KYBERNETIKA — VOLUME 48 (2012), NUMBER 4, PAGES 795–808

MODIFIED POWER DIVERGENCE ESTIMATORS IN NORMAL MODELS – SIMULATION AND COMPARATIVE STUDY

Iva Frýdlová, Igor Vajda and Václav Kůs

Point estimators based on minimization of information-theoretic divergences between empirical and hypothetical distribution induce a problem when working with continuous families which are measure-theoretically orthogonal with the family of empirical distributions. In this case, the ϕ-divergence is always equal to its upper bound, and the minimum ϕ-divergence estimates are trivial. Broniatowski and Vajda [3] proposed several modifications of the minimum divergence rule to provide a solution to the above mentioned problem. We examine these new estimation methods with respect to consistency, robustness and efficiency through an extended simulation study. We focus on the well-known family of power divergences parametrized by $\alpha \in \mathbb{R}$ in the Gaussian model, and we perform a comparative computer simulation for several randomly selected contaminated and uncontaminated data sets, different sample sizes and different ϕ-divergence parameters.

Keywords: minimum ϕ-divergence estimation, subdivergence, superdivergence, PC simulation, relative efficiency, robustness

Classification: 62B05, 62H30

1. INTRODUCTION

As was already mentioned in many previous publications, e. g. in Liese and Vajda [6], the well-known information-theoretic divergence measures, introduced in the 60ties by A. Renyi and I. Csiszar, cannot be directly applied in statistical estimation, since the divergence between the theoretical absolutely continuous probability measure and the discrete empirical probability measure takes on infinite values.

In 2006 Liese and Vajda [6] and independently Broniatowski and Keziou [1] established a general suprema representation of ϕ-divergences, which can be used for minimum divergence estimation. Another modification, referred to as *dual ϕ-divergence estimators*, was introduced by Broniatowski and Keziou [2] in 2009. These modifications were studied and extended by Broniatowski and Vajda [3]. They altered the traditional ϕ-divergences into *subdivergences* and *superdivergences* and defined *maximum subdivergence estimators with escort parameter θ* and *minimum superdivergence estimators*. Recently, Toma and Leoni-Aubin [7] and Toma and Broniatowski [8] explored theoretically this class of the so called dual ϕ-divergence estimators with respect to robustness through the influence function approach. They deal with the asymptotic relative efficiency of some robust hypothesis tests based on dual divergence saddlepoint approximations in the framework of parametric models for both non-contaminated and contaminated data (e. g. scale normal or Cauchy models).

We follow up the work of Broniatowski and Vajda [3] and focus on the important special cases of the so-called *power subdivergences*, *power superdivergences*, and the corresponding estimators. We describe the basic formulas for these estimators and the relationships between them. The main interest of our research (see also [5]) is to examine these modifications in practical use as to the consistency, robustness and efficiency of the estimators. We focus on the well-known family of power divergences parametrized by $\alpha \in \mathbb{R}$ under the normal model. We run a comparative computer simulation for several randomly selected contaminated and uncontaminated data sets, and we study the behavior of estimators for different sample sizes and different ϕ-divergence parameters.

2. BASIC CONCEPTS

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $\tilde{\mathcal{P}}$ be a set of all probability measures on $(\mathcal{X}, \mathcal{A})$. If $P \in \tilde{\mathcal{P}}$ is dominated by a σ-finite measure λ on $(\mathcal{X}, \mathcal{A})$, then $p = \mathrm{d}P/\mathrm{d}\lambda$ is the Radon–Nikodym density of P with respect to measure λ. Now, let $P, Q \in \tilde{\mathcal{P}}$. A *$\phi$-divergence* is a function $D_\phi : \tilde{\mathcal{P}} \times \tilde{\mathcal{P}} \to [0, \infty]$ defined by

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

where $\phi : (0, \infty) \to \mathbb{R}$ is a convex function, $\{P, Q\} \ll \lambda$, $p = \mathrm{d}P/\mathrm{d}\lambda$ and $q = \mathrm{d}Q/\mathrm{d}\lambda$. For this formula to be well defined, we put

$$q \, \phi\left(\frac{p}{q}\right) = \begin{cases} q \, \phi(0) & \text{if} \quad p = 0, \\ p \, \phi(\infty)/\infty & \text{if} \quad q = 0, \end{cases}$$

where $\phi(0) := \lim_{t \to 0_+} \phi(t)$ and $\phi(\infty)/\infty := \lim_{t \to \infty} \frac{\phi(t)}{t}$, while "$0 \cdot \infty = 0$".

From now on, we shall consider only ϕ which are twice differentiable, strictly convex functions with $\phi(1) = 0$ and endowed with well defined continuous extension to $t = 0_+$ denoted by $\phi(0)$. Let $\mathbf{\Phi}$ be the class of all such functions. As to the probability measures, we deal with P and Q which are either measure-theoretically equivalent $P \equiv Q$ (i. e. $pq > 0$ with respect to $\lambda = P + Q$ a.s.), or measure-theoretically orthogonal $P \perp Q$ (i. e. $pq = 0$ λ-a.s.).

For each generating function ϕ it holds that $\phi(0) + \phi(\infty)/\infty > 0$ and $\phi(1) \leq D_\phi(P, Q) \leq \phi(0) + \phi(\infty)/\infty$ for every $P, Q \in \tilde{\mathcal{P}}$. Moreover,

(i) $D_\phi(P, Q) = \phi(1)$ if $P = Q$,

(ii) if ϕ is strictly convex at 1, $D_\phi(P, Q) = \phi(1)$ iff $P = Q$,

(iii) $D_\phi(P, Q) = \phi(0) + \phi(\infty)/\infty$ if $P \perp Q$ (i. e. P, Q are singular),

(iv) if ϕ is strictly convex at 1 and $\phi(0) + \phi(\infty)/\infty < \infty$, $D_\phi(P, Q) = \phi(0) + \phi(\infty)/\infty$ iff $P \perp Q$.

The proofs of these properties can be be found in [9].

In the sequel, we use the power divergences

$$D_\alpha(P,Q) := D_{\phi_\alpha}(P,Q), \quad \text{for all } \alpha \in \mathbb{R}, \tag{2}$$

where

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, \quad \alpha \neq 0, \alpha \neq 1, \tag{3}$$

with the limiting cases $\phi_0(t) = -\ln t + t - 1$ and $\phi_1(t) = t\ln t - t + 1$. The function ϕ_α satisfies the relations

$$\phi_\alpha(0) = \begin{cases} \frac{1}{\alpha} & \text{if} \quad \alpha > 0\,, \\ \infty & \text{if} \quad \alpha \leq 0\,, \end{cases} \quad \text{and} \quad \phi_\alpha(\infty)/\infty = \begin{cases} \frac{1}{1-\alpha} & \text{if} \quad \alpha < 1\,, \\ \infty & \text{if} \quad \alpha \geq 1\,, \end{cases}$$

which implies

$$0 \leq D_\alpha(P,Q) \leq \begin{cases} \frac{1}{\alpha(1-\alpha)} & \text{if} \quad 0 < \alpha < 1\,, \\ \infty & \text{otherwise}\,, \end{cases} \tag{4}$$

for D_α divergences given by the formula

$$D_\alpha(P,Q) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left(\int p^\alpha q^{1-\alpha}\,\mathrm{d}\lambda - 1\right)\,, & \alpha \neq 0, \alpha \neq 1, \\ \int \ln \frac{q}{p}\,\mathrm{d}Q = I(P,Q)\,, & \alpha = 0, \\ \int \ln \frac{p}{q}\,\mathrm{d}P = I(Q,P)\,, & \alpha = 1. \end{cases} \tag{5}$$

Here, $I(P,Q)$ denotes the Kulback–Leibler informational divergence. If $0 < \alpha < 1$ then the right hand side equality in (4) takes place if and only if $P \perp Q$. Otherwise it takes place if $\alpha \leq 0$ and $Q \not\ll P$ or if $\alpha \geq 1$ and $P \not\ll Q$.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations governed by $P_{\theta_0} \in \mathcal{P}$, where $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ is a parametric family of probability measures on $(\mathcal{X}, \mathcal{A})$, and we assume that for every $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$, it holds $P_{\theta_1} \neq P_{\theta_2}$. Moreover, we assume the family $\mathcal{P}$ to be nonatomic (continuous), i.e. for all $\theta \in \Theta$ and $x \in \mathcal{X}$ we require $P_\theta(\{x\}) = 0$. Let the sample $X_1, X_2, \ldots, X_n$ be represented by the empirical probability measure $P_n = \frac{1}{n}\sum_{i=1}^n P_{X_i}$, where P_{x_i} are the Dirac probability measures concentrated at realizations $x_i \in \mathbb{R}$ of the random sample X_i, $i = 1, 2, \ldots, n$.

For $\phi \in \mathbf{\Phi}$, the parameter θ_0 is the unique minimizer of the ϕ-divergence $D_\phi(\theta, \theta_0)$ with respect to θ, and since the empirical probability measure P_n converges weakly to P_{θ_0}, it is reasonable to define the minimum ϕ-divergence estimator as follows. For $\phi \in \mathbf{\Phi}$, we say that an estimator $\widehat{\theta}_n : \mathcal{X}^n \to \Theta$ of the true parameter $\theta_0 \in \Theta$ is *minimum ϕ-divergence estimator* if

$$\widehat{\theta}_n = \mathrm{argmin}_\theta D_\phi(P_\theta, P_n).$$

The problem we encounter with these estimators is that the continuous family $\mathcal{P}$ and the family of empirical distributions $\mathcal{P}_{\mathrm{emp}}$ are measure-theoretically orthogonal, i.e. $P_\theta \perp P_n$ for every $P_\theta \in \mathcal{P}$ and $P_n \in \mathcal{P}_{\mathrm{emp}}$. This implies that for every $P_\theta \in \mathcal{P}$ and $P_n \in \mathcal{P}_{\mathrm{emp}}$

$$D_\phi(P_\theta, P_n) = \phi(0) + \phi(\infty)/\infty \tag{6}$$

and the above defined estimates are trivial. To face this problem, it is possible to use some nonparametric density estimator like we did in [4] by implementing the histogram. However, these methods bring another unpleasant obstructions such as the bandwidth selection in case of the histogram type estimator.

In the next section, we give several modifications of the minimum divergence principle to avoid these complications.

3. POWER SUBDIVERGENCE AND POWER SUPERDIVERGENCE ESTIMATORS

Throughout this section, we present the results of Broniatowski and Vajda [3] which are the main subject of our computer simulations in Section 4. We consider the probability measures $P_\theta \in \mathcal{P}$ and $Q \in \mathcal{Q}$ for $\mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{\mathrm{emp}}$ and the corresponding ϕ-divergences $D_\phi(P_\theta, Q)$ well defined for all pairs $(P_\theta, Q) \in \mathcal{P} \times \mathcal{Q}$. Consider the family of finite expectations

$$\underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, Q) = \int \phi'(p_\theta/p_{\tilde{\theta}})\, \mathrm{d}P_\theta + \int \phi^{\#}(p_\theta/p_{\tilde{\theta}})\, \mathrm{d}Q, \quad (P_\theta, Q) \in \mathcal{P} \times \mathcal{Q}\,, \tag{7}$$

parametrized by $\phi \in \mathbf{\Phi}$ and $\tilde{\theta} \in \Theta$, where $\phi^{\#}(t) = \phi(t) - t\phi'(t)$, ϕ' denotes the derivative of ϕ, $p_\theta = \mathrm{d}P_\theta/\mathrm{d}\lambda$ and $p_{\tilde{\theta}} = \mathrm{d}P_{\tilde{\theta}}/\mathrm{d}\lambda$ are the probability density functions, $P_{\tilde{\theta}} \in \mathcal{P}$. For (7) to be correctly defined, we assume that the integrals exist and have finite values.

Now, the *maximum subdivergence estimators* (briefly, the max$\underline{\mathrm{D}}_\phi$-estimators) are defined as

$$\begin{aligned} \tilde{\theta}_{\phi,\theta,n} &= \mathrm{argmax}_{\tilde{\theta}}\ \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n) \\ &= \mathrm{argmax}_{\tilde{\theta}} \left[\int \phi'\left(\frac{p_\theta}{p_{\tilde{\theta}}}\right) \mathrm{d}P_\theta + \frac{1}{n}\sum_{i=1}^{n} \phi^{\#}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)\right] \end{aligned} \tag{8}$$

with the so called escort parameter $\theta \in \mathbf{\Theta}$. Further, we define the *minimum superdivergence estimators* (briefly, the min $\bar{\mathrm{D}}_\phi$-estimators) as

$$\begin{aligned} \theta_{\phi,n} &= \mathrm{argmin}_\theta\ \sup_{\tilde{\theta}}\ \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n) \\ &= \mathrm{argmin}_\theta\ \sup_{\tilde{\theta}} \left[\int \phi'\left(\frac{p_\theta}{p_{\tilde{\theta}}}\right) \mathrm{d}P_\theta + \frac{1}{n}\sum_{i=1}^{n} \phi^{\#}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)\right]. \end{aligned} \tag{9}$$

The maximum subdivergence and minimum superdivergence estimators are both Fisher consistent estimators, see [3] for details.

If we restrict ourselves to the subclasses of these estimators determined by the power divergences (2) and (3), we have for $\alpha > 0$ the formulas

$$\tilde{\theta}_{\alpha,\theta,n} = \mathrm{argmin}_{\tilde{\theta}}\ M_{\alpha,\theta}(\tilde{\theta}, P_n) \tag{10}$$

and

$$\theta_{\alpha,n} = \mathrm{argmax}_\theta \mathrm{inf}_{\tilde{\theta}}\ M_{\alpha,\theta}(\tilde{\theta}, P_n) \quad \text{or} \quad \theta_{\alpha,n} = \mathrm{argmax}_\theta\ M_{\alpha,\theta}(\tilde{\theta}_{\alpha,\theta,n}, P_n)\,, \tag{11}$$

where

$$M_{\alpha,\theta}(\tilde{\theta}, P_n) = \begin{cases} \frac{1}{1-\alpha} \int \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha \mathrm{d}P_{\tilde{\theta}} + \frac{1}{\alpha n} \sum_{i=1}^n \left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)^\alpha & \text{if } \alpha > 0,\ \alpha \neq 1, \\ -\int \ln \frac{p_\theta}{p_{\tilde{\theta}}} \,\mathrm{d}P_\theta + \frac{1}{n} \sum_{i=1}^n \frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)} & \text{if } \alpha = 1, \end{cases} \tag{12}$$

and for $\alpha = 0$ the formulas

$$\tilde{\theta}_{0,\theta,n} = \mathrm{argmax}_{\tilde{\theta}}\ \Sigma_{i=1}^n \ln p_{\tilde{\theta}}(X_i) \quad \text{and} \quad \theta_{0,n} = \mathrm{argmax}_\theta\ \Sigma_{i=1}^n \ln p_\theta(X_i).$$

It is obvious that in the case of $\alpha = 0$ the estimators $\tilde{\theta}_{0,\theta,n}$, $\theta_{0,n}$ coincide with the MLE's, hence the classes of max$\underline{\mathrm{D}}_\phi$-estimators and of min$\bar{\mathrm{D}}_\phi$-estimators are both extensions of the MLE. Moreover, the divergences $\underline{\mathrm{D}}_{\phi,\tilde{\theta}}$ also differ from the original ϕ-divergences in general and the above mentioned problem (6) is bypassed by this modification.

3.1. Application to the normal model

Let the observation space $(\mathcal{X}, \mathcal{A})$ be $(\mathbb{R}, \mathcal{B})$ and $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R},\ \sigma > 0\}$ be the normal family with parameters of location μ and scale σ. We are interested in max$\underline{\mathrm{D}}_\alpha$-estimates $(\tilde{\mu}_{\alpha,\mu,\sigma,n}, \tilde{\sigma}_{\alpha,\mu,\sigma,n})$ with divergence power parameters $\alpha \geq 0$ and escort parameters $(\mu, \sigma) \in \mathbb{R} \times (0, \infty)$.

For $\alpha = 0$ these estimators reduce to

$$(\tilde{\mu}_{0,\mu,\sigma,n}, \tilde{\sigma}_{0,\mu,\sigma,n}) = \left(\frac{1}{n} \sum_{i=1}^n X_i,\ \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \tilde{\mu}_{0,\mu,\sigma,n})^2} \right), \tag{13}$$

which are the maximum likelihood estimators in the family of normal distributions. For $\alpha > 0$, $\alpha \neq 1$ the function (12) becomes

$$M_{\alpha,\mu,\sigma}(\tilde{\mu}, \tilde{\sigma}, P_n) = \frac{1}{1-\alpha} \int \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^\alpha \mathrm{d}P_{\tilde{\mu},\tilde{\sigma}} + \frac{1}{\alpha n} \sum_{i=1}^n \left(\frac{p_{\mu,\sigma}(X_i)}{p_{\tilde{\mu},\tilde{\sigma}}(X_i)}\right)^\alpha, \tag{14}$$

where

$$\left(\frac{p_{\mu,\sigma}(x)}{p_{\tilde{\mu},\tilde{\sigma}}(x)}\right)^\alpha = \left(\frac{\tilde{\sigma}}{\sigma}\right)^\alpha \exp\left\{ \frac{\alpha\,(x-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{\alpha\,(x-\mu)^2}{2\sigma^2} \right\} \tag{15}$$

and

$$\int \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^\alpha \mathrm{d}P_{\tilde{\mu},\tilde{\sigma}} = \exp\left\{ \frac{-\alpha(1-\alpha)(\mu-\tilde{\mu})^2}{2[\alpha\tilde{\sigma}^2 + (1-\alpha)\sigma^2]} - \ln \frac{\sqrt{\alpha\tilde{\sigma}^2 + (1-\alpha)\sigma^2}}{\tilde{\sigma}^\alpha \sigma^{1-\alpha}} \right\}. \tag{16}$$

For $\alpha = 1$

$$\begin{aligned} M_{1,\mu,\sigma}(\tilde{\mu}, \tilde{\sigma}, P_n) &= \lim_{\alpha \to 1} M_{\alpha,\mu,\sigma}(\tilde{\mu}, \tilde{\sigma}, P_n) \\ &= \frac{-(\mu-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{1}{2}\left[-\ln\left(\frac{\sigma}{\tilde{\sigma}}\right)^2 + \left(\frac{\sigma}{\tilde{\sigma}}\right)^2 - 1 \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left(\frac{\tilde{\sigma}}{\sigma}\right) \exp\left\{ \frac{(X_i-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{(X_i-\mu)^2}{2\sigma^2} \right\}. \end{aligned} \tag{17}$$

Now, the final full-length formulas for numerical computations of the $\max\underline{\mathrm{D}}_\phi$-estimators and $\min\bar{\mathrm{D}}_\phi$-estimators of normal location and scale can be obtained easily from (10) and (11) as

$$\begin{aligned}\tilde{\theta}_{\alpha,\theta,n} &= \mathrm{argmin}_{\tilde{\theta}} \quad \frac{1}{1-\alpha} \ \exp\left\{\frac{-\alpha(1-\alpha)(\mu-\tilde{\mu})^2}{2[\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2]} - \ln\frac{\sqrt{\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2}}{\tilde{\sigma}^\alpha\sigma^{1-\alpha}}\right\} \\ &\qquad + \frac{1}{\alpha n}\left(\frac{\tilde{\sigma}}{\sigma}\right)^\alpha \sum_{i=1}^n \exp\left\{\frac{\alpha\,(X_i-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{\alpha\,(X_i-\mu)^2}{2\sigma^2}\right\} \end{aligned} \tag{18}$$

and

$$\begin{aligned}\theta_{\alpha,n} &= \mathrm{argmax}_\theta \mathrm{inf}_{\tilde{\theta}} \quad \frac{1}{1-\alpha} \ \exp\left\{\frac{-\alpha(1-\alpha)(\mu-\tilde{\mu})^2}{2[\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2]} - \ln\frac{\sqrt{\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2}}{\tilde{\sigma}^\alpha\sigma^{1-\alpha}}\right\} \\ &\qquad + \frac{1}{\alpha n}\left(\frac{\tilde{\sigma}}{\sigma}\right)^\alpha \sum_{i=1}^n \exp\left\{\frac{\alpha\,(X_i-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{\alpha\,(X_i-\mu)^2}{2\sigma^2}\right\} \end{aligned} \tag{19}$$

for $\alpha > 0$, $\alpha \neq 1$. To obtain the estimators for $\alpha = 1$ we only straightforwardly insert (17) into (10) and (11).

4. COMPARATIVE SIMULATION STUDY IN THE NORMAL MODEL

In this section we consider both estimates in the normal family with scale parameter fixed at $\sigma = 1$ denoted as $\mathcal{P}_1 = \{P_\mu : \mu \in \mathbb{R}\}$, and in the normal family with the location parameter fixed at $\mu = 0$ denoted by $\mathcal{P}_0 = \{P_\sigma : \sigma \in (0,\infty)\}$. This section is to present the performances of the newly proposed estimators briefly described in the previous section. We study the following two families of estimators:

(i) *The power subdivergence and power superdivergence estimators of location* for the family $\mathcal{P}_1$ given for $\alpha = 0$ by

$$\mu_{0,n} = \tilde{\mu}_{0,\mu,n} = \bar{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

and for $\alpha > 0$ by

$$\tilde{\mu}_{\alpha,\mu,n} = \mathrm{argmin}_{\tilde{\mu}}\ M_{\alpha,\mu}(\tilde{\mu}, P_n)\,, \qquad \mu_{\alpha,n} = \mathrm{argmax}_\mu \mathrm{inf}_{\tilde{\mu}}\ M_{\alpha,\mu}(\tilde{\mu}, P_n)\,,$$

with $M_{\alpha,\mu}(\tilde{\mu}, P_n)$ given by (14) under parameter $\sigma = 1$ fixed, i.e.

$$\begin{aligned} M_{\alpha,\mu}(\tilde{\mu}, P_n) &= \frac{1}{1-\alpha}\left(\exp\{\alpha(\tilde{\mu}-\mu)(\tilde{\mu}-\mu)/2\}\right)^{\alpha-1} \\ &\quad + \frac{1}{\alpha n}\sum_{i=1}^n \exp\{\alpha(\tilde{\mu}-\mu)(\tilde{\mu}+\mu-2X_i)/2\}\,. \end{aligned} \tag{20}$$

For this maximum subdivergence estimator we used three different strategies for the choice of escort parameter μ, i.e. either μ fixed and independent of the data sample, or data based escort μ, which is not robust estimator (MLE), or data based escort μ, which is robust itself (MEDian).

(ii) *The power subdivergence and power superdivergence estimators of scale* for the family $\mathcal{P}_0$ given for $\alpha = 0$ by

$$\sigma^2_{0,n} = \tilde{\sigma}^2_{0,\sigma,n} = \boldsymbol{S}^2_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

and for $\alpha > 0$ by

$$\tilde{\sigma}_{\alpha,\sigma,n} = \text{argmin}_{\tilde{\sigma}}\ M_{\alpha,\sigma}(\tilde{\sigma}, P_n)\,, \qquad \sigma_{\alpha,n} = \text{argmax}_{\sigma}\text{inf}_{\tilde{\sigma}}\ M_{\alpha,\sigma}(\tilde{\sigma}, P_n)\,,$$

with $M_{\alpha,\sigma}(\tilde{\sigma}, P_n)$ given by (14) under the parameter $\mu = 0$ fixed, i. e.

$$\begin{aligned} M_{\alpha,\sigma}(\tilde{\sigma}, P_n) &= \frac{1}{1-\alpha}\frac{\tilde{\sigma}^{\alpha}\sigma^{1-\alpha}}{\sqrt{\alpha\tilde{\sigma}^2 + (1-\alpha)\sigma^2}} \\ &\quad + \frac{1}{\alpha n}\sum_{i=1}^{n}\left(\frac{\tilde{\sigma}}{\sigma}\right)^{\alpha}\exp\left\{\frac{\alpha X_i^2}{2}\left(\frac{1}{\tilde{\sigma}^2} - \frac{1}{\sigma^2}\right)\right\}. \end{aligned} \tag{21}$$

Again, for the maximum subdivergence estimator we used three basic different strategies for the choice of escort parameter σ, i. e. either σ fixed and independent of the data sample, or data based escort σ, which is not robust estimator (MLE), or data based escort σ, which is robust itself (MAD variant).

Apart from section 2, here, $X_1, \ldots, X_n$ are observations from the convex mixtures $P_\varepsilon = (1-\varepsilon)P + \varepsilon\tilde{Q}$, where P is the standard normal model $N(0,1)$ with location $\mu = 0$ and scale $\sigma = 1$, and $\tilde{Q}$ is successively normal $N(0,9)$, $N(0,100)$, logistic $Lo(0,1)$, and Cauchy $C(0,1)$ distributions. The contamination ε takes on the values 0, 0.01, 0.05, 0.1, 0.2, and 0.3. The sample sizes are considered successively $n =$ 20, 50, 100, 200, 500.

In the case of $\min\bar{\mathrm{D}}_\alpha$-estimators $\mu_{\alpha,n}$ and $\sigma_{\alpha,n}$, we consider power parameters $\alpha = 0$, 0.01, 0.05, 0.1, 0.2, and 0.5. In the case of $\max\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\mu}_{\alpha,\mu,n}$, we consider the same values of power parameter α and we select the escort parameters $\mu =$ 0, 0.1, 0.2, 0.5, 1 independently of the sample $X_1, \ldots, X_n$, and then also the data based escort parameter $\mu = \bar{\boldsymbol{X}}_n$ (MLE) or $\mu = \text{MED} = \text{med}_n(X_i)$ (the sample median). For $\max\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\sigma}_{\alpha,\sigma,n}$ we consider the same values of power parameter α and the escort parameters $\sigma^2 =$ 0.2, 0.4, 0.5, 0.6, 0.8, 1, 1.2, 1.5, 2, and similarly the data based choices $\sigma^2 = \boldsymbol{S}^2_n$ (MLE) or $\sigma =$MAD$= \text{med}_n(|X_j|)$ (MAD estimate of scale for known location parameter equal to 0, otherwise $\sigma =$MAD$= \text{med}_n(|X_j - \text{med}_n(X_i)|)$), or $\sigma = 1.483\,\text{MAD}$, where the constant 1.483 is a calibration constant for the normally distributed sample ensuring the Fisher consistency of the MAD estimate.

To evaluate the behavior of power superdivergence (or power subdivergence) estimators we generated K different data samples (K=1000) and we obtained K different estimates (further indexed by $^{(k)}$). We computed means and standard deviations

$$m(\mu) = \frac{1}{K}\sum_{k=1}^{K}\mu^{(k)}_{\alpha,n}\,, \qquad s(\mu) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mu^{(k)}_{\alpha,n} - m(\mu))^2}\,,$$

$$m(\sigma) = \frac{1}{K}\sum_{k=1}^{K} \sigma_{\alpha,n}^{(k)}\,, \qquad s(\sigma) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\sigma_{\alpha,n}^{(k)} - m(\sigma))^2},$$

of the min$\bar{\mathrm{D}}_\alpha$-estimators (or max$\underline{\mathrm{D}}_\alpha$-estimators) as well as the standard maximum likelihood estimators $\bar{\boldsymbol{X}}_n^{(k)}$ and $\boldsymbol{S}_n^{(k)}$ in the normal model. Thus, we obtained the empirical relative efficiencies

$$\mathrm{eref}(\mu) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\bar{\boldsymbol{X}}_n^{(k)})^2}{\frac{1}{K}\sum_{k=1}^{K}(\mu_{\alpha,n}^{(k)})^2} \qquad \text{and} \qquad \mathrm{eref}(\sigma) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{S}_n^{(k)} - 1)^2}{\frac{1}{K}\sum_{k=1}^{K}(\sigma_{\alpha,n}^{(k)} - 1)^2},$$

which compare the performance of the subdivergence (superdivergence) estimator with that of the MLE. If the value of empirical relative efficiency is greater than 1, we can say that the subdivergence (or superdivergence) estimator performs better than MLE, and if it is less than 1, we conclude the contrary.

4.1. Results for power subdivergence estimators of location

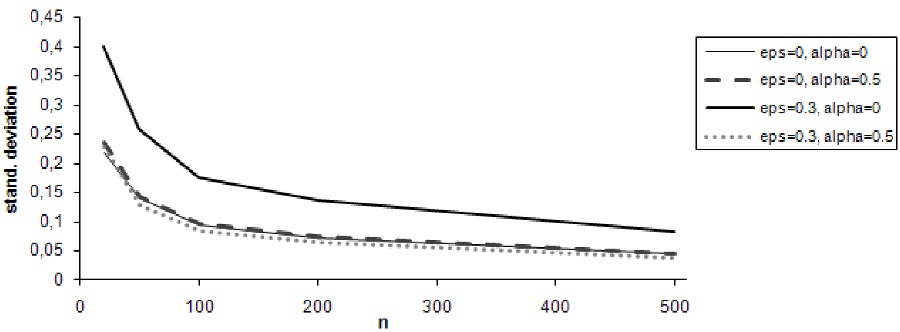

Fig. 1. Standard deviation $s(\tilde{\mu})$ of max$\underline{\mathrm{D}}_\alpha$-estimators with escort parameter $\mu = 0$ with respect to sample size n for data distributed by $(1-\varepsilon)N(0,1) + \varepsilon N(0,9)$.

When $\alpha = 0$, we can conclude that the estimates coincide with MLE, i. e. $\mathrm{eref}(\tilde{\mu}) = 1$, as was expected. In case of escort parameter $\mu = 0$, the max$\underline{\mathrm{D}}_\alpha$-estimators for uncontaminated data still more or less copy the behavior of MLE even for values of $\alpha > 0$, but as the contamination increases, we observe that the means and standard deviations of max$\underline{\mathrm{D}}_\alpha$-estimators move apart from the MLE taking on lower values than those of the MLE. In case of $m(\tilde{\mu})$ the difference is only slight (yet favourable), but in case of $s(\tilde{\mu})$ the difference is apparent (cf. Figure 1) and causes a fair increase in empirical relative efficiency, especially when the outliers get farther away as for the Cauchy contamination where the robustness of estimator escorted by $\mu = 0$ is rather stunning compared to MLE. The behavior of $\mathrm{eref}(\tilde{\mu})$ for different values of power parameter α shows that the robustness tendency is growing stronger with α increasing. Since the dependence on sample size n is almost constant for $n > 50$, we present in Figure 2 the values of $\mathrm{eref}(\tilde{\mu})$ only for $n = 500$ as a function of contamination ε for different levels of α which show the rising efficiency of $\max \underline{\mathrm{D}}_\alpha$-estimator (compared to MLE with $\alpha = 0$) with increasing contamination.

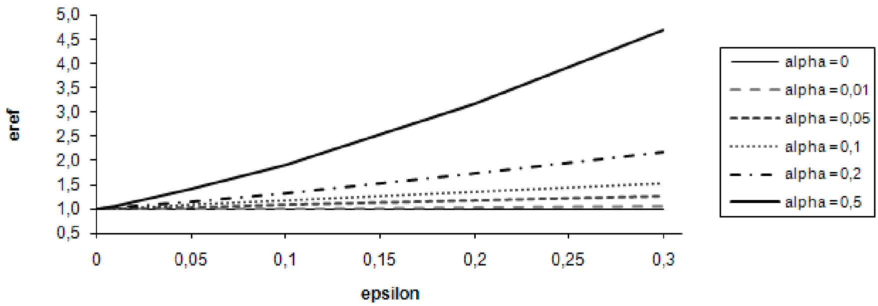

Fig. 2. Empirical relative efficiency $\mathrm{eref}(\tilde{\mu})$ of $\max \underline{\mathrm{D}}_\alpha$-estimators with escort parameter $\mu = 0$ with respect to contamination parameter ε for data distributed by $(1 - \varepsilon)N(0,1) + \varepsilon N(0,9)$.

All that was stated in the previous paragraph holds for $\mu = 0$. However, the situation changes to the worse for the parameter μ tending to 1. The consistency, efficiency, even the robustness tendencies slowly vanish, and apart from the case of $\mu = 0$, the $\max \underline{\mathrm{D}}_\alpha$-estimators do not possess the useful properties we would desire.

In accordance with the fact that the best results we obtained were for $\mu = 0$ which is the true parameter, some very good results were received for the escort parameter $\mu = \bar{\boldsymbol{X}}_n$. For the contaminations by $N(0,9)$, $N(0,100)$ and $Lo(0,1)$, we received perfect match with MLE for all values of ε. Nevertheless, an outstanding behavior was noticed in the case of contamination by Cauchy distribution, where the power subdivergence estimator shows significant resistance to distant outliers both in $s(\tilde{\mu})$ and $\mathrm{eref}(\tilde{\mu})$ (cf. Table 1) in comparison with the MLE standard estimator.

Even though these results are very encouraging, they are not as good as for fixed choice of escort parameter $\mu = 0$. Better results were obtained for estimators escorted by sample median, the well known robust estimate of location. Although the efficiencies $\mathrm{eref}(\tilde{\mu})$ again are not as good as for $\mu = 0$, it is clear that the choice of the escort sample median produces much more robust estimates than those escorted by MLE. Unlike the MLE escorted estimators, the median escorted estimators give us excellent, robust results for all cases with positive contamination, especially for Cauchy contamination (cf. Figure 3). For uncontaminated data, we receive almost a perfect match with the maximum likelihood estimator with only negligible differences from MLE.

To see whether the subdivergence estimators are at least as robust as sample median itself, we make a comparison of median escorted subdivergence estimators with the median itself by plugging it into the empirical relative efficiency formula instead of MLE. We can conclude that for the uncontaminated data the median is never better than subdivergence estimator for $\alpha = 0$, i.e. the MLE. When α is increasing, the estimates

α/n	20			50			100			200			500		
	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$\mathrm{eref}(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$\mathrm{eref}(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$\mathrm{eref}(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$\mathrm{eref}(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$\mathrm{eref}(\tilde{\mu})$
0.00	−0.41	24.2	1.00	0.01	25.1	1.00	0.21	6.93	1.00	−0.15	13.1	1.00	−0.42	19.0	1.00
0.01	0.30	8.75	7.66	−0.09	7.14	12.4	0.02	3.57	3.78	−0.02	4.66	7.94	0.00	3.24	34.4
0.05	−0.03	5.16	22.1	−0.06	5.06	24.6	0.02	3.57	3.78	−0.13	2.35	31.2	−0.05	2.18	76.1
0.10	−0.03	5.16	22.1	−0.07	4.27	34.6	−0.08	2.97	5.43	−0.16	2.16	36.8	−0.05	1.93	96.7
0.20	−0.10	3.60	45.2	−0.07	3.23	60.5	0.03	2.10	10.9	−0.14	1.98	43.6	−0.08	1.63	135
0.50	−0.07	2.49	94.7	−0.07	2.06	148	0.01	2.02	11.8	−0.15	1.65	62.9	−0.10	1.49	163

Tab. 1. Properties of $\max\underline{D}_\alpha$-estimators with escort parameter $\mu = \bar{\boldsymbol{X}}_n$ for generated mixture $0.7N(0,1) + 0.3C(0,1)$.

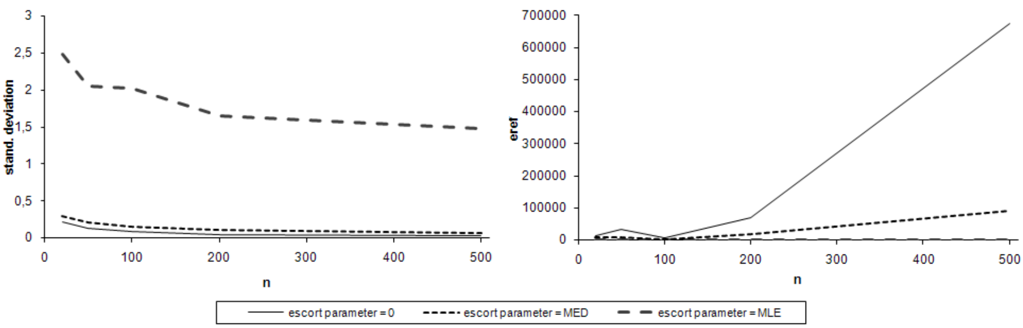

Fig. 3. Standard deviation $s(\tilde{\mu})$ and empirical relative efficiency $\mathrm{eref}(\tilde{\mu})$ of $\max\underline{D}_\alpha$-estimators with escort parameters $\mu = 0$, $\mu = \bar{\boldsymbol{X}}_n$ and $\mu = \mathrm{med}_n(X_i)$ with respect to sample size n for data distributed by $0.7N(0,1) + 0.3C(0,1)$ and divergence parameter $\alpha = 0.5$.

worsen a little (which corresponds with the previous results). For the contaminated data this trend reverses and we receive the best results for $\alpha = 0.5$. In the case of contamination by $N(0,9)$ or $Lo(0,1)$ (cf. Figure 4) we obtain estimates which are more robust than the median itself for all levels of contamination. In case of contamination by $N(0,100)$, or $C(0,1)$ the subdivergence estimators perform better than median only up to five, respectively ten percent contamination. However, even in those unfavourable cases, the subdivergence estimators are not substantially worse and their performance is comparable with the escorting sample median itself.

4.2. Results for power subdivergence estimators of scale

As expected, for $\alpha = 0$ we get the exact MLE, hence $\mathrm{eref}(\tilde{\sigma})$ is always equal to 1. For the uncontaminated data, the subdivergence estimators of scale more or less correspond with the MLE's, but they do not outperform them. When increasing α, the standard deviation $s(\tilde{\sigma})$ increases a little, which causes a certain loss of efficiency. For the contamination

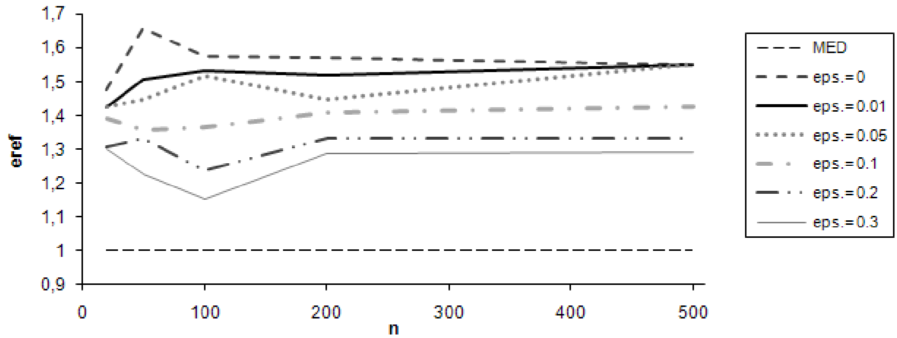

Fig. 4. Empirical relative efficiency $\text{eref}(\tilde{\mu})$ of the sample median and $\max\underline{D}_\alpha$-estimators with escort parameter $\mu = \text{med}_n(X_i)$ for data distributed by $(1-\varepsilon)N(0,1) + \varepsilon Lo(0,1)$ and divergence parameter $\alpha = 0.5$.

$\varepsilon > 0$ and escort parameters $\sigma^2 = 1,\ 1.2,\ 1.5,\ 2$ we observe a loss of consistency of $\max\underline{D}_\alpha$-estimators, likewise for the MLE. However, with increasing contamination the performance of $\max\underline{D}_\alpha$-estimators is better than MLE. The best results are obtained mostly for the escort parameter $\sigma^2 \in \langle 0.4, 0.6 \rangle$ depending on the contamination model. Here, the estimates preserve the consistency even for highly contaminated data with favourable values of $m(\tilde{\sigma})$ and $s(\tilde{\sigma})$ resulting in high empirical relative efficiency (cf. Table 2). The best values of $m(\tilde{\sigma})$ were obtained for power parameter $\alpha = 0.5$, however, the best empirical relative efficiencies for small data samples were achieved at $\alpha = 0.2$. For data contaminated by $N(0,100)$ or $C(0,1)$, the subdivergence estimators perform better then MLE even for very small level of contamination $\varepsilon = 0.01$ and all values of escort parameter σ.

α/n	20			50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\text{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\text{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\text{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\text{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\text{eref}(\tilde{\sigma})$
0.00	1.29	0.37	1.00	1.34	0.26	1.00	1.33	0.18	1.00	1.34	0.13	1.00	1.34	0.08	1.00
0.01	1.26	0.32	1.29	1.30	0.21	1.33	1.29	0.15	1.29	1.30	0.10	1.30	1.30	0.07	1.29
0.05	1.20	0.24	2.38	1.21	0.15	2.76	1.21	0.10	2.58	1.21	0.07	2.71	1.21	0.04	2.67
0.10	1.16	0.21	3.29	1.16	0.13	4.45	1.16	0.09	4.32	1.15	0.06	4.75	1.15	0.04	4.79
0.20	1.13	0.21	3.85	1.11	0.12	6.45	1.11	0.08	7.21	1.11	0.06	8.76	1.11	0.04	9.45
0.50	1.12	0.24	3.24	1.09	0.14	6.30	1.08	0.10	8.78	1.07	0.07	13.0	1.07	0.04	16.9

Tab. 2. Properties of $\max\underline{D}_\alpha$-estimators with escort parameter $\sigma^2 = 0.5$ for generated mixture $0.9N(0,1) + 0.1N(0,9)$.

As in the location case, we tried to escort the subdivergence estimator with the MLE $\sigma = \boldsymbol{S}_n$. However, we received only a perfect match with maximum likelihood

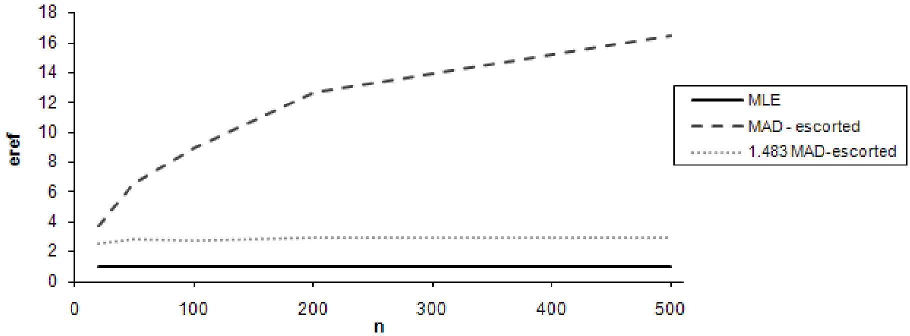

Fig. 5. Empirical relative efficiency $\mathrm{eref}(\tilde{\sigma})$ of the MLE and $\max\underline{\mathrm{D}}_\alpha$-estimators with escort parameter $\sigma = 1.483\,\mathrm{MAD}$ and $\sigma = \mathrm{MAD}$ with respect to sample size n for data distributed by $0.9N(0,1) + 0.1N(0,9)$ and divergence parameter $\alpha = 0.5$.

estimator, showing no robustness whatsoever. This motivated us to plug in a simple robust estimate of scale called median absolute deviation (MAD) multiplied by the constant 1.483 ensuring the Fisher consistency in normal model. For uncontaminated data this choice works as well as escorting by MLE, and for contaminated data the corresponding $\max\underline{\mathrm{D}}_\alpha$-estimators perform better than MLE (see Figure 5 in details). Even much better results were received by escorting with median absolute deviation where we left out the calibration constant 1.483 (cf. Figure 5). Moreover, we compared the subdivergence estimators with MAD itself by plugging it into the $\mathrm{eref}(\tilde{\sigma})$ formula instead of MLE, and we see that for the values of $\alpha \geq 0.05$ the subdivergence estimators perform even better than this robust MAD estimator (cf. Table 3). This outstanding behavior in some cases unfortunately vanishes with huge contamination.

α/n	20			50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\mathrm{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\mathrm{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\mathrm{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\mathrm{eref}(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$\mathrm{eref}(\tilde{\sigma})$
MAD	0.75	0.19	1.00	0.74	0.12	1.00	0.73	0.09	1.00	0.73	0.06	1.00	0.73	0.04	1.00
0.00	1.29	0.37	0.44	1.34	0.26	0.47	1.33	0.18	0.57	1.34	0.13	0.60	1.34	0.08	0.62
0.01	1.26	0.33	0.55	1.30	0.22	0.61	1.29	0.15	0.72	1.30	0.11	0.77	1.30	0.07	0.78
0.05	1.20	0.27	0.88	1.21	0.18	1.11	1.21	0.12	1.32	1.21	0.09	1.49	1.21	0.05	1.53
0.10	1.15	0.25	1.16	1.16	0.16	1.66	1.16	0.11	2.07	1.16	0.08	2.47	1.16	0.05	2.63
0.20	1.11	0.23	1.47	1.11	0.15	2.44	1.11	0.10	3.38	1.11	0.07	4.42	1.11	0.05	5.06
0.50	1.08	0.23	1.62	1.07	0.15	3.13	1.07	0.10	5.10	1.07	0.07	7.63	1.07	0.05	10.2

Tab. 3. Comparison of MAD estimator and $\max\underline{\mathrm{D}}_\alpha$-estimators using escort parameter $\sigma = \mathrm{MAD}$ for the generated mixture $0.9N(0,1) + 0.1N(0,9)$.

4.3. Results for power superdivergence estimators

For the power superdivergence estimators of location we received a perfect match with maximum likelihood estimator for all mixtures except for the mixture $(1-\varepsilon)N(0,1) + \varepsilon C(0,1)$. In this particular case, the $\min\bar{\mathrm{D}}_\alpha$-estimators show favourable robustness and high efficiency with higher contamination (cf. Figure 6). However, these robustness tendencies are not as strong as for the case of power subdivergence estimators escorted by $\mu = 0$, $\mu = \bar{\boldsymbol{X}}_n$, or $\mu = \mathrm{MED}$.

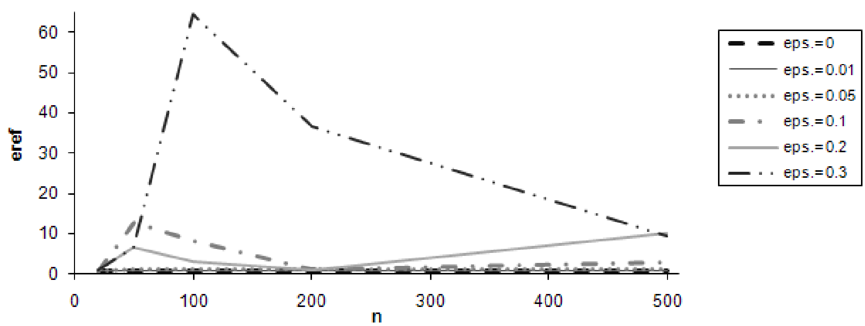

Fig. 6. Empirical relative efficiency $\mathrm{eref}(\tilde{\sigma})$ of the $\min\bar{\mathrm{D}}_\alpha$-estimators with respect to sample size n for data distributed by $(1-\varepsilon)N(0,1) + \varepsilon C(0,1)$ and divergence parameter $\alpha = 0.5$.

When estimating the scale parameter, we also receive estimates that coincide very well with the MLE, even in the case of contamination by Cauchy distribution. Therefore, these estimators do not possess any reasonable robustness and we do not consider them interesting for future research.

Another feature discouraging from the usage of superdivergence estimators is that the related numerical computations are extremely time consuming, which is caused by double optimization. This price is too high to pay for the above mentioned robustness, and it strongly discourages the users from further utilization.

ACKNOWLEDGEMENT

This work was supported by the grants GA ČR P202/10/0618, SGS 12/197/OHK4/3T/14, MSMT INGO-II LG12020, and by the research program of the Ministry of Education of Czech Republic under the contract MSM 6840770039.

(Received July 16, 2011)

REFERENCES

[1] M. Broniatowski and A. Keziou: Minimization of ϕ-divergences on sets of signed measures. Studia Sci. Math. Hungar. *43* (2006), 403–442.

[2] M. Broniatowski and A. Keziou: Parametric estimation and tests through divergences and the duality technique. J. Multivariate Anal. *100* (2009), 16–36.

[3] M. Broniatowski and I. Vajda: Several Applications of Divergence Criteria in Continuous Families. Research Report No. 2257. Institute of Information Theory and Automation, Prague 2009.

[4] I. Frýdlová: Minimum Kolmogorov Distance Estimators. Diploma Thesis. Czech Technical University, Prague 2004.

[5] I. Frýdlová: Modified Power Divergence Estimators and Their Performances in Normal Models. In: Proc. FernStat2010, Faculty of Social and Economic Studies UJEP, Ústí n. L. 2010, 28–33.

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

[7] A. Toma and S. Leoni-Aubin: Robust tests based on dual divergence estimators and saddlepoint approximations. J. Multivariate Anal. *101* (2010), 1143–1155.

[8] A. Toma and M. Broniatowski: Dual divergence estimators and tests: Robustness results. J. Multivariate Analysis *102* (2011), 20–36.

[9] I. Vajda: Theory of Statistical Inference and Information. Kluwer, Boston 1989.

Iva Frýdlová, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Trojanova 13, 120 00 Praha 2. Czech Republic.
e-mail: frydlova.iva@seznam.cz

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: –

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