Václav Kůs; Domingo Morales; Igor Vajda

Extensions of the parametric families of divergences used in statistical inference

Kybernetika, Vol. 44 (2008), No. 1, 95--112

Persistent URL: http://dml.cz/dmlcz/135836

KYBERNETIKA — VOLUME 44 (2008), NUMBER 1, PAGES 95 – 112

EXTENSIONS OF THE PARAMETRIC FAMILIES OF DIVERGENCES USED IN STATISTICAL INFERENCE

Václav Kůs, Domingo Morales and Igor Vajda

We propose a simple method of construction of new families of ϕ-divergences. This method called convex standardization is applicable to convex and concave functions $\psi(t)$ twice continuously differentiable in a neighborhood of $t = 1$ with nonzero second derivative at the point $t = 1$. Using this method we introduce several extensions of the LeCam, power, χ^a and Matusita divergences. The extended families are shown to connect smoothly these divergences with the Kullback divergence or they connect various pairs of these particular divergences themselves. We investigate also the metric properties of divergences from these extended families.

Keywords: divergences, metric divergences, families of f-divergences

AMS Subject Classification: 62B05, 62H30

1. INTRODUCTION

Statistical inference widely uses the divergences of probability distributions P, Q with densities $p = \mathrm{d}P/\mathrm{d}\mu$ and $q = \mathrm{d}Q/\mathrm{d}\mu$ on a measurable observation space $(\mathcal{X}, \mathcal{A})$ given by the formula

$$D_\phi(P, Q) = \int_{\mathcal{X}} q\, \phi\left(\frac{p}{q}\right) \mathrm{d}\mu \tag{1}$$

for $\phi(t)$ convex in the interval $(0, \infty)$ being equal to zero and strictly convex at $t = 1$. These divergences are called ϕ-divergences and the best known examples are the *total variation* (L_1-distance)

$$V(P, Q) = \int |p - q|\, \mathrm{d}\mu, \tag{2}$$

the *squared Hellinger distance*

$$H^2(P, Q) = \int \left(\sqrt{p} - \sqrt{q}\right)^2 \mathrm{d}\mu, \tag{3}$$

the χ^2*-divergence*

$$\chi^2(P, Q) = \int \frac{(p - q)^2}{q}\, \mathrm{d}\mu \tag{4}$$

and the *I-divergence* (information divergence, Kullback divergence)

$$I(P,Q) = \int p \ln \frac{p}{q} \, \mathrm{d}\mu. \tag{5}$$

Note that the integrands of (1) – (5) are extended from the domain $p > 0$, $q > 0$ to $p \geq 0$, $q \geq 0$ by preserving the convexity and continuity of the function $q\,\phi(p/q)$ of two variables (if the continuity cannot be preserved then the lower semicontinuity is required) and that this extension is unique. For this detail about definition and for the basic properties of ϕ-divergences we refer to Vajda [21] and Liese and Vajda [10, 11].

The divergences (3) – (5) can be found (eventually in a slightly rescaled form) in the class of the so-called *power divergences*

$$D_\alpha(P,Q) = D_{\varphi_\alpha}(P,Q), \qquad \alpha \in \mathbb{R} \tag{6}$$

where

$$\varphi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, \qquad t > 0 \tag{7}$$

for $\alpha \neq 0, \alpha \neq 1$ and the corresponding limits

$$\varphi_1(t) = t \ln t - t + 1, \quad \varphi_0(t) = -\ln t + t - 1 \tag{8}$$

satisfy the assumptions imposed on ϕ in (1). The particular case $D_{1/2}(P,Q) = 2H^2(P,Q)$ is *Hellinger divergence*, $D_2(P,Q) = \chi^2(P,Q)/2$ is *Pearson divergence* and $D_1(P,Q) = I(P,Q)$ is *Kullback divergence*. Further,

$$D_{-1}(P,Q) = \chi^2(Q,P)/2 \tag{9}$$

is a reversed Pearson divergence $D_2(Q,P)$ known also as a *Neyman divergence* and

$$D_0(P,Q) = I(Q,P) \tag{10}$$

is a reversed Kullback divergence $D_1(Q,P)$. Since the Kullback divergence was in fact introduced in the joint paper [6] of Kullback and Leibler, it seems to be convenient to call (10) a *Leibler divergence*.

Statistical applications of the divergences (1) – (10) were studied e. g. in the monographs of Read and Cressie [19], Vajda [21] and Pardo [18], and in the papers of Morales et al. [13, 14, 15], Vajda and van der Meulen [23], Beirlant et al. [1], Györfi and Vajda [3] and others cited there. Divergences $D_\phi(P,\hat{P})$ or $D_\alpha(P,\hat{P})$ between a hypothetic distribution P and an observations-based empirical distribution $\hat{P}$ are basic tools for the minimum divergence estimation of parameters of P and for the minimum divergence testing of statistical hypotheses about P.

The cited books and papers usually verified practical value of methods and results established for general ϕ-divergences by applying them to or testing them on special simply parametrized families such as the power divergences (6). In the present paper we propose a number of new simply parametrized families found by a special extension procedure. These families connect smoothly pairs of well known and

extensively applied divergences of different characteristic properties. The smooth transition of properties may be used in applications (e. g. in the statistical minimum divergence estimation and testing) by selecting divergences with most desirable properties. These properties are usually carefully weighted compromises between the properties of the connected pairs of ϕ-divergences, but on the trajectories connecting some pairs one sometimes meets ϕ-divergences with qualitatively new properties diametrically different from the properties of both members of the connected pair (see e. g. the application of the family (iii) below in robust statistical inference).

Our extension procedure is a *convex standardization* of convex or concave functions $\psi : (0, \infty) \to \mathbb{R}$ twice continuously differentiable in a neighborhood of $t = 1$ with the second derivative $\psi''(1) \neq 0$. The standard convex form of ψ (briefly, a *convex standard* of ψ) is defined by the formula

$$\phi(t) = \frac{\psi(t) - \psi(1) - \psi'(1)\,(t-1)}{\psi''(1)}, \qquad t > 0. \tag{11}$$

This function belongs the class $\mathbf{\Phi}$ of all convex functions $\phi : (0, \infty) \to \mathbb{R}$ twice continuously differentiable in a neighborhood of $t = 1$ with $\phi(1) = \phi'(1) = 0$ and $\phi''(1) = 1$. Obviously, functions $\phi \in \mathbf{\Phi}$ are strictly convex at $t = 1$ and also in the neighborhood of $t = 1$. If $\psi : (0, \infty) \to \mathbb{R}$ is twice continuously differentiable on $(0, \infty)$ and

$$\psi''(t) = 0 \qquad \text{for no } t > 0 \tag{12}$$

then it satisfies the conditions assumed in the convex standardization formula (11). Indeed, then ψ is either strictly convex or strictly concave on $(0, \infty)$ and $\psi''(1) \neq 0$. Examples of some functions ψ together with their convex standards $\phi \in \mathbf{\Phi}$ and corresponding ϕ-divergences are given in Table 1 below.

Table 1. ψ-functions for standard ϕ-divergences.

	Kullback (D_1)	**Leibler** (D_0)	**Pearson**(D_2)	**Neyman** (D_{-1})
$\boldsymbol{\psi(t)}$	$t \ln t$	$\ln t$	t^2	$\frac{1}{t}$
$\boldsymbol{\phi(t)}$	$t \ln t - t + 1$	$-\ln t + t - 1$	$(t-1)^2/2$	$\frac{(t-1)^2}{2\,t}$

	Le Cam (LC^2)	**Hellinger** $(D_{1/2})$	**Power** (D_α)
$\boldsymbol{\psi(t)}$	$\frac{1}{1+t}$	$\sqrt{t}$	$t^\alpha,\ \alpha \neq 0, 1$
$\boldsymbol{\phi(t)}$	$\frac{(t-1)^2}{t+1}$	$2\,(\sqrt{t}-1)^2$	$\frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}$

Remark 1. Sometimes it is convenient to multiply the function $\phi(t)$ of (11) by a positive constant factor. The modified function ϕ remains to be convex with $\phi(1) = \phi'(1) = 0$ and leads to a ϕ-divergence modified by the same factor.

Let us note that the method of convex standardization was first used in the research report [22] but majority of the families introduced in this paper together with their basic properties are new. They contain as special cases some or all of the well known classical ϕ-divergences presented in Table 1.

2. EXTENDED POWER DIVERGENCES

In this section we apply the convex standardization to the family of functions

$$\psi_{\alpha,\beta}(t) = \frac{1}{\beta t^\alpha + 1 - \beta}, \qquad t > 0 \tag{13}$$

for parameters (α, β) from a suitable subset $A \subset \mathbb{R}^2$. Obviously, the functions $\psi_{0,\beta}$ and $\psi_{\alpha,0}$ are constant on $(0, \infty)$ and therefore do not satisfy the assumption $\psi''(1) \neq 0$ of (11). Further, if $\alpha \neq 0$ then $\beta > 1$ or $\beta < 0$ lead for some $t = t_{\alpha,\beta} > 0$ to

$$\xi_{\alpha,\beta}(t) = 0 \qquad \text{where} \qquad \xi_{\alpha,\beta}(t) = \beta t^\alpha + 1 - \beta. \tag{14}$$

Hence $\alpha \neq 0$ and $\beta \in (0, 1]$ are necessary conditions for the parameters α, β to guarantee that $\xi_{\alpha,\beta}(t) > 0$ for every $t > 0$. As

$$\psi''_{\alpha,\beta}(t) = \frac{\alpha\beta t^{\alpha-2}[2\alpha\beta t^\alpha - (\alpha - 1)\xi_{\alpha,\beta}(t)]}{\xi^3_{\alpha,\beta}(t)}, \tag{15}$$

the condition (12) is fulfilled if $(\alpha, \beta) \in A_{-1} \cup A_1 \cup I$ for the subsets

$$A_{-1} = [-1, 0)\times(0, 1]-\{(-1, 1)\}, \quad A_1 = (0, 1]\times(0, 1] \quad \text{and} \quad I = \{(\alpha, 1) : \alpha \neq -1, 0\}$$

of $\mathbb{R}^2$.

Hence we get from (11) for all $(\alpha, \beta) \in A_{-1} \cup A_1 \cup I$ the convex standards

$$\phi_{\alpha,\beta}(t) = \frac{1}{\alpha(2\alpha\beta - \alpha + 1)} \left[\frac{1 - t^\alpha}{\beta t^\alpha + 1 - \beta} + \alpha(t - 1)\right], \qquad t > 0 \tag{16}$$

belonging to the set $\mathbf{\Phi}$. In the sequel we need also the sets

$$B_{-1} = [-1, 0) \times [0, 1] - \{(-1, 1)\}, \qquad B_1 = (0, 1] \times [0, 1] - \{(1, 0)\}$$

and

$$C = [-1, 1] \times [0, 1] - \{(-1, 1), (1, 0)\}$$

satisfying the inclusions

$$A_{-1} \cup A_1 \subset B_{-1} \cup B_1 \subset C.$$

The sets figuring in these inclusions are illustrated in Figures 1 – 3. The following properties hold for every $t > 0$.

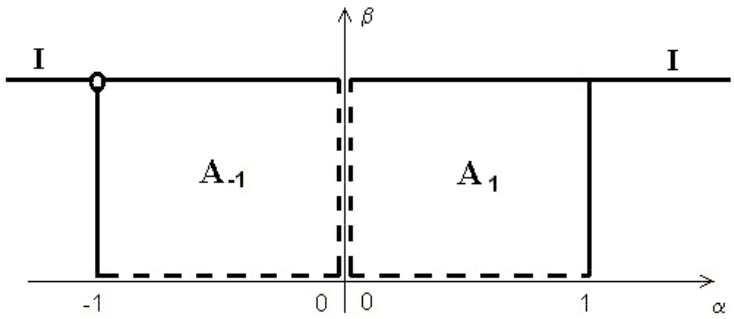

Fig. 1. The set $A_{-1} \cup A_1 \cup I$.

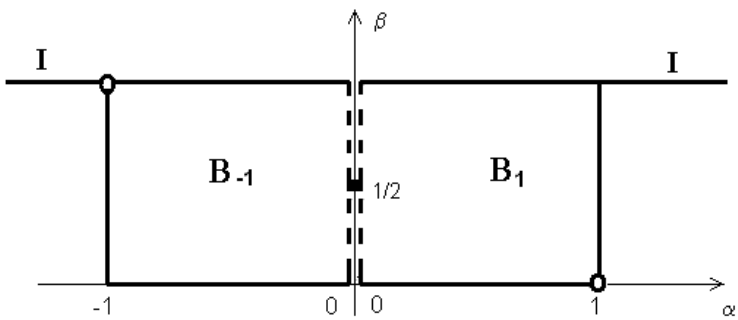

Fig. 2. The set $B_{-1} \cup B_1$ and the point $(0, 1/2)$.

(i) $\phi_{\alpha,\beta}(t)$ is continuous in the variables $(\alpha, \beta) \in A_{-1} \cup A_1 \cup I$.

(ii) $\phi_{\alpha,\beta}$ of (16) can be applied to all $(\alpha, \beta) \in B_{-1} \cup B_1$ and the extensions defined for all $\alpha \in [-1, 0) \cup (0, 1)$ by

$$\phi_{\alpha,0}(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha - 1)}, \qquad t > 0, \tag{17}$$

belong to $\mathbf{\Phi}$.

(iii) The values $\phi_{\alpha,\beta}(t)$ remain to be symmetric about $(0, 1/2) \in \mathbb{R}^2$ in the extended variables $(\alpha, \beta) \in B_{-1} \cup B_1 \subset \mathbb{R}^2$ in the sense

$$\phi_{\alpha,\beta}(t) = \phi_{-\alpha,1-\beta}(t). \tag{18}$$

It follows from (i) – (iii) that the extended $\phi_{\alpha,\beta}(t)$ defined by (16) for all parameters $(\alpha, \beta) \in B_{-1} \cup B_1$ remains to be continuous in these parameters.

(iv) For every $\beta \in [0, 1]$ there exists a limit $\phi_{0,\beta}(t)$ of $\phi_{\alpha,\widetilde{\beta}}(t)$ for $(\alpha, \widetilde{\beta}) \in B_{-1} \cup B_1$ tending to the point $(0, \beta) \in \mathbb{R}^2$. This limit is not depending on β and obviously

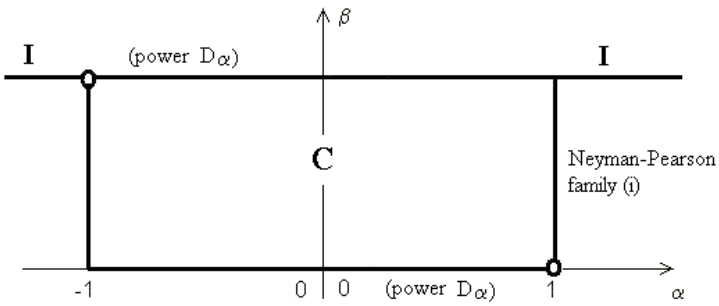

Fig. 3. The set C.

satisfies the relation

$$\phi_{0,\beta}(t) = \lim_{\alpha \to 0} \phi_{\alpha,\beta}(t) = \varphi_0(t), \qquad t > 0 \tag{19}$$

where $\varphi_0(t)$ was given in (8).

The formulas (16), (17) and (19) define the functions $\phi_{\alpha,\beta}(t)$ of variable $t > 0$ for all parameters $(\alpha, \beta) \in C$ where $C \subset \mathbb{R}^2$ is given in Figure 3. It follows from (iv) and from what has been said before (iv) that $\phi_{\alpha,\beta}(t)$ is continuous on C in the parameters (α, β). Further, since $\phi_{0,\beta}(t)$ is symmetric about $1/2$ in the variable $\beta \in [0, 1]$ in the sense

$$\phi_{0,\beta}(t) = \phi_{0,1-\beta}(t),$$

the symmetry (18) extends to all $(\alpha, \beta) \in C$.

Note that $\phi_{\alpha,\beta}(t)$ cannot be continuously extended to the corners $(-1, 1)$ and $(1, 0)$ of the set C. Namely, it follows from (16) and (18) that

$$\lim_{\alpha \to -1} \phi_{\alpha,1}(t) = \lim_{\alpha \to 1} \phi_{\alpha,0}(t) = \varphi_1(t) \tag{20}$$

and

$$\lim_{\beta \uparrow 1} \phi_{-1,\beta}(t) = \lim_{\beta \downarrow 0} \phi_{1,\beta}(t) = \varphi_2(t) \tag{21}$$

where $\varphi_1(t), \varphi_2(t)$ are different functions given by (7), (8).

We can conclude from (16), (17) and (19) that the following assertion holds.

Proposition 1. The class of convex functions

$$\{\phi_{\alpha,\beta} : (\alpha, \beta) \in C\} \subset \mathbf{\Phi} \tag{22}$$

contains the mutually equal subclasses

$$\{\phi_{\alpha,0} = \varphi_\alpha : \alpha \in [-1, 1)\} = \{\phi_{\alpha,1} = \varphi_{-\alpha} : \alpha \in (-1, 1]\} \tag{23}$$

of power divergence functions given by (7) and (8). Therefore (22) is an extension of the class $\{\varphi_\alpha : \alpha \in [-1, 1)\}$ of power divergence functions, and the corresponding $\phi_{\alpha,\beta}$-divergences

$$\mathcal{D}_{\alpha,\beta}(P, Q) = D_{\phi_{\alpha,\beta}}(P, Q) = \int q\phi_{\alpha,\beta}\left(\frac{p}{q}\right) \mathrm{d}\mu, \quad (\alpha, \beta) \in C \tag{24}$$

are extensions of the power divergences $D_\alpha(P, Q)$, $\alpha \in [-1, 1)$.

Let us now consider the extension

$$C^* = C \cup ((-\infty, -1) \cup [1, \infty) \times \{0\}) \cup ((-\infty, -1] \cup (1, \infty) \times \{1\})$$

of the set C and the functions $\phi_{\alpha,0} = \varphi_\alpha$ for $(\alpha, 0) \in C^* \setminus C$ and $\phi_{\alpha,1} = \varphi_{-\alpha}$ for $(\alpha, 1) \in C^* \setminus C$. Then the following assertion obviously holds which, together with Proposition 1, justifies the title of the present section.

Proposition 2. The class of convex functions

$$\{\phi_{\alpha,\beta} : (\alpha, \beta) \in C^*\} \subset \Phi \tag{25}$$

contains all power divergence functions φ_α, $\alpha \in \mathbb{R}$ and the class of divergences

$$\{\mathcal{D}_{\alpha,\beta}(P, Q) : (\alpha, \beta) \in C^*\} \tag{26}$$

defined by (21) contains all power divergences $D_\alpha(P, Q)$, $\alpha \in \mathbb{R}$.

In addition to the family (23) of power divergence functions and the corresponding power divergences $D_\alpha(P, Q)$, $\alpha \in [-1, 1)$, one can find some other interesting one-parameter families of functions in the class (22) and the corresponding families of divergences in the class (24).

(i) Neyman–Pearson family. One such family of functions is

$$\phi_{1,\beta}(t) = \frac{1}{2}\,\frac{(t-1)^2}{\beta t + 1 - \beta}, \qquad \beta \in (0, 1] \tag{27}$$

with the corresponding family of divergences

$$\mathcal{D}_{1,\beta}(P, Q) = \frac{1}{2}\int \frac{(p-q)^2}{\beta p + (1-\beta)q}\,\mathrm{d}\mu, \qquad \beta \in (0, 1]. \tag{28}$$

Since

$$\mathcal{D}_{1,1/2}(P, Q) = LC^2(P, Q) = \int \frac{(p-q)^2}{p+q}\,\mathrm{d}\mu \quad \text{(cf. Table 1)} \tag{29}$$

this family contains the *Le Cam divergence* (the squared Le Cam distance, see Chapter 4.2 in Le Cam [9]). This divergence belongs to the class of divergences (1) defined for functions $\phi(t) = |t-1|^\alpha/(t+1)^{\alpha-1}$ for $\alpha \geq 1$. As proved by Kafka et al. [5], the

$(1/\alpha)$th roots of the corresponding divergences are metrics in the space of probability distributions P, Q. By (23), the upper extremes in the classes (27), (28) are

$$\phi_{1,1} = \varphi_{-1} \qquad \text{and} \qquad \mathcal{D}_{1,1}(P,Q) = D_{-1}(P,Q) = \chi^2(Q,P)/2.$$

The lower extremes $\phi_{1,0}$ and $\mathcal{D}_{1,0}(P,Q)$ are undefined in (22), (24) since the point $(1,0)$ is not in C. This point is in the above introduced extended set C^*, but the corresponding

$$\phi_{1,0} = \varphi_1 \qquad \text{and} \qquad \mathcal{D}_{1,0}(P,Q) = D_1(P,Q) = I(P,Q)$$

considered in (25) and (26) are not continuous extensions of this family. By (21), such extensions are

$$\phi_{1,0} = \varphi_2 \qquad \text{and} \qquad \mathcal{D}_{1,0}(P,Q) = D_2(P,Q) = \chi^2(P,Q)/2. \tag{30}$$

These extensions together with (27), (28) define a complete *Neyman–Pearson family* $\{\phi_{1,\beta} : \beta \in [0,1]\}$ and $\{\mathcal{D}_{1,\beta}(P,Q) : \beta \in [0,1]\}$ respectively. This name of the family comes out of the fact that it smoothly connects the Pearson divergence $\mathcal{D}_{1,0}(P,Q) = D_2(P,Q)$ with the Neyman divergence $\mathcal{D}_{1,1}(P,Q) = D_{-1}(P,Q)$. By (29), this connection is passing through the Le Cam divergence (29).

Note that here and in the sequel the smoothness means the continuity of the functions $\phi_{1,\beta}$ in the parameter $\beta \in [0,1]$. By the Lebesgue dominated convergence theorem for integrals, this implies a similar continuity of the divergences

$$\mathcal{D}_{1,\beta}(P,Q) = \int q\phi_{1,\beta}\left(\frac{p}{q}\right) \mathrm{d}\mu$$

provided P, Q satisfy some assumptions. Simple and relatively mild assumptions are that $P = (p_1, \ldots, p_K)$, $Q = (q_1, \ldots, q_K)$ are positive discrete probability distributions. Then for any $\phi_\beta \in \mathbf{\Phi}$ with ϕ_β continuous in a real parameter β, the ϕ_β-divergence $D_{\phi_\beta}(P,Q)$ is the sum

$$D_{\phi_\beta}(P,Q) = \sum_{k=1}^{K} q_k \phi_\beta\left(\frac{p_k}{q_k}\right) \tag{31}$$

which is continuous in β too.

(ii) Kullback–Pearson family. Another interesting one-parameter family of functions in the class (22) with the corresponding family divergences in the class (24) is the *Kullback–Pearson family* parametrized by $\gamma \in [0,\infty]$. For $\gamma \in [0,\infty)$ it is obtained from the extended power divergence family (22) by the rule

$$\phi_\gamma(t) = \lim_{\alpha \uparrow 1} \phi_{\alpha,(1-\alpha)\gamma}(t), \qquad t > 0$$

where $\phi_{\alpha,(1-\alpha)\gamma}$ is given by (16) for $\alpha \in (0,1)$ and $\beta = (1-\alpha)\gamma \in [0,1)$. Extension to $\gamma = \infty$ is obtained from the continuity rule,

$$\phi_\infty(t) = \lim_{\gamma \to \infty} \phi_\gamma(t), \qquad t > 0.$$

If $\gamma \in [0, \infty)$ then substituting for β in (16) and taking the limit for $\alpha \uparrow 1$ we obtain for every $t > 0$

$$\phi_\gamma(t) = \frac{t \ln t + t - 1 + \gamma(t-1)^2}{2\gamma + 1} = \frac{\varphi_1(t) + 2\gamma\varphi_2(t)}{2\gamma + 1} \tag{32}$$

so that the extremes

$$\phi_0 = \varphi_1 \qquad \text{and} \qquad \phi_\infty = \varphi_2 \tag{33}$$

are the Kullback and Pearson divergence functions specified in (6), (7). Since φ_1, φ_2 belong to the class $\mathbf{\Phi}$, it is clear that all ϕ_γ, $\gamma \in [0, \infty]$ belong to $\mathbf{\Phi}$ too. The corresponding divergences are

$$D_{\phi_\gamma}(P, Q) = \frac{D_1(P, Q) + 2\gamma D_2(P, Q)}{2\gamma + 1} = \frac{I(P, Q) + \gamma\chi^2(P, Q)}{2\gamma + 1} \tag{34}$$

if $\gamma \in [0, \infty)$ and

$$D_{\phi_\infty}(P, Q) = \chi^2(P, Q)/2. \tag{35}$$

The Kullback–Pearson mixed family $\{D_{\phi_\gamma}(P, Q) : \gamma \in [0, \infty]\}$ smoothly connects the Kullback divergence $D_1(P, Q) = I(P, Q)$ with the Pearson divergence $D_2(P, Q) = \chi^2(P, Q)/2$ in a linear manner. This differs from the nonlinear connection in the power divergence subfamily $\{D_\alpha(P, Q) : \alpha \in [1, 2]\}$. The advantage of the linearity in some computations is obvious.

(iii) Leibler–Neyman family. Another interesting subfamily of (22) is the *Leibler–Neyman family* $\{\phi_{\alpha,(1-2\alpha)^2} : \alpha \in [0, 1]\}$ where the extremes are

$$\phi_{0,1} = \varphi_0, \qquad \phi_{1,1} = \varphi_{-1}$$

and for $\alpha = 1/2$ we obtain $\phi_{1/2,0} = \varphi_{1/2}$. Therefore the corresponding Leibler–Neyman family of divergences $\{\mathcal{D}_{\alpha,(1-2\alpha)^2}(P, Q) : \alpha \in [0, 1]\}$ smoothly connects the above introduced Leibler divergence $\mathcal{D}_{0,1}(P, Q) = D_0(P, Q)$ with the Neyman divergence $\mathcal{D}_{1,1}(P, Q) = D_{-1}(P, Q)$ and passes through the Hellinger divergence

$$\mathcal{D}_{1/2,0}(P, Q) = D_{1/2}(P, Q).$$

Such connection is impossible in the class of power divergences $\{D_\alpha(P, Q) : \alpha \in [-1, 0]\}$ because $\alpha = 1/2$ is out of the interval $[-1, 0]$.

To illustrate statistical applicability of the Leibler–Neyman family of divergences, consider empirical relative frequencies $\widehat{P} = (\widehat{p}_1, \ldots, \widehat{p}_K)$ of n i.i.d. observations in K disjoint $\mathcal{A}$-measurable cells covering the assumed observation space $\mathcal{X}$. Let $P = (p_1, \ldots, p_K)$ be a hypothetic probability distribution on these cells. Then

$$T_{0,n} = 2n\mathcal{D}_{0,1}(P, \widehat{P}) = 2n \sum_{k=1}^{K} \widehat{p}_k \ln \frac{\widehat{p}_k}{p_k}$$

is the log-likelihood ratio statistics,

$$T_{1,n} = 2n\mathcal{D}_{1,1}(P, \widehat{P}) = n \sum_{k=1}^{K} \frac{(\widehat{p}_k - p_k)^2}{p_k}$$

is the Pearson statistic and

$$T_{1/2,n} = 2n\mathcal{D}_{1/2,0}(P, \widehat{P}) = 8n \left(1 - \sum_{k=1}^{K} \sqrt{\widehat{p}_k p_k} \right) \tag{36}$$

is the Freeman–Tukey statistic (see e. g. Read and Cressie [19]) which differs from the previous two non-robust statistics by being robust (in the sense of Lindsay [12], cf. also Kůs [8]). We see that the family of Leibler–Neyman statistics

$$T_{\alpha,n} = 2n\mathcal{D}_{\alpha,(1-2\alpha)^2}(P, \widehat{P}) = 2n \sum_{k=1}^{K} \widehat{p}_k \phi_{\alpha,(1-2\alpha)^2} \left(\frac{p_k}{\widehat{p}_k} \right), \quad \alpha \in [0, 1]$$

smoothly connects the famous efficient but nonrobust statistics $T_{0,n}$ and $T_{1,n}$ by passing through the robust but less efficient statistic $T_{1/2,n}$. Similar connection of $T_{0,n}$ and $T_{1,n}$ is impossible in the class

$$\{2nD_\alpha(P, \widehat{P}) : \alpha \in [-1, 0]\}$$

of power divergence statistics. It smoothly connects the statistics

$$2nD_{-1}(P, \widehat{P}) = T_{1,n} \qquad \text{and} \qquad 2nD_0(P, \widehat{P}) = T_{0,n}$$

too but without containing $T_{1/2,n}$ or any other statistic robust in the above mentioned sense.

The interesting subfamilies of (22) are not exhausted by those listed above. For example the simple family $\{\phi_{1/2,\beta} : \beta \in [0, 1]\}$ leads to the divergences $\mathcal{D}_{1/2,\beta}(P, Q)$ for which the family of statistics

$$\left\{ U_{\beta,n} = 2n\mathcal{D}_{1/2,\beta}(P, \widehat{P}) : \beta \in [0, 1] \right\} \tag{37}$$

smoothly connects the Freeman–Tukey statistic $U_{0,n} = T_{1/2,n}$ given in (36) with the power divergence statistic

$$U_{1,n} = 2nD_{-1/2}(P, \widehat{P}) = \frac{8n}{3} \left(\sum_{k=1}^{K} \sqrt{\frac{\widehat{p}_k^3}{p_k}} - 1 \right)$$

shown in Read and Cressie [19] to be locally most powerful among all power divergence statistics $2nD_\alpha(P, \widehat{P})$, $\alpha \in \mathbb{R}$. Thus in the relatively simple and relatively narrow class (37) one can easily find reasonable compromises between the power and robustness of the power divergence statistics.

3. EXTENDED ABSOLUTE POWER DIVERGENCES

Vajda [20] used the absolute power functions $\phi(t) = |t-1|^\alpha$ defined on $(0,\infty)$ for $\alpha \geq 1$ and belonging to Φ to introduce the family of absolute power divergences

$$\chi^\alpha(P,Q) = \int \frac{|p-q|^\alpha}{p^{\alpha-1}} \mathrm{d}\mu, \qquad \alpha \geq 1. \tag{38}$$

This family contains the total variation $\chi^1(P,Q) = V(P,Q)$ and the Pearson divergence $\chi^2(P,Q)$ as particular cases. The divergences (38) help to generalize the Cramer–Rao inequality and to introduce the Fisher information of orders $\alpha > 1$ where the order $\alpha = 2$ means the classical Fisher information. As demonstrated e. g. in Hobza et al. [4], the Fisher informations of some orders $\alpha \neq 2$ are useful when the classical Fisher information is trivial or does not exist.

In this section the class $\{\chi^\alpha(P,Q) : \alpha > 0\}$ is extended by applying the convex standardization (11) to the functions

$$\psi_{\alpha,\beta}(t) = |t+\beta-1|^\alpha, \qquad t > 0 \tag{39}$$

for suitable parameters α, β. Since the function ϕ given in (11) is standardized in the sense that $\phi''(1) = 1$, the divergences (38) will be rescaled in the extended class. Obviously, $\psi_{\alpha,\beta}$ of this section differs from that of (13). Therefore we must pay attention to what section we have in mind when speaking about $\psi_{\alpha,\beta}$ (and about $\phi_{\alpha,\beta}$ resulting from $\psi_{\alpha,\beta}$ in the convex standardization (11)). The present function $\psi_{\alpha,\beta}$ is trivial for $\alpha = 0$ and not defined in the whole interval $(0,\infty)$ for $\alpha < 0$, $\beta < 1$. A detailed analysis shows that this function satisfies the assumptions of (11) only for $(\alpha,\beta) \in A$ where A is the union of four subsets of the plane $\mathbb{R}^2$ (see Figure 4), namely

$$\begin{aligned} A_1 &= (-\infty,0) \times [1,\infty), & A_2 &= (0,1) \times [1,\infty), \\ A_3 &= (1,\infty) \times [0,\infty), & A_4 &= (1,\infty) \times (-\infty,0). \end{aligned}$$

Thus we get from (11) the family of functions

$$\phi_{\alpha,\beta}(t) = \frac{|t+\beta-1|^\alpha - |\beta|^\alpha - \alpha\,\mathrm{sign}(\beta)\,|\beta|^{\alpha-1}(t-1)}{\alpha(\alpha-1)}, \quad (\alpha,\beta) \in A \tag{40}$$

where we drop out the positive factor $|\beta|^{\alpha-2}$ from denominator of (40) (see Remark 1 in Section 1).

The limits

$$\phi_{\alpha,0}(t) = \frac{|t-1|^\alpha}{\alpha(\alpha-1)} \qquad \text{for} \qquad \alpha \in (1,\infty), \tag{41}$$

$$\phi_{0,\beta}(t) = \frac{t-1}{\beta} - \ln\frac{t+\beta-1}{\beta} \qquad \text{for} \qquad \beta \in [1,\infty) \tag{42}$$

and

$$\phi_{1,\beta}(t) = |t+\beta-1| \ln \frac{|t+\beta-1|}{\beta^{\mathrm{sgn}(\beta)}} - \mathrm{sgn}(\beta)(t-1) \tag{43}$$

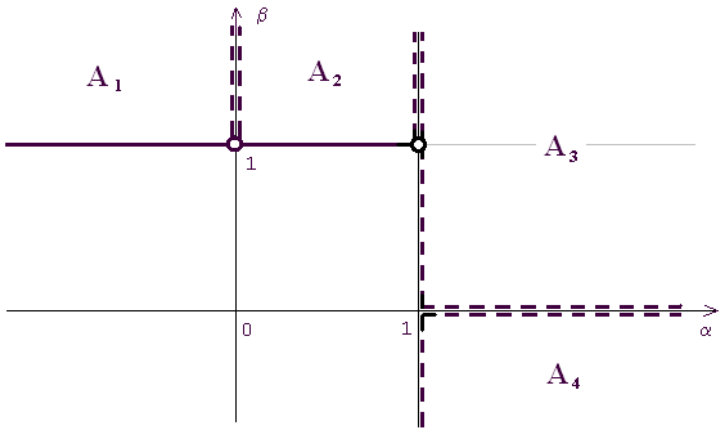

Fig. 4. The set $A_1 \cup A_2 \cup A_3 \cup A_4$.

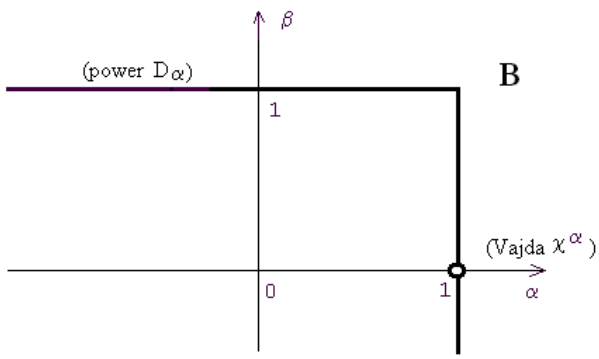

Fig. 5. The set B.

for $\beta \neq 0$, with the convention $0 \ln 0 = 0$, define continuous extension of the family (40) on the set

$$B = \overline{A} - \{(1,0)\} = \mathbb{R}^2 - ((-\infty, 1) \times (-\infty, 1) \cup \{(1,0)\})$$

where $\overline{A}$ is the closure of A (see Figure 5). The family

$$\{\phi_{\alpha,\beta} : (\alpha, \beta) \in B\} \tag{44}$$

with $\phi_{\alpha,\beta}$ given by (40) – (43) defines the family of $\phi_{\alpha,\beta}$-divergences

$$\chi^{\alpha}_{\beta}(P,Q) = \int q \phi_{\alpha,\beta}\left(\frac{p}{q}\right) \mathrm{d}\mu, \qquad (\alpha, \beta) \in B \tag{45}$$

called *extended absolute power divergence family*. This terminology is justified by the following easily verifiable assertion.

Proposition 3. The subfamily $\{\chi_0^\alpha(P,Q) : \alpha > 1\} \subset \{\chi_\beta^\alpha(P,Q) : (\alpha,\beta) \in B\}$ contains all rescaled absolute power divergences $\chi^\alpha(P,Q)/[\alpha(\alpha-1)], \alpha > 1$, see (38). The subfamily $\{\chi_1^\alpha(P,Q) = D_\alpha(P,Q) : \alpha \in \mathbb{R}\}$ contains all power divergences $D_\alpha(P,Q) : \alpha \in \mathbb{R}$, see (6).

Some symmetric or symmetrized divergences from the class (45) define metrics in the space $\mathcal{P}$ of all probability distributions on $(\mathcal{X},\mathcal{A})$. For example, from (42) we get the symmetrized divergence

$$\chi_2^0(P,Q) + \chi_2^0(Q,P) = I(P,(P+Q)/2) + I(Q,(P+Q)/2).$$

The sum of the Kullback divergences on the right-hand side is the ϕ-divergence $D_\phi(P,Q)$ for

$$\phi(t) = t\ln\frac{2t}{t+1} + \ln\frac{2}{t+1}, \qquad t > 0. \tag{46}$$

As proved in Österreicher and Vajda [17], this ϕ-divergence is a squared metric distance on $\mathcal{P}$. Further, one can deduce from Csiszár and Fischer [2] that

$$\rho_\alpha(P,Q) = \left(\chi_1^\alpha(P,Q) + \chi_1^\alpha(Q,P)\right)^\alpha, \qquad \alpha \in (0,1/2]$$

is a family of metrics on $\mathcal{P}$. The particular metric

$$\rho_{1/2}(P,Q) = 2H(P,Q)$$

is twice the Hellinger distance on $\mathcal{P}$. Finally, (40) implies

$$\phi_{-1,2}(t) = \frac{(t-1)^2}{8(t+1)}, \qquad t > 0.$$

Therefore the extended absolute power divergence

$$\chi_2^{-1}(P,Q) = \frac{1}{8}\mathcal{D}_{1,1/2}(P,Q)$$

is nothing but a rescaled Le Cam divergence. Consequently this divergence is a squared metric on $\mathcal{P}$. An open problem is whether this list of metrics obtained in the family (45) is exhaustive. The metric properties of ϕ-divergences are very desirable because they extend the toolbox of mathematical methods applicable in their analysis and statistical implementations.

4. EXTENDED MATUSITA DIVERGENCES

In this last section we extend the family of Matusita divergences

$$M_\alpha(P,Q) = \int |p^\alpha - q^\alpha|^{1/\alpha}\,\mathrm{d}\mu, \qquad \alpha \in (0,1] \tag{47}$$

defined by the functions

$$\phi_\alpha(t) = |t^\alpha - 1|^{1/\alpha}, \qquad t > 0. \tag{48}$$

Since all ϕ-divergences are reflexive and $M_\alpha(P,Q)$ are symmetric in P,Q, we see that $\rho_\alpha(P,Q) = (M_\alpha(P,Q))^\alpha$ are metrics on the space $\mathcal{P}$ of probability distributions P,Q under consideration. Our extension will contain further divergences with metric properties.

As before, we apply the convex standardization (11), in this case to the functions

$$\psi_{\alpha,\beta}(t) = |t^\alpha + \beta - 1|^{1/\alpha} \tag{49}$$

for suitable real parameters α and β. The same argument as in the previous section leads to the domain $A = A_1 \cup A_2 \cup A_3 \cup A_4 \subset \mathbb{R}^2$ for (α,β) where

$$\begin{aligned} A_1 &= (-\infty,0)\times(1,\infty), & A_2 &= (0,1)\times(1,\infty) \\ A_3 &= (1,\infty)\times(1,\infty), & A_4 &= (0,1)\times(-\infty,1) \end{aligned}$$

(see Figure 6). From (11) we obtain the family

$$\phi_{\alpha,\beta}(t) = \frac{|t^\alpha + \beta - 1|^{1/\alpha} - \operatorname{sgn}(\beta)|\beta|^{1/\alpha-1}(t+\beta-1)}{(\alpha-1)(\beta-1)|\beta|^{1/\alpha-2}}$$

for $(\alpha,\beta) \in A$. Consider for any fixed $t > 0$ the continuous extensions

$$\phi_{\alpha,1}(t) = \frac{t^{1-\alpha}}{\alpha(\alpha-1)} + \frac{t}{\alpha} - \frac{1}{\alpha-1} = t\varphi_\alpha(1/t), \qquad \alpha \neq 0,1 \tag{50}$$

$$\phi_{0,\beta}(t) = \frac{\beta}{\beta-1}\left[t - \beta t^{1/\beta} + \beta - 1\right], \qquad \beta \neq 0,1 \tag{51}$$

and

$$\phi_{1,\beta}(t) = \frac{|\beta|}{\beta-1}\left[t\ln(t)\operatorname{sgn}(t+\beta-1) - |t+\beta-1|\ln\left|\frac{t+\beta-1}{\beta}\right|\right], \qquad \beta \neq 0,1\,. \tag{52}$$

These extensions lead to the family

$$\{\phi_{\alpha,\beta}(t) : (\alpha,\beta) \in B\} \tag{53}$$

for

$$B = \overline{A} - \{(0,0),(1,0)\} = (0,1)\times(-\infty,1) \cup \mathbb{R}\times[1,\infty]$$

where $\overline{A}$ is the closure of A (see Figure 7). Note that the set B differs from $B \subset \mathbb{R}^2$ of Section 3. The functions $\phi_{\alpha,1}$ of (50) are *adjoint* to the power divergence functions φ_α of (7) in the sense that $\phi_{\alpha,1}(t) = t\varphi_\alpha(1/t)$, $t > 0$. Moreover, the continuous extensions to the corner points $(\alpha,\beta) = (0,1)$ and $(\alpha,\beta) = (1,1)$ of B are

$$\phi_{0,1}(t) = t\varphi_0(1/t) = \varphi_1(t) \qquad \text{and} \qquad \phi_{1,1}(t) = t\varphi_1(1/t) = \varphi_0(t).$$

The family of functions $\phi_{\alpha,\beta}$ defines the family of $\phi_{\alpha,\beta}$-divergences

$$M_{\alpha,\beta}(P,Q) = \int q\phi_{\alpha,\beta}\left(\frac{p}{q}\right)\mathrm{d}\mu, \qquad (\alpha,\beta) \in B \tag{54}$$

called *extended Matusita family* because the following easily verifiable assertion holds.

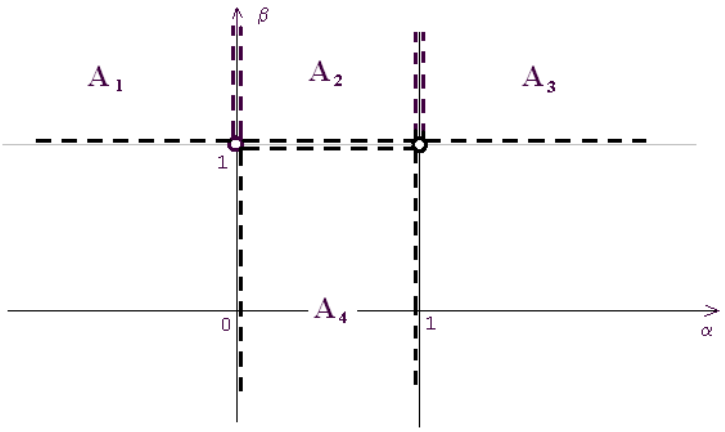

Fig. 6. The set $A_1 \cup A_2 \cup A_3 \cup A_4$.

Proposition 4. The family (54) extends the Matusita divergences in the sense that the subfamily $\{M_{\alpha,0}(P,Q) : \alpha \in (0,1)\} \subset \{M_{\alpha,\beta}(P,Q) : (\alpha,\beta) \in B\}$ coincides with the family $\{M_\alpha(P,Q) : \alpha \in (0,1)\}$ of Matusita divergences. Moreover, $\{(M_{\alpha,0}(P,Q))^{1/\alpha}\colon\ \alpha \in (0,1)\}$ is the class of metrics on the space $\mathcal{P}$ of distributions P, Q.

More interesting continuous subfamily of (53) than $\{\phi_{\alpha,0} : \alpha \in (0,1)\}$ seems to be $\{\phi_{\alpha,2} : \alpha \in \mathbb{R}\}$ given by the explicit formulas

$$\phi_{\alpha,2}(t) = \frac{4}{\alpha - 1}\left[\left(\frac{t^\alpha + 1}{2}\right)^{1/\alpha} - \frac{t+1}{2}\right] \tag{55}$$

or

$$\phi_{1,2}(t) = 2t \ln \frac{2t}{t+1} + 2 \ln \frac{2}{t+1} \quad \text{and} \quad \phi_{0,2}(t) = 4(1 - \sqrt{t}) + 2(t-1) \tag{56}$$

if $\alpha(1-\alpha) \neq 0$ or $\alpha(1-\alpha) = 0$ respectively. For $\alpha > 0$ the functions of this subfamily are related by

$$\phi_{\alpha,2}(t) = \frac{2^{2-1/\alpha}}{\alpha} f_\alpha(t)$$

to the functions $f_\alpha \in \Phi$, $\alpha > 0$ introduced by Österreicher and Vajda [17]. These authors proved that the latter functions lead to f_α-divergences with the roots

$$\rho_\alpha(P,Q) = (D_{f_\alpha}(P,Q))^{\min\{\alpha,1/2\}} \tag{57}$$

being metrics in the space $\mathcal{P}$ of probability measures P, Q. Therefore the roots

$$\left\{(M_{\alpha,2}(P,Q))^{\min\{\alpha,1/2\}} : \alpha > 0\right\} \tag{58}$$

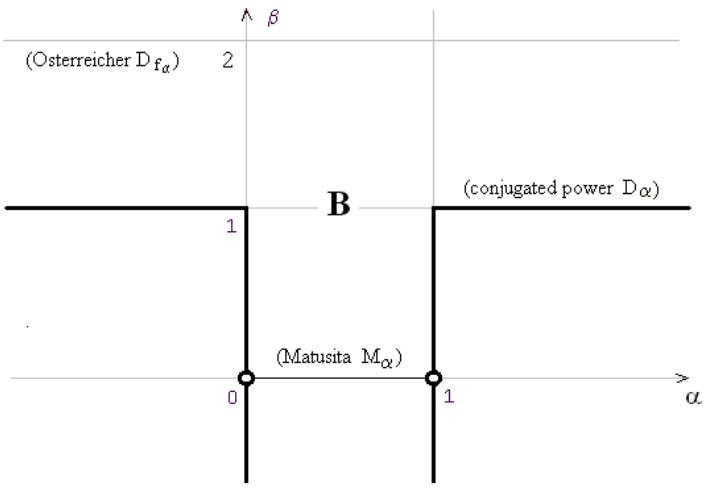

Fig. 7. The set B.

of the extended Matusita divergences $M_{\alpha,2}(P,Q)$, $\alpha > 0$ are metrics on $\mathcal{P}$ too. This means in particular that the Kullback–type divergence

$$M_{1,2}(P,Q) = 2[I(P,(P+Q)/2) + I(Q,(P+Q)/2)]$$

(cf. (56) and (46)) is a squared metric distance on $\mathcal{P}$, which was already mentioned in the previous section. However, we see from the second formula in (56) that $M_{\alpha,2}(P,Q)$ for $\alpha = 0$ is twice the Hellinger divergence $2H^2(P,Q)$ so that the extremal case $(M_{0,2}(P,Q))^{1/2}$ is metric on $\mathcal{P}$ too. Further, according to (55)

$$\phi_{-1,2}(t) = \frac{2(t-1)^2}{t+1}.$$

Therefore $M_{-1,2}(P,Q)$ is twice the Le Cam divergence $LC^2(P,Q)$ of (29). This means that

$$(M_{-1,2}(P,Q))^{1/2} = \sqrt{2}LC(P,Q)$$

is metric on $\mathcal{P}$ too. This suggests the conjecture that the square roots $(M_{\alpha,2}(P,Q))^{1/2}$ of all extended Matusita divergences $M_{\alpha,2}(P,Q)$, $\alpha \in \mathbb{R}$ are metrics on $\mathcal{P}$. If true, this conjecture means a new result for $\alpha < 0$, $\alpha \neq -1$, and a stronger result than the metricity of the roots in (58) for $0 < \alpha < 1/2$.

In any case, the new result is the possibility of a smooth divergence connection of the family of metric divergences found by Österreicher [16] and later extended by Österreicher and Vajda [17] with the famous Hellinger and Le Cam divergences which are not in the family of Österreicher and Vajda. Such a possibility was not demonstrated in the previous literature.

5. CONCLUSIONS

Distances or pseudo-distances between hypothetical and empirical probability distributions play a fundamental role in statistical inference. They are widely applied in the minimum distance estimation and testing. The parametric families of divergences introduced in this paper enable, among others, smooth connections of various pairs of ϕ_1-divergences and ϕ_2-divergences leading separately to minimum distance statistical methods with different (sometimes diametrally different) properties. For example, we may face a low bias and a high mean squared error of a minimum ϕ_1-divergence estimator but a high bias and a low mean squared error of a minimum ϕ_2-divergence estimator. Smooth ϕ-divergence connection of the ϕ_1- and ϕ_2-divergences usually leads to a smooth transition of properties of the corresponding estimators. Thus among the ϕ-divergences smoothly connecting these extremal divergences one can find one candidate leading to an estimator with desirably tuned compromise between the bias and mean squared error. Similar compromise choices among various statistical procedures are typical for the statistics – well known examples are the compromises between efficiency and robustness.

We believe that the families of divergences proposed above will be helpful in the research of optimal practically applicable statistical procedures. But concrete applications are left for future studies as they exceed the scope of the present paper.

ACKNOWLEDGEMENT

This work was partially supported by the Grant Agency of the Czech Republic under grant 102/07/1131, by the Ministry of Education, Youth and Sports of the Czech Republic under project 1M 0572 and also by MPO FI-IM3/136 and MTM2006-05693 grants.

(Received January 5, 2007.)

REFERENCES

[1] J. Beirlant, L. Devroye, L. Győrfi, and I. Vajda: Large deviations of divergence measures of partitions. J. Statist. Plann. Inference *93* (2001), 1–16.

[2] I. Csiszár and J. Fisher: Informationsentfernungen im Raum der Narcheinlichkeitsverteilungen. Publ. Math. Inst. Hungar. Acad. Sci. *7* (1962), 159–180.

[3] L. Győrfi and I. Vajda: Asymptotic distributions for goodness-of-fit statistics in a sequence of multinomial models. Statist. Probab. Lett. *56* (2002), 57–67.

[4] T. Hobza, I. Molina, and I. Vajda: On convergence of Fisher's information in continuous models with quantized observations. Test *4* (2005), 151–179.

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

[6] S. Kullback and R. Leibler: On information and sufficiency. Ann. Math. Statist. *22* (1951), 79–86.

[7] S. Kullback: Statistics and Information Theory. Wiley, New York 1957.

[8] V. Kůs: Blended ϕ-divergences with examples. Kybernetika *39* (2003), 43–54.

[9] L. Le Cam: Asymptotic Methods in Statistical Decision Theory. Springer, New York 1986.

[10] F. Liese and I. Vajda: Convex Statistical Distances. Teubner, Leipzig 1987.

[11] F. Liese and I. Vajda: On divergences and informations in statistics and information theory. IEEE Trans. Inform. Theory *52* (2006), 4394–4412.
[12] B. G. Lindsay: Efficiency versus robustness: The case of minimum Hellinger distance and other methods. Ann. Statist. *22* (1994), 1081–1114.
[13] D. Morales, L. Pardo, and I. Vajda: Some new statistics for testing hypotheses in parametric models. J. Multivariate Anal. *62* (1997), 137–168.
[14] D. Morales, L. Pardo, and I. Vajda: Limit laws for disparities of spacings. Nonparametric Statistics *15* (2003), 325–342.
[15] D. Morales, L. Pardo, and I. Vajda: On the optimal number of classes in the Pearson goodness-of-fit tests. Kybernetika *41* (2005), 677–698.
[16] F. Österreicher: On a class of perimeter-type distances of probability distributions. Kybernetika *32* (1996), 389–393.
[17] F. Österreicher and I. Vajda: A new class of metric divergences on probability spaces and its applicability in statistics. Ann. Inst. Statist. Math. *55* (2003), 639–653.
[18] L. Pardo: Statistical Inference Based on Divergence Measures. Chapman&Hall, London 2006.
[19] T. C. R. Read and N. A. Cressie: Goodness-of-fit Statistics for Discrete Multivariate Data. Springer, Berlin 1988.
[20] I. Vajda: χ^a-divergence and generalized Fisher information. In: Trans. 6th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Academia, Prague 1973, pp. 872–886.
[21] I. Vajda: Theory of Statistical Inference and Information. Kluwer, Boston 1989.
[22] I. Vajda and V. Kůs: Relations Between Divergences, Total Variations and Euclidean Distances. Research Report No. 1853, Institute of Information Theory, Prague 1995.
[23] I. Vajda and E. C. van der Meulen: Optimization of Barron density stimates. IEEE Trans. Inform. Theory *47* (2001), 1867–1883.

Václav Kůs, Department of Mathematics, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2. Czech Republic.
e-mail: kus@fjfi.cvut.cz

Domingo Morales, Operational Research Center, Miguel Hernández University of Elche, Elche. Spain.
e-mail: d.morales@umh.es

Igor Vajda, Institute of Information Theory and Automation — Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: vajda@utia.cas.cz