Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Igor Vajda:

Modifications of Divergence Criteria for Applications in Continuous Families

No. 2230 November 2008

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42) (2) 688 4903
E-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

Modifications of Divergence Criteria for Applications in Continuous Families

Igor Vajda

ÚTIA AV ČR
⟨vajda@utia.cas.cz⟩

1. INTRODUCTION

Let $\phi : (0, \infty) \mapsto \mathbb{R}$ be twice differentiable strictly convex function with $\phi(1) = 0$ and (possibly infinite) continuous extension to $t = 0+$ denoted by $\phi(0)$, and let $\mathbf{\Phi}$ be the class of all such functions. We use also the related functions

$$\phi^{\#}(t) = \phi(t) - t\phi'(t) \quad \text{and} \quad \phi^*(t) = t\phi(1/t) \tag{1}$$

where ϕ' denotes the derivative, $\phi^{\#}$ is nonincreasing, ϕ^* belongs to $\mathbf{\Phi}$ and the star operation is idempotent in the sense $(\phi^*)^* = \phi$.

Let P, Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ with densities p, q w.r.t. a dominating σ-finite measure λ. Following Liese and Vajda (1987 or 2006), for every $\phi \in \mathbf{\Phi}$ we define *ϕ-divergence* of P and Q by

$$D_\phi(P, Q) = \begin{cases} \int \phi\,(p/q)\,\mathrm{d}Q & \text{if} \quad pq > 0 \quad \lambda\text{-a. s.} \\ \\ \phi(0) + \phi^*(0) & \text{if} \quad pq = 0 \quad \lambda\text{-a. s.} \end{cases} \tag{2}$$

Here the condition $pq > 0 \;\; \lambda$-a. s. means that P, Q are measure-theoretically equivalent (in symbols $P \equiv Q$) and $pq = 0 \;\; \lambda$-a. s. means that P, Q are measure-theoretically orthogonal (in symbols $P \perp Q$).

We shall deal mainly with the power divergences

$$D_\alpha(P, Q) := D_{\phi_\alpha}(P, Q) \quad \text{of real orders } \alpha \in \mathbb{R} \tag{3}$$

for the power functions $\phi_\alpha \in \mathbf{\Phi}$ defined by

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha - 1)} \quad \text{if} \quad \alpha(\alpha - 1) \neq 0 \tag{4}$$

and otherwise by the corresponding limits

$$\phi_0(t) = -\ln t + t - 1, \qquad \phi_1(t) = \phi_0^*(t) = t \ln t - t + 1. \tag{5}$$

For $P \equiv Q$ we get from (2) and (4) or (5)

$$D_\alpha(P,Q) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[\int (p/q)^\alpha \, \mathrm{d}Q - 1\right] & \text{if} \quad \alpha(\alpha-1) \neq 0 \\ \int \ln(p/q)\, \mathrm{d}P = D_0(Q,P) & \text{if} \quad \alpha = 1 \end{cases} \tag{6}$$

and for $P \perp Q$ similarly

$$D_\alpha(P,Q) = \begin{cases} 1/\alpha(1-\alpha) & \text{if} \quad 0 < \alpha < 1 \\ \infty & \text{otherwise.} \end{cases} \tag{7}$$

The special cases $D_2(P,Q)$ or $D_1(P,Q)$ are sometimes called Pearson or Kullback divergences and $D_{-1}(P,Q) = D_2(Q,P)$ or $D_0(P,Q) = D_1(Q,P)$ reversed Pearson or Kullback divergences, respectively.

The ϕ-divergences and power divergences will be applied in the *standard statistical estimation model* with i.i.d. observations $X_1, \dots, X_n$ governed by P_{θ_0} from a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures on $(\mathcal{X}, \mathcal{A})$ indexed by a set of parameters $\Theta \subset \mathbb{R}^d$. The family is assumed to be dominated with densities

$$p_\theta = \mathrm{d}P_\theta/\mathrm{d}\lambda \quad \text{for all } \theta \in \Theta \tag{8}$$

and to satisfy the relations

$$P_\theta(\{x\}) = 0 \qquad \text{for all} \;\; x \in \mathcal{X}, \;\; \theta \in \Theta \tag{9}$$

and

$$P_\theta \neq P_{\theta_0} \quad \text{and} \quad P_\theta \equiv P_{\theta_0} \quad \text{for all } \theta, \, \theta_0 \in \Theta \;\; \text{with} \;\; \theta \neq \theta_0. \tag{10}$$

Here (9) means that the family $\mathcal{P}$ is *continuous* (nonatomic). The first property in (10) means the identifiability of true parameter θ_0 and the second property means the measure-theoretic equivalence of all pairs from the family $\mathcal{P}$. In this model the parameter θ_0 is assumed to be estimated on the basis of observations $X_1, \dots, X_n$ by measurable functions $\hat{\theta}_n : \mathcal{X}^n \mapsto \Theta$ called estimates. Collection of estimates for various sample sizes n is an estimator.

The assumed strict convexity of $\phi(t)$ at $t = 1$ together with the identifiability of θ_0 assumed in (10) means that $D_\phi(P_{\hat{\theta}}, P_{\theta_0}) \geq 0$ for all $\hat{\theta}, \, \theta \in \Theta$ with the equality iff $\hat{\theta} = \theta_0$. In other words, the unknown parameter θ_0 is the unique minimizer of the function $D_\phi(P_{\hat{\theta}}, P_{\theta_0})$ of variable $\hat{\theta} \in \Theta$,

$$\theta_0 = \mathrm{argmin}_{\hat{\theta}} \mathcal{D}(P_{\hat{\theta}}, P_{\theta_0}) \qquad \text{for every } \theta_0 \in \Theta. \tag{11}$$

Further, the observations $X_1, \dots, X_n$ are in a statistically sufficient manner represented by the empirical probability measure

$$P_n = \frac{1}{n} \sum_{i=1}^{n} P_{X_i} \tag{12}$$

where P_x denotes the Dirac probability measure with all mass concentrated at $x \in \mathcal{X}$. The empirical measures P_n are known to converge weakly to P_{θ_0} as $n \to \infty$. Therefore the minimizer

$$\hat{\theta}_n = \hat{\theta}_{n,\phi} = \mathrm{argmin}_{\hat{\theta} \in \Theta} D_\phi \left(P_{\hat{\theta}}, P_n \right) \tag{13}$$

is intuitively expected to estimate θ_0 consistently in the usual sense of the convergence $\hat{\theta}_n \to \theta_0$ for $n \to \infty$. However, the reality is different: the problem is that for the continuous family $\mathcal{P}$ under consideration and the discrete family $\mathcal{P}_{\mathrm{emp}}$ of empirical distributions (12)

$$P_{\hat{\theta}} \bot P_n \Longrightarrow D_\phi(P_{\hat{\theta}}, P_n) = \phi(0) + \phi^*(0) \quad \text{when } P_{\hat{\theta}} \in \mathcal{P} \text{ and } P_n \in \mathcal{P}_{\mathrm{emp}}. \tag{14}$$

This means that the estimates $\hat{\theta}_n$ proposed in (13) are trivial, with the arg min extending over the whole space Θ.

In this paper we list and motivate several modifications of the minimum divergence rule (13) which allow to bypass the problem (14). Some of them are new and some known from the previous literature. The estimators corresponding to the listed modifications will be studied in more detail in a subsequent paper.

2. SUBDIVERGENCES AND SUPERDIVERGENCES

In the rest of the paper we consider the probability measures

$$P \in \mathcal{P} \quad \text{and} \quad Q \in \mathcal{Q} \quad \text{for} \quad \mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{\mathrm{emp}} \quad \text{(cf. (14))} \tag{15}$$

These measures are either measurable-theoretically equivalent (if $Q \in \mathcal{P}$) or measurable-theoretically orthogonal (if $Q \in \mathcal{P}_{\mathrm{emp}}$). Therefore the ϕ-divergences $D_\phi(P, Q)$ are well defined by (1) for all $(P, Q) \in \mathcal{P} \otimes \mathcal{Q}$.

We often use also the likelihood ratios $\boldsymbol{\ell}_{\theta,\hat{\theta}} = p_\theta / p_{\hat{\theta}}$ well defined a. s. on $\mathcal{X}$ in the statistical model under consideration. In the rest of the paper we suppose that

$$\left\{ \phi\left(\boldsymbol{\ell}_{\theta,\hat{\theta}}\right),\ \phi'\left(\boldsymbol{\ell}_{\theta,\hat{\theta}}\right),\ \phi^{\#}\left(\boldsymbol{\ell}_{\theta,\hat{\theta}}\right) \right\} \subset \mathbb{L}_1(Q) \quad \text{for all } \theta,\, \hat{\theta} \in \Theta \ \text{ and } \ Q \in \mathcal{Q} \tag{16}$$

where $\mathbb{L}_1(Q)$ denotes in this paper the set of all absolutely Q-integrable functions $L : \mathcal{X} \mapsto \mathbb{R}$ so that (16) automatically holds if $Q \in \mathcal{P}_{\mathrm{emp}}$. Further, we put for brevity

$$Q \cdot L = \int L \, \mathrm{d}Q \quad \text{for } L \in \mathbb{L}_1(Q). \tag{17}$$

Finally, for all pairs $\theta,\, \hat{\theta} \in \Theta$ we consider the functions $L_\phi(\theta,\, \hat{\theta}) = L_\phi(\theta,\, \hat{\theta}, x)$ of variable $x \in \mathcal{X}$ defined by the formula

$$L_\phi(\theta,\, \hat{\theta}) = P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\hat{\theta}}) + \phi^{\#}(\boldsymbol{\ell}_{\theta,\hat{\theta}}) \quad \text{(cf. (1))}$$

which are due to (16) Q-integrable for all $Q \in \mathcal{Q}$. Therefore

$$d_{\phi,\hat{\theta}}\left(P_\theta, Q\right) = Q \cdot L_\phi(\hat{\theta},\, \theta) = P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\hat{\theta}}) + Q \cdot \phi^{\#}(\boldsymbol{\ell}_{\theta,\hat{\theta}}), \quad \hat{\theta} \in \Theta \tag{18}$$

is for all $P_\theta \in \mathcal{P}$ and $Q \in \mathcal{Q}$ a family of finite expectations. If $Q \in \mathcal{P}$ then, by a supremal representation of ϕ-divergences established independently by Broniatowski & Keziou (2006) and Liese & Vajda (2006), each ϕ-divergence $D_\phi(P_\theta, Q)$ is maximum of the subdivergences $d_{\phi,\hat{\theta}}(P_\theta, Q)$ over $\hat{\theta} \in \Theta$, in symbols

$$D_\phi(P_\theta, Q) = \sup_{\hat{\theta}\in\Theta} d_{\phi,\hat{\theta}}(P_\theta, Q) \quad \text{for all } P_\theta, Q \in \mathcal{P}. \tag{19}$$

This to some extent justifies to interpret (18) as a family of ***subdivergences*** of $P_\theta \in \mathcal{P}$ and $Q \in \mathcal{Q}$. Obviously, if $Q = P_{\theta_0} \in \mathcal{P}$ then the subdivergence formula (18) reduces to

$$d_{\phi,\hat{\theta}}(P_\theta, P_{\theta_0}) = P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\hat{\theta}}) + P_{\theta_0}\phi^{\#}(\boldsymbol{\ell}_{\theta,\hat{\theta}}), \quad \hat{\theta} \in \Theta \tag{20}$$

and if $Q = P_n \in \mathcal{P}_{\text{emp}}$ then it reduces to

$$\begin{aligned} d_{\phi,\hat{\theta}}(P_\theta, P_n) &= P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\hat{\theta}}) + P_n \cdot \phi^{\#}(\boldsymbol{\ell}_{\theta,\hat{\theta}}) & (21)\\ &= P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\hat{\theta}}) + \frac{1}{n}\sum_{i=1}^{n} \phi^{\#}(\boldsymbol{\ell}_{\theta,\hat{\theta}}(X_i)), \quad \hat{\theta} \in \Theta. & (22)\end{aligned}$$

The supremum of all subdivergences of $d_{\phi,\hat{\theta}}(P_\theta, Q)$, $\hat{\theta} \in \Theta$,

$$\mathcal{D}_\phi(P_\theta, Q) = \sup_{\hat{\theta}\in\Theta} d_{\phi,\hat{\theta}}(P_\theta, Q) \quad \text{(cf. (18) - (21))} \tag{23}$$

is well defined for all $P_\theta \in \mathcal{P}$, $Q \in \mathcal{Q}$ and represents a ***superdivergence*** of P_θ and Q. If $Q \in \mathcal{P}$ then it is seen from (19) that the superdivergence $\mathcal{D}_\phi(P_\theta, Q)$ coincides with the ϕ-divergence $D_\phi(P_\theta, Q)$, i.e.,

$$D_\phi(P_\theta, Q) = \mathcal{D}_\phi(P_\theta, Q) \quad \text{for all } P_\theta, Q \in \mathcal{P}. \tag{24}$$

If $Q = P_{\theta_0}$ then (24) reduces to the formula $D_\phi(P_\theta, P_{\theta_0}) = \mathcal{D}_\phi(P_\theta, P_{\theta_0})$ for all $\theta \in \Theta$. This superdivergence representation of of the ϕ-divergence function $D_\phi(P_\theta, P_{\theta_0})$ of variable $\theta \in \Theta$ justifies the replacement of the meaningless minimum divergence estimates (13) by the meaningful ***minimum superdivergence estimates***

$$\begin{aligned} \hat{\theta}_n &= \hat{\theta}_{n,\phi} = \text{argmin}_{\hat{\theta}} \mathcal{D}_\phi(P_{\hat{\theta}}, P_n) \qquad \text{(cf. (23))} & (25)\\ &= \text{argmin}_{\hat{\theta}} \sup_{\theta\in\Theta} \left[P_{\hat{\theta}} \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) + P_n \cdot \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta})\right] \qquad \text{(cf. (21))} & \\ &= \text{argmin}_{\hat{\theta}} \sup_{\theta\in\Theta} \left[P_{\hat{\theta}} \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) + \frac{1}{n}\sum_{i=1}^{n} \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}(X_i)) \quad \text{(cf. (22))}\right]. & (26)\end{aligned}$$

We see that this approach to the estimation of unknown parameter θ_0 bypasses the problem mentioned in (14) by replacing the $\hat{\theta}$-insensitive divergence $D_\phi(P_{\hat{\theta}}, P_n)$ in the argmin formula (13) by the $\hat{\theta}$-sensitive superdivergence $\mathcal{D}_\phi\left(P_{\hat{\theta}}, P_n\right)$. Then, by the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}(X_i)) \xrightarrow{\text{a.s.}} P_{\theta_0}\phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}) \quad \text{for all } \hat{\theta},\, \theta \in \Theta.$$

This means that under a mild additional regularity the supremum in (26) tends for each argument $\hat{\theta} \in \Theta$ to $\sup[P_{\hat{\theta}} \cdot \phi'(L_{\hat{\theta},\theta}) + P_{\theta_0}\phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta})]$ which is by (20), (23)and (24) the divergence $D_\phi\left(P_{\hat{\theta}}, P_{\theta_0}\right)$. In other words, the nontrivial functions $\mathcal{D}_\phi\left(P_{\hat{\theta}}, P_n\right)$ of variable $\hat{\theta} \in \Theta$ tend to the nontrivial function $D_\phi\left(P_{\hat{\theta}}, P_{\theta_0}\right)$ of the same variable, while the trivial constant function $D_\phi\left(P_{\hat{\theta}}, P_n\right)$ of the variable $\hat{\theta} \in \Theta$ does not do so. Moreover, the limit function $D_\phi\left(P_{\hat{\theta}}, P_{\theta_0}\right) = \mathcal{D}_\phi\left(P_{\hat{\theta}}, P_{\theta_0}\right)$ preserves the optimality condition (11), i.e.,

$$\theta_0 = \text{argmin}_{\hat{\theta}}\mathcal{D}_\phi\left(P_{\hat{\theta}}, P_{\theta_0}\right) \quad \text{for each } \theta_0 \in \Theta \tag{27}$$

which means that the minimum superdivergence estimator (25) is Fisher consistent (cf. e.g. Hampel et al. (1986)).

Note that the minimum superdivergence estimators (26) were first introduced under the name *minimum ϕ-divergence estimators* by Liese and Vajda (2006) and independently by Broniatowski and Keziou (2007) under the name *minimum dual ϕ-divergence estimators*.

Example 1. Restrict ourselves to the logarithmic function $\phi(t) = -\ln t + t - 1$ introduced in (5). For this function we get $\phi'(t) = (t-1)/t$ and $\phi^{\#}(t) = -\ln t$ so that the integrability (16) takes place if the likelihood ratios $\boldsymbol{\ell}_{\hat{\theta},\theta} = p_{\hat{\theta}}/p_\theta$ satisfy for all θ, $\hat{\theta}$, $\theta_0 \in \Theta$ the condition $P_{\theta_0} \cdot \boldsymbol{\ell}_{\hat{\theta},\theta} < \infty$. Further, $P_{\hat{\theta}} \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) = 0$ so that

$$d_{\phi,\hat{\theta}}\left(P_\theta, P_n\right) = P_n \cdot \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}) = P_n \cdot [\ln p_\theta - \ln p_{\hat{\theta}}]. \tag{28}$$

Hence we see from (26) that the minimum superdivergence estimator is in this case the MLE

$$\hat{\theta}_n = \text{argmin}_{\hat{\theta}} \sup_{\theta\in\Theta}\left[\left(\frac{1}{n}\sum_{i=1}^{n}[\ln p_\theta(X_i) - \ln p_{\hat{\theta}}(X_i)]\right)\right] = \text{argmax}_{\hat{\theta}}\sum_{i=1}^{n}\ln p_{\hat{\theta}}(X_i). \tag{29}$$

3. SUBDIVERGENCE DEFICITS

In this section we study for arbitrary convex function $\phi \in \boldsymbol{\Phi}$ the family of differences

$$\mathbb{D}_{\phi,\hat{\theta}}(P_\theta, Q) = \mathcal{D}_\phi(P_\theta, Q) - d_{\phi,\hat{\theta}}\left(P_\theta, Q\right), \quad \hat{\theta} \in \Theta \tag{30}$$

between the ϕ-superdivergences (23) and ϕ-subdivergences (18) of probability measures $P_\theta \in \mathcal{P}$ and $Q \in \mathcal{Q}$. These differences can be interpreted as ***deficits of subdivergences*** on the space $\mathcal{P} \otimes \mathcal{Q}$ parametrized by $\hat{\theta} \in \Theta$. By (21), the subdivergence deficit formula (30) can be rewritten into the equivalent form

$$\mathbb{D}_{\phi,\hat{\theta}}(P_\theta, Q) = \mathcal{D}_\phi(P_\theta, Q) - \left[P_\theta \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) + Q \cdot \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta})\right], \quad \hat{\theta} \in \Theta \tag{31}$$

where $\boldsymbol{\ell}_{\hat{\theta},\theta}$ are the likelihood ratios $p_{\hat{\theta}}/p_{\theta}$.

From Theorem 1 below we see that the subdivergence deficits $\mathbb{D}_{\phi,\hat{\theta}}(P_\theta, P_{\theta_0})$ as functions of the parameter $\hat{\theta} \in \Theta$ satisfy the optimality condition

$$\theta_0 = \text{argmin}_{\hat{\theta}} \mathbb{D}_{\phi,\hat{\theta}}(P_\theta, P_{\theta_0}) \quad \text{for all} \;\; \theta,\, \theta_0 \in \Theta. \tag{32}$$

Therefore the ***minimum subdivergence deficit estimators*** of θ_0 defined for all possible statistical parameters $\theta \in \Theta$ by the formula

$$\begin{aligned} \hat{\theta}_{n,\theta} &= \hat{\theta}_{n,\theta,\phi} = \text{argmin}_{\hat{\theta}} \mathbb{D}_{\phi}(P_\theta, P_n) \\ &= \text{argmax}_{\hat{\theta}} \left[P_{\hat{\theta}} \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) + P_n \cdot \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}) \right] \qquad \text{(cf. (31))} \\ &= \text{argmax}_{\hat{\theta}} \left[P_{\hat{\theta}} \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) + \frac{1}{n} \sum_{i=1}^{n} \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}(X_i)) \quad \text{(cf. (22))} \right] \end{aligned} \tag{33}$$

are Fisher consistent.

Theorem 1. For arbitrary θ, $\hat{\theta}$, $\theta_0 \in \Theta$ it holds

$$\mathbb{D}_{\phi,\hat{\theta}}(P_\theta, P_{\theta_0}) \geq 0 \tag{34}$$

and the equality takes place iff $\hat{\theta} = \theta_0$.

Proof. By the Taylor theorem for convex functions ϕ it holds for all θ, $\hat{\theta}$, $\theta_0 \in \Theta$

$$\phi\left(\frac{p_\theta}{p_{\theta_0}}\right) \geq \phi\left(\frac{p_\theta}{p_{\hat{\theta}}}\right) + \phi'\left(\frac{p_\theta}{p_{\hat{\theta}}}\right)\left(\frac{p_\theta}{p_{\theta_0}} - \frac{p_\theta}{p_{\hat{\theta}}}\right) \qquad \lambda\text{-a.s.}$$

which is under (16) equivalent to

$$P_{\theta_0} \cdot \left[\phi\left(\frac{p_\theta}{p_{\theta_0}}\right) - \frac{p_\theta}{p_{\theta_0}} \phi'\left(\frac{p_\theta}{p_{\hat{\theta}}}\right) - \phi\left(\frac{p_\theta}{p_{\hat{\theta}}}\right) - \frac{p_\theta}{p_{\hat{\theta}}} \phi'\left(\frac{p_\theta}{p_{\hat{\theta}}}\right) \right] \geq 0$$

or to

$$P_{\theta_0} \cdot \phi\left(\frac{p_\theta}{p_{\theta_0}}\right) - P_\theta \cdot \phi'\left(\frac{p_\theta}{p_{\hat{\theta}}}\right) - P_{\theta_0} \cdot \phi^{\#}\left(\frac{p_\theta}{p_{\hat{\theta}}}\right) \geq 0,$$

i.e. to

$$D_\phi(P_\theta, P_{\theta_0}) - \left[P_\theta \cdot \phi'(\boldsymbol{\ell}_{\hat{\theta},\theta}) + P_{\theta_0} \cdot \phi^{\#}(\boldsymbol{\ell}_{\hat{\theta},\theta}) \right] \geq 0.$$

By (24) $D_\phi(P_\theta, P_{\theta_0}) = \mathcal{D}_\phi(P_\theta, P_{\theta_0})$ so that (34) follows from the definition of $\mathbb{D}_{\phi,\hat{\theta}}(P_\theta, P_{\theta_0})$ in (31), this means that (34) holds for all θ, $\hat{\theta}$, $\theta_0 \in \Theta$. The equalities above take place iff

$$\frac{p_\theta}{p_{\hat{\theta}}} = \frac{p_\theta}{p_{\theta_0}} \qquad \lambda\text{-a.s.}$$

which is equivalent to $p_{\hat{\theta}} = p_{\theta_0}$ λ-a.s., or to $P_{\hat{\theta}} = P_{\theta_0}$. The desired equivalence with the equality $\hat{\theta} = \theta_0$ follows from the identifiability of the true parameter θ_0 assumed in (10). □

The minimum subdivergence deficit estimating formula (33) can be rewritten into the more detailed form

$$\hat{\theta}_{n,\theta} = \mathrm{argmax}_{\hat{\theta}} \left[P_{\hat{\theta}} \cdot \phi' \left(\frac{p_\theta(X_i)}{p_{\hat{\theta}}(X_i)} \right) + \frac{1}{n} \sum_{i=1}^{n} \phi^{\#} \left(\frac{p_\theta(X_i)}{p_{\hat{\theta}}(X_i)} \right) \right] \tag{35}$$

for $\phi^{\#}$ defined in (1). The estimators (35) were first introduced by Broniatowski and Keziou (2007) under the name *dual ϕ-divergence estimators*.

Example 2. If $\phi(t) = -\ln t + t - 1$ then it follows easily from the formulas of Example 1 that for every $\theta \in \Theta$

$$\hat{\theta}_{n,\theta} = \mathrm{argmax}_{\hat{\theta}} \left[\left(\frac{1}{n} \sum_{i=1}^{n} [\ln p_{\hat{\theta}}(X_i) - \ln p_\theta(X_i)] \right) \right] = \mathrm{argmax}_{\hat{\theta}} \sum_{i=1}^{n} \ln p_{\hat{\theta}}(X_i) \tag{36}$$

so for this choice of ϕ all estimates $\hat{\theta}_{n,\theta}$, $\theta \in \Theta$ minimizing the subdivergence deficits are the MLE's.

4. DECOMPOSABLE PSEUDODISTANCES

The ϕ-divergences $D_\phi(P,Q)$, $\phi \in \mathbf{\Phi}$ can be characterized by the *information processing property*, i. e. by the complete invariance w.r.t. the statistically sufficient transformations of the observation space $(\mathcal{X}, \mathcal{A})$. This property is useful but probably not unavoidable in the minimum distance estimation based on similarity between theoretical and empirical distributions. Hence we admit in the rest of the paper general *pseudodistances* $\mathcal{D}(P,Q)$ of probability measures $P \in \mathcal{P}$ and $Q \in \mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{\mathrm{emp}}$ restricted only by the reflexivity condition

$$\mathcal{D}(P,Q) \geq 0 \quad \text{with} \;\; \mathcal{D} = 0 \quad \text{iff} \;\; P = Q \tag{37}$$

on the subdomain $\mathcal{P} \otimes \mathcal{P}$. For such functionals of $(P,Q) \in \mathcal{P} \otimes \mathcal{Q}$ the information processing property may not hold.

An additional restriction considered in this section will be the *decomposability* on the statistical family $\mathcal{P}$, i.e. the existence of mappings $\hat{\mathcal{D}}$, $\mathcal{D}^0 : \mathcal{P} \mapsto \mathbb{R}$ and δ, $\rho : \mathbb{R} \mapsto \mathbb{R}$ such that

$$\mathcal{D}(P,Q) = \mathcal{D}^0(Q) + \hat{\mathcal{D}}(P) + Q \cdot \rho(p), \quad P, Q \in \mathcal{P}, \;\; p = \mathrm{d}P/\mathrm{d}\lambda. \tag{38}$$

In the rest of this paper the decomposable pseudodistances are called briefly ***decodistances*** .

Note that using in this section the symbol $\mathcal{D}(P,Q)$ we do not indicate any connection of the present concepts with the superdivergences $\mathcal{D}_\phi(P,Q)$ of Section 2. The only known connection is between some special pseudodistances $\mathcal{D}(P,Q)$ and ϕ-divergences $D_\phi(P,Q)$ mentioned after formula (45) below.

The class of all decodistances

$$\mathcal{D}(P_{\hat\theta},P_{\theta_0}) = \mathcal{D}^0(P_{\theta_0}) + \hat{\mathcal{D}}(P_{\hat\theta}) + P_{\theta_0}\cdot\rho(p_{\hat\theta}),\quad \hat\theta,\theta_0\in\Theta \text{ (cf. (38))} \tag{39}$$

defines the ***minimum decodistance estimators*** of the true parameter θ_0 by the formula

$$\begin{aligned}\hat\theta_n &= \operatorname{argmin}_{\hat\theta}\left[\hat{\mathcal{D}}(P_{\hat\theta}) + P_n\cdot\rho(p_{\hat\theta})\right]\\ &= \operatorname{argmin}_{\hat\theta}\left[\hat{\mathcal{D}}(P_{\hat\theta}) + \frac{1}{n}\sum_{i=1}^n\rho(p_{\hat\theta}(X_i)\right].\end{aligned} \tag{40}$$

Due to the reflexivity of $\mathcal{D}(P_{\hat\theta},P_{\theta_0})$ the true parameter $\theta_0\in\Theta$ is identifiable by this decodistance in the sense

$$\theta_0 = \operatorname{argmin}_{\hat\theta}\mathcal{D}(P_{\hat\theta},P_{\theta_0}). \tag{41}$$

Hence the minimum decodistance estimators are Fisher consistent. Further, the decomposability of $\mathcal{D}(P,Q)$ leads to the additive structure of the criterion function in (40) which opens the possibility to apply the methods of the asymptotic theory of M-estimators (cf. Hampel et al. (1986), van der Vaar and Wellner (1996) and Mieske and Liese (2008)).

The next two subsections deal with two different classes of decodistances and with the related minimum decodistance estimators. Here we present a simple example allowing to interpret the estimators minimizing the subdivergence deficits as the minimum decodistance estimators.

Example 3. Let us consider the special class of pseudodistances of probability measures $P\in\mathcal{P}$ and $Q\in\mathcal{Q}=\mathcal{P}\cup\mathcal{P}_{\rm emp}$ parametrized by $\theta\in\Theta$ and given for each θ by the formula

$$\mathcal{D}_{\phi,\theta}(P,Q) = \mathbb{D}_{\phi,\theta}\left(P_{\hat\theta},Q\right)\qquad\text{when}\ \ P=P_{\hat\theta} \tag{42}$$

By Theorem 1 of previous section, these pseudodistances are reflexive. Further, by (42) and (31),

$$\mathcal{D}_{\phi,\theta}(P,Q) = \mathcal{D}_\phi(P,Q) - \left[P_\theta\cdot\phi'(\boldsymbol{\ell}_{\theta,\hat\theta}) + Q\cdot\phi^{\#}(\boldsymbol{\ell}_{\theta,\hat\theta})\right]\qquad\text{when}\ \ P=P_{\hat\theta}.$$

For every $\hat\theta\in\Theta$ and $P=P_{\hat\theta}$ with the density $p={\rm d}P/{\rm d}\lambda$ it holds

$$\mathcal{D}_{\phi,\theta}(P,Q) = \mathcal{D}^0_{\phi,\theta}(Q) + \hat{\mathcal{D}}_{\phi,\theta}(P) + Q\cdot\rho_{\phi,\theta}(p)$$

as it is required in the decomposability condition (38), with the components finite and given by

$$\mathcal{D}^0_{\phi,\theta}(Q) = \mathcal{D}_\phi(P,Q),\quad \hat{\mathcal{D}}_{\phi,\theta}(P) = -P_\theta\cdot\phi'(\boldsymbol{\ell}_{\theta,\hat\theta}),\quad \rho_{\phi,\theta}(p) = -\phi^{\#}(\boldsymbol{\ell}_{\theta,\hat\theta}). \tag{43}$$

Thus the pseudodistances $\mathcal{D}_{\phi,\theta}(P,Q)$ of probability measures $P=P_{\hat\theta}\in\mathcal{P}$ and $Q\in\mathcal{P}\cup\mathcal{P}_{\rm emp}$ are for all $\theta\in\Theta$ decodistances and the corresponding special *minimum decodistance estimators* coincide with the *minimum subdivergence deficit estimators* of Section 3.

5. DISTURBED POWER DIVERGENCES

In this section we study a special class of integral pseudodistances of the above introduced probability measures $P \in \mathcal{P}$ and $Q \in \mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{\text{emp}}$ called *ψ-pseudodistances*. They are defined by

$$\mathcal{D}_\psi(P,Q) = \int \psi(p,q)\,\mathrm{d}\lambda, \quad p = \mathrm{d}P/\mathrm{d}\lambda, q = \mathrm{d}Q/\mathrm{d}\lambda \tag{44}$$

for nonnegative functions $\psi(s,t)$ of arguments $s,t \geq 0$ reflexive in the sense $\psi(s,t) = 0$ iff $s = t$. The ϕ-divergences $D_\phi(P,Q)$ are special ψ-pseudodistances for the functions

$$\psi(s,t) = \phi(s/t)\,t - \phi'(1)(t-1), \quad s,t > 0 \tag{45}$$

since they are nonnegative and reflexive, and $\mathcal{D}_\psi(P,Q) = D_\phi(P,Q)$ for all P,Q.

Obviously, the ψ-pseudodistances for the functions ψ decomposable in the sense

$$\psi(s,t) = \hat{\psi}(s) + \psi^0(t) + \rho(s)\,t, \quad s,t \geq 0 \tag{46}$$

for some $\hat{\psi}, \psi_0, \rho : [0,\infty) \to \mathbb{R}$ are decodistances in sense of the previous section satisfying the decomposability condition

$$\mathcal{D}_\psi(P,Q) = \mathcal{D}^0_\psi(Q) + \hat{\mathcal{D}}_\psi(P) + Q \cdot \rho(p) \quad \text{(cf. (38))} \tag{47}$$

for

$$\mathcal{D}^0_\psi(Q) = \int \psi^0(q)\,\mathrm{d}\lambda \quad \text{and} \quad \hat{\mathcal{D}}_\psi(P) = \int \hat{\psi}(p)\,\mathrm{d}\lambda. \tag{48}$$

EXAMPLE 3. The quadratic function $\psi(s,t) = (s-t)^2$ defines the L_2-distance

$$\mathcal{D}_\psi(P,Q) = \|p-q\|_2^2 = \int (p-q)^2\,\mathrm{d}\lambda$$

which is reflexive and also decomposable in the sense of (47), (48) for

$$\mathcal{D}^0_\psi(Q) = \int q^2\,\mathrm{d}\lambda, \;\; \hat{\mathcal{D}}_\psi(P) = \int p^2\,\mathrm{d}\lambda \quad \text{and} \quad \rho(p) = -2p.$$

In other words the L_2-distance is an example of decodistance. For this distance the minimum decodistace estimator defined by (40) is the L_2-estimator

$$\hat{\theta}_n = \operatorname{argmax}_{\hat{\theta}} \left[\frac{2}{n} \sum_{i=1}^{n} p_{\hat{\theta}}(X_i) - \int p_{\hat{\theta}}^2\,\mathrm{d}\lambda \right] \tag{49}$$

which is known to be robust but not efficient.

To build a smooth bridge between the robustness and efficiency, one needs to replace the reflexive and decomposable functions ψ by classes $\{\psi_\alpha : \alpha \geq 0\}$ of reflexive functions decomposable in the sense

$$\psi_\alpha(s,t) = \psi^0_\alpha(t) + \hat{\psi}_\alpha(s) + \rho_\alpha(s)\,t \quad \text{for all } \alpha \geq 0 \quad \text{(cf. (46))} \tag{50}$$

and satisfying the conditions

$$\lim_{\alpha\downarrow 0}\psi_\alpha(s,t)=\psi_0(s,t)\quad\text{and}\quad\int\hat{\psi}_0(p_{\hat{\theta}})\,\mathrm{d}\lambda=\text{const},\quad\rho_0(s)=-\ln s. \tag{51}$$

Then for all $\alpha\geq 0$

$$\mathcal{D}_{\psi_\alpha}(P,Q)=\mathcal{D}^0_{\psi_\alpha}(Q)+\hat{\mathcal{D}}_{\psi_\alpha}(P)+Q\cdot\rho_\alpha(p)\qquad\text{(cf. (38))} \tag{52}$$

with

$$\mathcal{D}^0_{\psi_\alpha}(Q)=\int\psi^0_\alpha(q)\,\mathrm{d}\lambda\quad\text{and}\quad\hat{\mathcal{D}}_{\psi_\alpha}(P)=\int\hat{\psi}_\alpha(p)\,\mathrm{d}\lambda \tag{53}$$

are decodistances which define in accordance with (40) the family of ***minimum decodistance estimators*** of θ_0 by the formula

$$\begin{aligned}\hat{\theta}_{n,\alpha}&=\operatorname{argmin}_{\hat{\theta}}\left[\hat{\mathcal{D}}_{\psi_\alpha}(P_{\hat{\theta}})+P_n\cdot\rho_\alpha(p_{\hat{\theta}}))\right] &(54)\\ &=\operatorname{argmin}_{\hat{\theta}}\left[\int\hat{\psi}_\alpha(p_{\hat{\theta}})\,\mathrm{d}\lambda+\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha(p_{\hat{\theta}}(X_i))\right],\quad\alpha\geq 0. &(55)\end{aligned}$$

Obviously, this family contains as a special case for $\alpha=0$ the efficient MLE

$$\hat{\theta}_{0,n}=\operatorname{argmin}_{\hat{\theta}}\left[\text{const}-\frac{1}{n}\sum_{i=1}^{n}\ln p_{\hat{\theta}}(X_i)\right]. \tag{56}$$

The next theorem presents one family of functions $\psi_\alpha(s,t)$, $\alpha\geq 0$ satisfying (50) – (56).

Theorem 2. Each of the functions defined on the domain s, $t>0$ by

$$\psi_\alpha(s,t)=\begin{cases}s^{1+\alpha}+\frac{t^{1+\alpha}}{\alpha}-\frac{(1+\alpha)\,s^\alpha t}{\alpha} & \text{if }\alpha>0\\ s+t\ln t-t\ln s & \text{if }\alpha=0\end{cases} \tag{57}$$

is nonnegative, reflexive and decomposable in the sense of (50) with

$$\psi^0_\alpha(t)=\begin{cases}\frac{t^{1+\alpha}}{\alpha}\\ t\ln t\end{cases},\quad\hat{\psi}_\alpha(t)=\begin{cases}s^{1+\alpha}\\ s\end{cases}\quad\text{and}\quad\rho_\alpha(t)=\begin{cases}\frac{(1+\alpha)\,s^\alpha t}{\alpha} & \text{if }\alpha>0\\ t\ln s & \text{if }\alpha=0\end{cases} \tag{58}$$

and the class (57) satisfies (51) for const $=1$.

Proof. For arbitrary arguments s, $t>0$ and fixed parameters a, $b>0$ with the property $1/a+1/b=1$ it holds

$$st\leq\frac{s^a}{a}+\frac{t^b}{b} \tag{59}$$

with the equality iff $s^a = t^b$. Indeed, from the strict concavity of the logarithmic function we deduce the inequality

$$\ln(st) = \frac{1}{a}\ln s^a + \frac{1}{b}\ln t^b \leq \ln\left(\frac{s^a}{a} + \frac{t^b}{b}\right)$$

and the stated condition for equality. Substituting $s \longrightarrow s^\alpha$, $a \rightarrow (1+\alpha)/\alpha$ and $b \rightarrow 1+\alpha$ for $\alpha > 0$ we get

$$s^\alpha t \leq \frac{s^{1+\alpha}}{(1+\alpha)/\alpha} + \frac{t^{1+\alpha}}{1+\alpha}$$

with the equality condition $s^{\alpha a} = t^b$, i.e. $s^{1+\alpha} = t^{1+\alpha}$. This implies that the function $\psi_\alpha(s,t)$ is nonnegative and reflexive. It is easy to see that it satisfies (50) and also (51) for $\psi_0(s,t)$ given in (57)and const $= 1$. □

By Theorem 2,

$$\begin{aligned} \mathcal{D}_{\psi_\alpha}(P,Q) &= \int \psi_\alpha(p,q)\,\mathrm{d}\lambda, \quad \alpha \geq 0 && (60) \\ &= \begin{cases} P \cdot p^\alpha + \frac{1}{\alpha} Q \cdot q^\alpha - \frac{(1+\alpha)}{\alpha} Q{\cdot}p^\alpha & \text{if } \alpha > 0 \\ 1 + Q \cdot \ln q - Q \cdot \ln p & \text{if } \alpha = 0. \end{cases} && (61) \end{aligned}$$

is a family of decodistances. Its relation to the family of the classical power divergences $D_\alpha(P,Q)$ defined by (3) is rigorously established in the next theorem. This theorem refers to the family of functions

$$\varphi_\alpha(s,t) = s^{1+\alpha}t^{-\alpha} + \frac{t}{\alpha} - \frac{1+\alpha}{\alpha}\, s^\alpha t^{1-\alpha} \tag{62}$$

of arguments $s, t > 0$ parametrized by $\alpha > 0$ with the limit

$$\varphi_0(s,t) = \lim_{\alpha \downarrow 0} \varphi_\alpha(s,t) = \psi_0(s,t) \tag{63}$$

given by the second row in (57). One can easily verify that the functions (62), (63) define the following mixed power divergences

$$(1+\alpha)\,[\alpha\, D_{1+\alpha}(P,Q) + (1-\alpha)\, D_\alpha(P,Q)] = \int \varphi_\alpha(p,q)\,\mathrm{d}\lambda, \qquad \alpha \geq 0. \tag{64}$$

Theorem 3. The decodistances $\mathcal{D}_{\psi_\alpha}(P,Q)$ of (60) are distorted versions of the mixed power divergences (64) in the sense that the decodistance density $\psi_\alpha(p,q)$ appearing in (60) is for every $\alpha > 0$ the product $w_\alpha(q)\varphi_\alpha(p,q)$ of the weight function $w_\alpha(q) = q^\alpha$ with the power divergence density $\varphi_\alpha(p,q)$ appearing in (64) and for $\alpha = 0$ it is the limit

$$\psi_0(p,q) = \lim_{\alpha \downarrow 0} w_\alpha(q)\varphi_\alpha(p,q) = p + q\ln q - q\ln p \quad \text{(cf. (57), (63))}. \tag{65}$$

Proof. Let $\alpha > 0$ and

$$\tilde{\phi}_\alpha = (1+\alpha)\left[\alpha\phi_{1+\alpha} + (1-\alpha)\phi_\alpha\right] \in \Phi$$

for ϕ_α given by (4). Then

$$\tilde{\phi}_\alpha(s) = s^{1+\alpha} + \frac{1}{\alpha} - \frac{(1+\alpha)}{\alpha}s^\alpha \tag{66}$$

and we get from the definition of ϕ-divergence in (1)

$$D_{\tilde{\phi}_\alpha}(P,Q) = \int q\tilde{\phi}_\alpha(p/q)\,\mathrm{d}\lambda.$$

By (66),

$$t\tilde{\phi}_\alpha(s/t) = \varphi_\alpha(s,t) \quad \text{(cf. (62))}$$

which proves the power divergence formula (64) for φ_α given by (62). The equality $\psi_\alpha(s,t) = t^\alpha\varphi_\alpha(s,t)$ is clear for all $s,t>0$ from the definitions of $\psi_\alpha(s,t)$ and $\varphi_\alpha(s,t)$ in (57), (62). Verification of the convergence (65) is easy. □

The next theorem deals with the continuity of the decodistances (61) at $\alpha = 0$. It assumes that for some $\beta > 0$

$$p^\beta,\ q^\beta, \ln p \in \mathbb{L}_1(Q) \quad \text{for all } P\in\mathcal{P},\ Q\in\mathcal{P}\cup\mathcal{P}_{\text{emp}}. \tag{67}$$

This assumption is equivalent to

$$p_\theta^\beta,\ p_{\theta_0}^\beta, \ln p_\theta \in \mathbb{L}_1(P_{\theta_0}) \quad \text{for all } \theta,\theta_0\in\Theta$$

because if $Q = P_n \in \mathcal{P}_{\text{emp}}$ then (67) automatically holds for all $\beta > 0$.

Theorem 4. If (67) holds for some $\beta > 0$ then the decodistances $\mathcal{D}_{\psi_\alpha}(P,Q)$ from (61) are well defined for all $0 \le \alpha \le \beta$ and satisfy the limit relation

$$\lim_{\alpha\downarrow 0}\mathcal{D}_{\psi_\alpha}(P,Q) = \mathcal{D}_{\psi_0}(P,Q) = Q\cdot\ln q - Q\cdot\ln p < \infty. \tag{68}$$

Proof. The convergences $P\cdot p^\alpha \to 1$ and $Q\cdot p^\alpha \to 1$ follow from the assumptions of integrability and from the monotone convergence theorem for integrals. The convergence of

$$\frac{1}{\alpha}\left(Q\cdot q^\alpha - Q\cdot p^\alpha\right) = Q\cdot\frac{q^\alpha - 1}{\alpha} - Q\cdot\frac{p^\alpha - 1}{\alpha}$$

to the meaningful above bounded limit $Q\cdot\ln q - Q\cdot\ln p$ follows from the monotone convergence as well. Indeed, for every fixed $t > 0$

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\frac{t^\alpha - 1}{\alpha} = \frac{1 - t^\alpha(1-\ln t)}{\alpha^2} > \frac{1 - t^\alpha t^{-\alpha}}{\alpha^2} = 0$$

so that the expressions $(q^\alpha - 1)/\alpha$ and $(q^\alpha - 1)/\alpha$ tend monotonically to $\ln q$ and $\ln p$. □

In accordance with (54) and (55), the decodistances

$$\hat{\mathcal{D}}_{\psi_\alpha}(P_{\hat\theta}) + P_n \cdot \rho_\alpha(p_{\hat\theta}) = \begin{cases} P_{\hat\theta} \cdot p_{\hat\theta}^\alpha + \frac{1}{\alpha} P_{\theta_0} \cdot p_{\theta_0}^\alpha - \frac{(1+\alpha)}{\alpha} P_{\theta_0} \cdot p_{\hat\theta}^\alpha & \text{if } \alpha > 0 \\ 1 + P_{\theta_0} \cdot \ln p_{\theta_0} - P_{\theta_0} \cdot \ln p_{\hat\theta} & \text{if } \alpha = 0 \end{cases} \quad \text{(cf. (61))} \tag{69}$$

define the family of estimators $\hat\theta_{n,\alpha}$ which minimize the functions

$$\tilde{\mathcal{D}}_\alpha(P_{\hat\theta}, P_n) = \begin{cases} P_{\hat\theta} \cdot p_{\hat\theta}^\alpha - \frac{(1+\alpha)}{\alpha} P_n \cdot p_{\hat\theta}^\alpha & \text{if } \alpha > 0 \\ 1 - P_n \cdot \ln p_{\hat\theta} & \text{if } \alpha = 0 \end{cases} \quad \text{(cf. (53), (58))}$$

i.e.,

$$\hat\theta_{n,\alpha} = \begin{cases} \operatorname{argmin}_{\hat\theta} \left(\int p_{\hat\theta}^{1+\alpha} \, \mathrm{d}\lambda - \frac{1+\alpha}{n\alpha} \sum_{i=1}^n p_{\hat\theta}^\alpha(X_i) \right) & \text{if } \alpha > 0 \\ \operatorname{argmax}_{\hat\theta} \sum_{i=1}^n \ln p_{\hat\theta}(X_i) & \text{if } \alpha = 0. \end{cases} \tag{70}$$

In view of the interpretation of the decodistances (69) in Theorem 2, the estimators (70) can be called ***minimum distorted power divergence estimators***.

Example 4. By (70),

$$\hat\theta_{n,1} = \operatorname{argmin}_{\hat\theta} \left(\int p_{\hat\theta}^2 \, \mathrm{d}\lambda - \frac{2}{n} \sum_{i=1}^n p_{\hat\theta}(X_i) \right)$$

so that this estimator coincides with the L_2-estimator $\hat\theta_n$ from Example 3. The family of estimators $\hat\theta_{n,\alpha}$ from (70) smoothly connects this robust estimator with the efficient MLE $\hat\theta_{n,0}$ when the parameter α decreases from 1 to 0.

Remark. The above mentioned robustness and efficiency properties expected in the family of the estimators (70) were in fact confirmed by Basu et al. (1998) who first introduced these estimators and related divergences. The verification of nonnegativity and reflexivity given in the proof of Theorem 1 seems to be new. In fact, the relation of the weighted power divergences (70) to the classical power divergences (3) of Liese and Vajda (1987) and Read and Cressie (1988) was not clarified by Basu et al. and the precise relation given in Theorem 2 is a new result.

6. DISTURBED RÉNYI DIVERGENCES

In this section we propose for probability measures $P \in \mathcal{P}$ and $Q \in \mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{\text{emp}}$ considered in the previous sections a family of decodistances $\mathcal{D}_\alpha(P, Q)$ for $\alpha > 0$ which are not of the integral type as $\mathcal{D}_\psi(P, Q)$ of (44) or $\mathcal{D}_{\psi_\alpha}(P, Q)$ of (61). Our proposal is based on the following theorem.

Theorem 5. The formula

$$\mathcal{D}_\alpha(P,Q) = \begin{cases} \ln\left(P\cdot p^\alpha\right) + \frac{1}{\alpha}\ln(Q\cdot q^\alpha) - \frac{1+\alpha}{\alpha}\ln(Q\cdot p^\alpha) & \text{if} \quad 0<\alpha<\beta \\ Q\cdot\ln q - Q\cdot\ln p & \text{if} \quad \alpha = 0. \end{cases} \tag{71}$$

well defines a family of decodistances satisfying (39) and the limit relation $\mathcal{D}_\alpha(P,Q) \to \mathcal{D}_0(P,Q)$ for $\alpha \downarrow 0$.

Proof. Under (16) the expressions $\ln(Q\cdot q^\alpha)$, $\ln(Q\cdot p^\alpha)$ and $Q\cdot\ln p$ appearing in (71) are finite so that the expressions $\mathcal{D}_\alpha(P,Q)$ are well defined by (71). Substituting for $\alpha>0$

$$s = \frac{p^\alpha}{\left(\int p^{\alpha a}\,\mathrm{d}\lambda\right)^{1/b}}, \quad t = \frac{q}{\left(\int q^b\,\mathrm{d}\lambda\right)^{1/b}} \quad \text{and} \quad a = \frac{1+\alpha}{\alpha}, \quad b = 1+\alpha$$

in the inequality (59) and integrating both sides we obtain the Hölder inequality

$$\int p^\alpha q\,\mathrm{d}\lambda \le \left(\int p^{1+\alpha}\,\mathrm{d}\lambda\right)^{\alpha/(1+\alpha)} \left(\int q^{1+\alpha}\,\mathrm{d}\lambda\right)^{1/(1+\alpha)}$$

with the equality iff $p^{\alpha a} = q^b$ λ-a.s., i.e. iff $p = q$ λ-a.s. Thus the Rényi type divergence

$$\mathcal{D}_\alpha(P,Q) = \ln\left[\left(\int p^{1+\alpha}\,\mathrm{d}\lambda\right)^{\alpha/(1+\alpha)} \left(\int q^{1+\alpha}\,\mathrm{d}\lambda\right)^{1/(1+\alpha)}\right] - \ln\int p^\alpha q\,\mathrm{d}\lambda \quad (\text{cf.}(71))$$

is pseudodistance which is equivalently given by (71) and thus satisfies the decomposability (38) for

$$\hat{\mathcal{D}}_\alpha(P) = \ln(P\cdot p^\alpha),\ \mathcal{D}_{0,\alpha}(Q) = \frac{1}{\alpha}\ln(Q\cdot q^\alpha),\ \delta_\alpha(s) = \frac{1+\alpha}{\alpha}s,\ \rho_\alpha(s) = s^\alpha. \tag{72}$$

Therefore the expressions $\mathcal{D}_\alpha(P,Q)$ are decodistances. The limit relation

$$\mathcal{D}_0(P,Q) = \lim_{\alpha\downarrow 0}\mathcal{D}_\alpha(P,Q)$$

can be proved in a similar manner as in the proof of Theorem 3. □

There is some similarity between the decodistances $\mathcal{D}_\alpha(P,Q)$, $\alpha>0$ of (71) and the Rényi divergences

$$\mathcal{R}_\alpha(P,Q) = \frac{1}{\alpha-1}\ln\left(Q\cdot(p/q)^\alpha\right), \alpha>0 \quad (\text{cf. Rényi (1961)}.$$

Namely, replacing in the decodistance formula

$$\mathcal{D}_\alpha(P,Q) = \ln\frac{Q\cdot(p^{1+\alpha}/q)}{Q\cdot p^\alpha} + \frac{1}{\alpha}\ln\frac{Q\cdot q^\alpha}{Q\cdot p^\alpha}$$

the ratios of expectations by the expectations of ratios, we get

$$\mathcal{D}_\alpha(P,Q) = \ln(Q\cdot(p/q)) + \frac{1}{\alpha}\ln(Q\cdot(q/p)^\alpha) = \mathcal{R}_{\alpha+1}(Q,P) \tag{73}$$

Therefore we call the special decodistances (71) ***distorted Rényi distances***.

If we substitute in the Rényi type divergences $\mathcal{D}_\alpha(P_{\hat\theta}, P_{\theta_0})$ with parameters $\hat\theta$, $\theta_0 \in \Theta$ the hypothetical distribution P_{θ_0} by the empirical distribution P_n, we get the divergences $\mathcal{D}_\alpha(P_{\hat\theta}, P_n)$ satisfying the integrability condition of Theorem 4 for all $P_{\hat\theta} \in \mathcal{P}$ and $\beta > 0$. From (40) and (72) we obtain the family of ***minimum distorted Rényi distance estimators***

$$\begin{aligned} \hat\theta_{n,\alpha} &= \operatorname{argmin}_{\hat\theta} \left[\ln(P_{\hat\theta} \cdot p_{\hat\theta}^{\alpha}) - \frac{1+\alpha}{\alpha n} \ln \sum_{i=1}^{n} p_{\hat\theta}^{\alpha}(X_i) \right] \quad \text{for} \quad \alpha > 0, & (74) \\ \hat\theta_{n,0} &= \operatorname{argmin}_{\hat\theta} \left[-\frac{1}{n} \sum_{i=1}^{n} \ln p_{\hat\theta}(X_i) \right]. & (75) \end{aligned}$$

The estimates $\hat\theta_{\alpha,n}$ of (74) tend under obvious regularity to the MLE (75),

$$\lim_{\alpha \downarrow 0} \hat\theta_{n,\alpha} = \hat\theta_{n,0}. \tag{76}$$

Example 5. The minimum distorted Rényi distance estimator $\hat\theta_{n,1}$ defined by (74) differs from the minimum distorted power divergence estimator $\hat\theta_{n,1}$ from Example 4 in that it minimizes the ratio

$$\int p_{\hat\theta}^2 \, \mathrm{d}\lambda \Big/ \prod_{i=1}^{n} \sqrt[n]{p_{\hat\theta}^2(X_i)}$$

instead of the difference

$$\int p_{\hat\theta}^2 \, \mathrm{d}\lambda - \frac{2}{n} \sum_{i=1}^{n} p_{\hat\theta}(X_i).$$

It is easy to verify, that the two estimators are different.

Acknowledgement This research was supported by the grants GA ČR 102/07/1131 and MŠMT 1M 0572.

References

[1] A. Basu, I. R. Harris, N.L. Hjort and M. C. Jones (1998). "Robust and efficient estimation by minimizing a density power divergence," *Biometrika*, vol. 85, No. 3, pp. 549–559.

[2] M. Broniatowski and A. Keziou (2006). "Minimization of ϕ-diveregnces on sets of signed measures," *Studia Scientiarum Mathematica Hungarica*, vol. 43, pp. 403–442.

[3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseuw and W. A. Stahel (1986). *Robust Statistics: The approach Based on Influence Functions*, New York: Willey.

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

[5] F. Liese and I. Vajda, (1987). "On divergences and informations in statistics and information theory," *IEEE Trans.actions on Information Theory*, vol. 52, No. 10, pp. 4394–4412.

[6] C. Miescke and F. Liese (2008). *Statistical Decision Theory*, Berlin: Springer.

[7] M. R. C. Read and N. A. C. Cressie (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Berlin: Springer.

[8] A. Rényi (1961). "On measures of entropy and information," *Proc. 4-th Berkeley Symp. on Probability and Statistics*, vol. 1, pp. 547-561. Berkeley: University of California Press.

[9] A. Toma, M. Broniatowski (2008). "Minimum divergence estimators and tests: Robustness results," submitted.

[10] A. W. van der Vaart and J. A. Wellner (1996). *Weak Convergence and Empirical Processes,* Berlin: Springer.