Ann. Inst. Statist. Math.
Vol. 55, No. 3, 639–653 (2003)

A NEW CLASS OF METRIC DIVERGENCES ON PROBABILITY SPACES AND ITS APPLICABILITY IN STATISTICS

FERDINAND ÖSTERREICHER[1] AND IGOR VAJDA[2]*

[1]*Institute of Mathematics, University of Salzburg, 5020 Salzburg, Austria*
[2]*Institute of Information Theory and Automation, Academy of Sciences, 18208 Prague, Czech Republic*

(Received July 17, 1997; revised September 2, 2002)

Abstract. The class I_{f_β}, $\beta \in (0,\infty]$, of f-divergences investigated in this paper is defined in terms of a class of entropies introduced by Arimoto (1971, *Information and Control*, **19**, 181–194). It contains the squared Hellinger distance (for $\beta = 1/2$), the sum $I(Q_1\|(Q_1+Q_2)/2)+I(Q_2\|(Q_1+Q_2)/2)$ of Kullback-Leibler divergences (for $\beta = 1$) and half of the variation distance (for $\beta = \infty$) and continuously extends the class of squared perimeter-type distances introduced by Österreicher (1996, *Kybernetika*, **32**, 389–393) (for $\beta \in (1,\infty]$). It is shown that $(I_{f_\beta}(Q_1,Q_2))^{\min(\beta,1/2)}$ are distances of probability distributions Q_1, Q_2 for $\beta \in (0,\infty)$. The applicability of I_{f_β}-divergences in statistics is also considered. In particular, it is shown that the I_{f_β}-projections of appropriate empirical distributions to regular families define distribution estimates which are in the case of an i.i.d. sample of size n consistent. The order of consistency is investigated as well.

Key words and phrases: Dissimilarities, metric divergences, minimum distance estimators.

1. Introduction

In this paper we consider the intervals $\overline{I\!R} = (-\infty,\infty]$, $I\!R_+ = [0,\infty)$, $I\!R_0 = (0,\infty)$ and $\overline{I\!R}_0 = (0,\infty]$. Let $(\mathcal{X},\mathcal{A})$ be a nondegenerate measurable space (i.e. $|\mathcal{A}| > 2$ and hence $|\mathcal{X}| > 1$) and let $\mathcal{Q}(\mathcal{X},\mathcal{A})$ be the set of probability distributions on $(\mathcal{X},\mathcal{A})$. Furthermore, let $\mathcal{F}$ be the set of convex functions $f : I\!R_+ \mapsto \overline{I\!R}$ which are finite on $I\!R_0$ and continuous on $I\!R_+$. In addition, let the function $f^* \in \mathcal{F}$ be defined by

$$f^*(u) = u \cdot f\left(\frac{1}{u}\right) \quad \text{for} \quad u \in I\!R_0.$$

Remark 1. By setting

$$0f\left(\frac{v}{0}\right) = \begin{cases} 0 & \text{for} \quad v = 0 \\ v \cdot f^*(0) & \text{for} \quad v > 0 \end{cases}$$

*Supported by the EC grant Copernicus 579.

for all $f \in \mathcal{F}$, it holds

$$x \cdot f^*\left(\frac{y}{x}\right) = y \cdot f\left(\frac{x}{y}\right) \qquad \text{for all} \quad x, y \in \mathbb{R}_+.$$

DEFINITION 1. (cf. Csiszár (1963) and Ali and Silvey (1966)) Let $Q_1, Q_2 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$. Then

$$I_f(Q_1, Q_2) = \int f\left(\frac{q_1}{q_2}\right) \cdot q_2 d\mu$$

is called f-divergence of Q_1 and Q_2. (As usual, q_1 and q_2 denote the Radon-Nikodym-derivatives of Q_1 and Q_2 with respect to a dominating σ-finite measure μ.)

Let

$$H(p_1, \ldots, p_m) = \sum_{i=1}^{m} g(p_i)$$

be an entropy of a discrete probability distribution $(p_1, \ldots, p_m)$. Then, according to Morales *et al.* (1996), $g(t)$, $t \in [0,1]$, must be a concave function with $g(0) = g(1) = 0$. In order to avoid trivial cases we assume $0 < g(t) < \infty$ for $t \in (0,1)$. Then for $m = 2$ the entropy H is given by the function $h(t) = H(t, 1-t)$, $t \in [0,1]$, which is nonnegative, concave, symmetric with respect to $t = \frac{1}{2}$ and satisfies $h(0) = h(1) = 0$ and $h(\frac{1}{2}) \in (0, \infty)$. Consequently,

$$f(u) = (1+u)[h(1/2) - h(u/(1+u))], \qquad u \in \mathbb{R}_+,$$

is convex and satisfies $f(1) = 0$, $f^*(u) \equiv f(u)$ and $f(0) = h(1/2) \in (0, \infty)$. Using this representation, we obtain a class of f-divergences I_{f_β}, $\beta \in (0, \infty]$, from a class of entropies due to Arimoto (1971), defined in terms of such concave functions h. Appropriate powers of our f-divergences are shown to be distances on a given space of probability distributions. We also demonstrate that f-divergences, providing a metric, enable a comfortable treatment of statistical applications such as distribution and parameter estimation.

The f-divergence I_{f_2} was introduced by Österreicher (1982), and applied by him and by Reschenhofer and Bomze (1991) in different areas of hypotheses testing. Furthermore, it was shown by Österreicher (1996) that for every $\beta \in (1, \infty)$ the square root of the I_{f_β}-divergence defines a distance on the set of probability distributions. This generalized a result achieved for $\beta = 2$ by Kafka *et al.* (1991).

From the former literature on the subject, the powers of f-divergences defining distances are known for the subsequent classes. For the class of Hellinger divergences of order $s \in (0,1)$ given by $f^{(s)}(u) = 1 + u - (u^s + u^{1-s})$, already Csiszár and Fischer (1962) have shown that the corresponding maximal power is $\min(s, 1-s)$. For the following two classes the maximal power coincides with their parameter. The class given in terms of $f_{(\alpha)}(u) = |1 - u^\alpha|^{1/\alpha}$, $\alpha \in (0,1]$, was introduced by Matusita (1964) and investigated by Boekee (1977), Liese and Vajda (1987) and many other authors. The previous class and this one have the special case $s = \alpha = \frac{1}{2}$ in common. This famous special case was already investigated by Matusita (1955). The class given by $\varphi_\alpha(u) = |1-u|^{1/\alpha}(1+u)^{1-1/\alpha}$, $\alpha \in (0,1]$, and investigated in Kafka *et al.* (1991), Example 3, contains the special case $\alpha = \frac{1}{2}$ introduced by Vincze (1981).

2. Preliminaries

Let us restate some results from Kafka *et al.* (1991) which are basic for the statement and proof of the main result of this paper. For further information on f-divergences we refer to the monograph by Liese and Vajda (1987) and the paper of Österreicher and Vajda (1993).

Provided

(f1) $f(1) = 0$ and f is strictly convex at 1 and

(f2) $f^*(u) \equiv f(u)$

it holds for any $Q_1, Q_2 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$

(M1) $I_f(Q_1, Q_2) \geq 0$ with equality iff $Q_1 = Q_2$,

(M2) $I_f(Q_1, Q_2) = I_f(Q_2, Q_1)$

respectively. If, in addition to (f1) and (f2), an $\alpha \in \mathbb{R}_0$ exists such that

(f3,α) the function $h(u) = \dfrac{(1-u^\alpha)^{1/\alpha}}{f(u)}$, $u \in [0,1)$, is nonincreasing

then, according to Kafka *et al.* (1991), Theorems 1 and 2, the power

$$\rho_\alpha(Q_1, Q_2) = [I_f(Q_1, Q_2)]^\alpha$$

of the f-divergence satisfies for all $Q_1, Q_2, Q_3 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$ the triangle inequality

(M3,α) $\rho_\alpha(Q_1, Q_3) \leq \rho_\alpha(Q_1, Q_2) + \rho_\alpha(Q_2, Q_3)$.

Remark 2. Note that by virtue of Jensen's inequality

$$\frac{f(u) + f^*(u)}{1+u} = \frac{1}{1+u} \cdot f(u) + \frac{u}{1+u} \cdot f\left(\frac{1}{u}\right) \geq f(1).$$

Therefore (f1) and (f2) imply $f(u) > 0$ for all $u \in \mathbb{R}_+ \backslash \{1\}$. The validity of (f3,$\alpha$) for any $\alpha \in \mathbb{R}_0$ implies $f(0) < \infty$ (cf. property (f3) in Lemma 1). (Provided that $\mathcal{A}$ is infinite, this property is—together with (f1) and (f2)—a necessary condition so that the associated f-divergence allows for the definition of a metric.) Moreover, it can be easily seen that if $0 < \beta < \alpha$ then (f3,β) follows from (f3,α).

The following remark is a consequence of Kafka *et al.* (1991), Propositions 5 and 6.

Remark 3. Let (f1) and (f2) hold true and let $\alpha_0 \in (0,1]$ be the maximal α for which (f3,α) is satisfied. Then the following statement concerning α_0 holds. If for some $k_0, k_1, c_0, c_1 \in \mathbb{R}_0$

$$f(0) \cdot (1+u) - f(u) \sim c_0 \cdot u^{k_0} \quad \text{for} \quad u \downarrow 0 \quad \text{and}$$
$$f(u) \sim c_1 \cdot |u-1|^{k_1} \quad \text{for} \quad u \uparrow 1$$

then $k_0 \leq 1$, $k_1 \geq 1$ and $\alpha_0 \leq \min(k_0, 1/k_1) \leq 1$.

3. Definition of f_β-divergences

Let us start with the following class of entropies due to Arimoto (1971)

$$h_\alpha(t) = \begin{cases} \dfrac{1}{1-\alpha}[1-(t^{1/\alpha}+(1-t)^{1/\alpha})^\alpha] & \text{if } \alpha \in \mathbb{R}_0\backslash\{1\} \\ -[t\ln t + (1-t)\ln(1-t)] & \text{if } \alpha = 1 \\ \min(t, 1-t) & \text{if } \alpha = 0, \end{cases}$$

where we make use of the convention $0\ln(0) = 0$ and define the corresponding class of convex functions f_β by $f_\beta(u) = (1+u)[h_{1/\beta}(1/2) - h_{1/\beta}(u/(1+u))]$, $u \in [0,\infty)$,

$$f_\beta(u) = \begin{cases} \dfrac{1}{1-1/\beta}[(1+u^\beta)^{1/\beta} - 2^{1/\beta-1}(1+u)] & \text{if } \beta \in \mathbb{R}_0\backslash\{1\} \\ (1+u)\ln(2) + u\ln(u) - (1+u)\ln(1+u) & \text{if } \beta = 1 \\ |1-u|/2 & \text{if } \beta = \infty. \end{cases}$$

Remark 4. As for the corresponding entropies, $\beta = 1$ and $\beta = \infty$ are limiting cases, i.e. it holds $\lim_{\beta\to 1} f_\beta(u) = f_1(u)$ and $\lim_{\beta\to\infty} f_\beta(u) = f_\infty(u)$. Furthermore, it holds $f_{1/2}(u) = (1-\sqrt{u})^2$. Therefore

$$I_{f_{1/2}}(Q_1, Q_2) = \int (\sqrt{q_2} - \sqrt{q_1})^2 d\mu = H^2(Q_1, Q_2)$$

is the squared Hellinger distance. In addition,

$$\begin{aligned} I_{f_1}(Q_1, Q_2) &= I\left(Q_1 \,\middle\|\, \frac{Q_1+Q_2}{2}\right) + I\left(Q_2 \,\middle\|\, \frac{Q_1+Q_2}{2}\right) \\ &= 2H\left(\frac{Q_1+Q_2}{2}\right) - [H(Q_1) + H(Q_2)] \end{aligned}$$

where I and H are the classical information divergence (f-divergence for $f(u) = u\ln u$), respectively Shannon's entropy and

$$I_{f_\infty}(Q_1, Q_2) = \frac{1}{2}V(Q_1, Q_2)$$

where V is the total variation (f-divergence for $f(u) = |1-u|$). The appeal of the special case

$$I_{f_2}(Q_1, Q_2) = 2\left[\int \sqrt{q_1^2 + q_2^2}\, d\mu - \sqrt{2}\right],$$

given by $f_2(u) = 2[\sqrt{1+u^2} - (1+u)/\sqrt{2}]$, is its geometric interpretation (cf. Österreicher (1992)).

Finally note that the limiting case $I_{f_1}(Q_1, Q_2)$, which is a symmetric form of the f-divergence $I(Q_1 \,\|\, (Q_1+Q_2)/2)$ considered by Lin (1991), is part of the identity

$$I(Q_1 \,\|\, Q_3) + I(Q_1 \,\|\, Q_3) = 2I((Q_1+Q_2)/2 \,\|\, Q_3) + I_{f_1}(Q_1 \,\|\, Q_2)$$

exploited by Csiszár (1975).

The introductory properties of the following lemma enable the class f_β to define f-divergences. With (f1)–(f3) it provides basic properties for the corresponding distances. The limiting case $\beta = \infty$ will be excluded from the rest of this section.

LEMMA 1. *The class of functions* f_β, $\beta \in (0, \infty)$ *satisfies* $f_\beta \in \mathcal{F}$, (f1) *and* (f2). *Furthermore*

(f3) $$f_\beta(0) \in (0, \infty),$$

in particular,

$$f_\beta(0) = \begin{cases} \dfrac{1}{1-1/\beta}[1 - 2^{1/\beta - 1}] & \text{if } \beta \neq 1 \\ \ln(2) & \text{if } \beta = 1 \end{cases}$$

(f4) $$f'_\beta(u) = \begin{cases} \dfrac{1}{1-1/\beta}[(1+u^\beta)^{1/\beta-1}u^{\beta-1} - 2^{1/\beta-1}] & \text{if } \beta \neq 1 \\ \ln(2) + \ln(u) - \ln(1+u) & \text{if } \beta = 1 \end{cases}$$

and hence $f'_\beta(1) = 0$

(f5) $$f''_\beta(u) = \beta(1+u^\beta)^{1/\beta-2}u^{\beta-2} > 0 \text{ and hence } f''_\beta(1) = \beta 2^{1/\beta-2}.$$

PROOF. The properties $f_\beta \in \mathcal{F}$, (f1) and (f2) hold according to the basic properties of entropies. The properties (f3)–(f5) are also obvious. □

The following remark is a consequence of Kafka *et al.* (1991), Propositions 5 and 6.

Remark 5. (f3) and the application of *Newton's Binomial formula* $(1+x)^\alpha \cdot = \sum_{i=0}^{\infty} \binom{\alpha}{i} x^i$ for $|x| < 1$ and $\alpha \in \mathbb{R}$ yield

$$f_\beta(0)(1+u) - f_\beta(u) = \begin{cases} \dfrac{1}{1-1/\beta}\left[u - \displaystyle\sum_{i=1}^{\infty} \binom{1/\beta}{i} u^{\beta i}\right] & \text{if } \beta \neq 1 \\ (1+u)\ln(1+u) - u\ln(u) & \text{if } \beta = 1, \end{cases}$$

and hence, for $u \downarrow 0$ the asymptotic equality

$$f_\beta(0)(1+u) - f_\beta(u) \sim \begin{cases} \dfrac{1}{1-\beta}u^\beta & \text{if } \beta < 1 \\ \dfrac{1}{1-1/\beta}u & \text{if } \beta > 1 \\ -u\ln(u) & \text{if } \beta = 1. \end{cases}$$

Since, owing to (f1), (f4) and (f5),

(f6) $$f_\beta(u) \sim \beta 2^{1/\beta-3}(u-1)^2$$

for $u \to 1$, the maximal $\alpha \in (0,\infty)$ satisfying (f3,α) with $f(u)$ replaced by $f_\beta(u)$—if there is any—must be $\alpha_0 \le \min(\beta, 1/2)$ (cf. Remark 3).

LEMMA 2. *The function*

$$\xi_\beta(u) = \begin{cases} \dfrac{(1-u^\beta)^{1/\beta}}{f_\beta(u)} & \text{if } \beta \in (0,1/2) \\ \dfrac{(1-\sqrt{u})^2}{f_\beta(u)} & \text{if } \beta \in [1/2,\infty) \end{cases}$$

defined for all $u \in [0,1)$ *satisfies*

$$\lim_{u\uparrow 1} \xi_\beta(u) = \begin{cases} 0 & \text{if } \beta \in (0,1/2) \\ \dfrac{2^{1-1/\beta}}{\beta} & \text{if } \beta \in [1/2,\infty) \end{cases}$$

and $\xi_{1/2}(u) \equiv 1$. *For* $\beta \in (0,1/2)\cup(1/2,\infty)$ *this function is strictly monotone decreasing.*

PROOF. The first statement is a consequence of (f1), (f4) and (f5). The second one is obvious. Now for the proof of the monotony of ξ_β for $\beta \in (0,1/2)$: Owing to $1 - 1/\beta < 0$ it holds

$$\xi'_\beta(u) = \frac{1}{1-1/\beta} \frac{(1-u^\beta)^{1/\beta-1} u^{\beta-1}}{f_\beta^2(u)} \varphi_\beta(u) < 0$$

for all $u \in (0,1)$ since

$$\begin{aligned} \varphi_\beta(u) &= -(1-1/\beta)[f_\beta(u) + (u^{1-\beta} - u) f'_\beta(u)] \\ &= 2[2^{1/\beta-2}(1+u^{1-\beta}) - (1+u^\beta)^{1/\beta-1}] > 0. \end{aligned}$$

The latter holds because of $\varphi_\beta(1) = 0$ and since for $\beta \in (0,1/2)$

$$u^\beta \varphi'_\beta(u)/2 = (1-\beta)\left[2^{1/\beta-2} - \left(1 + \frac{1}{u^\beta}\right)^{1/\beta-2}\right] < 0.$$

The proof of the monotony of ξ_β for $\beta \in (1/2,\infty)\backslash\{1\}$ can be taken almost literally from that of Lemma 2 in Österreicher (1996). □

THEOREM 1. *Let* $\beta \in (0,\infty)$. *Then* $\rho_\beta(Q_1,Q_2) = [I_{f_\beta}(Q_1,Q_2)]^{\min(\beta,1/2)}$ *defines a metric on the space* $\mathcal{Q}(\mathcal{X},\mathcal{A})$.

PROOF. This assertion is clear from Lemma 2 and from what is said after property (f3,α). □

Remark 6. As already mentioned in Remark 4, the f-divergence for the case $\beta = \infty$ equals half of the variation distance, i.e.

$$I_{f_\infty}(Q_1,Q_2) = \frac{1}{2}\int |q_1 - q_2| d\mu$$

and therefore is a metric. Owing to $|q_1 - q_2| = q_1 + q_2 - 2\min(q_1, q_2)$ and $\int q_1 d\mu = \int q_2 d\mu = 1$ it equals

$$I_{f_\infty}(Q_1, Q_2) = 1 - 2b_{1/2}(Q_1, Q_2)$$

where $b_{1/2}(Q_1, Q_2) = \frac{1}{2}[Q_1(q_2 > q_1) + Q_2(q_1 > q_2)]$ is the minimal Bayes risk with respect to the prior distribution $(\frac{1}{2}, \frac{1}{2})$, i.e. the weighted probability of error when testing the hypothesis Q_1 against the alternative Q_2.

4. Properties of f_β-divergences

All divergences $I_{f_\beta}(Q_1, Q_2)$, $\beta \in (0, \infty]$, satisfy (M1)–(M3,α) with $\alpha = \min(\beta, 1/2)$ for $\beta < \infty$ and $\alpha = 1$ for $\beta = \infty$. In addition to the triangle inequality considered in (M3,α), all these divergences satisfy the following weaker triangle inequality: For arbitrary $Q_1, Q_2, Q_3 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$

$$I_{f_\beta}(Q_1, Q_3) \le 2^{1/\alpha - 1}[I_{f_\beta}(Q_1, Q_2) + I_{f_\beta}(Q_2, Q_3)].$$

This is trivial for $\beta = \infty$. For $\beta < \infty$ Theorem 1 and the application of Jensen's inequality to the concave function $x \mapsto x^\alpha$ yields

$$\begin{aligned} \frac{1}{2} I^\alpha_{f_\beta}(Q_1, Q_3) &\le \frac{1}{2}[I^\alpha_{f_\beta}(Q_1, Q_2) + I^\alpha_{f_\beta}(Q_2, Q_3)] \quad \text{(cf. Theorem 1)} \\ &\le \left[\frac{I_{f_\beta}(Q_1, Q_2) + I_{f_\beta}(Q_2, Q_3)}{2}\right]^\alpha, \end{aligned}$$

which already implies the desired result.

Other properties of the divergences under consideration can be derived from the properties of general f-divergences presented in Liese and Vajda (1987). In particular, for arbitrary $Q_1, Q_2 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$ and $f_\beta(0)$ given by (f3)

$$I_{f_\beta}(Q_1, Q_2) \le 2 f_\beta(0),$$

where the equality holds if and only if $Q_1 \perp Q_2$.

We sharpen the last inequality in terms of $V(Q_1, Q_2)/2$. The following theorem thus enables estimating the Bayes risk $b_{1/2}(Q_1, Q_2)$ by means of the f_β-divergences for suitable large $\beta \in \mathbb{R}_0$ arbitrarily closely.

THEOREM 2. *Let $\beta \in (0, \infty]$. Then for all $Q_1, Q_2 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$*

$$\psi_\beta(V(Q_1, Q_2)/2) \le I_{f_\beta}(Q_1, Q_2) \le \psi_\beta(1) \cdot V(Q_1, Q_2)/2,$$

where the function $\psi_\beta : [0, 1] \mapsto \mathbb{R}$ defined by

$$\psi_\beta(x) = \begin{cases} \dfrac{1}{1 - 1/\beta}([(1+x)^\beta + (1-x)^\beta]^{1/\beta} - 2^{1/\beta}) & \text{if } \beta \in \mathbb{R}_0 \setminus \{1\} \\ (1+x)\ln(1+x) + (1-x)\ln(1-x) & \text{if } \beta = 1, \\ x & \text{if } \beta = \infty \end{cases}$$

is convex, strictly monotone increasing and satisfies $\psi_\beta(0) = 0$ and $\psi_\beta(1) = 2f_\beta(0)$. The maximal difference between the above upper and lower bound

$$d(\beta) = \sup_{x \in [0,1]} [\psi_\beta(1)x - \psi_\beta(x)]$$

satisfies the relation

$$\lim_{\beta\to\infty} d(\beta) = d(\infty) = 0.$$

PROOF. In order to achieve the first assertion of the theorem we use terminology and results of Corollary 1 and Theorem 2 from Feldman and Österreicher (1989) which we refer to below. Note, however, that the result follows equally well from (8.26) and Proposition 8.28 in Liese and Vajda (1987). Lower and upper bound follow from Corollary 1 applied to the function

$$\psi_\beta(x) = c_{f_\beta}(x) = (1+x) f_\beta\left(\frac{1-x}{1+x}\right),$$

where the latter follows, by virtue of Theorem 2 (d), from the validity of (f2) for the functions f_β. The properties of the function ψ_β follows from Theorem 2 (a)–(c), and the properties (f1) and (f2).

In order to prove the limiting property of $d(\beta)$ note that $\psi_\beta(1) = 2f_\beta(0) = \frac{1}{1-1/\beta}(2-2^{1/\beta})$ and furthermore, that, owing to Remark 5,

$$D_-\psi_\beta(1) = \lim_{u\downarrow 0} \frac{f_\beta(0)(1+u) - f_\beta(u)}{u} = \begin{cases} \infty & \text{for} \quad \beta \in (0,1] \\ \dfrac{1}{1-1/\beta} & \text{for} \quad \beta \in (1,\infty) \end{cases}.$$

Now let $\beta \in (1,\infty)$. Since, in addition, ψ_β is nonnegative, convex and satisfies $\psi_\beta(0) = 0$ it holds $\psi_\beta(x) \geq \max\{0, \psi_\beta(1) + D_-\psi_\beta(1)(x-1)\}$. Consequently

$$d(\beta) \leq \psi_\beta(1)\left(1 - \frac{\psi_\beta(1)}{D_-\psi_\beta(1)}\right) = \frac{1}{1-1/\beta}(2-2^{1/\beta})(2^{1/\beta}-1).$$

From this the assertion follows since the derived upper bound decreases to 0 as $\beta \uparrow \infty$. □

THEOREM 3. *Let $\beta \in (0,\infty)$. Then*

$$\psi_\beta(x) \sim \beta 2^{1/\beta-1}\cdot x^2 \quad \textit{for} \quad x \downarrow 0 \quad \textit{and}$$
$$\psi_\beta(x) \geq \beta 2^{1/\beta-1}\cdot x^2 \quad \forall x \in [0,1] \Leftrightarrow \beta \in (0, 3/2].$$

PROOF. The first assertion can be easily seen from (f6) and the definition of ψ_β in the proof of Theorem 2. Now, let us extend the definition of ψ_β to $[-1,1]$ and let

$$\varphi_\beta(x) = [\psi_\beta(x) - \beta 2^{1/\beta-1}\cdot x^2]/(\beta 2^{1/\beta}), \qquad x \in (-1,1)$$

and $a_\beta(x) = [(1+x)^\beta + (1-x)^\beta]/2$. Then

$$\varphi'_\beta(x) = \begin{cases} a_\beta^{1/\beta-1}(x)[(1+x)^{\beta-1} - (1-x)^{\beta-1}]/(2(\beta-1)) - x & \text{for} \quad \beta \neq 1 \\ [\ln(1+x) - \ln(1-x)]/2 - x & \text{for} \quad \beta = 1 \end{cases} \quad \text{and}$$

$$\varphi''_\beta(x) = a_\beta^{1/\beta-2}(x)(1-x^2)^{\beta-2} - 1$$

and hence $\varphi'_\beta(0) = \varphi''_\beta(0) = 0$.

In order to prove or disprove $\varphi_\beta(x) \geq 0$ for all $x \in [0,1]$ it suffices, owing to $\varphi'_\beta(0) = 0$, to show that φ_β is convex on $[0,1]$ and strictly concave on a suitable subinterval $[0,\delta]$ of $[0,1]$ respectively.

At first let $\beta \in [1/2,1]$. Then the application of Jensen's inequality to the concave function $u \mapsto u^\beta$, $u \in [0,\infty)$, yields $a_\beta(x) \leq 1$ and since—owing to $1/\beta - 2 \leq 0$—the function $u \mapsto u^{1/\beta-2}$ is decreasing, consequently $a_\beta^{1/\beta-2}(x) \geq 1$. Owing to $\beta - 2 < 0$ this implies

$$\varphi''_\beta(x) \geq (1-x^2)^{\beta-2} - 1 > 0 \quad \text{for all} \quad x \in (0,1).$$

Now let

$$\varphi^{(3)}_\beta(x) = a_\beta^{1/\beta-3}(x)(1-x^2)^{\beta-3} g_\beta(x)/2$$

be the third derivative of φ_β where

$$g_\beta(x) = (1-2\beta)((1+x)^\beta - (1-x)^\beta) + 3x((1+x)^\beta + (1-x)^\beta).$$

In the sequel we will show $\varphi^{(3)}_\beta(x) \geq 0$ for all $x \in [0,1)$ or $\varphi^{(3)}_\beta(x) < 0$ for all $x \in (0,\delta)$ for a suitable $\delta \in (0,1)$ since, owing to $\varphi''_\beta(0) = 0$, consequently φ_β is either convex on $[0,1]$ or strictly concave on $[0,\delta]$, respectively.

Next let $\beta \in (0,1/2)$. Then since $1 - 2\beta > 0$ and since the function $u \mapsto u^\beta$ is increasing, every factor of g_β and hence of $\varphi^{(3)}_\beta$ is nonnegative on $[0,1)$. Consequently φ_β is convex.

For the remaining case $\beta \in (1,\infty)$ we have to investigate the crucial term g_β in more detail. Note that the value of the first derivative

$$\begin{aligned} g'_\beta(x) = {} & \beta(1-2\beta)((1+x)^{\beta-1} + (1-x)^{\beta-1}) + 3((1+x)^\beta + (1-x)^\beta) \\ & + 3\beta x((1+x)^{\beta-1} - (1-x)^{\beta-1}) \end{aligned}$$

at the point 0 equals $g'_\beta(0) = 4(\frac{3}{2} - \beta)(\beta + 1)$.

For the case $\beta \in (1,3/2]$ this value is $g'_\beta(0) \geq 0$. Since, in addition, g_β will turn out to be strictly convex and satisfies $g_p(0) = 0$ it holds therefore $g_\beta(x) > 0$ and thus $\varphi^{(3)}_\beta(x) > 0$ for all $x \in (0,1)$. In fact, since $1 - 2\beta < 0$, $\beta > 1$ and since the functions $u \mapsto u^{\beta-2}$, $u \mapsto u^{\beta-1}$ are decreasing resp. increasing, the second derivative of g_β

$$\begin{aligned} g''_\beta(x) = {} & (1-2\beta)\beta(\beta-1)((1+x)^{\beta-2} - (1-x)^{\beta-2}) + 6\beta((1+x)^{\beta-1} - (1-x)^{\beta-1}) \\ & + 3\beta(\beta-1)x((1+x)^{\beta-2} + (1-x)^{\beta-2}), \end{aligned}$$

is positive on $(0,1)$.

Finally let $\beta \in (3/2,\infty)$. Then $g'_\beta(0) < 0$. Since, in addition, g'_β is continuous, there is a $\delta \in (0,1)$ such that $g'_\beta(x) < 0$ and hence $\varphi^{(3)}_\beta(x) < 0$ for all $x \in [0,\delta)$. This completes the proof. □

Remark 7. Let $k_\beta = \max\{k \geq 0 : kx^2 \leq \psi_\beta(x)\ \forall x \in [0,1]\}$ and apply Remark 3 in Feldman and Österreicher (1989) to the present class of f-divergences. Then, by virtue of the discriminant

$$D(f_\beta) = -\frac{4}{3}\frac{[f^{(3)}_\beta(1)]^2}{f''_\beta(1)} + f^{(4)}_\beta(1) = \left(\frac{3}{2} - \beta\right)(\beta+1)\beta 2^{1/\beta-3}$$

and owing to $k_\beta < 2f''_\beta(1) \Leftrightarrow D(f_\beta) < 0$ and $2f''_\beta(1) = \beta 2^{1/\beta - 1}$, it holds

$$k_\beta < \beta 2^{1/\beta - 1} \Leftrightarrow \beta > \frac{3}{2}.$$

This is equivalent to the second statement of Theorem 3. According to $D(f_\beta) = g'_\beta(0)\beta 2^{1/\beta - 5}$ this reestablishes the associated result achieved in terms of $g'_\beta(0)$ in the course of the preceding proof.

5. Minimum f_β-divergence estimates

In this section we consider a subfamily $\mathcal{P} \subset \mathcal{Q}(\mathcal{X}, \mathcal{A})$ of probability distributions and a random sample $(X_1, \ldots, X_n)$ with independent $\mathcal{X}$-valued components distributed according to a fixed element $P_0 \in \mathcal{P}$. By a distribution estimate of P_0 we mean a mapping $Q_n : \mathcal{X}^n \mapsto \mathcal{Q}(\mathcal{X}, \mathcal{A})$ such that all probabilities $Q_n(A)$, $A \in \mathcal{A}$, are measurable functions of the sample $(X_1, \ldots, X_n)$. Of course, we are mainly interested in estimates $P_n : \mathcal{X}^n \to \mathcal{P}$.

We suppose that the σ-algebra $\mathcal{A}$ is countably generated. Then, by Theorem 1.30 in Liese and Vajda (1987), each f-divergence satisfies the relation

$$I_f(Q_1, Q_2) = \sup_{\mathcal{D} \subset \mathcal{A}} \sum_{A \in \mathcal{D}} f\left(\frac{Q_1(A)}{Q_2(A)}\right) \cdot Q_2(A),$$

where the supremum extends over all finite partitions $\mathcal{D} \subset \mathcal{A}$ of the observation space $\mathcal{X}$. This implies that for any $Q \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$ and any distribution estimate Q_n, the f-divergence $I_f(Q, Q_n)$ is a random variable (measurable function of the sample).

Barron *et al.* (1992) introduced a consistent distribution estimate P_n of P_0 in total variation and in information divergence, i.e. satisfying the conditions

$$V(P_0, P_n) = o_p(1) \quad \text{or} \quad I(P_0, P_n) = o_p(1)$$

respectively, under certain weak assumptions about $\mathcal{P}$ (see Remark 4; the symbols $o_p(1)$ and $O_p(c_n^{-1})$ denote in this paper random variables satisfying the well known asymptotic relations for nondecreasing sequences c_n of positive real numbers). These authors also presented several statistical and information-theoretic arguments leading to these rather strong types of consistency. In fact, their arguments can be extended to motivate estimates consistent in f-divergences also for $f \in \mathcal{F}$ different from $f(u) = |1 - u|$ and $f(u) = u \ln u$ (see Berlinet *et al.* (1997)).

For dominated families $\mathcal{P}$ the distribution estimates Q_n consistent in total variation reduce to the density estimates consistent in the L_1-norm (cf. Definition 1 with $f(u) = |1 - u|$). As well known (cf. e.g. Devroye and Győrfi (1985)), typical density estimates consistent in the L_1-norm, such as the histogram estimates or kernel estimates with kernels of bounded support, are discontinuous. Densities of the estimates considered by Barron *et al.* (1992) are also discontinuous. This is an obvious drawback in cases where $\mathcal{P}$ consists of continuous densities. One approach of modern density estimation to cope with this problem is the wavelet smoothing of the density of Q_n (see e.g. Hall and Patil (1995)). Another possibility presented already by Beran (1977, 1978) and Tamura and Boos (1986), and developed in Győrfi *et al.* (1994, 1996), Cutler and Codero-Brana (1996) and Pak (1996), is the projection of Q_n onto $\mathcal{P}$ in an appropriately metrized

space $\mathcal{Q}(\mathcal{X}, \mathcal{A})$. The result is a smooth estimate $P_n \in \mathcal{P}$. Györfi *et al.* (1996) used the Kolmogorov distance on $\mathcal{Q}(\mathcal{X}, \mathcal{A})$, the remaining papers used the Hellinger distance.

These facts motivate the following definitions.

DEFINITION 2. Let us consider $f \in \mathcal{F}$ and $\mathcal{P} \subset \mathcal{Q}(\mathcal{X}, \mathcal{A})$. An arbitrary estimate $Q_n : \mathcal{X}^n \to \mathcal{Q}(\mathcal{X}, \mathcal{A})$ is said to be consistent in the f-divergence if the relation $I_f(P_0, Q_n) = o_p(1)$ is satisfied for all possible distributions $P_0 \in \mathcal{P}$. If, in addition, c_n is a nondecreasing sequence of positive, real numbers tending to infinity as $n \to \infty$, then the estimate Q_n is said to be consistent of the order of c_n^{-1} in the f-divergence if the previous relation holds with $O_p(c_n^{-1})$ instead of $o_p(1)$.

DEFINITION 3. Let us consider an arbitrary estimate $Q_n : \mathcal{X}^n \to \mathcal{Q}(\mathcal{X}, \mathcal{A})$, the same function $f \in \mathcal{F}$ and the same family $\mathcal{P}$. Then the estimate $P_{f,n} : \mathcal{X}^n \to \mathcal{P}$ is said to be an asymptotically minimum f-divergence estimate for Q_n if

$$I_f(P_{f,n}, Q_n) \leq \inf_{P \in \mathcal{P}} I_f(P, Q_n) + o_p(1). \tag{5.1}$$

If, in addition, $\tilde{c}_n$ is a sequence of real numbers with the same properties as in Definition 2, then $P_{f,n}$ is said to be an asymptotically minimum f-divergence estimate for Q_n of the order of $\tilde{c}_n^{-1}$ if (5.1) holds with $O_p(\tilde{c}_n^{-1})$ instead of $o_p(1)$.

We are interested in estimates $P_{\beta,n} = P_{f_\beta,n}$, $\beta \in (0, \infty]$, which are asymptotically minimum f_β-divergence estimates for a given distribution estimate Q_n. If $P_{f_\beta,n}$ is the f_β-projection of Q_n onto the subfamily $\mathcal{P}$ in the sense of Chapter 8 in Liese and Vajda (1987), then it is an asymptotically minimum f_β-divergence estimator for Q_n of any order.

Example 1. Let $(\mathcal{X}, \mathcal{A})$ be the Borel line, λ the Lebesgue measure, and

$$Q(A) = \lambda((0,1) \cap A), \qquad Q_n(A) = \frac{\lambda((X_{n1}, X_{nn}) \cap A)}{\lambda((X_{n1}, X_{nn}))}$$

for all $A \in \mathcal{A}$, where $(X_{n1}, \ldots, X_{nn})$ is the ordered sample corresponding to $(X_1, \ldots, X_n)$. Let $\mathcal{P}$ be the class of all shifts of the distribution Q. Then the estimate

$$P_{\infty,n}(A) = \lambda\left(\left(\frac{X_{n1} + X_{nn} - 1}{2}, \frac{X_{n1} + X_{nn} + 1}{2}\right) \cap A\right)$$

minimizes $I_{f_\infty}(P, Q_n)$ on $\mathcal{P}$, i.e. $P_{\infty,n}$ satisfies $I_{f_\infty}(P_{\infty,n}, Q_n) = \inf_{P \in \mathcal{P}} I_{f_\infty}(P, Q_n)$.

THEOREM 4. *Let $\mathcal{P} \subset \mathcal{Q}(\mathcal{X}, \mathcal{A})$ be arbitrary, $P_0 \in \mathcal{P}$ and let $\beta \in \overline{\mathbb{R}}_0$. If Q_n is a consistent estimate of P_0 in the f_β-divergence, then every corresponding asymptotically minimum f_β-divergence estimate $P_{\beta,n}$ is also a consistent estimate in the f_β-divergence. Let, in addition, Q_n be consistent of the order of c_n^{-1} and let $P_{\beta,n}$ satisfy* (5.1) *with $o_p(1)$ replaced by $O_p(\tilde{c}_n^{-1})$ and $\tilde{c}_n^{-1} = O_p(c_n^{-1})$. Then the previous statement holds with "consistent" replaced by "consistent of the order of c_n^{-1}".*

PROOF. Put $\alpha = \min(\beta, 1/2)$ and $\rho_\beta(Q_1, Q_2) = I_{f_\beta}^\alpha(Q_1, Q_2)$. By Theorem 1, ρ_β is a distance on $\mathcal{Q}(\mathcal{X}, A)$. Thus the triangle inequality implies

$$\rho_\beta(P_0, P_{\beta,n}) \leq \rho_\beta(P_0, Q_n) + \rho_\beta(P_{\beta,n}, Q_n).$$

By the definition of $P_{\beta,n}$ (cf. (5.1))

$$\rho_\beta(P_{\beta,n}, Q_n) = I^\alpha_{f_\beta}(P_{\beta,n}, Q_n) \le [I_{f_\beta}(P_0, Q_n) + o_p(1)]^\alpha,$$

so that

$$I_{f_\beta}(P_0, P_{\beta,n}) \le (I^\alpha_{f_\beta}(P_0, Q_n) + [I_{f_\beta}(P_0, Q_n) + o_p(1)]^\alpha)^{1/\alpha}.$$

The consistency statement is clear from this fact. Under the additional assumptions we similarly obtain

$$c_n I_{f_\beta}(P_0, P_{\beta,n}) \le ([c_n I_{f_\beta}(P_0, Q_n)]^\alpha + [c_n I_{f_\beta}(P_0, Q_n) + c_n O_p(\tilde{c}_n^{-1})]^\alpha)^{1/\alpha}.$$

Thus the statement concerning the rate of consistency is clear, too. □

Unfortunately, there are no estimators Q_n satisfying the consistency assumption of Theorem 4 when $\mathcal{P} = \mathcal{Q}(\mathcal{X}, \mathcal{A})$. In this respect we can formulate the following negative result.

THEOREM 5. *For every sequence of estimates Q_n and every $\beta \in (0, \infty]$ there exists a $Q_0 \in \mathcal{Q}(\mathcal{X}, \mathcal{A})$ and an $\varepsilon > 0$ such that*

$$\inf_n I_{f_\beta}(Q_0, Q_n) \ge \varepsilon \qquad a.s.\ (\textit{almost surely}).$$

PROOF. According to Remark 6 and the Theorem of Devroye and Györfi (1990), this assertion holds for $\beta = \infty$, i.e. for every sequence of estimates Q_n there exists a Q_0 and an $\varepsilon_0 > 0$ such that

$$\inf_n I_{f_\infty}(Q_0, Q_n) = \inf_n V(Q_0, Q_n)/2 \ge \varepsilon_0 \qquad \text{a.s.}$$

This fact and Theorem 2 imply $\inf_n I_{f_\beta}(Q_0, Q_n) \ge \psi_\beta(\varepsilon_0) > 0$ a.s. □

In spite of the negative result of Theorem 5, there exist estimates Q_n, as assumed in Theorem 4, i.e. consistent for all P_0 from a wide variety of families $\mathcal{P} \subset \mathcal{Q}(\mathcal{X}, \mathcal{A})$, and even consistent of the order of $c_n^{-1} = n^{-\alpha}$ for appropriate $\alpha \in (0, 1]$.

Example 2. Let $(\mathcal{X}, \mathcal{A})$ be the Borel line and let Q_n be the histogram estimate of Berlinet *et al.* (1995) for partitions of $\mathbb{R}$ into intervals of size $h_n = \text{const} \cdot n^{-1/2}$. If $\mathcal{P} \subset \mathcal{Q}(\mathcal{X}, \mathcal{A})$ is the family of all distributions with continuously differentiable densities then these authors showed that $P_0 \in \mathcal{P}$ implies $V(P_0, Q_n) = O_p(n^{-1/3})$, i.e. Q_n is consistent of the order of $c_n^{-1} = n^{-1/3}$ in the f_∞-divergence. Let $\beta \in \mathbb{R}_0$. Then by the upper bound in Theorem 2 Q_n is consistent of the same order in the f_β-divergence.

Example 3. Let $\mathcal{Q}(\mathcal{X}, \mathcal{A})$ be as in the previous example. Györfi *et al.* (1996) proved that, under relatively mild restrictions on $\mathcal{P} \subset \mathcal{Q}(\mathcal{X}, \mathcal{A})$, the Kolmogorov distance projection P_n of the classical empirical distribution

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \tag{5.2}$$

onto $\mathcal{P}$ is a consistent estimate of the order of $c_n^{-1} = n^{-1/2}$ in the f_∞-divergence. As in Example 2, we obtain the same order of consistency of P_n in every f_β-divergence, $\beta \in I\!R_0$.

Example 4. Let $\mathcal{Q}(\mathcal{X}, \mathcal{A})$ be arbitrary and let S be a finite measurable subset of $\mathcal{X}$. Furthermore, let $\mathcal{P}$ be the family of all distributions with support S then the empirical distribution Q_n given by (5.2) satisfies $Q_n \in \mathcal{P}$ and is a consistent estimate of the order of $c_n^{-1} = n^{-1/2}$ in the f_∞-divergence. This fact is a consequence of the central limit theorem. By Corollary 1 of Morales *et al.* (1995) Q_n is consistent of the order of $c_n^{-2} = n^{-1}$ in any f_β-divergence, $\beta \in I\!R_0$. Since $Q_n \in \mathcal{P}$ it holds $P_{\beta,n} = Q_n$ for all $\beta \in \overline{I\!R}_0$.

Concluding Remark. Let us briefly mention the applications of f_β-divergences, $\beta \in I\!R_0$, in the model of Example 4 with $\mathcal{P}$ containing only some of the distributions with support S. This is typically satisfied if $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametrized family of distributions supported by S. In this case the estimates $P_{\beta,n} \equiv P_{\theta_{\beta,n}}$ depend on β and in general differ from the relative frequency estimate Q_n given by (5.2). If the parameter is identifiable in $\mathcal{P}$ (i.e. if $P_{\theta_1} \neq P_{\theta_2}$ for all $\theta_1 \neq \theta_2$) then these distribution estimates $P_{\theta_{\beta,n}}$ define in a one-to-one manner point estimates $\theta_{\beta,n}$, $\beta \in I\!R_0$.

Estimates minimizing the f-divergences $I_{\varphi_\alpha}(P_\theta, Q_n)$, $\alpha \in I\!R$, for the functions $f = \varphi_\alpha \in \mathcal{F}$ defined by

$$\varphi_\alpha(u) = \begin{cases} u - 1 - \ln u & \text{if } \alpha = 0 \\ \dfrac{\alpha u + 1 - \alpha - u^\alpha}{\alpha(1-\alpha)} & \text{if } \alpha \in I\!R \backslash \{0,1\} \\ 1 - u + u \ln u & \text{if } \alpha = 1 \end{cases}$$

(cf. Section 2 in Liese and Vajda (1987) or Read and Cressie (1988)) were studied by Lindsay (1994). This class yields only for $\alpha = 1/2$ a metric divergence, namely the squared Hellinger distance (cf. e.g. Basu and Lindsay (1994)). The family of estimates $\theta^*_{\alpha,n}$, $\alpha \in I\!R$, obtained by replacing the functions $f_\beta \in \mathcal{F}$ by $\varphi_\alpha \in \mathcal{F}$ obviously differs from the family $\theta_{\beta,n}$, $\beta \in I\!R_0$, of estimates given above.

Lindsay (1994) introduced in the model under consideration the robustness of estimators against outliers (and inliers). Members f_β of our class satisfy, owing to $\lim_{u\to\infty} \frac{f_\beta(u)}{u} = f^*_\beta(0)$, (f2) and (f3)

$$f_\beta(0) = \lim_{u\to\infty} \frac{f_\beta(u)}{u} \in (0, \infty)$$

and therefore Assumption 10 in Lindsay (1994). (Note that Lindsay's functions G corresponds to our functions f via $G(\delta) \equiv f(\delta + 1)$.) Consequently his Proposition 12 yields that our class $I_{f_\beta}(Q, P)$, $\beta \in I\!R_0$, of metric divergences fulfils Lindsay's outlier stability property (21).

References

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another, *J. Roy. Statist. Soc. Ser. B*, **28**, 131–142.

Arimoto, S. (1971). Information-theoretical considerations on estimation problems, *Information and Control*, **19**, 181–194.

Barron, A. R., Győrfi, L. and van der Meulen, E. (1992). Distribution estimates consistent in total variation and in two types of information divergence, *IEEE Trans. Inform. Theory*, **38**, 1437–1454.

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: Efficiency, distribution and robustness, *Ann. Inst. Statist. Math.*, **46**, 683–705.

Beran, R. (1977). Minimum Hellinger distance estimates for parameteric models, *Ann. Statist.*, **5**, 445–463.

Beran, R. (1978). An efficient and robust adaptive estimator of location, *Ann. Statist.*, **6**, 292–313.

Berlinet, A., Devroye, L. and Győrfi, L. (1995). Asymptotic normality of L_1-error in density estimation, *Statistics*, **26**, 329–343.

Berlinet, A., Vajda, I. and van der Meulen, E. (1997). About the asymptotic accuracy of Barron density estimator, *IEEE Trans. Inform. Theory* (submitted).

Boekee, D. E. (1977). *A Generalization of the Fisher Information Measure*, Delft University Press, Delft.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Publ. Math. Inst. Hungar. Acad. Sci.*, **8**, 85–107.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Ann. Probab.*, **3**(1), 146–158.

Csiszár, I. and Fischer, J. (1962). Informationsentfernungen im Raum der Wahrscheinlichkeitsverteilungen, *Magyar Tud. Akad. Mat. Kutató Int. Kösl.*, **7**, 159–180.

Cutler, A. and Cordero-Brana, O. I. (1996). Minimum Hellinger distance estimation for finite mixture models, *J. Amer. Statist. Assoc.*, **91**, 1716–1723.

Devroye, L. and Győrfi, L. (1985). *Nonparametric Density Estimation: The L_1-View*, Wiley, New York.

Devroye, L. and Győrfi, L. (1990). No empirical measure can converge in total variation sense for all distributions, *Ann. Statist.*, **18**, 1496–1499.

Feldman, D. and Österreicher, F. (1989). A note on f-divergences, *Studia Sci. Math. Hungar.*, **24**, 191–200.

Győrfi, L., Vajda, I. and van der Meulen, E. (1994). Minimum Hellinger distance point estimates consistent under weak family regularity, *Math. Methods Statist.*, **3**, 25–45.

Győrfi, L., Vajda, I. and van der Meulen, E. (1996). Minimum Kolmogorov distance estimates of parameters and parametrized distributions, *Metrika*, **43**, 237–255.

Hall, P. and Patil, P. (1995). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators, *Ann. Statist.*, **23**(3), 905–928.

Kafka, P., Österreicher, F. and Vincze, I. (1991). On powers of f-divergences defining a distance, *Studia Sci. Math. Hungar.*, **26**, 415–422.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*, Teubner-Texte zur Mathematik, Band 95, Teubner, Leipzig.

Lin, J. (1991). Divergence measures based on the Shannon entropy, *IEEE Trans. Inform. Theory*, **37**, 145–151.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger and related methods, *Ann. Statist.*, **22**(2), 1081–1114.

Matusita, K. (1955). Decision rules based on the distance for problems of fit, two samples and estimation, *Ann. Math. Statist.*, **26**, 631–640.

Matusita, K. (1964). Distances and decision rules, *Ann. Inst. Statist. Math.*, **16**, 305–320.

Morales, D., Pardo, L. and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions, *J. Statist. Plann. Inference*, **47**, 347–369.

Morales, D., Pardo, L. and Vajda, I. (1996). Uncertainty of discrete stochastic systems: General theory and statistical inference, *IEEE Trans. Systems, Man and Cybernetics*, **26**, 681–697.

Österreicher, F. (1982). The construction of least favourable distributions is traceable to a minimal perimeter problem, *Studia Sci. Math. Hungar.*, **17**, 341–351.

Österreicher, F. (1992). The risk set of a testing problem—A vivid statistical tool, *Transactions of the Eleventh Prague Conference*, Vol. A, 175–188, Academia, Prague.

Österreicher, F. (1996). On a class of perimeter-type distances of probability distributions, *Kybernetika*, **32**, 389–393.

Österreicher, F. and Vajda, I. (1993). Statistical information and discrimination, *IEEE Trans. Inform. Theory*, **39**(3), 1036–1039.

Pak, R. J. (1996). Minimum Hellinger distance estimation in simple linear regression models; distribution and efficiency, *Statist. Probab. Lett.*, **26**, 263–269.

Read, T. C. R. and Cressie, N. A. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer, New York.

Reschenhofer, E. and Bomze, I. M. (1991). Length tests for goodness of fit, *Biometrika*, **78**, 207–216.

Tamura, R. D. and Boos, D. D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance, *J. Amer. Statist. Assoc.*, **81**, 223–229.

Vincze, I. (1981). On the concept and measure of information contained in an observation, *Contributions to Probability* (eds. J. Gani and V. F. Rohatgi), 207–214, Academic Press, New York.