Divergence-type errors of smooth Barron-type density estimators

Jan Beirlant
Department of Mathematics, Section of Applied Mathematics
Catholic University Leuven, Belgium
Alain Berlinet and Gérard Biau*
Laboratory of Probability and Statistics
University Montpellier II, France
Igor Vajda
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Czech Republic

Abstract

Barron-type estimators are histogram-based distribution estimators that have been proved to have good consistency properties according to several information theoretic criteria. However they are not continuous. In this paper, we examine a new class of continuous distribution estimators obtained as a combination of Barron-type estimators with the frequency polygon. We prove the consistency of these estimators in expected information divergence and expected χ^2-divergence. For one of them we evaluate the rate of convergence in expected χ^2-divergence.

Key Words: Barron-type density estimators, consistency in expected information divergence, consistency in expected χ^2-divergence, rate of convergence.

AMS subject classification: 62G07.

1 Introduction

Barron (1988) introduced a class of nonparametric density estimators $f_{n,B}$ which are convex combinations

$$(1 - a_n)\tilde{f}_n + a_n\, g, \quad 0 < a_n < 1,$$

Jan Beirlant is supported by the University Montpellier II. Igor Vajda is supported by the GACR grant 102/99/1137 and by the University Montpellier II.

*Correspondence to: Gérard Biau, Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie – Paris VI, Boîte 158, 175 rue du Chevaleret, 75013 Paris, France. Email: biau@ccr.jussieu.fr

of an *a priori* given *reference density* g and the classical histogram estimator f_n modified with the help of g by the rule

$$\tilde{f}_n(x) = f_n(x)\, g(x), \quad x \in \mathbb{R}.$$

If the reference density dominates the density f of the observations, then the estimators $f_{n,B}$ are consistent in the sense of the *information divergence* error criterion $D(f, f_{n,B})$, which is considerably stronger than the usual L_1 error criterion $\|f - f_{n,B}\|_1$. This consistency, as well as the consistency with respect to even a stronger *chi-squared* error criterion $\chi^2(f, f_{n,B})$, have been studied by Barron, Györfi and van der Meulen (1992), Györfi, Liese, Vajda and van der Meulen (1998), Berlinet, Vajda and van der Meulen (1998), and Vajda and van der Meulen (2001), where it is also explained why stronger criteria are needed for applications in information transfer and communication.

Let us describe the situation in a more formal way. Denote by $\mathbb{P}_\nu$ the set of all probability measures on a measurable space $(X, \mathcal{A})$ dominated by a given probability measure ν on X. If $\phi : [0, +\infty[\to \mathbb{R}$ is continuous and convex then for every μ, $\mu^* \in \mathbb{P}_\nu$ the *ϕ-divergence* (Liese and Vajda 1987) is defined by the formula

$$D_\phi(\mu, \mu^*) = \int_X f^* \, \phi\left(\frac{f}{f^*}\right) \mathrm{d}\nu\,, \tag{1.1}$$

where f and f^* are the distribution densities,

$$f = \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \quad \text{and} \quad f^* = \frac{\mathrm{d}\mu^*}{\mathrm{d}\nu}\,.$$

In (1.1),

$$0\, \phi\left(\frac{f}{0}\right) = f \lim_{t\to\infty} \frac{\phi(t)}{t} \quad \text{for every } f \geq 0.$$

Therefore

$$D_\phi(\mu, \mu^*) = \int_{\{f^*>0\}} f^* \, \phi\left(\frac{f}{f^*}\right) \mathrm{d}\nu + \mu\big(\{f^* = 0\}\big) \lim_{t\to\infty} \frac{\phi(t)}{t}\,,$$

with the standard convention $0 \times \infty = 0$. In particular, the *information divergence* is given by $D(\mu, \mu^*) = D_{t \ln t}(\mu, \mu^*)$, *i.e.*,

$$D(\mu, \mu^*) = \int_{\{f^*>0\}} f \ln \frac{f}{f^*} \, \mathrm{d}\nu + \mu\big(\{f^* = 0\}\big) \times \infty\,. \tag{1.2}$$

The χ^2-*divergence* is defined by $\chi^2(\mu,\mu^*) = D_{t^2-1}(\mu,\mu^*)$, *i.e.*,

$$\chi^2(\mu,\mu^*) = \int_{\{f^*>0\}} \frac{f^2}{f^*}\, \mathrm{d}\nu - 1 + \mu\big(\{f^*=0\}\big) \times \infty\,. \tag{1.3}$$

Note that $\chi^2(\mu,\mu^*)$ can also be obtained with $\phi(t) = (t-1)^2$, which leads to

$$\chi^2(\mu,\mu^*) = \int_{\{f^*>0\}} \frac{(f-f^*)^2}{f^*}\, \mathrm{d}\nu + \mu\big(\{f^*=0\}\big) \times \infty\,.$$

It follows from (1.2) and (1.3) that $D(\mu,\mu^*) < \infty$ or $\chi^2(\mu,\mu^*) < \infty$ implies $\mu\big(\{f^*=0\}\big) = 0$. The condition $\mu\big(\{f^*=0\}\big) = 0$ is known to be equivalent to

$$\mu \ll \mu^*. \tag{1.4}$$

In case $\mu^* = \nu$, the density f^* is identical to 1 and in this manner we obtain the simpler formula

$$D_\phi(\mu,\nu) = \int_X \phi(f)\, \mathrm{d}\nu\,. \tag{1.5}$$

In particular,

$$D(\mu,\nu) = \int_X f\ln f\, \mathrm{d}\nu = \mathbf{E}_\nu(f\ln f) \tag{1.6}$$

and

$$\chi^2(\mu,\nu) = \int_X f^2\, \mathrm{d}\nu - 1 = \mathbf{E}_\nu(f^2) - 1. \tag{1.7}$$

Here and throughout, $\mathbf{E}_\nu$ denotes the expectation over the random argument distributed on X by ν.

Note that (1.4) does not imply $D(\mu,\mu^*) < \infty$ or $\chi^2(\mu,\mu^*) < \infty$, as one can easily see from (1.5)–(1.7): take for instance the Borel line $\big(\mathbb{R},\mathcal{B}(\mathbb{R})\big)$ for $(X,\mathcal{A})$ (here, $\mathcal{B}(\mathbb{R})$ denotes the collection of Borel sets in $\mathbb{R}$), the standard normal distribution for ν, and the Cauchy distribution for μ, leading to the density

$$f(x) = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x) = \sqrt{\frac{2}{\pi}}\,\frac{\exp(x^2/2)}{1+x^2}, \quad x\in\mathbb{R}.$$

Consider now the common problem of estimating the unknown probability distribution $\mu \in \mathbb{P}_\nu$ defined on the measurable space $(X,\mathcal{A})$, based on independent observations $X_1,\dots,X_n$ distributed by μ. We can represent μ by its probability density f, and estimators $\hat\mu_n$ of μ with values in $\mathbb{P}_\nu$ by their densities f_n. Barron et al. (1992) studied the existence of estimators

$\hat{\mu}_n \in \mathbb{P}_\nu$ consistent in information divergence and in expected information divergence, *i.e.*,

$$\lim_{n\to\infty} D(\mu, \hat{\mu}_n) = 0 \quad a.s.$$

and

$$\lim_{n\to\infty} \mathbf{E}D(\mu, \hat{\mu}_n) = 0,$$

where the symbol $\mathbf{E}$ denotes the expectation with respect to the random vector $(X_1, \ldots, X_n)$ figuring in the estimator $\hat{\mu}_n$. They proved that if ν is a probability measure with $D(\mu, \nu) < \infty$, then the histogram-based estimator of Barron (1988), as well as some modifications of this estimator (referred to as *Barron-type estimators* in the sequel), are consistent both in information and expected information divergence. For the Barron estimator itself these authors discussed the rate of convergence in expected information divergence. More recently, Györfi et al. (1998) presented arguments in favor of the χ^2-divergence, and proved similar results for the Barron-type estimators with the information divergence replaced by the χ^2-divergence. The consistency and rate of convergence statements given in the paper by Györfi et al. are stronger than in the paper by Barron et al. Indeed, by using the inequality $\ln t \le t - 1$ one easily obtains that

$$D(\mu, \mu^*) = \int_X f \ln \frac{f}{f^*}\, \mathrm{d}\nu \le \int_X \frac{f^2}{f^*}\, \mathrm{d}\nu - 1 = \chi^2(\mu, \mu^*). \tag{1.8}$$

On the other hand, there are examples where $D(\mu, \mu^*) \to 0$ and $\chi^2(\mu, \mu^*) \to \infty$. Therefore the χ^2-divergence is strictly topologically stronger, *i.e.*, the consistency defined by the χ^2-divergence implies that in the information divergence but not *vice versa*. Motivations for the consistency of estimators $\hat{\mu}_n$ in the sense of $D(\mu, \hat{\mu}_n)$ can be found in Barron et al. (1992) and also in Devroye, Györfi and Lugosi (1996). Similar motivations for the stronger error criterion $\chi^2(\mu, \hat{\mu}_n)$ have been given in Györfi et al. (1998) and Berlinet et al. (1998).

The class of Barron-type estimators studied by Barron et al. (1992) and Györfi et al. (1998) can be obtained as follows. Consider a sequence of integers $\{m_n\}_{n\ge1}$, $1 \le m_n \le n$, associated with a sequence of measurable partitions $\{\mathcal{P}_n\}_{n\ge1} = \{A_{n1}, \ldots, A_{nm_n}\}_{n\ge1}$ where the A_{nj}'s are rectangles with

$$\nu(A_{nj}) = h_n = \frac{1}{m_n}, \quad 1 \le j \le m_n.$$

Let $A_n(x)$ stand for the rectangle A_{nj} of $\mathcal{P}_n$ such that $x \in A_{nj}$. For a given sequence $\{a_n\}_{n\geq 1}$ $(0 < a_n < 1)$ one defines an estimator $f_{n,B}$ of $f = \mathrm{d}\mu/\mathrm{d}\nu$ by

$$f_{n,B}(x) = (1 - a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n, \quad x \in X, \tag{1.9}$$

where μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$. Naturally, it is assumed that

$$\lim_{n\to\infty} a_n = 0. \tag{1.10}$$

The estimators $f_{n,B}$ are defined with respect to the dominating probability measure ν (which is supposed to be known) and can be regarded as modified histograms with respect to ν. Observe that if we assume that $(X, \mathcal{A}) = \big(\mathbb{R}^d, B(\mathbb{R}^d)\big)$, $d \geq 1$, and that ν is absolutely continuous with respect to the Lebesgue measure λ, with $g = \mathrm{d}\nu/\mathrm{d}\lambda$, then the Barron-type distribution estimators can be derived from the density estimators

$$f_{n,B}(x) = \left[(1 - a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\right] g(x), \quad x \in \mathbb{R}^d. \tag{1.11}$$

In this framework, the probability density g is sometimes called the *reference density*. For the choice

$$a_n = \frac{1}{nh_n + 1} \tag{1.12}$$

the estimator $f_{n,B}$ in (1.9) is precisely the density estimator first introduced by Barron (1988).

In the present paper we assume $(X, \mathcal{A}) = \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$. We restrict ourselves to dominating probability measures ν with distributions G and densities g with respect to the Lebesgue measure λ. The discontinuities in the estimate $f_{n,B}$ appear as a drawback when the underlying density can be assumed to be smooth. Also they can limit the usefulness of $f_{n,B}$ as a graphical tool. Here, we consider a smooth version of the Barron-type density estimators replacing the histogram by the frequency polygon, defined by connecting the mid-bin values of the histogram with straight lines (see Scott 1985, and Scott 1992, Chapters 3 and 4). From the analysis below it turns out that this extension is not so easily handled. The complexity of the information divergence and chi-squared criterion implies difficulties when dealing with estimates which are not piecewise multiple of a constant density g. Currently the authors are working on extensions of the present results to splines and other smoothers in the uni- and multivariate case. This is far from being straightforward.

The paper is organized as follows. In Section 2, the new class of distribution estimators is presented and some information theoretic issues are discussed. Section 3 is devoted to the proof of the consistency in expected information divergence and expected χ^2-divergence of the estimators under study. Finally, the rate of convergence in expected χ^2-divergence of one of these estimators is evaluated in Section 4.

Convention Measures μ, $\mu^* \in \mathbb{P}_\nu$ as well as the estimators $\hat{\mu}_n$ of μ will be represented by their densities f, f^* and $\hat{f}_n$ with respect to ν. In such cases we shall write $D(f, f^*)$ and $\chi^2(f, f^*)$ instead of $D(\mu, \mu^*)$ and $\chi^2(\mu, \mu^*)$, and similarly for $\mu^* = \hat{\mu}_n$ and $f^* = \hat{f}_n$.

2 Smooth Barron-type density estimators

2.1 Construction

Let G be the distribution function associated with a Lebesgue density g of ν on $\mathbb{R}$. Consider the function

$$k(u) = f\big(G^{-1}(u)\big), \quad u \in]0,1[\,, \tag{2.1}$$

where G^{-1} is the quantile function, defined on $]0,1[$ by $G^{-1}(u) = \inf\{x : G(x) \geq u\}$. Denoting by λ the Lebesgue measure on $]0,1[$, the substitution $x = G^{-1}(u)$ leads to

$$\int_{]0,1[} k \, \mathrm{d}\lambda = \int_{\mathbb{R}} f \, \mathrm{d}\nu = 1,$$

from which we see that k is a probability density on $]0,1[$. Without loss of generality we can assume that ν is the Lebesgue measure on $]0,1[$ and that f coincides with the probability density k on $]0,1[$ defined in (2.1). Indeed the mapping $G : \mathbb{R} \to]0,1[$ transforms ν into the Lebesgue measure on $]0,1[$, A_{nj}'s $\in \mathcal{P}_n$ into intervals $\big](j-1)/m_n, j/m_n]$, the density f into k, and preserves the value of $D(\mu, \mu^*)$ and $\chi^2(\mu, \mu^*)$, whatever may be the probability measure $\mu^* \in \mathbb{P}_\nu$.

Thus, throughout this paper we let X be the unit interval $]0,1[$ and $\mathcal{A}$ be the σ-field of all Borel sets in $]0,1[$. Further, we let ν be the Lebesgue measure on $\big(]0,1[,\mathcal{B}(]0,1[)\big)$, μ any probability measure on $\big(]0,1[,\mathcal{B}(]0,1[)\big)$ satisfying $\mu \in \mathbb{P}_\nu$ and f the density of μ with respect to ν. Finally, we let $\{\mathcal{P}_n\}_{n\geq 1} = \{A_{n1},\ldots,A_{nm_n}\}_{n\geq 1}$ be the sequence of uniform partitions of $]0,1]$, defined by

$$A_{nj} =]t_{n(j-1)}, t_{nj}] \quad \text{for } t_{nj} = \frac{j}{m_n}, \quad 1 \leq j \leq m_n.$$

In this manner, the Barron-type density estimators defined in (1.9) then read as follows:

$$f_{n,B}(u) = (1-a_n)\frac{\mu_n\big(A_n(u)\big)}{h_n} + a_n, \quad u \in]0,1[\,. \tag{2.2}$$

Define now

$$c_{nj} = \frac{t_{n(j-1)} + t_{nj}}{2}, \quad 1 \leq j \leq m_n,$$

so that A_{nj} and t_{nj} satisfy the relations

$$A_{nj} =]t_{n(j-1)}, t_{nj}] = \Big]c_{nj} - \frac{h_n}{2}, c_{nj} + \frac{h_n}{2}\Big] \quad \text{and} \quad \nu(A_{nj}) = h_n.$$

If we denote by f_n the standard histogram density on $]0,1[$, namely

$$f_n(u) = \frac{\mu_n(A_{nj})}{h_n} \quad \text{for } u \in A_{nj}, \quad 1 \leq j \leq m_n, \tag{2.3}$$

then the resulting *smooth Barron-type density estimators* $f_{n,P}$ studied in the present paper read as follows:

$$f_{n,P}(u) = (1-a_n)p_n(u) + a_n, \quad u \in]0,1[\,, \tag{2.4}$$

where

$$p_n(u) = \begin{cases} f_n(c_{n1}) & \text{if } 0 < u \leq c_{n1} \\ f_n(c_{nj})\frac{c_{n(j+1)}-u}{h_n} + f_n(c_{n(j+1)})\frac{u-c_{nj}}{h_n} & \text{if } c_{nj} < u \leq c_{n(j+1)} \\ f_n(c_{nm_n}) & \text{if } c_{nm_n} < u < 1, \end{cases}$$

where $1 \leq j \leq m_n - 1$. If we put

$$A_{nj} = A_{nj}^- \cup A_{nj}^+, \quad \text{where } A_{nj}^- =]t_{n(j-1)}, c_{nj}] \text{ and } A_{nj}^+ =]c_{nj}, t_{nj}],$$

then

$$p_n(u) = \begin{cases} \frac{\mu_n(A_{n1})}{h_n} & \text{if } u \in A_{n1}^- \\ \frac{\mu_n(A_{nj})}{h_n} \frac{c_{n(j+1)}-u}{h_n} + \frac{\mu_n(A_{n(j+1)})}{h_n} \frac{u-c_{nj}}{h_n} & \text{if } u \in A_{nj}^+ \cup A_{n(j+1)}^- \\ \frac{\mu_n(A_{nm_n})}{h_n} & \text{if } u \in A_{nm_n}^+, \end{cases} \tag{2.5}$$

where $1 \leq j \leq m_n - 1$. The estimator p_n is called the *frequency polygon* (see Scott 1985, Scott 1992, Chapters 3 and 4, and Beirlant, Berlinet and Györfi 1998) and is defined by broken lines joining the two neighboring points $(c_{nj}, \mu_n(A_{nj})/h_n)$, $(c_{n(j+1)}, \mu_n(A_{n(j+1)})/h_n)$. Remark that both p_n and $f_{n,P}$ integrate to one. Remark also that the probability measure μ is absolutely continuous with respect to the probability measure associated with $f_{n,P}$, so that $D(f, f_{n,P})$ and $\chi^2(f, f_{n,P})$ are well defined (see (1.2) and (1.3)). Unlike the standard histogram, the frequency polygon is continuous (Figure 1). Consequently this will also be true for the estimators $f_{n,P}$ under study (see Figure 2 that depicts examples of Barron and piecewise linear Barron density estimators).

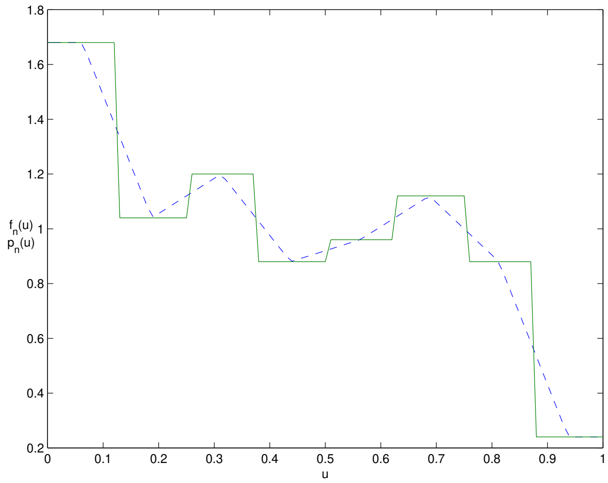

Figure 1: Histogram (f_n, continuous line) and polygon (p_n, dashed line) estimates of 100 *data in* $]0,1[$.

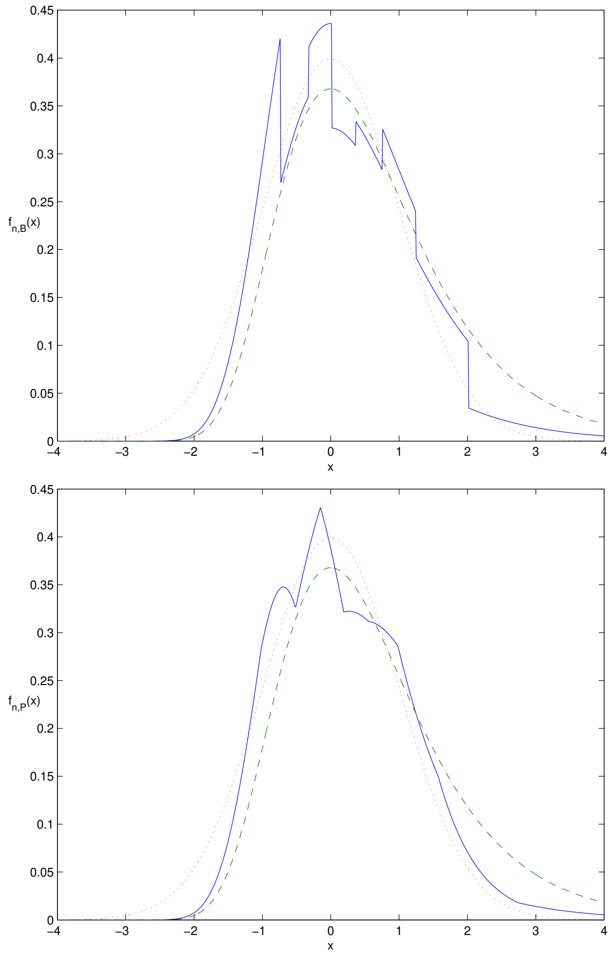

Figure 2: Barron ($f_{n,B}$, top) and smooth Barron ($f_{n,P}$, bottom) density estimates of $n = 100$ Gaussian (dotted line) data ($m_n = 8$). The reference density (see (1.11)) is Gumbel (dashed line).

2.2 Approximability

Consider the probability densities f_n^* and p_n^* with respect to the Lebesgue measure ν defined on the unit interval $]0, 1[$ as follows:

$$f_n^*(u) = \frac{\mu(A_{nj})}{h_n} \quad \text{for } u \in A_{nj},\ 1 \le j \le m_n, \tag{2.6}$$

and

$$p_n^*(u) = \begin{cases} \frac{\mu(A_{n1})}{h_n} & \text{if } u \in A_{n1}^- \\ \frac{\mu(A_{nj})}{h_n}\frac{c_{n(j+1)}-u}{h_n} + \frac{\mu(A_{n(j+1)})}{h_n}\frac{u-c_{nj}}{h_n} & \text{if } u \in A_{nj}^+ \cup A_{n(j+1)}^- \\ \frac{\mu(A_{nm_n})}{h_n} & \text{if } u \in A_{nm_n}^+, \end{cases} \tag{2.7}$$

where $1 \le j \le m_n - 1$. It is easy to see that f_n^* (*resp.* p_n^*) is the histogram-shaped (*resp.* the polygon-shaped) density with respect to the Lebesgue measure based on the true probability distribution μ (compare (2.6) and (2.7) with (2.3) and (2.5)). For similar reasons as in the original discontinuous Barron-type estimators (see Barron et al. 1992, and Györfi et al. 1998), where the authors were interested in

$$D(f, f_n^*) = \int_{]0,1[} f \ln \frac{f}{f_n^*}\, \mathrm{d}\nu \quad \text{and} \quad \chi^2(f, f_n^*) = \int_{]0,1[} \frac{f^2}{f_n^*}\, \mathrm{d}\nu - 1,$$

here we are interested in

$$D(f, p_n^*) = \int_{]0,1[} f \ln \frac{f}{p_n^*}\, \mathrm{d}\nu \tag{2.8}$$

as well as in

$$\chi^2(f, p_n^*) = \int_{]0,1[} \frac{f^2}{p_n^*}\, \mathrm{d}\nu - 1. \tag{2.9}$$

2.2.1 The information divergence case

Let us first examine the asymptotic behavior of formula (2.8). According to Csiszár's (1973) approximation theorem for the uniform partitions $\{\mathcal{P}_n\}_{n\ge1}$ under study, the two conditions

$$\text{(i)} \lim_{n\to\infty} m_n = \infty \quad \text{and} \quad \text{(ii)}\ D(\mu, \nu) < \infty$$

suffice to ensure the convergence

$$\lim_{n\to\infty} D(f, f_n^*) = 0. \tag{2.10}$$

Here, we assume (i) and (ii) and we want to prove that also

$$\lim_{n\to\infty} D(f, p_n^*) = 0. \tag{2.11}$$

Our approach is based on the fact that the function $t \mapsto a/t$, $t > 0$, is convex for $a \geq 0$. Therefore, by (2.7),

$$
\begin{aligned}
&\int_{A_{nj}^+ \cup A_{n(j+1)}^-} f \ln \frac{f}{p_n^*}\ \mathrm{d}\nu \\
&\leq \int_{A_{nj}^+ \cup A_{n(j+1)}^-} \left[\frac{c_{n(j+1)} - u}{h_n} f(u) \ln \frac{f(u)}{f_n^*(c_{nj})} \right. \\
&\qquad\qquad \left. + \frac{u - c_{nj}}{h_n} f(u) \ln \frac{f(u)}{f_n^*(c_{n(j+1)})} \right] \nu(\mathrm{d}u) \\
&= \int_{A_{nj}^+} f \ln \frac{f}{f_n^*(c_{nj})}\ \mathrm{d}\nu + \int_{A_{nj}^+} \frac{u - c_{nj}}{h_n} f(u) \ln \frac{f_n^*(c_{nj})}{f_n^*(c_{n(j+1)})}\ \nu(\mathrm{d}u) \\
&\quad + \int_{A_{n(j+1)}^-} f \ln \frac{f}{f_n^*(c_{n(j+1)})}\ \mathrm{d}\nu \\
&\quad + \int_{A_{n(j+1)}^-} \frac{c_{n(j+1)} - u}{h_n} f(u) \ln \frac{f_n^*(c_{n(j+1)})}{f_n^*(c_{nj})}\ \nu(\mathrm{d}u) \\
&= \int_{A_{nj}^+} f \ln \frac{f}{f_n^*}\ \mathrm{d}\nu + d_{nj}^+ \ln \frac{\mu(A_{nj})}{\mu(A_{n(j+1)})} \\
&\quad + \int_{A_{n(j+1)}^-} f \ln \frac{f}{f_n^*}\ \mathrm{d}\nu + d_{n(j+1)}^- \ln \frac{\mu(A_{n(j+1)})}{\mu(A_{nj})}\,,
\end{aligned}
$$

where

$$
d_{nj}^+ = \int_{A_{nj}^+} \frac{u - c_{nj}}{h_n} f(u)\ \nu(\mathrm{d}u) \leq \frac{\mu(A_{nj}^+)}{2} \tag{2.12}
$$

and

$$
d_{nj}^- = \int_{A_{nj}^-} \frac{c_{nj} - u}{h_n} f(u)\ \nu(\mathrm{d}u) \leq \frac{\mu(A_{nj}^-)}{2}\,. \tag{2.13}
$$

Since $p_n^*(u) = f_n^*(u)$ for $u \in A_{n1}^- \cup A_{nm_n}^+$, we obtain

$$
\begin{aligned}
D(f, p_n^*) &= \int_{]0,1[} f \ln \frac{f}{p_n^*}\ \mathrm{d}\nu \\
&= \int_{A_{n1}^- \cup A_{nm_n}^+} f \ln \frac{f}{p_n^*}\ \mathrm{d}\nu + \sum_{j=1}^{m_n - 1} \int_{A_{nj}^+ \cup A_{n(j+1)}^-} f \ln \frac{f}{p_n^*}\ \mathrm{d}\nu \\
&\leq \int_{A_{n1}^- \cup A_{nm_n}^+} f \ln \frac{f}{f_n^*}\ \mathrm{d}\nu
\end{aligned}
$$

$$+\sum_{j=1}^{m_n-1}\left[\int_{A^+_{nj}\cup A^-_{n(j+1)}} f\ln\frac{f}{f^*_n}\,\mathrm{d}\nu + (d^-_{n(j+1)} - d^+_{nj})\ln\frac{\mu(A_{n(j+1)})}{\mu(A_{nj})}\right]$$

$$= D(f, f^*_n) + \sum_{j=1}^{m_n-1}(d^-_{n(j+1)} - d^+_{nj})\ln\frac{\mu(A_{n(j+1)})}{\mu(A_{nj})}\,.$$

Thus, from (2.10), (2.11) holds if

$$\lim_{n\to\infty}\sum_{j=1}^{m_n-1}(d^-_{n(j+1)} - d^+_{nj})\ln\frac{\mu(A_{n(j+1)})}{\mu(A_{nj})} = 0. \tag{2.14}$$

This is an interesting new analytic problem. Next, we discuss some sufficient conditions for (2.14) to hold.

(a) Let us assume that

$$\lim_{n\to\infty}\max_{1\le j\le m_n-1}\left|\ln\frac{\mu(A_{n(j+1)})}{\mu(A_{nj})}\right| = 0. \tag{2.15}$$

By (2.12) and (2.13)

$$|d^-_{n(j+1)} - d^+_{nj}| \le d^-_{n(j+1)} + d^+_{nj} \le \frac{\mu(A^+_{nj}) + \mu(A^-_{n(j+1)})}{2}\,,$$

and hence

$$\sum_{j=1}^{m_n-1}|d^-_{n(j+1)} - d^+_{nj}| \le \frac{1-\mu(A^-_{n1}) - \mu(A^+_{nm_n})}{2} \le \frac{1}{2}\,.$$

Consequently (2.15) implies (2.14).

(b) If f is nondecreasing on $]0,1[$ then

$$\mu(A_{n1}) \le \mu(A_{n2}) \le \ldots \le \mu(A_{nm_n})$$

and

$$0 \le d^+_{n1} \le d^-_{n2} \le d^+_{n2} \le \ldots \le d^-_{nm_n}.$$

Therefore in this case

$$\sum_{j=1}^{m_n-1}(d^-_{n(j+1)} - d^+_{nj})\ln\frac{\mu(A_{n(j+1)})}{\mu(A_{nj})} \le \frac{1}{2}\sum_{j=1}^{m_n-1}\mu(A_{n(j+1)})\ln\frac{\mu(A_{n(j+1)})}{\mu(A_{nj})}\,.$$

Thus (2.14) holds if

$$\lim_{n\to\infty} \sum_{j=1}^{m_n-1} \mu(A_{n(j+1)}) \ln \frac{\mu(A_{n(j+1)})}{\mu(A_{nj})} = 0.$$

This type of conditions on discrete distributions seems to be new in the information theory.

Another sufficient condition in this case is

$$\lim_{n\to\infty} \left[\max_{1\le j\le m_n-1} (d^-_{n(j+1)} - d^+_{nj}) \right] \ln \frac{\mu(A_{nm_n})}{\mu(A_{n1})} = 0.$$

(c) An interesting case arises if f is *α-Hölderian* ($\alpha > 0$) on $]0,1[$, *i.e.*, for some $K \ge 0$,

$$|f(x_1) - f(x_2)| \le K|x_1 - x_2|^\alpha \quad \text{for all } x_1, x_2 \in]0,1[\,. \tag{2.16}$$

By definition,

$$d^-_{n(j+1)} - d^+_{nj} = \frac{1}{h_n} \int_0^{h_n/2} v \Big[f(c_{n(j+1)} - v) - f(c_{nj} + v) \Big] \; \nu(\mathrm{d}v)$$

so that, by (2.16),

$$|d^-_{n(j+1)} - d^+_{nj}| \le \frac{K}{h_n} \int_0^{h_n/2} v(h_n - 2v)^\alpha \; \nu(\mathrm{d}v) = \frac{K h_n^{\alpha+1}}{4(\alpha+1)(\alpha+2)}\,. \tag{2.17}$$

Further,

$$\mu(A_{n(j+1)}) - \mu(A_{nj}) = \int_0^{h_n} \Big[f(t_{nj} + v) - f(t_{nj} - v) \Big] \; \nu(\mathrm{d}v)\,,$$

so that similarly

$$|\mu(A_{n(j+1)}) - \mu(A_{nj})| \le 2^\alpha K \int_0^{h_n} v^\alpha \; \nu(\mathrm{d}v) = \frac{2^\alpha K h_n^{\alpha+1}}{\alpha+1}\,. \tag{2.18}$$

From (2.17), (2.14) holds if

$$\lim_{n\to\infty} h_n^{\alpha+1} \sum_{j=1}^{m_n-1} \left| \ln \frac{\mu(A_{n(j+1)})}{\mu(A_{nj})} \right| = 0.$$

By the mean value theorem, for $1 \le j \le m_n - 1$,

$$\left| \ln \frac{\mu(A_{n(j+1)})}{\mu(A_{nj})} \right| \le \frac{|\mu(A_{n(j+1)}) - \mu(A_{nj})|}{\min\{\mu(A_{n(j+1)}), \mu(A_{nj})\}},$$

so that, by (2.18),

$$\sum_{j=1}^{m_n-1} \left| \ln \frac{\mu(A_{n(j+1)})}{\mu(A_{nj})} \right| \le \frac{2^\alpha K h_n^{\alpha+1}}{\alpha+1} \sum_{j=1}^{m_n-1} \max\left\{ \frac{1}{\mu(A_{n(j+1)})}, \frac{1}{\mu(A_{nj})} \right\}.$$

Hence (2.14) holds if

$$\lim_{n\to\infty} h_n^{2\alpha+2} \sum_{j=1}^{m_n} \frac{1}{\mu(A_{nj})} = 0$$

(the summation extends over j such that $\mu(A_{nj}) > 0$). From this one can deduce sufficient conditions for (2.14). By Jensen's inequality with respect to the uniform distribution on A_{nj} for the convex function $t \mapsto 1/t$, $t > 0$,

$$\frac{h_n}{\mu(A_{nj})} \le \frac{1}{h_n} \int_{A_{nj}} \frac{1}{f} \, \mathrm{d}\nu$$

so that

$$h_n^{2\alpha+2} \sum_{j=1}^{m_n} \frac{1}{\mu(A_{nj})} \le h_n^{2\alpha} \int_{]0,1[} \frac{1}{f} \, \mathrm{d}\nu \, .$$

Therefore if $1/f$ is ν-integrable, then (2.14) holds.

2.2.2 The χ^2-divergence case

Let us now turn to the analysis of the asymptotic behavior of expression (2.9). Generally speaking, the straightforward extension of Csiszár's (1973) approximation theorem to the ϕ-divergences with $\phi(t) \neq t \ln t$ is impossible. Györfi et al. (1998) extended Csiszár's results to the χ^2-divergence under some additional conditions. They proved that, provided

(i) $\chi^2(\mu, \nu) < \infty$ and

(ii) the sequence of partitions $\{\mathcal{P}_n\}_{n\ge1}$ is (μ,ν)-approximating, then the convergence

$$\lim_{n\to\infty} \chi^2(f, f_n^*) = 0 \tag{2.19}$$

holds. Recall that the sequence of partitions $\{\mathcal{P}_n\}_{n\geq 1}$ is (μ,ν)-approximating if

(i) for every $g \in L_1(\nu)$, $\lim_{n\to\infty} \mathbf{E}_\nu(g\,/\,\mathcal{P}_n) = g$ ν-*a.s.* (here $\mathbf{E}_\nu(.\,/\,\mathcal{P}_n)$ is the conditional expectation with respect to the subfield of $\mathcal{B}\big(]0,1[\big)$ generated by $\mathcal{P}_n$) and

(ii) the random sequence $\mathbf{E}_\nu(f^2\,/\,\mathcal{P}_n)/\mathbf{E}_\nu(f\,/\,\mathcal{P}_n)$ is uniformly ν-integrable.

See Györfi et al. (1998) for more details and observe that the (μ,ν)-approximation property implies that $\lim_{n\to\infty} m_n = \infty$.

Here, we need to discuss the property

$$\lim_{n\to\infty} \chi^2(f, p_n^*) = 0. \tag{2.20}$$

We assume that (i) and (ii) hold. By using the convexity of the function $t \mapsto a/t$, $t > 0$, for $a \geq 0$, we get, in a similar way as in the information divergence case, that for every $1 \leq j \leq m_n - 1$,

$$\begin{aligned}
&\int_{A_{nj}^+\cup A_{n(j+1)}^-} \frac{f^2}{p_n^*}\,\mathrm{d}\nu \\
&\quad\leq \int_{A_{nj}^+\cup A_{n(j+1)}^-} \left[\frac{c_{n(j+1)}-u}{h_n}\frac{f^2(u)}{f_n^*(c_{nj})} + \frac{u-c_{nj}}{h_n}\frac{f^2(u)}{f_n^*(c_{n(j+1)})}\right]\nu(\mathrm{d}u) \\
&\quad= \int_{A_{nj}^+} \frac{f^2}{f_n^*(c_{nj})}\,\mathrm{d}\nu + \int_{A_{nj}^+} \frac{u-c_{nj}}{h_n} f^2(u) \frac{f_n^*(c_{nj}) - f_n^*(c_{n(j+1)})}{f_n^*(c_{nj})f_n^*(c_{n(j+1)})}\,\nu(\mathrm{d}u) \\
&\qquad+ \int_{A_{n(j+1)}^-} \frac{f^2}{f_n^*(c_{n(j+1)})}\,\mathrm{d}\nu \\
&\qquad+ \int_{A_{n(j+1)}^-} \frac{c_{n(j+1)}-u}{h_n} f^2(u) \frac{f_n^*(c_{n(j+1)}) - f_n^*(c_{nj})}{f_n^*(c_{nj})f_n^*(c_{n(j+1)})}\,\nu(\mathrm{d}u) \\
&\quad= \int_{A_{nj}^+\cup A_{n(j+1)}^-} \frac{f^2}{f_n^*}\,\mathrm{d}\nu + (d_{n(j+1)}^- - d_{nj}^+)\frac{\mu(A_{n(j+1)}) - \mu(A_{nj})}{\mu(A_{nj})\mu(A_{n(j+1)})},
\end{aligned}$$

where

$$d_{nj}^+ = \int_{A_{nj}^+} (u - c_{nj}) f^2(u)\,\nu(\mathrm{d}u) \quad \text{and} \quad d_{nj}^- = \int_{A_{nj}^-} (c_{nj} - u) f^2(u)\,\nu(\mathrm{d}u).$$

Therefore

$$\chi^2(f, p_n^*) \leq \chi^2(f, f_n^*) + \sum_{j=1}^{m_n-1} (d_{n(j+1)}^- - d_{nj}^+) \frac{\mu(A_{n(j+1)}) - \mu(A_{nj})}{\mu(A_{nj})\mu(A_{n(j+1)})}$$

so that, with (2.19), (2.20) holds if

$$\lim_{n\to\infty} \sum_{j=1}^{m_n-1} (d_{n(j+1)}^- - d_{nj}^+) \frac{\mu(A_{n(j+1)}) - \mu(A_{nj})}{\mu(A_{nj})\mu(A_{n(j+1)})} = 0. \tag{2.21}$$

Condition (2.21) is the analogue of (2.14) for the information divergence case. Since

$$\sum_{j=1}^{m_n-1} (d_{n(j+1)}^- + d_{nj}^+) \leq \int_{]0,1[} f^2 \, \mathrm{d}\nu = \chi^2(\mu, \nu) + 1 < \infty,$$

(2.21) holds if

$$\lim_{n\to\infty} \max_{1\leq j\leq m_n-1} \frac{|\mu(A_{n(j+1)}) - \mu(A_{nj})|}{\mu(A_{nj})\mu(A_{n(j+1)})} = 0.$$

Observe that

$$d_{n(j+1)}^- - d_{nj}^+ = \int_0^{h_n/2} v \left[f^2(c_{n(j+1)} - v) - f^2(c_{nj} + v) \right] \nu(\mathrm{d}v).$$

Consequently, if f^2 satisfies the Hölderian condition (2.16), one gets

$$|d_{n(j+1)}^- - d_{nj}^+| \leq \frac{K^\alpha h_n^{\alpha+2}}{4(\alpha+1)(\alpha+2)}\,.$$

At the same time, as already mentioned for the information divergence case,

$$|\mu(A_{n(j+1)}) - \mu(A_{nj})| \leq \frac{2^\alpha K h_n^{\alpha+1}}{\alpha+1}.$$

Therefore (2.21) holds if

$$\lim_{n\to\infty} h_n^{2\alpha+3} \sum_{j=1}^{m_n-1} \frac{1}{\mu(A_{nj})\mu(A_{n(j+1)})} = 0. \tag{2.22}$$

Here, by the Cauchy-Schwarz inequality,

$$\sum_{j=1}^{m_n-1} \frac{1}{\mu(A_{nj})\mu(A_{n(j+1)})} \leq \sqrt{\sum_{j=1}^{m_n-1} \frac{1}{\mu^2(A_{nj})} \sum_{j=1}^{m_n-1} \frac{1}{\mu^2(A_{n(j+1)})}}$$
$$\leq \sum_{j=1}^{m_n} \frac{1}{\mu^2(A_{nj})}\,.$$

Similarly as in the information divergence case, by Jensen's inequality,

$$\frac{h_n^2}{\mu^2(A_{nj})} \leq \frac{1}{h_n} \int_{A_{nj}} \frac{1}{f^2}\, \mathrm{d}\nu.$$

Therefore (2.22) holds if

$$\lim_{n\to\infty} h_n^{2\alpha} \int_{]0,1[} \frac{1}{f^2}\, \mathrm{d}\nu = 0,$$

i.e., if $1/f^2$ is ν-integrable or, equivalently, if $1/f \in L_2(\nu)$. We see that (2.20) is less easy to satisfy than (2.11). But this is natural, because the consistency in χ^2-divergence is stronger than in information divergence. We finally remark that the Hölder condition on f^2 follows from (2.16).

3 Consistency

In this section we use the results of previous section to establish consistency of the smooth Barron-type density estimators $f_{n,P}$ (2.4) in expected information divergence (Theorem 3.1) and in expected χ^2-divergence (Theorem 3.2). Similar results for the Barron-type density estimators $f_{n,B}$ (2.2) can be found in Barron et al. (1992) for the information divergence case and in Györfi et al. (1998) for the χ^2-divergence case.

Theorem 3.1. *Let $\mu \in \mathbb{P}_\nu$ be a probability measure on $]0,1[$ with $D(\mu,\nu) < \infty$ and with probability density $f = \mathrm{d}\mu/\mathrm{d}\nu$. If (2.14) holds, then the density estimators $f_{n,P}$ of f defined in (2.4) with*

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} nh_n = \infty \tag{3.1}$$

and

$$\limsup_{n\to\infty} \frac{1}{na_nh_n} \leq 1 \tag{3.2}$$

satisfy the relation

$$\lim_{n\to\infty} \mathbf{E}D(f, f_{n,P}) = \lim_{n\to\infty} \mathbf{E} \int_{]0,1[} f \ln \frac{f}{f_{n,P}}\, \mathrm{d}\nu = 0. \tag{3.3}$$

Proof We have

$$D(f, f_{n,P}) = \int_{]0,1[} f \ln \frac{f}{f_{n,P}}\, \mathrm{d}\nu\,,$$

and thus

$$\begin{aligned} D(f, f_{n,P}) &= \int_{]0,1[} f \ln \frac{f}{p_n^*}\, \mathrm{d}\nu + \int_{]0,1[} f \ln \frac{p_n^*}{f_{n,P}}\, \mathrm{d}\nu \\ &= D(f, p_n^*) + \int_{]0,1[} f \ln \frac{p_n^*}{f_{n,P}}\, \mathrm{d}\nu\,, \end{aligned}$$

where p_n^* is the frequency polygon on $]0,1[$ based on the true probability distribution μ $\big($see (2.7)$\big)$. From the previous section, we know that the hypotheses $D(\mu,\nu) < \infty$, $\lim_{n\to\infty} h_n = 0$ and (2.14) suffice to ensure the consistency of $D(f, p_n^*)$ towards 0. Since $D(f, f_{n,P})$ is a ϕ-divergence, $D(f, f_{n,P}) \geq 0$. Thus, for (3.3) to be true it suffices to show that

$$\limsup_{n\to\infty} \mathbf{E} \int_{]0,1[} f \ln \frac{p_n^*}{f_{n,P}}\, \mathrm{d}\nu \leq 0.$$

By Jensen's inequality for the concave function $t \mapsto \ln at$, t and $a > 0$,

$$\mathbf{E} \int_{]0,1[} f \ln \frac{p_n^*}{f_{n,P}}\, \mathrm{d}\nu \leq \int_{]0,1[} f \ln \left[p_n^* \mathbf{E}\left(\frac{1}{f_{n,P}} \right) \right] \mathrm{d}\nu. \tag{3.4}$$

Moreover

$$\begin{aligned} \mathbf{E}\left(\frac{1}{f_{n,P}} \right) &= \mathbf{E}\left(\frac{1}{(1-a_n)p_n + a_n} \right) \\ &\leq \mathbf{E}\left(\frac{1}{\min\{(1-a_n)/nh_n, a_n\}[nh_n p_n + 1]} \right), \end{aligned}$$

and as a consequence

$$\mathbf{E}\left(\frac{1}{f_{n,P}} \right) \leq \frac{1}{\min\{(1-a_n)/nh_n, a_n\}}\, \mathbf{E}\left(\frac{1}{nh_n p_n + 1} \right). \tag{3.5}$$

In order to provide an upper bound for the right hand side of (3.5), we need the following lemma and corollary. The proofs of these are deferred to the appendix.

Lemma 3.1. *If (X_1, X_2) has a multinomial distribution with parameters n, p_1, p_2, and if $\alpha \in [0,1]$, then*

$$\mathbf{E}\left(\frac{1}{\alpha(X_1+1)+(1-\alpha)(X_2+1)}\right) \le \frac{n}{n+1}\,\frac{1}{\alpha n p_1 + (1-\alpha) n p_2}\,.$$

Corollary 3.1. *Let p_n and p_n^* be defined as in (2.5) and (2.7). Then*

$$\mathbf{E}\left(\frac{1}{nh_n p_n + 1}\right) \le \frac{n}{n+1}\,\frac{1}{nh_n p_n^*}\,.$$

Thus, in accordance with Corollary 3.1 and inequality (3.5), we obtain

$$\begin{aligned}\mathbf{E}\left(\frac{1}{f_{n,P}}\right) &\le \frac{1}{\min\{(1-a_n)/nh_n, a_n\}}\,\frac{n}{n+1}\,\frac{1}{nh_n p_n^*} \\ &= \frac{1}{\min\{(1-a_n), na_nh_n\}}\,\frac{n}{n+1}\,\frac{1}{p_n^*}\,.\end{aligned}$$

This, together with (3.4), leads to

$$\mathbf{E}\int_{]0,1[} f \ln \frac{p_n^*}{f_{n,P}}\,\mathrm{d}\nu \le \ln\left(\frac{1}{\min\{(1-a_n), na_nh_n\}}\right) + \ln\left(\frac{n}{n+1}\right).$$

Now, conditions (1.10), (3.1) and (3.2) imply that

$$\limsup_{n\to\infty} \mathbf{E}\int_{]0,1[} f \ln \frac{p_n^*}{f_{n,P}}\,\mathrm{d}\nu \le \limsup_{n\to\infty} \ln\left(\frac{1}{\min\{(1-a_n), na_nh_n\}}\right) \le 0$$

and the proof is complete. ■

Theorem 3.2. *Let $\mu \in \mathbb{P}_\nu$ be a probability measure on $]0,1[$ with $\chi^2(\mu,\nu) < \infty$ and with probability density $f = \mathrm{d}\mu/\mathrm{d}\nu$. Assume that condition (2.21) is satisfied and that the sequence $\{\mathcal{P}_n\}_{n\ge 1}$ of uniform partitions on $]0,1[$ is (μ,ν)-approximating. Then the density estimators $f_{n,P}$ of f defined in (2.4) with*

$$\lim_{n\to\infty} h_n = 0, \qquad \lim_{n\to\infty} nh_n = \infty \tag{3.6}$$

and

$$\limsup_{n\to\infty} \frac{1}{na_nh_n} \le 1 \tag{3.7}$$

satisfy the relation

$$\lim_{n\to\infty} \mathbf{E}\chi^2(f, f_{n,P}) = \lim_{n\to\infty} \mathbf{E}\int_{]0,1[} \frac{f^2}{f_{n,P}}\,\mathrm{d}\nu - 1 = 0. \tag{3.8}$$

Proof By definition

$$\chi^2(f, f_{n,P}) + 1 = \int_{]0,1[} \frac{f^2}{f_{n,P}} \, \mathrm{d}\nu = \mathbf{E}_\nu\Big(\frac{f^2}{(1-a_n)p_n + a_n}\Big),$$

so that

$$\begin{aligned}
\mathbf{E}\chi^2(f, f_{n,P}) + 1 &= \mathbf{E}\,\mathbf{E}_\nu\Big(\frac{f^2}{(1-a_n)p_n + a_n}\Big) \\
&= \mathbf{E}_\nu\Big[f^2\mathbf{E}\Big(\frac{1}{(1-a_n)p_n + a_n}\Big)\Big] \\
&\le \mathbf{E}_\nu\Big[\frac{f^2}{\min\{(1-a_n)/nh_n, a_n\}}\ \mathbf{E}\Big(\frac{1}{nh_np_n+1}\Big)\Big] \\
&\le \frac{1}{\min\{(1-a_n)/nh_n, a_n\}}\ \frac{n}{n+1}\ \mathbf{E}_\nu\Big(\frac{f^2}{nh_np_n^*}\Big) \\
&= \frac{n}{n+1}\ \frac{\chi^2(f, p_n^*) + 1}{\min\{(1-a_n), na_nh_n\}},
\end{aligned}$$

where the second inequality follows from Corollary 3.1. Finally, (3.8) follows from $\chi^2(\mu, \nu) < \infty$ and (2.21) (see the results of the previous section about the consistency of $\chi^2(f, p_n^*)$ towards 0), as well as from (1.10), (3.6) and (3.7). ∎

Remark 3.1. Notice that the estimator $f_{n,P}$ defined with a_n as in (1.12) satisfies conditions (3.2) and (3.7) considered in Theorem 3.1 and Theorem 3.2.

4 Rate of convergence

In this last section we study the rate of convergence to zero of $\mathbf{E}\chi^2(f, f_{n,P})$ for the smooth Barron density estimator defined by (2.4) and (1.12). Analogous results for the asymptotic rates of the original discontinuous Barron estimator $f_{n,B}$ (see (2.2)) were presented in Györfi et al. (1998). Note that (1.8) implies $\mathbf{E}D(f, f_{n,P}) \le \mathbf{E}\chi^2(f, f_{n,P})$ so that the bounds derived in this section for $\mathbf{E}\chi^2(f, f_{n,P})$ remain valid for $\mathbf{E}D(f, f_{n,P})$ as well.

Theorem 4.1. *Let $\mu \in \mathbb{P}_\nu$ be a probability measure on $]0,1[$ with probability density $f = \mathrm{d}\mu/\mathrm{d}\nu$. Let $f_{n,P}$ be the density estimator of f defined by (2.4) and (1.12). Let p_n^* be defined as in (2.7). Then*

$$\mathbf{E}\chi^2(f, f_{n,P}) \le \chi^2(f, p_n^*) + \frac{1}{nh_n}\ \frac{n}{n+1}\ \Big(1 + \chi^2(f, p_n^*)\Big). \tag{4.1}$$

If $1/f$ is bounded above and if f is twice continuously differentiable with f' and f'' uniformly bounded then, provided

$$\lim_{n\to\infty} h_n = 0 \quad \text{and} \quad \lim_{n\to\infty} nh_n = \infty,$$

we have

$$\mathbf{E}\chi^2(f, p_n^*) = \mathrm{O}(h_n^3), \tag{4.2}$$

so that

$$\mathbf{E}\chi^2(f, f_{n,P}) = \mathrm{O}(h_n^3) + \mathrm{O}\left(\frac{1}{nh_n}\right).$$

In this case the optimal choice for h_n is $n^{-1/4}$ for which

$$\mathbf{E}\chi^2(f, f_{n,P}) = \mathrm{O}(n^{-3/4}).$$

Proof We first prove inequality (4.1). To this end, we write

$$\begin{aligned}\mathbf{E}\chi^2(f, f_{n,P}) &= \chi^2(f, p_n^*) + \mathbf{E}_\nu\left\{f^2\left[\mathbf{E}\left(\frac{1}{(1-a_n)p_n + a_n}\right) - \frac{1}{p_n^*}\right]\right\} \\ &= \chi^2(f, p_n^*) + \mathbf{E}_\nu\left\{f^2\left[\mathbf{E}\left(\frac{nh_n+1}{nh_np_n+1}\right) - \frac{1}{p_n^*}\right]\right\}.\end{aligned}$$

From Corollary 3.1, we obtain

$$\begin{aligned}\mathbf{E}\left(\frac{nh_n+1}{nh_np_n+1}\right) - \frac{1}{p_n^*} &\le \frac{n}{n+1}\,\frac{nh_n+1}{nh_np_n^*} - \frac{1}{p_n^*} \\ &\le \frac{1}{nh_n}\,\frac{n}{n+1}\,\frac{1}{p_n^*}.\end{aligned}$$

The equality

$$\frac{1}{nh_n}\,\frac{n}{n+1}\,\mathbf{E}_\nu\left(\frac{f^2}{p_n^*}\right) = \frac{1}{nh_n}\,\frac{n}{n+1}\,\left(1 + \chi^2(f, p_n^*)\right)$$

leads to the first statement of the theorem. Let us now show (4.2). To this end, write

$$p_n^*(u) = \begin{cases} \frac{\mu(A_{n1})}{h_n} & \text{if } u \in A_{n1}^- \\ \frac{\mu(A_{nj})}{h_n} + \frac{u-c_{nj}}{h_n}\left(\frac{\mu(A_{n(j+1)})}{h_n} - \frac{\mu(A_{nj})}{h_n}\right) & \text{if } u \in A_{nj}^+ \cup A_{n(j+1)}^- \\ \frac{\mu(A_{nm_n})}{h_n} & \text{if } u \in A_{nm_n}^+, \end{cases}$$

where $1 \le j \le m_n - 1$. As in Györfi et al. (1998), one proves that when f' is uniformly bounded, there exists a positive real constant Γ_1 such that

$$\int_{A_{n1}^-} \frac{(f - p_n^*)^2}{p_n^*}\,\mathrm{d}\nu \le \Gamma_1 h_n^2 \int_0^{h_n/2} \frac{1}{f}\,\mathrm{d}\nu.$$

Since $1/f$ is bounded above, this last inequality leads to

$$\int_{A_{n1}^-} \frac{(f - p_n^*)^2}{p_n^*}\,\mathrm{d}\nu = \mathrm{O}(h_n^3). \tag{4.3}$$

In a similar way, we obtain that

$$\int_{A_{nm_n}^+} \frac{(f - p_n^*)^2}{p_n^*}\,\mathrm{d}\nu = \mathrm{O}(h_n^3). \tag{4.4}$$

Now, using an exact second order Taylor expansion gives, for $1 \le j \le m_n$,

$$\begin{aligned}\frac{\mu(A_{nj})}{h_n} &\\ &= \frac{1}{h_n} \int_{c_{nj}-h_n/2}^{c_{nj}+h_n/2} \left[f(c_{nj}) + (u - c_{nj}) f'(c_{nj}) + \frac{1}{2}(u - c_{nj})^2 f''(\alpha_{nj}) \right] \nu(\mathrm{d}u)\,,\end{aligned}$$

where α_{nj} is in between u and c_{nj}. As a consequence, since f'' is uniformly bounded,

$$\frac{\mu(A_{nj})}{h_n} = f(c_{nj}) + \mathrm{O}(h_n^2),$$

uniformly for $j \in \{1, \ldots, m_n\}$. Thus, for u in $A_{nj}^+ \cup A_{n(j+1)}^-$, $1 \le j \le m_n - 1$, we obtain

$$p_n^*(u) = f(c_{nj}) + \frac{u - c_{nj}}{h_n} \left[f(c_{n(j+1)}) - f(c_{nj}) + \mathrm{O}(h_n^2) \right] + \mathrm{O}(h_n^2),$$

and thus

$$p_n^*(u) = f(c_{nj}) + \frac{u - c_{nj}}{h_n} \left[h_n f'(c_{nj}) + \mathrm{O}(h_n^2) \right] + \mathrm{O}(h_n^2). \tag{4.5}$$

Therefore, for u in $A_{nj}^+ \cup A_{n(j+1)}^-$, $1 \le j \le m_n - 1$,

$$\begin{aligned} f(u) - p_n^*(u) &= (u - c_{nj}) f'(c_{nj}) + \frac{u - c_{nj}}{h_n} \left[h_n f'(c_{nj}) + \mathrm{O}(h_n^2) \right] + \mathrm{O}(h_n^2) \\ &= \mathrm{O}(h_n^2). \end{aligned} \tag{4.6}$$

Since f' is uniformly bounded, it is clear from (4.5) that for u in $A_{nj}^{+} \cup A_{n(j+1)}^{-}$,

$$\frac{1}{p_n^*(u)} = \frac{1}{f(c_{nj}) + \mathrm{o}(1)} \quad \text{as } n \to \infty,$$

uniformly for $j \in \{1, \ldots, m_n\}$. Consequently, thanks to (4.6), we obtain

$$\sum_{j=1}^{m_n - 1} \int_{A_{nj}^{+} \cup A_{n(j+1)}^{-}} \frac{\left(f - p_n^*\right)^2}{p_n^*} \, \mathrm{d}\nu \leq \Gamma_2 h_n^3 \sum_{j=1}^{m_n - 1} \frac{h_n}{f(c_{nj}) + \mathrm{o}(1)}$$

for some positive real constant Γ_2. Denoting by ζ an upper bound for $1/f$ leads, for large enough n, to

$$0 \leq \sum_{j=1}^{m_n - 1} \frac{h_n}{f(c_{nj}) + \mathrm{o}(1)} \leq \sum_{j=1}^{m_n - 1} \frac{h_n}{f(c_{nj}) - \zeta^{-1}/2}.$$

Consequently, for large enough n,

$$0 \leq \sum_{j=1}^{m_n - 1} \frac{h_n}{f(c_{nj}) + \mathrm{o}(1)} \leq 2\zeta.$$

Therefore

$$\sum_{j=1}^{m_n - 1} \int_{A_{nj}^{+} \cup A_{n(j+1)}^{-}} \frac{\left(f - p_n^*\right)^2}{p_n^*} \, \mathrm{d}\nu = \mathrm{O}(h_n^3). \tag{4.7}$$

Inequality (4.7), together with (4.3) and (4.4), leads to the desired result. ■

Remark 4.1. Similarly as in Györfi et al. (1998), the convergence of $\mathbf{E}\chi^2(f, f_{n,P})$ holds without the assumption considered in Theorem 3.2 that the partitions are (μ, ν)-approximating.

Remark 4.2. In comparison with the results obtained in Györfi et al. (1998), we showed that the smooth Barron estimator possesses a faster rate of convergence than the histogram-based Barron estimator. However the conditions for such a result are stronger.

Remark 4.3. From the proof of Theorem 4.1, it can be seen that the rate of convergence of $\mathbf{E}\chi^2(f, f_{n,P})$ is governed by the rate of the histogram part near the endpoints 0 and 1 $\big($see (4.3) and $(4.4)\big)$.

5 Appendix

5.1 Proof of Lemma 3.1

We have

$$\begin{aligned}
&\mathbf{E}\left(\frac{1}{\alpha(X_1+1)+(1-\alpha)(X_2+1)}\right) \\
&\quad= \sum_{\substack{0\le k_1,k_2\le n\\ k_1+k_2\le n}} \frac{1}{\alpha(k_1+1)+(1-\alpha)(k_2+1)}\ \frac{n!}{k_1!k_2!(n-k_1-k_2)!} \\
&\qquad\qquad \times p_1^{k_1}p_2^{k_2}(1-p_1-p_2)^{n-k_1-k_2} \\
&\quad= \frac{1}{p_1p_2(n+1)(n+2)} \sum_{\substack{1\le l_1,l_2\le n+1\\ l_1+l_2\le n+2}} \frac{l_1l_2}{\alpha l_1+(1-\alpha)l_2}\ \frac{(n+2)!}{l_1!l_2!(n+2-l_1-l_2)!} \\
&\qquad\qquad\qquad \times p_1^{l_1}p_2^{l_2}(1-p_1-p_2)^{n+2-l_1-l_2}\,.
\end{aligned}$$

Consider now the continuous function $f:\mathbb{R}^+\times\mathbb{R}^+\to\mathbb{R}$, defined by

$$f(x,y)=\frac{xy}{\alpha x+(1-\alpha)y}\,,$$

setting $f(0,0)=0$. We have

$$\mathbf{E}\left(\frac{1}{\alpha(X_1+1)+(1-\alpha)(X_2+1)}\right)=\frac{1}{p_1p_2(n+1)(n+2)}\mathbf{E}\Big(f(Y_1,Y_2)\Big), \tag{5.1}$$

where (Y_1,Y_2) follows a multinomial distribution with parameters $n+2, p_1, p_2$. The function f is twice continuously differentiable on $\mathbb{R}^{+*}\times\mathbb{R}^{+*}$. Straightforward calculations show that the Hessian matrix of f at the point $(x,y)\in\mathbb{R}^{+*}\times\mathbb{R}^{+*}$ is given by

$$\begin{pmatrix} \dfrac{\partial^2 f}{\partial^2 x} & \dfrac{\partial^2 f}{\partial x\partial y} \\ \dfrac{\partial^2 f}{\partial x\partial y} & \dfrac{\partial^2 f}{\partial^2 y}\end{pmatrix}(x,y) \;=\; \frac{2\alpha(1-\alpha)}{\{\alpha x+(1-\alpha)y\}^3}\begin{pmatrix} -y^2 & xy \\ xy & -x^2\end{pmatrix}.$$

Thus, the function f is seen to be negative semi-definite on $\mathbb{R}^{+*}\times\mathbb{R}^{+*}$ and it is therefore concave (by continuity) on $\mathbb{R}^+\times\mathbb{R}^+$ (Rockafellar 1970).

Applying Jensen's inequality in (5.1) leads to

$$\begin{aligned}
&\mathbf{E}\left(\frac{1}{\alpha(X_1+1)+(1-\alpha)(X_2+1)}\right)\\
&\quad\le \frac{1}{p_1p_2(n+1)(n+2)}\frac{(n+2)^2p_1p_2}{(n+2)\alpha p_1+(n+2)(1-\alpha)p_2},
\end{aligned}$$

i.e.,

$$\mathbf{E}\left(\frac{1}{\alpha(X_1+1)+(1-\alpha)(X_2+1)}\right)\le\frac{n}{n+1}\,\frac{1}{\alpha np_1+(1-\alpha)np_2}.$$

5.2 Proof of Corollary 3.1

From (2.5) and (2.7), we easily obtain

$$nh_np_n(u)+1=$$

$$\begin{cases} n\mu_n(A_{n1})+1 & \text{if } u\in A_{n1}^- \\ (1-\frac{u-c_{nj}}{h_n})\big(n\mu_n(A_{nj})+1\big)+\frac{u-c_{nj}}{h_n}\big(n\mu_n(A_{n(j+1)})+1\big) & \text{if } u\in A_{nj}^+\cup A_{n(j+1)}^- \\ n\mu_n(A_{nm_n})+1 & \text{if } u\in A_{nm_n}^+, \end{cases}$$

and

$$nh_np_n^*(u)=$$

$$\begin{cases} n\mu(A_{n1}) & \text{if } u\in A_{n1}^- \\ (1-\frac{u-c_{nj}}{h_n})n\mu(A_{nj})+\frac{u-c_{nj}}{h_n}n\mu(A_{n(j+1)}) & \text{if } u\in A_{nj}^+\cup A_{n(j+1)}^- \\ n\mu(A_{nm_n}) & \text{if } u\in A_{nm_n}^+, \end{cases}$$

where the current index j runs from 1 to m_n-1. Applying Lemma 3.1 to the cells $A_{n1},\dots,A_{nm_n}$ leads to the desired result.

Acknowledgments

The authors thank an anonymous referee for his helpful suggestions.

References

Barron, A.R. (1988). The convergence in information of probability density estimators. *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe, Japan, June 19–24.

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.

Beirlant, J., Berlinet, A. and Györfi, L. (1998). On piecewise linear density estimates. *Statistica Neerlandica*, **53**, 287–308.

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.

Csiszár, I. (1973). Generalized entropy and quantization problems. *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, Academia, Prague, 159–174.

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

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

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner–Verlag, Leipzig.

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton University Press, Princeton.

Scott, D.W. (1985). Frequency polygons: theory and application. *Journal of the American Statistical Association*, **80**, 348–354.

Scott, D.W. (1992). *Multivariate Density Estimation – Theory, Practice, and Visualization*. Wiley, New York.

Vajda, I. and van der Meulen, E.C. (2001). Optimization of Barron density estimates. *IEEE Transactions on Information Theory*, **47**, 1867–1883.