E}\big\{|Y_1|\big\}}\,. \tag{9}$$

Let $p(x)$ stand for $\mu\big(A(x)\big)$. Clearly,

$$\begin{cases} \mathbf{E}\big\{K_\theta(x,X_1)\big\} = p(x) \\ \mathbf{E}\big\{|Y_1|\big\} = 2p(x)\big(1-p(x)\big)\,. \end{cases}$$

By assumption, and using the fact that $\ell \ge 2$, we obtain, still holding x fixed,

$$\alpha \le p(x) \le 1-\alpha\,.$$

Note that $0<\alpha\le 1/2$. By (9)

$$\frac{\mathbf{E}\big\{|f_{n-m,\theta}-f_{n,\theta}|\big\}}{\mathbf{E}\big\{|f_{n,\theta}-\mathbf{E}f_{n,\theta}|\big\}} \le \frac{m}{n-m}+4\sqrt{\frac{m}{n}}+\frac{m}{(n-m)h+1}\,\frac{\sqrt{8}\,|h-p(x)|}{2\sqrt{n}\,p(x)\big(1-p(x)\big)}\,.$$

Moreover

$$\frac{1}{p(x)\big(1-p(x)\big)} \le \frac{1}{\alpha(1-\alpha)}\,.$$

On the other hand,

$$|h-p(x)| \le \max\left(1,\frac{p(x)}{h}\right) h \le rh\,.$$

Putting all pieces together, we obtain

$$\begin{aligned}\frac{m}{(n-m)h+1}\,\frac{\sqrt{8}\,|h-p(x)|}{2\sqrt{n}\,p(x)\big(1-p(x)\big)} &\le \frac{mh}{(n-m)h+1}\,\frac{\sqrt{2}\,r}{\sqrt{n}\,\alpha(1-\alpha)} \\ &\le \frac{\sqrt{2}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\,.\end{aligned}$$

This implies that for any fixed θ

$$\begin{aligned}&\mathbf{E}\left\{\int |f_{n-m,\theta} - f_{n,\theta}| \,\mathrm{d}x\right\}\\ &\le \left(\frac{m}{n-m} + 4\sqrt{\frac{m}{n}} + \frac{\sqrt{2}\,mr}{(n-m)\sqrt{n}\,\alpha(1-\alpha)}\right)\mathbf{E}\left\{\int |f_{n,\theta} - \mathbf{E}f_{n,\theta}| \,\mathrm{d}x\right\}.\end{aligned}$$

This completes the proof of the lemma. ■

Acknowledgments. The authors greatly thank both referees for their comments and suggestions. They are also indebted to the *UMR Biométrie et Analyse des Systèmes, ENSAM-INRA, Montpellier, France* for providing facilities.

References

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

[2] Liese, F. and Vajda, I., 1987, *Convex Statistical Distances*. Teubner–Verlag, Leipzig.

[3] Barron, A. R., 1988, The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe: Japan, June 19-24.

[4] Barron, A. R., Györfi, L. and van der Meulen, E. C., 1992, Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, **38**,1437–1454.

[5] Berlinet, A., Györfi, L. and van der Meulen, E. C., 1997, Asymptotic normality of relative entropy in multivariate density estimation. *Publications de l'Institut de Statistique de l'Université de Paris*, **41**,3–27.

[6] Györfi, L., Liese, F., Vajda, I. and van der Meulen, E. C., 1998, Distribution estimates consistent in χ^2-divergence. *Statistics*, **32**,31–57.

[7] Berlinet, A., Vajda, I. and van der Meulen, E. C., 1998, About the asymptotic accuracy of Barron density estimates. *IEEE Transactions on Information Theory*, **44**,999–1009.

[8] Berlinet, A. and Brunel, E., 2000, Choix optimal du nombre de classes pour l'estimateur de Barron de la densité. *Comptes Rendus de l'Académie des Sciences de Paris*, **331**,713–716.

[9] Berlinet, A. and Brunel, E., 2004, Cross-validated density estimates based on Kullback-Leibler information. *Journal of Nonparametric Statistics*, **16**,493–513.

[10] Yatracos, Y.G., 1985, Rates of convergence of minimun distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, **13**,768–774.

[11] 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*, **16**,264–280.

[12] Scheffé, H., 1947, A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, **18**,434–438.

[13] Keziou, A., 2003, Dual representation of ϕ-divergences and applications. *Comptes Rendus de l'Académie des Sciences de Paris*, **336**,857–862.

[14] Sauer, N., 1972, On the density of families of sets. *Journal of Combinatorial Theory, Series A*, **13**,145–147.

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

[16] Berlinet, A. and Biau, G., 2004, Iterated modified histograms as dynamical systems. *Journal of Nonparametric Statistics*, **16**,385–401.