November 3, 2008

Divergence criteria for improved selection rules

A. Berlinet[1] and I. Vajda[2]

Abstract

At the basis of combinatorial methods in density estimation introduced by Devroye and Lugosi is the so-called Scheffé selection rule. We show by an examples that this rule based on L_1 errors may not bring the selection closer to optimality than tossing of a coin. As in any estimation problem, the choice of a criterion is at the heart of the matter. The optimality of the Scheffé estimate is perceived differently by different ϕ–divergence criteria. We show that the L_1 oracle inequality satisfied by the Scheffé estimate can be extended to ϕ–divergences. It can be also extended to estimates associated with selection rules based on ϕ–divergences. As the L_1 rule, the new rules are applicable to any selection problem in density estimation.

AMS 1991 subject classification: 62 G 05.

Key Words: Density estimation. Nonparametric estimation. Selection of estimates. Information divergence. Optimality. Combinatorial methods.

1 Introduction and basic concepts

In the book of Devroye and Lugosi (2001), the authors considered the statistical model with $\mathbb{R}^d$-valued observations $X_1, \ldots, X_n$ i.i.d. by a probability density f on $\mathbb{R}^d$ and two estimates

$$f_n^{(i)} = f_n^{(i)}(\cdot; X_1, \ldots, X_n), \quad i \in \{1, 2\} \tag{1}$$

of this density. They were interested in the problem how to select for each realization of $X_1, \ldots, X_n$ the better of these estimates in the sense of L_1-error. Obviously, the optimal but practically unachievable selection is

$$f_n^{(0)} = \begin{cases} f_n^{(1)} & \text{if } \displaystyle\int |f_n^{(1)} - f| < \int |f_n^{(2)} - f|, \\ f_n^{(2)} & \text{otherwise.} \end{cases} \tag{2}$$

They proposed a practically achievable approximation to this selection called *Scheffé estimate* which is selected by by the rule

$$f_n^{*} = \begin{cases} f_n^{(1)} & \text{if } \left| \displaystyle\int_{A_n} f_n^{(1)} - \mu_n(A_n) \right| < \left| \displaystyle\int_{A_n} f_n^{(2)} - \mu_n(A_n) \right|, \\ f_n^{(2)} & \text{otherwise} \end{cases} \tag{3}$$

[1] I3M, UMR CNRS 5149, University of Montpellier II, 34095 Montpellier Cedex, France

[2] Institute of Information Theory and Automation, ASCR, 182 08 Prague, Czech Republic

where

$$A_n = A\left(f_n^{(1)}; f_n^{(2)}\right) = \left\{x : f_n^{(1)}(x) > f_n^{(2)}(x)\right\} \tag{4}$$

is the so-called *Scheffé set* for $f_n^{(1)}$, $f_n^{(2)}$ and

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}\left(X_i \in A\right), \quad A \in \mathcal{B}^d, \tag{5}$$

is the empirical probability measure on the σ-algebra of Borel sets $\mathcal{B}^d$. Chapter 6 of the cited book contains a number of arguments in favour of the Scheffé selection rule (3). However, the next example demonstrates that the favorization of the Scheffé rule is problematic in some cases. As above, $\boldsymbol{I}\left(\cdot\right)$ denotes the indicator function.

Example 1. Consider f uniform on the closed interval $[0,1] \subset \mathbb{R}$ and the corresponding ordered sample $X_{n:1}, \ldots, X_{n:n}$. For the estimates

$$f_n^{(1)} = \boldsymbol{I}(X_{n:1} \le x \le X_{n:1}+1) \text{ and } f_n^{(2)} = \boldsymbol{I}(X_{n:n}-1 \le x \le X_{n:n})$$

of f we get

$$A_n = (X_{n:n},\ X_{n:1}+1\ |, \quad \mu_n(A_n) = 0$$

$$\int_{A_n} f_n^{(1)} = X_{n:1}+1-X_{n:n} \text{ and } \int_{A_n} f_n^{(2)} \ = \ 0$$

so that

$$\left|\int_{A_n} f_n^{(1)} - \mu_n(A_n)\right| = |X_{n:1}+1-X_{n:n}| > X_{n:1}$$

exceeds with probability 1 the absolute deviation

$$\left|\int_{A_n} f_n^{(2)} - \mu_n(A_n)\right| = 0.$$

Consequently the Scheffé rule selects the estimate $f_n^{(2)}$ achieving the L_1-error $\int |f_n^{(2)} - f| = 2\,(1-X_{n:n})$ whereas the estimate $f_n^{(1)}$ achieves the error $\int |f_n^{(1)} - f| = 2X_{n:1}$, so that is strictly better in the L_1-sense with the probability $\Pr(X_{n:1}+X_{n:n} < 1) = 1/2$.

The book of Devroye and Lugosi (2001) presents a systematic theory dealing with properties and applications of the Scheffé selection f_n^*. This theory is based on Theorem 6.1 which compares the errors

$$\int |f_n^{(0)} - f| \ = \min\left\{\int |f_n^{(1)} - f|, \int |f_n^{(2)} - f|\right\} \quad \text{and} \quad \int |f_n^* - f|.$$

This fundamental theorem can be given the form of the inequality

$$\int |f_n^* - f| \le 3\int |f_n^{(0)} - f| + 4\left|\int_{A_n} f - \mu_n(A_n)\right| \tag{6}$$

where A_n is the Scheffé set for $f_n^{(1)}$, $f_n^{(2)}$. This inequality states that the selection f_n^* can achieve the error level $3\int |f_n^{(0)} - f|$ up to the universal error term appearing on the right. This inequality was applied not only in Chapters 7 − 17 of the Devroye-Lugosi book, but also in subsequent papers, among them in Berlinet, Biau and Rouviere (2005 a,b).

The latter papers observed that the L_1-error criterion $\int |f - g|$ for the estimates g being formally probability densities is a special case of the more general ϕ-divergence criterion $D_\phi(f, g)$ defined for arbitrary probability densities f, g by the formula

$$D_\phi(f, g) = \int g\,\phi\left(\frac{f}{g}\right). \tag{7}$$

Here $\phi(t)$ is nonnegative and convex in the domain $t \in (0, \infty)$, strictly convex and vanishing at the point $t = 1$ (for details about formula (7) and the basic properties of ϕ-divergences used below, see Csiszár (1967a) or Liese and Vajda (1987, 2006).

The L_1-error is the ϕ-divergence for $\phi(t) = |t - 1|$, called *total variation* and denoted by $V(f, g)$, i. e.

$$V(f, g) = \int |f - g| = 2 \sup_{A \in \mathcal{B}^d} \left| \int_A f - \int_A g \right|. \tag{8}$$

Other examples are the *squared Hellinger distance*

$$H^2(f, g) = 2 \int \left(\sqrt{f} - \sqrt{g}\right)^2 \quad \text{for } \phi(t) = 2\left(\sqrt{t} - 1\right)^2, \tag{9}$$

the *squared Le Cam distance*

$$LC^2(f, g) = \frac{1}{2} \int \frac{(f - g)^2}{f + g} \quad \text{for } \phi(t) = \frac{(t - 1)^2}{2(t + 1)}, \tag{10}$$

and the *information divergence*

$$I(f, g) = \int f \ln \frac{f}{g} \quad \text{for } \phi(t) = t \ln t. \tag{11}$$

Natural motivation for the alternative ϕ-divergence error criteria is the need to work with estimates converegent in the topologies stronger than that induced by the total variation (cf. Csiszár 1967b and Österreicher and Vajda (2003). This paper introduces a new motivation achieved in Example 3 below by extending the framework of Example 1 through admitting non-uniform densities with unit supports on $\mathbb{R}$. In this extended setting Example 3 demonstrates that for some densities f the alternative ϕ-divergence error criteria exhibit with positive probabilities optimality of the estimate $g = f^{(1)}$ at the same time when the L_1-error exhibits the optimality of $g = f^{(2)}$.

Since the optimality of the Scheffé estimates f_n^* is perceived differently by different ϕ-divergence error criteria, it is important to see whether or how the fundamental Devroye–Lugosi inequality (6) can be extended from the total variation criteria

$$V(f, f_n^*) = \int |f_n^* - f| \quad \text{and} \quad V(f, f_n^{(0)}) = \int |f_n^{(0)} - f| \tag{12}$$

to the more general ϕ-divergence criteria

$$D_\phi(f, f_n^*) = \int f_n^* \, \phi\left(\frac{f}{f_n^*}\right) \quad \text{and} \quad D_\phi(f, f_n^{(0)}) = \int f_n^{(0)} \, \phi\left(\frac{f}{f_n^{(0)}}\right). \tag{13}$$

This problem is solved in Section 3.

Section 4 introduces a replacement of the Scheffé L_1-based selection rule by a more general ϕ-divergence selection rule and solves a problem parallel to that of Section 3, namely whether or how the Devroye–Lugosi inequality (6) can be extended to these estimates and to the more general ϕ-divergence criteria.

For the obvious reasons, in this paper the attention is restricted to the estimates (1) which are a.s. probability densities themselves.

2 Metric divergence criteria of errors

Let us start with the following basic properties of ϕ-divergences needed in the sequel:

(i) The *range of values* is

$$0 \le D_\phi(f, g) \le \phi(0) + \phi^*(0) \tag{14}$$

where $\phi(0)$, $\phi^*(0)$ are smooth extensions of $\phi(t)$, $\phi^*(t) = t\,\phi(1/t)$ to the point $t = 0$. In (14) $D_\phi(f, g) = 0$ if and only if $f = g$ a.s. and $D_\phi(f, g) = \phi(0) + \phi^*(0)$ if (for finite $\phi(0) + \phi^*(0)$ if and only if) $f \perp g$ (disjoint supports).

(ii) The *symmetry* $D_\phi(f, g) = D_\phi(g, f)$ for all f, g holds if and only if $\phi = \phi^*$ fot the adjoint function ϕ^* defined in **(i)**.

(iii) The *monotonicity property* deals with relations between ϕ-divergences

$$D_\phi(\mu, \nu) \equiv D_\phi(f, g)$$

of the distributions

$$\mu(A) = \int_A f, \quad \nu(A) = \int_A g, \quad A \in \mathcal{B}^d$$

and ϕ-divergences of restrictions of these distributions on sub-σ-algebras $\mathcal{S} \subset \mathcal{B}^d$ of the Borel σ-algebra $\mathcal{B}^d$ defined by formula

$$D_\phi(\mu, \nu | \mathcal{S}) = D_\phi(f_\mathcal{S}, g_\mathcal{S}) = \int g_\mathcal{S} \, \phi\left(\frac{f_\mathcal{S}}{g_\mathcal{S}}\right)$$

for $\mathcal{S}$-measurable versions $f_\mathcal{S}$, $g_\mathcal{S}$ of densities f, g. It states that the ordering

$$D_\phi(f, g | \mathcal{S}) \equiv D_\phi(\mu, \nu | \mathcal{S}) \le D_\phi(\mu, \nu) \equiv D_\phi(f, g) \tag{15}$$

holds. If the equality in (15) takes place then we say that $\mathcal{S}$ *preserves the ϕ-divergence* $D_\phi(f, g)$. It is known (see e.g. Corollary 1.29 in Liese and Vajda (1987)) that if a sub-σ-algebra $\mathcal{S}$ is sufficient for the pair $\{f, g\}$ then the equality takes place in (15), i.e. the sufficient $\mathcal{S}$ allways preserves the ϕ-divergence $D_\phi(f,g)$.

(iv) Finally, the *spectral representation* says that if a sub-σ-algebra $\mathcal{S} \subset \mathcal{B}^d$ is generated by a finite or countable $\mathcal{B}^d$-measurable partition $\mathcal{P}$ of $\mathbb{R}^d$ (spectrum of $\mathcal{S}$, in symbols we write $\mathcal{S} = \mathcal{S}(\mathcal{P})$) then

$$D_\phi(f,g|\mathcal{S}) = \sum_{A\in\mathcal{P}} \int_A g\,.\phi\left(\frac{\int_A f}{\int_A g}\right). \tag{16}$$

Example 2. Consider for every $A \in \mathcal{B}^d$ the partition $\mathcal{P} = (A, A^c)$ of $\mathbb{R}^d$ and the $\mathcal{P}$-generated (or, more simply, A-generated) algebra

$$\mathcal{S}_A := \mathcal{S}\left(A, A^c\right) \subset \mathcal{B}^d \tag{17}$$

consisting of the sets $\mathbb{R}^d, A, A^c, \emptyset$. Then the general spectral representation (16) implies

$$V(f,g|\mathcal{S}_A) = \sum_{B\in\{A,A^c\}} \left|\int_B f - \int_B g\right| = 2\left|\int_A f - \int_A g\right|. \tag{18}$$

From (8) and (18) we see that the fundamental Devroye–Lugosi inequality (6) can be given the form

$$V(f,f_n^*) \le 3V(f,f_n^{(0)}) + 2V\left(\mu,\mu_n|\mathcal{S}_{A_n}\right) \tag{19}$$

for the Scheffé set A_n of the estimates $f_n^{(1)}$ and $f_n^{(2)}$.

If A in (18) is the Scheffé set $A(f;g)$ of f and g then the absolute difference on the right of (18) can be replaced by the ordinary difference. Moreover, it is seen from (8) that then $\mathcal{S}_A$ preserves $V(f,g)$ so that the formula (18) can be extended and specified as follows

$$V\left(f,g|\mathcal{S}_A\right) = V(f,g) = 2\left(\int_A f - \int_A g\right). \tag{20}$$

The following sections extend the Devroye–Lugosi theorem (6), or equivalently (19), to the error criteria $D(f,g)$ for probability densities f, g on $(\mathbb{R}^d, \mathcal{B}^d)$ satisfying similar metric properties as the total variation criterion $V(f,g)$ namely

the *reflexivity*

$$D(f,g) = 0 \quad \text{if and only if } f = g \text{ a.s.}, \tag{21}$$

the *symmetry*

$$D(f,g) = D(g,f) \quad \text{for all } f,\ g \tag{22}$$

and the *triangle inequality*

$$D(f,g) \le D(f,h) + D(h,g) \quad \text{for all } f,\ g,\ h. \tag{23}$$

We restrict ourselves to the *metric divergence criteria* defined as powers

$$D(f,g) = D_\phi(f,g)^\pi, \quad \pi > 0$$

of ϕ-divergences $D_\phi(f,g)$ satisfying (21)-(23). Theseϕ-divergences achieve finite uppr bounds

$$\phi(0) + \phi^*(0) = 2\phi(0) < \infty \tag{24}$$

(see **(ii)** for the equality and Csiszár (1967b) for the finiteness).

To provide a sufficiently rich class of such criteria, let us introduce the class of ϕ_α-divergences

$$\mathcal{D}_\alpha(f,g) = D_{\phi_\alpha}(f,g), \quad \alpha \in \mathbb{R}. \tag{25}$$

Here the convex functions $\phi_\alpha(t)$ are given in the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{\mid \alpha \mid}{\alpha(\alpha-1)} \left(2^{\alpha-1}(t+1) - (t^{1/\alpha}+1)^\alpha\right) \tag{26}$$

if $\alpha(\alpha-1) \neq 0$, and by the corresponding limits

$$\phi_0(t) = \ \mid t-1 \mid /2, \tag{27}$$

$$\phi_1(t) = t\ln t + (t+1)\ln\frac{2}{t+1} \tag{28}$$

otherwise. The subclass of these divergences for $\alpha \geq 0$ was proposed (with a different parametrization) by Österreicher and Vajda (2003). The extension to $\alpha < 0$ was proposed recently by Vajda (2008). It is easy to verify for all f, g the formulas

$$\mathcal{D}_0(f,g) = \frac{1}{2}V(f,g) \qquad \text{(total variation, (8))}, \tag{29}$$

$$\mathcal{D}_2(f,g) = \frac{1}{2}H^2(f,g) \qquad \text{(Hellinger, (9))}, \tag{30}$$

$$\mathcal{D}_{-1}(f,g) = \frac{1}{4}LC^2(f,g) \quad \text{(Le Cam, (10))} \tag{31}$$

and

$$\mathcal{D}_1(f,g) \ = \ I\left(f,(f+g)/2\right) + I(g,(f+g)/2). \tag{32}$$

In the Appendix we demonstrate that the powers

$$D(f,g) := \mathcal{D}_\alpha(f,g)^{1/\max\{2,\alpha\}}, \quad \alpha \in \mathbb{R} \tag{33}$$

of the divergences (25) satisfy (21) – (23), i. e. that they are metric divergence criteria.

3 Scheffé selection rule

This section extends the fundamental Devroye–Lugosi inequality (6) for the Scheffé estimates f_n^* from the total variation error criteria (12) to the more general ϕ-divergence criteria (13). The next example provides a motivation for this extension.

Example 3. Main result of this section is the following theorem. This theorem and its proof refer to the lower and upper error bounds

$$L_\phi(V) \le D_\phi(f,g) \le U_\phi(V) \tag{34}$$

achieved for a given convex ϕ by the ϕ-divergences $D_\phi(f,g)$ on the class of densities f, g satisying the total variation condition

$$V(f,g) = V, \quad 0 \le V \le 2.$$

By Proposition 8.27 in Liese and Vajda (1987), the upper bound is for general ϕ given by the formula

$$U_\phi(V) = V \cdot c_\phi \quad \text{where} \quad c_\phi = \frac{\phi(0) + \phi^*(0)}{2} \quad \text{(cf. (14))} \tag{35}$$

and the lower bound $L_\phi(V)$ is convex and strictly increasing in the variable V from the minimum $L_\phi(0) = 0$ to the maximum $L_\phi(2) = \phi(0) + \phi^*(0) = 2c_\phi$. Hence the strictly increasing and concave inverse function

$$L_\phi^{-1}(D) : [0, 2c_\phi] \longrightarrow [0,2] \tag{36}$$

allways exists. For ϕ such that the powers $D_\phi(f,g)^\pi$ are metrics on the space of densities f, g (24) implies

$$c_\phi = \phi(0) < \infty. \tag{37}$$

Theorem 1. Let f be an estimated distribution on $\mathbb{R}^d$, $f_n^{(0)}$, $f_n^{(1)}$and $f_n^{(2)}$ the estimates considered in (1), (2) with the corresponding Scheffé set A_n and f_n^* the Scheffé estimate resulting from the selection rule (3). Then for every metric divergence criterion $D(f,g) = D_\phi(f,g)^\pi$

$$D\left(f_n^*, f\right) \le D\left(f_n^{(0)}, f\right) + 2^\pi c_\phi^\pi \left[L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi} \right) + 2 \left| \int_{A_n} f - \mu_n(A_n) \right| \right]^\pi \tag{38}$$

where L_ϕ^{-1} and c_ϕ are given by (36) and (37)

Proof. Consider the random variables

$$\mathcal{E}_{ij} = \boldsymbol{I}\left(f_n^* = f_n^{(i)},\, f_n^{(0)} = f_n^{(j)}\right) \text{ where } \sum_{i,j=1}^{2} \mathcal{E}_{ij} = 1. \tag{39}$$

By the triangle inequality and symmetry of $D(f,g)$, and by the definition of $\mathcal{E}_{ii}$,

$$\begin{aligned} D\left(f_n^*, f\right) &\le D\left(f_n^{(0)}, f\right) + \sum_{i,j=1}^{2} D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{ij} \\ &= D\left(f_n^{(0)}, f\right) + D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{21} + D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{12}. \end{aligned} \tag{40}$$

It suffices to prove that for $i \ne j$

$$D\left(f_n^*, f_n^{(0)}\right) \mathcal{E}_{ij} \le 2^\pi c_\phi^\pi \left[L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi} \right) + 2 \left| \int_{A_n} f - \mu_n(A_n) \right| \right]^\pi \mathcal{E}_{ij}. \tag{41}$$

We restrict ourselves to $\mathcal{E}_{21}$. For $\mathcal{E}_{12}$ the proof is similar. By the definition of $\mathcal{E}_{21}$ and (35), (36),

$$\begin{aligned} D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{21} &= D\left(f_n^{(1)}, f_n^{(2)}\right)\mathcal{E}_{21} \le \left[c_\phi V\left(f_n^{(1)}, f_n^{(2)}\right)\right]^\pi \mathcal{E}_{21} \\ &= c_\phi^\pi V\left(f_n^{(1)}, f_n^{(2)}|\mathcal{S}_{A_n}\right)^\pi \mathcal{E}_{21} \\ &\le c_\phi^\pi \left[V\left(f_n^{(1)}, f|\mathcal{S}_{A_n}\right) + V\left(f_n^{(2)}, f|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21} \\ &\le c_\phi^\pi \left[V\left(f_n^{(1)}, f\right) + V\left(\mu_n^{(2)}, \mu_n|\mathcal{S}_{A_n}\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21} \\ &\le 2^\pi c_\phi^\pi \left[V\left(f_n^{(0)}, f\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21} \\ &\le 2^\pi c_\phi^\pi \left[L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi}\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right)\right]^\pi \mathcal{E}_{21}. \end{aligned}$$

where we bounded the sum of the total variations in the third line above by

$$\begin{aligned} &V\left(f_n^{(1)}, f\right) + V\left(\mu_n^{(1)}, \mu_n|\mathcal{S}_{A_n}\right) + V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right) \\ &\le V\left(f_n^{(1)}, f\right) + V\left(\mu_n^{(1)}, \mu|\mathcal{S}_{A_n}\right) + 2V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right) \\ &\le 2V\left(f_n^{(1)}, f\right) + 2V\left(\mu_n, \mu|\mathcal{S}_{A_n}\right). \end{aligned}$$

This completes the proof.

The next corollary reformulates the result of Theorem 1 in a simpler but slightly weaker form.

Corollary 1. For $0 < \pi \le 1$, under the assumptions and notations of Theorem 1,

$$D\left(f_n^*, f\right) \le 2^{1-\pi} c_\phi^\pi \left[3L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi}\right) + 4\left|\int_{A_n} f - \mu_n(A_n)\right|\right]^\pi \tag{42}$$

Proof. Clear from (38) by taking into account the inequalities

$$D\left(f_n^{(0)}, f\right) \le U_\phi\left(V\left(f_n^{(0)}, f\right)\right)^\pi = \left[c_\phi V\left(f_n^{(0)}, f\right)\right]^\pi \le \left[c_\phi L_\phi^{-1}\left(D\left(f_n^{(0)}, f\right)^{1/\pi}\right)\right]^\pi$$

obtained from (34), (36) and also the inequality

$$\psi_\pi(a) + \psi_\pi(b) \le 2^{1-\pi}\psi_\pi(a+b)$$

obtained from Jensen's inequality for the concave function $\psi_\pi(x) = x^\pi$.

The next example demonstrates that Theorem 1 generalizes the Devroye and Lugosi inequality (6).

Example 4. Put $D(f,g) = \mathcal{D}_0(f,g) = V(f,g)/2$ (cf. (29)). Then $c_0 = \phi_0(0) = 1/2$,

$$L_0(V) = U_0(V) = \frac{V}{2}, \quad 0 \le V \le 2$$

and $L_0^{-1}(D) = 2D$. Hence Theorem 1 implies

$$\mathcal{D}_0\left(f_n^*, f\right) \le \mathcal{D}_0\left(f_n^{(0)}, f\right) + 2\mathcal{D}_0\left(f_n^{(0)}, f\right) + 2\left|\int_{A_n} f - \mu_n(A_n)\right|$$

or, equivalently,

$$V\left(f_n^*; f\right) \le 3V\left(f_n^{(0)}, f\right) + 4\left|\int_{A_n} f - \mu_n(A_n)\right|$$

which coincides with (6) and (19).

The next example illustrates contributions of Theorem 1 and its Corollary 1 beyond the framework of Devroye and Lugosi.

Example 5. Put $D(f,g) = \mathcal{D}_{-1}(f,g)^{1/2}$, i.e. take the Le Cam error criterion $LC(f,g)/2$ (cf. (31)). Then parts (ii) and (iii) of Theorem A1 in the Appendix imply that $c_{-1} = 1/8$, $U_{-1}(V) = V/16$ and

$$\begin{aligned} L_{-1}(V) &= \frac{1}{2}\left(\frac{1}{2} - \left[\frac{1}{1+V/2} + \frac{1}{1-V/2}\right]^{-1}\right) \\ &= \frac{1}{2}\left[\frac{1}{2} - \frac{1-(V/2)^2}{2}\right] = \left(\frac{V}{4}\right)^2. \end{aligned}$$

Therefore $L_{-1}^{-1}(D) = 4\sqrt{D}$ and for the Scheffé selection f_n^* of Devroye and Lugosi we get from Theorem 1 the relation

$$D\left(f_n^*, f\right) \le D\left(f_n^{(0)}, f\right) + \left[\frac{2}{8}\left(4D\left(f_n^{(0)}, f\right) + 2\left|\int_{A_n} f - \mu_n(A_n)\right|\right)\right]^{1/2}$$

i. e.

$$LC\left(f_n^*, f\right) \le LC\left(f_n^{(0)}, f\right) + \sqrt{\frac{1}{2}LC\left(f_n^{(0)}, f\right) + \frac{1}{8}\left|f_n - \mu_n(A_n)\right|}$$

where A_n is the Scheffé set of the initial estimates $f_n^{(1)}$ and $f_n^{(2)}$. Corollary 1 implies for the same f_n^* and A_n as before the alternative inequality

$$D\left(f_n^*, f\right) \le \left(2\,\frac{3}{8}\,4D\left(f_n^{(0)}, f\right) + 2\,\frac{4}{8}\left|\int_{A_n} f - \mu_n(A_n)\right|\right)^{1/2}$$

i. e.

$$LC\left(f_n^*, f\right) \le \sqrt{\frac{3}{2}\,LC\left(f_n^{(0)}, f\right) + \frac{1}{4}\left|\int_{A_n} f - \mu_n(A_n)\right|}.$$

We see that the rate of convergence of the Le Cam error $LC\left(f_n^*, f\right)$ to zero garanteed by our theory for the Scheffé estimate is strictly below the rate of the Le Cam error $LC\left(f_n^{(0)}, f\right)$ achieved by the ideal estimate $f_n^{(0)}$. One can deduce from the known properties of the lower bound $L_\phi(V)$ and its inverse $L_\phi^{-1}(D)$ that similar result can be expected also for other divergence errors $D_\phi\left(f_n^*, f\right)$ with ϕ strictly convex everywhere.

4 Divergence selection rule

This section is a continuation of Section 3 where the estimation errors are still evaluated by metric divergence criteria of the type $D(f,g) = D_\phi(f,g)^\pi, \pi > 0$ but the Scheffé selection (3) of Devroye and Lugosi (2001) is replaced by a more general selection. One arrives quite naturally at such generalization if he applies the same criteria also to the definition of the optimal estimate $f_n^{(0)}$ and its practical approximation f_n^*. In other words, the generalization consists in the replacement of the L_1-based definition (2) by the divergence based definition

$$f_n^{(0)} = \begin{cases} f_n^{(1)} & \text{if } D\left(f_n^{(1)}, f\right) < D\left(f_n^{(2)}, f\right) \\ f_n^{(2)} & \text{otherwise.} \end{cases} \tag{43}$$

and the L_1-based Scheffé selection rule (3) by the *divergence selection rule*

$$f_n^* = \begin{cases} f_n^{(1)} & \text{if } D\left(\mu_n^{(1)}, \mu_n|\mathcal{S}_n\right) < D\left(\mu_n^{(2)}, \mu_n|\mathcal{S}_n\right) \\ f_n^{(2)} & \text{otherwise.} \end{cases} \tag{44}$$

This rule uses the empirical distribution μ_n defined by (5), the estimates

$$\mu_n^{(i)}(B) = \int_E f_n^{(i)}, \quad B \in \mathcal{B}^d,\ i \in \{1,2\}$$

of the probability distribution $\mu \sim f$, and the sub-σ-algebra $\mathcal{S}_n \subset \mathcal{B}^d$ preserving the divergence $D\left(\mu_n^{(1)}, \mu_n^{(2)}\right)$, i.e. satisfying the equality

$$D\left(\mu_n^{(1)}, \mu_n^{(2)}\right) = D\left(\mu_n^{(1)}, \mu_n^{(2)}|\mathcal{S}_n\right) \quad \text{(cf. (15)).} \tag{45}$$

Next follows the main result of this section where $\mu \sim f$ is the estimated distribution and $\mathcal{S}_n$ is sub-σ-algebra preserving the divergence $D(f_n^{(1)}, f_n^{(2)})$ of the estimates $f_n^{(1)}, f_n^{(2)}$ in (43) (e.g. the intersection of all sub-σ-algebras $\mathcal{S} \subset \mathcal{B}^d$ preserving this divergence).

Theorem 2. The estimate f_n^* resulting from the metric divergence selection rule (44) satisfies the inequality

$$D\left(f_n^*, f\right) \leq 3D\left(f_n^{(0)}, f\right) + 2D\left(\mu, \mu_n|\mathcal{S}_n\right). \tag{46}$$

Proof. We can start with the equality (40) valid in the present situation as well. It suffices to prove that for $i \neq j$

$$D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{ij} \leq 2\left[D\left(f_n^{(0)}, f\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{ij}$$

where $\mathcal{E}_{ij}$ is defined by (39) for $f_n^*, f_n^{(0)}$ given by (43), (44). Using repeatedly the triangle inequality and relations (45) and (15) we obtain

$$\begin{aligned}
D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{21} &= D\left(f_n^{(0)}, f_n^{(2)}\right)\mathcal{E}_{21} = D\left(f_n^{(1)}, f_n^{(2)}|\mathcal{S}_n\right)\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f|\mathcal{S}_n\right) + D\left(f_n^{(2)}, f|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f\right) + D\left(\mu_n^{(2)}, \mu_n|\mathcal{S}_n\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f\right) + D\left(\mu_n^{(2)}, \mu|\mathcal{S}_n\right) + 2D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[D\left(f_n^{(1)}, f\right) + D\left(\mu_n^{(1)}, \mu|\mathcal{S}_n\right) + 2D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&\leq \left[2D\left(f_n^{(1)}, f\right) + 2D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21} \\
&= 2\left[D\left(f_n^{(0)}, f\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{21}.
\end{aligned}$$

In the same manner we obtain

$$D\left(f_n^*, f_n^{(0)}\right)\mathcal{E}_{12} \leq 2\left[D\left(f_n^{(0)}, f\right) + D\left(\mu_n, \mu|\mathcal{S}_n\right)\right]\mathcal{E}_{12}$$

which completes the proof of (46).

The next corollary presents a different expression of the error term in (46).

Corollary 2. The estimate f_n^* resulting from the selection rule (44) employing a metric divergence $D(f,g) = D_\phi(f,g)^\pi$ satisfies the inequality

$$D\left(f_n^*, f\right) \leq 3D\left(f_n^{(0)}, f\right) + 2^{\pi+1}c_\phi^\pi \sup_{B\in\mathcal{S}_n}\left|\int_A f - \mu_n(B)\right|^\pi \tag{47}$$

where $f_n^{(0)}, f$ and $\mathcal{S}_n$ are the same as in Theorem 2 and $c_\phi = \phi(0) < \infty$.

Proof. By Proposition 8.27 in Liese and Vajda (1987) and (8),

$$D_\phi\left(\mu_n, \mu|\mathcal{S}_n\right) \leq c_\phi V\left(\mu_n, \mu|\mathcal{S}_n\right) \quad \text{and} \quad V\left(\mu_n, \mu|\mathcal{S}_n\right) = 2\sup_{A\in\mathcal{S}_n}|\mu(A) - \mu_n(A)|$$

and the rest follows from Theorem 2 and (37)..

As in the previous section, our first step is to verify that Theorem 1 generalizes the Devroye–Lugosi result (6).

Example 6. Putting $D(f,g) = V(f,g)$ in Theorem 2 and using the fact that by (20) the sub-σ-algebra $\mathcal{S}_{A_n}$ preserves the total variation $V(f_n^{(1)}, f_n^{(2)})$ of the estimates $f_n^{(1)}, f_n^{(2)}$, we get

$$V\left(f_n^*, f\right) \leq 3V\left(f_n^{(0)}, f\right) + 2V\left(\mu, \mu_n|\mathcal{S}_{A_n}\right).$$

This coincides with the equivalent form (19) of the Devroye-Lugosi inequality (6).

Most important from the point of view of applications is the complexity of the sub-σ-algebra $\mathcal{S}_n \subset \mathcal{B}^d$ which appears in the right-hand error terms of (46) and (47). It depends on the complexity of the used divergence error criterion $D(f,g)$ and the complexity of the estimates $f_n^{(1)}$, $f_n^{(2)}$. In the previous example we have seen that if $D(f,g)$ is as simple as the total variation $V(f,g)$, then $\mathcal{S}_n$ is the simple σ-algebra $\mathcal{S}_{A_n}$ generated by just one set – the Scheffé set A_n of the estimates $f_n^{(1)}$, $f_n^{(2)}$ – irrespectively of how complex these estimates are. In the following example we shall see the opposite extreme, namely simple estimates $f_n^{(1)}$, $f_n^{(2)}$ leading to a simple σ-algebra $\mathcal{S}_n = \mathcal{S}_{B_n}$ generated by just one set B_n specified by these estimates, irrespectively of how complex the divergence criterion $D(f,g)$ is. More precisely, B_n does not depend on this criterion at all.

Example 7. Let the sample $X_1, \dots, X_n$ be governed by a bell-shaped density f on $\mathbb{R}$ and consider the sample mean and variance

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_n)^2,$$

and also the following *central cover set*

$$B_n = \{x : |x - \mu_n| < 3\sigma_n\}. \tag{48}$$

Let f be estimated by Cauchy type densities

$$f_n^{(1)}(x) = \frac{\sigma_n}{\pi\left[\sigma_n^2 + (x-\mu_n)^2\right]}$$

and

$$f_n^{(2)}(x) = \boldsymbol{I}(x \in B_n)\,\frac{b\sigma_n}{\pi[\sigma_n^2 + (x-\mu_n)^2]} \tag{49}$$

where

$$b = \left[1 - 2\left(\frac{1}{2} - \frac{1}{\pi}\operatorname{arctg} 3\right)\right]^{-1} = \frac{\pi}{2\operatorname{arctg} 3}.$$

In (49) we used the fact that the condition $\boldsymbol{I}(x \in B_n)$ cuts away from $f_n^{(1)}(x)$ two tail probabilities of the size

$$\begin{aligned}\int_{-\infty}^{\mu_n - 3\sigma_n} f_n^{(1)} &= \int_{-\infty}^{-3} \frac{\mathrm{d}x}{\pi[1+x^2]} \\ &= \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg}(-3) = \frac{1}{2} - \frac{1}{\pi}\operatorname{arctg} 3\end{aligned}$$

so that the $f_n^{(1)}$-probability of the sample central cover set is 1/b. The likelihood ratio $f_n^{(2)}/f_n^{(1)}$ is piecewise constant,

$$\frac{f_n^{(2)}(x)}{f_n^{(1)}(x)} = \begin{cases} b & \text{if } x \in B_n \\ 0 & \text{otherwise,} \end{cases}$$

where b is the normalizing factor used in (49). Therefore the sub-σ-algebra $\mathcal{S}_{B_n} = \{\mathbb{R}, B_n, B_n^c, \emptyset\} \subset \mathcal{B}$ generated by the central cover set B_n of (48) is sufficient for the family $\{f_n^{(1)}, f_n^{(2)}\}$. By what was said in Section 2, this means that $\mathcal{S}_{B_n}$ preserves for every convex ϕ the ϕ-divergence $D_\phi(f_n^{(1)}, f_n^{(2)})$. In other words, the sub-σ-algebra $\mathcal{S}_n$ considered in Theorem 2 and Corollary 2 is $\mathcal{S}_{B_n}$. Hence, by Theorem 1 and formula (16), for every metric divergence criterion $D(f,g) = D_\phi(f,g)^\pi$ with $\pi > 0$

$$D\left(f_n^*, f\right) \le 3D\left(f_n^{(0)}, f\right) + 2\left[\sum_{B\in\{B_n, B_n^c\}} \int_B f\,\phi\left(\frac{\mu_n(B)}{\int_B f}\right)\right]^\pi . \tag{50}$$

By Corollary 2, simpler but in general weaker variant of the result (50) is the inequality

$$D\left(f_n^*, f\right) \le 3D\left(f_n^{(0)}, f\right) + 2^{\pi+1}\phi^\pi(0)\left|\int_{B_n} f - \mu_n(B_n)\right|^\pi . \tag{51}$$

Next follows a theorem which generalizes and makes precise the phenomena observed in the last example.

Theorem 3. If the metric divergence criterion $D(f,g)$ is a ϕ-divergence power with $\phi(t)$ strictly convex in the whole domain $t > 0$ then a sub-σ-algebra $\mathcal{S}_n \subset \mathcal{B}^d$ preserves $D(f_n^{(1)}, f_n^{(2)})$ in the sense

$$D\left(f_n^{(1)}, f_n^{(2)}|\mathcal{S}_n\right) = D\left(f_n^{(1)}, f_n^{(2)}\right)$$

if and only if $\mathcal{S}_n$ is sufficient for $\{f_n^{(1)}, f_n^{(2)}\}$.

Proof. Let $D(f_n^{(1)}, f_n^{(2)}) = D_\phi(f_n^{(1)}, f_n^{(2)})^\pi$ for some $\pi > 0$. By the Corollary 2 above, the metricity of $D_\phi(f,g)^\pi$ implies $D_\phi(f_n^{(1)}, f_n^{(2)}) \le 2\phi(0) < \infty$. Hence, by Corollary 1.29 in Liese and Vajda (1987), the equality $D_\phi(f_n^{(1)}, f_n^{(2)}) = D_\phi(f_n^{(1)}, f_n^{(2)}|\mathcal{S}_n)$ takes place if and only if $\mathcal{S}_n$ is sufficient.

From this theorem we see that functions ϕ strictly convex everywhere define the most complex divergence criteria for which the σ-algebra $\mathcal{S}_n$ is simple only if the estimates $f_n^{(1)}$, $f_n^{(2)}$ are simple enough. Example 4 illustrated such situation.

5 Appendix

For practical applications of the results of Sections 3 and 4 one needs concrete metric divergence criteria $D(f,g) = D_\phi(f,g)^\pi$ with known and simple upper and lower bound $U_\phi(V)$ and $L_\phi(V)$ introduced in (34). For this purpose he can use the criteria from theclass

$$D(f,g) = \mathcal{D}_\alpha(f,g)^{\pi(\alpha)} \quad \text{for} \quad \pi(\alpha) = \frac{1}{\max\{2,\alpha\}} = \begin{cases} \frac{1}{2} & \text{when } -\infty < \alpha \le 2 \\ \\ \frac{1}{\alpha} & \text{when } \alpha > 2. \end{cases} \tag{52}$$

introduced in (25) − (28). The following theorem summarizes basic relevant properties of the divergences $\mathcal{D}_\alpha(f,g)$. For the proof we refer to Vajda (2008).

Theorem A1.

(i) $\mathcal{D}_\alpha(f,g)$ are ϕ_α-divergences with functions $\phi_\alpha(t)$ strictly convex in the domain $t > 0$ when $\alpha \neq 0$.

(ii) The lower bounds of the divergences $\mathcal{D}_\alpha(f,g)$, $\alpha \in \mathbb{R}$ under the constraint $V(f,g) = V$ are given for $0 \leq V \leq 2$ by the formulas

$$L_\alpha(V) = \frac{|\alpha|}{\alpha(\alpha-1)}\left(2^\alpha - \left[\left(1+\frac{V}{2}\right)^{1/\alpha} + \left(1-\frac{V}{2}\right)^{1/\alpha}\right]^\alpha\right) \tag{53}$$

if $\alpha(\alpha-1) \neq 0$ and otherwise by the corresponding limits

$$L_0(V) = V/2, \quad L_1(V) = \left(1+\frac{V}{2}\right)\ln\left(1+\frac{V}{2}\right) + \left(1-\frac{V}{2}\right)\ln\left(1-\frac{V}{2}\right). \tag{54}$$

(iii) The upper bounds of the divergences $\mathcal{D}_\alpha(f,g)$, $\alpha \in \mathbb{R}$ under the constraint $V(f,g) = V$ are $U_\alpha(V) = c_\alpha V$ where $c_\alpha > 0$ is continuous in the variable $\alpha \in \mathbb{R}$, given by the formula

$$c_\alpha = \phi_\alpha(0) = \begin{cases} \dfrac{2^{\alpha-1}}{|\alpha|+1} & \text{when } \alpha < 0 \\ \ln 2 & \text{when } \alpha = 1 \\ \dfrac{2^{\alpha-1}-1}{\alpha-1} & \text{when } \alpha \geq 0,\ \alpha \neq 1. \end{cases} \tag{55}$$

(iv) The powers $\mathcal{D}_\alpha(f,g)^{\pi(\alpha)}$ given in (52) are metrics in the space of probability densities f, g.

Remark. Putting $\alpha = 0$ in (iv) of Theorem A1 one obtains among other the triangle inequality

$$\sqrt{\mathcal{D}_0(f,g)} \leq \sqrt{\mathcal{D}_0(f,h)} + \sqrt{\mathcal{D}_0(h,g)}$$

for the divergence $\mathcal{D}_0(f,g) = V(f,g)/2$ which is weaker than the classical triangle inequality

$$\mathcal{D}_0(f,g) \leq \mathcal{D}_0(f,h) + \mathcal{D}_0(h,g) \tag{56}$$

obtained by applying the L_1-norm argument to the total variation $V(f,g)$. Using the continuity of the divergences $\mathcal{D}_\alpha(f,g)$ in the variable $\alpha \in \mathbb{R}$ we can deduce from (56) that more sophisticated arguments than those used to prove Theorem A1 lead to stronger triangle inequalities also for the remaining divergences $\mathcal{D}_\alpha(f,g)$, $\alpha \in \mathbb{R}$, in particular for those with α close to 0.

Acknowledgement. This research was supported by the grants GA ČR 102/07/1131 and MŠMT 1M052.

References

A. Berlinet, G. Biau and L. Rouviere (2005a): Parameter selection in modified histogram estimates. *Statistics* **39,** 91-105.

A. Berlinet, G. Biau and L. Rouviere (2005b): Optimal L_1 bandwith selection for variable kernel estimates. *Statistics and Probability Letters* **74,**.116-127.

I. Csiszár (1967a): Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungarica* **2**, 299–318.

I. Csiszár (1967b): On topological properties of f-divergences. *Studia Sci. Math. Hungarica* **2**, 329–339.

L. Devroye and G. Lugosi (2001): *Combinatorial Methods in Density Estimation.* Springer, Berlin.

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

F. Liese and I. Vajda (2006): On divergences and informations in statistics and information theory. *IEEE Trans. Inform. Theory* **52**, 10, 4394–4412.

F. Österreicher and I. Vajda (2003): A new class of metric divergences on probability spaces and its statistical applications. *Ann. Inst. Statist. Math.* **55**, 639–653.

I. Vajda (2008): On metric f-divergences of probability measures. *Kybernetika* (submitted).