Michel Broniatowski

Several applications of divergence criteria in continuous families

Kybernetika, Vol. 48 (2012), No. 4, 600--636

Persistent URL: `http://dml.cz/dmlcz/143054`

KYBERNETIKA — VOLUME *48* (2012), NUMBER 4, PAGES 600–636

SEVERAL APPLICATIONS OF DIVERGENCE CRITERIA IN CONTINUOUS FAMILIES

Michel Broniatowski and Igor Vajda

This paper deals with four types of point estimators based on minimization of information-theoretic divergences between hypothetical and empirical distributions. These were introduced

(i) by Liese and Vajda [9] and independently Broniatowski and Keziou [3], called here *power superdivergence estimators,*

(ii) by Broniatowski and Keziou [4] , called here *power subdivergence estimators,*

(iii) by Basu et al. [2], called here *power pseudodistance estimators,* and

(iv) by Vajda [18] called here *Rényi pseudodistance estimators.*

These various criterions have in common to eliminate all need for grouping or smoothing in statistical inference. The paper studies and compares general properties of these estimators such as Fisher consistency and influence curves, and illustrates these properties by detailed analysis of the applications to the estimation of normal location and scale.

Keywords: divergence,parametric estimation, robustness

Classification: 62B10, 62F10, 62F35

1. MOTIVATION

The aim of this paper is to introduce a general setting for various statistical criterions which have been intensively used in the statistical literature in the recent decades. Many authors have considered alternatives to the Maximum Likelihood (ML) paradigm for parametric inference, mostly for reasons linked with robustness; it is well known that ML estimators may lack robustness for simple classical models, such as the standard normal one under simple sampling. Huber's proposal in the 70's, followed by many others, have led to a wide literature, among others in relation with the so-called minimum divergence approach, which is known to include ML as a special case, in connection with Kullback–Leibler divergence. Invariance of divergence based methods with respect to smooth changes in the scale of the variables have also been a strong argument in favor of these new statistical criterions. These methods have in common with ML that they strongly rely on the structure of the model; as such, robustness with respect to misspecification or with respect to outliers cannot be guaranteed, although efficiency may hold; trade off between these two basic properties of statistical inference is still a challenge, and attempts to derive a global class of statistical tools, including ML as a benchmark, is a necessary step in this respect.

Meanwhile a strong obstacle for the definition of such tools has been observed: for example the celebrated χ^2 criterion cannot be defined between the empirical measure of the data, which has finite support, and the plausible candidates in a continuous model; this leads to the well known problem of grouping; in a similar way, most divergence based criterions rely on smoothing techniques, so that the empirical measure is substituted by some non parametric estimate of the underlying density prior to inference with respect to a continuous model. This has been a serious obstacle for these methods.

Following parallel lines Liese and Vajda [9] and Broniatowski and Keziou [3] developed a variational form of divergence criterions which avoids the aforementioned obstacle, therefore making grouping or smoothing unnecessary for parametric inference. The present paper extends these results outside of the restricted field of divergences, and includes other kinds of statistical criterions, such as the L2 one or the proposals by Basu et al. [2] as special cases.

This paper is restricted to the case when the sampling is i.i.d. under a given model with continuous distributions. We have focussed on some classical properties of the proposed estimators (mainly Fisher consistency and robustness through the Influence Function approach); this is clearly very partial, and we did not develop all the asymptotic properties of those estimates. Some reason for this is the following: divergence based approach in inference is not connected only with properties under i.i.d. sampling. It can be seen that according to the sampling, and without referring to any robustness property, there is a strong connection between the sampling scheme and the statistical criterion to be used, whatever the model; this is well known for the ML paradigm: mimicking the observed data set under i.i.d. sampling according to the most likely member of the model turns out to optimize the Kullback–Leibler divergence between the model and the empirical measure, which leads to ML estimators. A similar development can be achieved for other sampling schemes (sample survey ones, weighted bootstrap and others), each one leading to a divergence based procedure. It is also the role of this paper to introduce various notions for future work in this direction.

2. BASIC CONCEPTS AND RESULTS

Let $\phi : (0, \infty) \mapsto \mathbb{R}$ be a 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. For every $\phi \in \mathbf{\Phi}$ we consider the adjoint function

$$\phi^*(t) = t\phi(1/t) \quad \text{where} \quad \phi^* \in \mathbf{\Phi},\ (\phi^*)^* = \phi. \tag{1}$$

For every $\phi \in \mathbf{\Phi}$ we consider *ϕ-divergence* of probability measures P and Q on a measurable space $(\mathcal{X}, \mathcal{A})$ with densities p, q w.r.t. a dominating σ-finite measure λ. In this paper we deal with P, Q which are either measure-theoretically equivalent (i. e. satisfying $pq > 0$ λ-a. s., in symbols $P \equiv Q$) or measure-theoretically orthogonal (i. e. satisfying $pq = 0$ λ-a. s., in symbols $P \perp Q$). Thus, by Liese and Vajda [8] or [9], for all P, Q under consideration

$$D_\phi(P,Q) = \begin{cases} \int \phi\,(p/q)\,\mathrm{d}Q & \text{if} \quad P \equiv Q, \\ \\ \phi(0) + \phi^*(0) & \text{if} \quad P \perp Q, \end{cases} \tag{2}$$

where the range of values is

$$0 \le D_\phi(P,Q) \le \phi(0) + \phi^*(0) \tag{3}$$

and $D_\phi(P,Q) = 0$ iff $P = Q$ or $D_\phi(P,Q) = \phi(0) + \phi^*(0)$ if (for $\phi(0) + \phi^*(0) < \infty$ iff) $P \perp Q$. Another important property is the skew symmetry

$$D_\phi(Q,P) = D_{\phi^*}(P,Q). \tag{4}$$

We shall deal mainly with the power divergences

$$D_\alpha(P,Q) := D_{\phi_\alpha}(P,Q) \quad \text{of real powers } \alpha \in \mathbb{R} \tag{5}$$

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

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha t + \alpha - 1}{\alpha(\alpha - 1)} \quad \text{if} \quad \alpha(\alpha - 1) \neq 0 \tag{6}$$

and otherwise by the corresponding limits

$$\phi_0(t) = -\ln t + t - 1, \qquad \phi_1(t) = \phi_0^*(t) = t \ln t - t + 1. \tag{7}$$

It is easy to verify for all $\alpha \in \mathbb{R}$ the relation

$$\phi_\alpha^* = \phi_{1-\alpha} \text{ so that } \; D_\alpha(Q,P) = D_{1-\alpha}(P,Q).$$

For $P \equiv Q$ we get from (2) and (5) – (7)

$$D_\alpha(P,Q) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \left[\int (p/q)^\alpha \,\mathrm{d}Q - 1 \right] & \text{if} \quad \alpha(\alpha-1) \neq 0, \\ \\ \int \ln(p/q)\,\mathrm{d}P = D_0(Q,P) & \text{if} \quad \alpha = 1, \end{cases} \tag{8}$$

and for $P \perp Q$ similarly

$$D_\alpha(P,Q) = \begin{cases} 1/\alpha(1-\alpha) & \text{if} \quad 0 < \alpha < 1, \\ \infty & \text{otherwise.} \end{cases} \tag{9}$$

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 reverse 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 parameter θ_0 is assumed to be identifiable and the family $\mathcal{P}$ measure-theoreticaly equivalent in the sense

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

Further, the family is assumed to be *continuous* (nonatomic) in the sense

$$P_\theta(\{x\}) = 0 \quad \text{ for all } \; x \in \mathcal{X},\; \theta \in \Theta \tag{11}$$

and dominated by a σ-finite measure λ with densities

$$p_\theta = \mathrm{d}P_\theta/\mathrm{d}\lambda \quad \text{for all } \theta \in \Theta. \tag{12}$$

In this model the parameter θ_0 is assumed to be estimated on the basis of observations $X_1, \ldots, X_n$ by measurable functions $\theta_n : \mathcal{X}^n \mapsto \Theta$ called estimates. Collection of estimates for various sample sizes n is an estimator. Estimators are denoted in this paper by the same symbols θ_n as the corresponding estimates.

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

$$\theta_0 = \mathrm{argmin}_\theta D(P_\theta, P_{\theta_0}) \quad \text{ for every } \theta_0 \in \Theta. \tag{13}$$

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

$$P_n = \frac{1}{n} \sum_{i=1}^{n} P_{X_i}, \tag{14}$$

where P_x denotes the Dirac probability measure with all mass concentrated at $x \in \mathcal{X}$. The empirical probability measures P_n are known to converge weakly to P_{θ_0} as $n \to \infty$. Therefore by plugging in (13) the measures P_n for P_{θ_0} one intuitively expects to obtain the estimator

$$\theta_n = \theta_{n,\phi} := \mathrm{argmin}_\theta D_\phi\left(P_\theta, P_n\right), \tag{15}$$

which estimates θ_0 consistently in the usual sense of the convergence $\theta_n \to \theta_0$ 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 (14)

$$P_\theta \perp P_n \Rightarrow D_\phi(P_\theta, P_n) = \phi(0) + \phi^*(0) \quad \text{when } P_\theta \in \mathcal{P} \text{ and } P_n \in \mathcal{P}_{\mathrm{emp}}. \tag{16}$$

This means that the estimates θ_n proposed in (15) are trivial, with the argmin $= \Theta$.
In the following two sections we list and motivate several modifications of the minimum divergence rule (15) which allow to bypass the problem (16). Some of them are new and some known from the previous literature. We illustrate the general forms of these estimators by applying them to the basic standard statistical families and investigate

their robustness. The model of robust statistics is richer than the standard statistical model defined by the triplet

$$(\mathcal{X}, \mathcal{A}, \mathcal{Q}) \quad \text{with} \quad \mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{\text{emp}}$$

introduced above. Namely in addition to the hypothesis that the observations $X_1, \ldots, X_n$ are i.i.d. by $P_{\theta_0} \in \mathcal{P}$ the model of robust statistics admits the alternative that the observations are distributed by a probability measure $P_0 \notin \mathcal{P}$ with density

$$\frac{\mathrm{d}P_0}{\mathrm{d}\lambda} = p_0.$$

Throughout this paper we assume that P_0 is measure-theoretically equivalent with the probability measures from $\mathcal{P}$ and we consider the probability measures

$$P \in \mathcal{P} \quad \text{and} \quad Q \in \mathcal{Q} = \mathcal{P}^+ \cup \mathcal{P}_{\text{emp}} \quad \text{where} \quad \mathcal{P}^+ = \mathcal{P} \cup \{P_0\}. \tag{17}$$

Measures P, Q are either measure-theoretically equivalent (if $Q \in \mathcal{P}^+$) or measure-theoretically orthogonal (if $Q \in \mathcal{P}_{\text{emp}}$). Therefore the ϕ-divergences $D_\phi(P, Q)$ are well defined by (2) for all pairs P, Q considered in this paper. Further, we denote by $\mathbb{L}_1(Q)$ the set of all absolutely Q-integrable functions $f : \mathcal{X} \mapsto \mathbb{R}$ and put for brevity

$$Q \cdot f = \int f \, \mathrm{d}Q \quad \text{for } f \in \mathbb{L}_1(Q). \tag{18}$$

In the rest of this section we introduce basic concepts and results of the robust statistics needed in the sequel. Let us consider the Dirac probability measures $\delta_x \in \mathcal{P}_{\text{emp}}$, $x \in \mathcal{X}$ and denote by $C(\mathcal{Q})$ the set of the convex mixtures

$$Q_{x,\varepsilon} = (1 - \varepsilon)Q + \varepsilon\delta_x \quad \text{for all} \quad x \in \mathcal{X}, \; Q \in \mathcal{Q} \; \text{ and } 0 \le \varepsilon \le 1. \tag{19}$$

Further, consider a mapping $M(Q, \theta) : C(\mathcal{Q}) \otimes \Theta \to \mathbb{R}$ differentiable in $\theta \in \Theta$ for each $Q \in C(\mathcal{Q})$ with the derivatives

$$\Psi(Q, \theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} M(Q, \theta) \tag{20}$$

and let $T(Q) \in \Theta$ solve the equation $\Psi(Q, \theta) = 0$ in the variable $\theta \in \Theta$ for $Q \in C(\mathcal{Q})$. The following definition and theorem deal with the general M-estimators

$$\theta_n = \text{argmin}_\theta M(P_n, \theta) \quad \text{i.e.} \quad \theta_n = T(P_n) \;\; \text{for} \;\; P_n \in \mathcal{P}_{\text{emp}}.$$

Both the definition and theorem are variants of the well known classical results of robust statistics, see e.g. Hampel et al. [7].

Definition 2.1. If for some $Q \in \mathcal{P}^+$ the limits

$$\text{IF}(x; T, Q) = \lim_{\varepsilon \downarrow 0} \frac{T(Q_{\varepsilon,x}) - T(Q)}{\varepsilon} \tag{21}$$

exist for all $x \in \mathcal{X}$ then (21) is called influence function of the estimator θ_n on $\mathcal{X}$ at Q.

In the following theorem we consider the functions

$$\boldsymbol{\psi}(x,\theta) = \Psi(\delta_x,\theta) \tag{22}$$

and assume the existence of the derivatives

$$\mathring{\psi}(x,\theta) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} \boldsymbol{\psi}(x,\theta) \quad \text{on } \ \mathcal{X} \otimes \Theta \quad (\text{with }^{\mathrm{t}} \text{ for } \textit{transpose}) \tag{23}$$

as well as the expectations

$$\boldsymbol{I}(Q) = Q \cdot \mathring{\psi}(x, T(Q)), \quad Q \in \mathcal{P}^+. \tag{24}$$

Following e. g. Hampel et al. [7] we have

Theorem 2.1. If the influence function (21) exists then it is given by the formula

$$\mathrm{IF}(x;T,Q) = -\boldsymbol{I}(Q)^{-1}\, \boldsymbol{\psi}(x, T(Q)) \tag{25}$$

for the inverse matrix (24).

The estimator $\theta_n = T(P_n)$ is said to be *Fisher consistent* if

$$T(P_\theta) = \theta \quad \text{for all } \theta \in \Theta. \tag{26}$$

In the following Corollary and in the sequel, we put

$$\mathrm{IF}(x;T,\theta) = \mathrm{IF}(x;T,P_\theta) \quad \text{and} \quad \boldsymbol{I}(\theta) = \boldsymbol{I}(P_\theta) \quad (\text{cf. } (24)). \tag{27}$$

Corollary 2.1. The influence function of a Fisher consistent estimator at $Q = P_\theta$ is

$$\mathrm{IF}(x;T,\theta) = -\boldsymbol{I}(\theta)^{-1}\, \boldsymbol{\psi}(x,\theta). \tag{28}$$

3. SUBDIVERGENCES AND SUPERDIVERGENCES

Throughout this section we use the likelihood ratios $\boldsymbol{\ell}_{\theta,\tilde\theta} = p_\theta/p_{\tilde\theta}$ well defined a. s. on $\mathcal{X}$ in the statistical model under consideration, the nonincreasing functions

$$\phi^{\#}(t) = \phi(t) - t\phi'(t) \quad \text{for every } \phi \in \boldsymbol{\Phi}, \tag{29}$$

where ϕ' denotes the derivative of ϕ, and we restrict ourselves to the families $\mathcal{P}$ such that

$$\left\{\phi\left(\boldsymbol{\ell}_{\theta,\tilde\theta}\right),\ \phi'\left(\boldsymbol{\ell}_{\theta,\tilde\theta}\right),\ \phi^{\#}\left(\boldsymbol{\ell}_{\theta,\tilde\theta}\right)\right\} \subset \mathbb{L}_1(Q) \quad \text{for all } \theta,\, \tilde\theta \in \Theta \ \text{ and } \ Q \in \mathcal{Q}. \tag{30}$$

Obviously, this assumption automatically holds for all $Q = P_n \in \mathcal{P}_{\mathrm{emp}}$. Finally, for all pairs $\theta,\, \tilde\theta \in \Theta$ we consider the functions $L_\phi(\theta,\, \tilde\theta) = L_\phi(\theta,\, \tilde\theta,\, x)$ of the variable $x \in \mathcal{X}$ defined by the formula

$$L_\phi(\theta,\, \tilde\theta) = P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\tilde\theta}) + \phi^{\#}(\boldsymbol{\ell}_{\theta,\tilde\theta}).$$

Due to (30), the functions $L_\phi(\theta,\,\tilde\theta)$ are Q-integrable for all $Q \in \mathcal{Q}$. Consider the family of finite expectations

$$\underline{\mathrm{D}}_{\phi,\tilde\theta}\,(P_\theta, Q) = Q \cdot L_\phi(\,\theta, \tilde\theta) = P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\tilde\theta}) + Q \cdot \phi^{\#}(\boldsymbol{\ell}_{\theta,\tilde\theta}), \quad (P_\theta,\ Q) \in \mathcal{P} \otimes \mathcal{Q} \tag{31}$$

parametrized by $(\phi, \tilde\theta) \in \boldsymbol{\Phi} \otimes \boldsymbol{\Theta}$. Broniatowski and Keziou [3] and Liese and Vajda [9] independently established a general supremal representation of ϕ-divergences $D_\phi\,(P,Q)$ which implies the following result.

Theorem 3.1. For each $(P_\theta, P_{\theta_0}) \in \mathcal{P} \otimes \mathcal{P}$ and $\phi \in \boldsymbol{\Phi}$, the ϕ-divergence $D_\phi\,(P_\theta, P_{\theta_0})$ is maximum of the finite expectations $\underline{D}_{\phi,\tilde\theta}\,(P_\theta, P_{\theta_0})$ over $\tilde\theta \in \Theta$ attained at the unique point $\tilde\theta = \theta_0$. In other words,

$$D_\phi\,(P_\theta, P_{\theta_0}) \geq \underline{\mathrm{D}}_{\phi,\tilde\theta}\,(P_\theta, P_{\theta_0}) \quad \text{for all } \theta, \theta_0 \in \Theta, \tag{32}$$

where the equality holds iff $\tilde\theta = \theta_0$.

P r o o f. For the sake of completeness we present the simple proof of Liese and Vajda. For fixed $s > 0$, the strictly convex function $\phi(t)$ is strictly above the straight line $\phi(s) + \phi'(s)(t-s)$ except $t = s$, i. e.

$$\phi(t) \geq \phi(s) + \phi'(s)(t-s)$$

with the equality only for $t = s$. Putting in this inequality $t = \boldsymbol{\ell}_{\theta,\theta_0}$, $s = \boldsymbol{\ell}_{\theta,\tilde\theta}$ and integrating both sides over P_{θ_0} we get (32) including the iff condition for the equality. □

Theorem 3.1 implies the formula

$$D_\phi\,(P_\theta, Q) = \max_{\tilde\theta \in \Theta} \underline{\mathrm{D}}_{\phi,\tilde\theta}\,(P_\theta, Q) \quad \text{for all } (P_\theta, Q) \in \mathcal{P} \otimes \mathcal{P} \tag{33}$$

which justifies us to interpret $\underline{\mathrm{D}}_{\phi,\tilde\theta}\,(P_\theta, Q)$ as ***subdivergences*** of P_θ, Q with parameters $(\phi, \tilde\theta) \in \boldsymbol{\Phi} \otimes \boldsymbol{\Theta}$.
Now we introduce the family of suprema

$$\bar{\mathrm{D}}_\phi\,(P_\theta, Q) := \sup_{\tilde\theta \in \Theta} \underline{\mathrm{D}}_{\phi,\tilde\theta}\,(P_\theta, Q) \quad \text{for all } (P_\theta, Q) \in \mathcal{P} \otimes \mathcal{Q} \tag{34}$$

parametrized by $\phi \in \boldsymbol{\Phi}$. This family extends the ϕ-divergences $D_\phi\,(P,Q)$ from the domain $\mathcal{P} \otimes \mathcal{P}$ to $\mathcal{P} \otimes \mathcal{Q}$ allowing therefore $\mathcal{Q}$ to be discrete. Indeed, by Theorem 3.1,

$$\bar{\mathrm{D}}_\phi\,(P_\theta, Q) = D_\phi\,(P_\theta, Q) \quad \text{for all } (P_\theta, Q) \in \mathcal{P} \otimes \mathcal{P}. \tag{35}$$

This justifies us to interpret $\bar{\mathrm{D}}_\phi\,(P_\theta, Q)$ as ***superdivergences*** of $(P_\theta, Q) \in \mathcal{P} \otimes \mathcal{Q}$ with parameters $\phi \in \boldsymbol{\Phi}$.
Note that (35) need not hold for $Q \notin \mathcal{P}$ because if $Q = P_n \in \mathcal{P}_{\mathrm{emp}}$ then the superdivergence values $\bar{\mathrm{D}}_\phi\,(P_\theta, P_n)$ differ from the constant divergence values $D_\phi\,(P_\theta, P_n) \equiv$

$\phi(0) + \phi^*(0)$ (cf. (16)).
The subdivergences $\underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n)$ and superdivergences $\bar{\mathrm{D}}_\phi(P_\theta, P_n)$ can replace the divergences $D_\phi(P_\theta, P_n)$ as optimality criteria in definition of M-estimators. Let us consider the families of functionals $\tilde{T}_{\phi,\theta} : \mathcal{Q} \mapsto \Theta$ and $T_\phi : \mathcal{Q} \mapsto \Theta$ defined by

$$\tilde{T}_{\phi,\theta}(Q) = \mathrm{argmax}_{\tilde{\theta}}\ \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, Q) \quad \text{for } (\phi,\theta) \in \mathbf{\Phi} \otimes \mathbf{\Theta} \tag{36}$$

and

$$T_\phi(Q) = \mathrm{argmin}_\theta\ \bar{\mathrm{D}}_\phi(P_\theta, Q) \quad \text{for } \phi \in \mathbf{\Phi} \tag{37}$$

respectively. Replacing the general argument Q by P_n defined by (14) we obtain the ***maximum subdivergence estimators*** (briefly, the max$\underline{\mathrm{D}}_\phi$-estimators)

$$\begin{aligned}\tilde{\theta}_{\phi,\theta,n} &= \tilde{T}_{\phi,\theta}(P_n) = \mathrm{argmax}_{\tilde{\theta}}\ \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n) && (38)\\ &= \mathrm{argmax}_{\tilde{\theta}}\ \left[P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\tilde{\theta}}) + P_n \cdot \phi^{\#}(\boldsymbol{\ell}_{\theta,\tilde{\theta}})\right] \quad \text{(cf. (31))} \\ &= \mathrm{argmax}_{\tilde{\theta}} \left[P_\theta \cdot \phi'\left(\frac{p_\theta}{p_{\tilde{\theta}}}\right) + \frac{1}{n}\sum_{i=1}^n \phi^{\#}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)\right] && (39)\end{aligned}$$

with escort parameters $\theta \in \mathbf{\Theta}$, and the ***minimum superdivergence estimators*** (briefly, the min$\bar{\mathrm{D}}_\phi$-estimators)

$$\begin{aligned}\theta_{\phi,n} &= T_\phi(P_n) = \mathrm{argmin}_\theta\ \bar{\mathrm{D}}_\phi(P_\theta, P_n) = \mathrm{argmin}_\theta \mathrm{sup}_{\tilde{\theta}}\ \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n) \ \text{(cf. (34))} && (40)\\ &= \mathrm{argmin}_\theta \mathrm{sup}_{\tilde{\theta}} \left[P_\theta \cdot \phi'(\boldsymbol{\ell}_{\theta,\tilde{\theta}}) + P_n \cdot \phi^{\#}(\boldsymbol{\ell}_{\theta,\tilde{\theta}})\right] \quad \text{(cf. (31))} \\ &= \mathrm{argmin}_\theta \mathrm{sup}_{\tilde{\theta}} \left[P_\theta \cdot \phi'\left(\frac{p_\theta}{p_{\tilde{\theta}}}\right) + \frac{1}{n}\sum_{i=1}^n \phi^{\#}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)\right]. && (41)\end{aligned}$$

Theorem 3.2. The max$\underline{\mathrm{D}}_\phi$-estimators are as well as the min$\bar{\mathrm{D}}_\phi$-estimators are Fisher consistent.

Proof. By (33) and (35),

$$\tilde{T}_{\phi,\theta}(P_{\theta_0}) = \mathrm{argmax}_{\tilde{\theta}}\ \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_{\theta_0}) \quad \text{for } (\phi,\theta) \in \mathbf{\Phi} \otimes \mathbf{\Theta} \tag{42}$$

and

$$T_\phi(P_{\theta_0}) = \mathrm{argmin}_\theta\ \bar{\mathrm{D}}_\phi(P_\theta, P_{\theta_0}) \quad \text{for } \phi \in \mathbf{\Phi} \tag{43}$$

which completes the proof. □

The min$\bar{\mathrm{D}}_\phi$-estimators $\theta_{\phi,n}$ in (40) were proposed independently by Liese and Vajda [9] under the name **modified ϕ-divergence estimators** and Broniatowski and Keziou [3] under the name **minimum dual ϕ-divergence estimators**. The max$\underline{\mathrm{D}}_\phi$-estimators $\tilde{\theta}_{\phi,\theta,n}$ in (38) were proposed by Broniatowski and Keziou [4] and called **dual ϕ-divergence estimators** by them. Both types of these estimators were in the cited papers motivated by the mentioned Fisher consistency and by the property easily verifiable from (39) and (41), namely that $\phi(t) = -\ln t$ implies

$$\tilde{\theta}_{\phi,\theta,n} = \mathrm{argmax}_{\tilde{\theta}}\ \Sigma_{i=1}^n \ln p_{\tilde{\theta}}(X_i) \quad \text{and} \quad \theta_{\phi,n} = \mathrm{argmax}_\theta\ \Sigma_{i=1}^n \ln p_\theta(X_i), \tag{44}$$

where the left equality holds for all escort parameters $\theta \in \Theta$. In other words, the logarithmic choice $\phi(t) = -\ln t$ reduces all the variants of the max$\underline{\mathrm{D}}_\phi$-estimator as well as the min$\bar{\mathrm{D}}_\phi$-estimator to the MLE. It is challenging to investigate the extent to which the max$\underline{\mathrm{D}}_\phi$-estimators $\tilde{\theta}_{\phi,\theta,n}$ and the min$\bar{\mathrm{D}}_\phi$-estimator $\theta_{\phi,n}$ as extensions of the MLE are efficient and robust under various specifications of ϕ, θ and ϕ respectively.
In this paper we restrict ourselves to special subclasses of the power divergences $D_\alpha(P,Q)$ $:= D_{\phi_\alpha}(P,Q)$ defined by (6) – (8). For the power functions ϕ_α from (6), (7) we get the functions

$$\mathring{\phi}_\alpha(t) := t\phi'_\alpha(t) = \begin{cases} \frac{t^\alpha - t}{\alpha - 1} & \text{for } \alpha \neq 1, \\ \lim_{\alpha \to 1} \frac{t^\alpha - t}{\alpha - 1} = t \ln t & \text{for } \alpha = 1, \end{cases} \tag{45}$$

and

$$\phi^{\#}_\alpha(t) = \phi_\alpha(t) - \mathring{\phi}_\alpha(t) = \begin{cases} \frac{1}{\alpha}(1 - t^\alpha) & \text{for } \alpha \neq 0, \\ \lim_{\alpha \to 0} \frac{1}{\alpha}(1 - t^\alpha) = -\ln t & \text{for } \alpha = 0. \end{cases} \tag{46}$$

They lead to the max$\underline{\mathrm{D}}_\alpha$-estimators (briefly, ***power subdivergence estimators***)

$$\tilde{\theta}_{\alpha,\theta,n} = \mathrm{argmax}_{\tilde{\theta}} \left[P_{\tilde{\theta}} \cdot \mathring{\phi}_\alpha \left(\frac{p_\theta}{p_{\tilde{\theta}}} \right) + P_n \cdot \phi^{\#}_\alpha \left(\frac{p_\theta}{p_{\tilde{\theta}}} \right) \right] \tag{47}$$

with power parameters $\alpha \in \mathbb{R}$ and escort parameters $\theta \in \Theta$ and to the min$\bar{\mathrm{D}}_\alpha$-estimators (briefly, ***power superdivergence estimators***)

$$\theta_{\alpha,n} = \mathrm{argmin}_\theta \mathrm{sup}_{\tilde{\theta}} \left[P_{\tilde{\theta}} \cdot \mathring{\phi}_\alpha \left(\frac{p_\theta}{p_{\tilde{\theta}}} \right) + P_n \cdot \phi^{\#}_\alpha \left(\frac{p_\theta}{p_{\tilde{\theta}}} \right) \right] \tag{48}$$

with power parameters $\alpha \in \mathbb{R}$. If the argmaxima in (47) exist then

$$\theta_{\alpha,n} = \mathrm{argmin}_\theta \left[P_{\tilde{\theta}_{\alpha,\theta,n}} \cdot \mathring{\phi}_\alpha \left(\frac{p_\theta}{p_{\tilde{\theta}_{\alpha,\theta,n}}} \right) + P_n \cdot \phi^{\#}_\alpha \left(\frac{p_\theta}{p_{\tilde{\theta}_{\alpha,\theta,n}}} \right) \right]. \tag{49}$$

The next two subsections deal correspondingly with the max$\underline{\mathrm{D}}_\alpha$-estimators and min$\bar{\mathrm{D}}_\alpha$-estimators. In both sections are considered the power parameters $\alpha \geq 0$. Since $\phi_0(t) = -\ln t$, we see from (44) that

$$\tilde{\theta}_{0,\theta,n} = \mathrm{argmax}_{\tilde{\theta}} \ \Sigma_{i=1}^n \ln p_{\tilde{\theta}}(X_i) \ \text{ and } \ \theta_{0,n} = \mathrm{argmax}_\theta \ \Sigma_{i=1}^n \ln p_\theta(X_i) \tag{50}$$

are the MLE's. If $\alpha > 0$ then by (45) – (48),

$$\tilde{\theta}_{\alpha,\theta,n} = \mathrm{argmin}_{\tilde{\theta}} \ M_{\alpha,\theta}(P_n, \tilde{\theta}) \tag{51}$$

and

$$\theta_{\alpha,n} = \mathrm{argmax}_\theta \mathrm{inf}_{\tilde{\theta}} \ M_{\alpha,\theta}(P_n, \tilde{\theta}) \equiv \mathrm{argmax}_\theta \ M_{\alpha,\theta}(P_n, \tilde{\theta}_{\alpha,\theta,n}), \tag{52}$$

where

$$M_{\alpha,\theta}(Q,\tilde{\theta}) = \frac{1}{1-\alpha} P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha + \frac{1}{\alpha} Q \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha \quad \text{if } \alpha > 0,\, \alpha \neq 1$$

$$= P_\theta \cdot \ln \frac{p_{\tilde{\theta}}}{p_\theta} + Q \cdot \frac{p_\theta}{p_{\tilde{\theta}}} \quad \text{if } \alpha = 1 \tag{53}$$

for all $Q \in \mathcal{Q}$.
Throughout both subsections we restrict ourselves to the densities p_θ twice differentiable with respect to $\theta \in \Theta \subset \mathbb{R}^d$, we put

$$s_\theta = \frac{\mathrm{d}}{\mathrm{d}\theta} \ln p_\theta \quad \text{and} \quad \mathring{s}_\theta = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} s_\theta \tag{54}$$

and suppose that the functions $M_{\alpha,\theta}(Q,\tilde{\theta})$ of (53) are twice differentiable in the vector variable $\tilde{\theta}$, with the differentiation and integration interchangeable in (53). Moreover, we suppose that the derivatives

$$\Psi_{\alpha,\theta}(Q,\tilde{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\tilde{\theta}} M_{\alpha,\theta}(Q,\tilde{\theta}) = P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} - Q \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} \tag{55}$$

admit solutions of the equations $\Psi_{\alpha,\theta}(Q,\tilde{\theta}) = 0$ in the variable $\tilde{\theta} \in \Theta$ for $Q \in Q$.

3.1. Power subdivergence estimators

In this subsection we study the max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta,n}$ with the divergence power parameters $\alpha \geq 0$ and the escort parameters $\theta \in \Theta$. As said above, for $\alpha = 0$ they coincide with the MLE's (50). Therefore we restrict ourselves to $\alpha > 0$ and to the definition formula (51), (53).
By assumptions, the argminima

$$\tilde{T}_{\alpha,\theta}(Q) = \mathrm{argmin}_{\tilde{\theta}}\; M_{\alpha,\theta}(Q,\tilde{\theta}), \quad \alpha > 0, \quad Q \in \mathcal{Q} \quad \text{cf. (36))} \tag{56}$$

solve the equations $\Psi_{\alpha,\theta}(Q,\tilde{\theta}) = 0$ in the variable $\tilde{\theta} \in \Theta$ and, in particular, $\tilde{\theta}_{\alpha,\theta,n} = \tilde{T}_{\alpha,\theta}(P_n)$ are for all $\alpha > 0$ solutions of the equations

$$P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} - \frac{1}{n}\sum_{i=1}^n \left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)^\alpha s_{\tilde{\theta}}(X_i) = 0 \tag{57}$$

in the variable $\tilde{\theta} \in \Theta$.
The next Theorem explicits the influence functions of the max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta,n}$. Its proof is provided in Broniatowski and Vajda [6]. For a complete discussion on scale and location robustness of max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta,n}$ see Toma and Broniatowski [13]. Let us quote that the family of max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta,n}$estimators are shown to be robust in many cases under heavy tailed models, as is also seen by simulation for various sample sizes and various types of contamination.

Theorem 3.1.1. The influence functions of the max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta,n}$ under consideration are at P_{θ_0} given by the formula

$$\mathrm{IF}(x;\tilde{T}_{\alpha,\theta},\theta_0) = \boldsymbol{I}_{\alpha,\theta}(\theta_0)^{-1}\left[\left(\frac{p_\theta(x)}{p_{\theta_0}(x)}\right)^\alpha s_{\theta_0}(x) - P_{\theta_0}\cdot\left(\frac{p_\theta}{p_{\theta_0}}\right)^\alpha s_{\theta_0}\right] \quad \text{if } \alpha > 0 \tag{58}$$

$$\mathrm{IF}(x;\tilde{T}_{0,\theta},\theta_0) = \boldsymbol{I}(\theta_0)^{-1}s_{\theta_0}(x) \quad \text{otherwise,} \tag{59}$$

where

$$\boldsymbol{I}_{\alpha,\theta}(\theta_0) = P_{\theta_0}\cdot\left(\frac{p_\theta}{p_{\theta_0}}\right)^\alpha s^{\mathrm{t}}_{\theta_0}s_{\theta_0} \qquad \text{if } \alpha > 0, \tag{60}$$

$$\boldsymbol{I}(\theta_0) = P_{\theta_0}\cdot s^{\mathrm{t}}_{\theta_0}s_{\theta_0} \qquad \text{if } \alpha = 0. \tag{61}$$

If the escort parameter θ coincides with the true parameter θ_0 then

$$\mathrm{IF}(x;\tilde{T}_{\alpha,\theta_0},\theta_0) = \boldsymbol{I}(\theta_0)^{-1}s_{\theta_0}(x) \quad \text{for all } \alpha \geq 0.$$

The following examples provide some insight on the Fisher consistency of Power subdivergence estimators. This property may be lossed under misspecification, due to the fact that they may not be free with respect to the value assumed for some known parameter in the model, showing a loss of consistency under misspecification. We also evaluate the influence functions of the scale and location parameters.

Example 3.1.1. Power subdivergence estimators in normal family. Let the observation space $(\mathcal{X},\mathcal{A})$ be the Borel line $(\mathbb{R},\mathcal{B})$ and $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R},\ \sigma > 0\}$ the normal family with parameters of location μ and scale σ (i.e. variances σ^2). We are interested in the max$\underline{\mathrm{D}}_\alpha$-estimates $(\tilde{\mu}_{\alpha,\mu,\sigma,n},\tilde{\sigma}_{\alpha,\mu,\sigma,n})$ with power parameters $\alpha \geq 0$ and escort parameters $(\mu,\sigma) \in \mathbb{R}\otimes(0,\infty)\}$.
If $\alpha = 0$ then these estimators reduce for all escort parameters μ,σ to the well known MLE's

$$(\tilde{\mu}_{0,\mu,\sigma,n},\tilde{\sigma}_{0,\mu,\sigma,n}) = \left(\frac{1}{n}\sum_{i=1}^n X_i,\ \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \tilde{\mu}_{0,n})^2}\right) \tag{62}$$

For $0 < \alpha < 1$ the function (53) takes on the form

$$M_{\alpha,\mu,\sigma}(Q,\tilde{\mu},\tilde{\sigma}) = \frac{1}{1-\alpha}P_{\tilde{\mu},\tilde{\sigma}}\cdot\left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^\alpha + \frac{1}{\alpha}Q\cdot\left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^\alpha \tag{63}$$

where

$$\left(\frac{p_{\mu,\sigma}(x)}{p_{\tilde{\mu},\tilde{\sigma}}(x)}\right)^\alpha = \left(\frac{\tilde{\sigma}}{\sigma}\right)^\alpha \exp\left\{\frac{\alpha\,(x-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{\alpha\,(x-\mu)^2}{2\sigma^2}\right\}, \tag{64}$$

and

$$P_{\tilde{\mu},\tilde{\sigma}}\cdot\left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^\alpha = \exp\left\{-\frac{\alpha(1-\alpha)(\mu-\tilde{\mu})^2}{2[\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2]} - \ln\frac{\sqrt{\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2}}{\tilde{\sigma}^\alpha\sigma^{1-\alpha}}\right\}. \tag{65}$$

Using the likelihood ratio function (64) and the score function

$$s_{\mu,\sigma}(x) = \left(\frac{x-\mu}{\sigma^2}, \frac{1}{\sigma} \left[\left(\frac{x-\mu}{\sigma} \right)^2 - 1 \right] \right) \tag{66}$$

one obtains for all $\alpha > 0$ the derivative

$$\Psi_{\alpha,\mu,\sigma}(Q,\tilde{\mu},\tilde{\sigma}) = \left(\frac{\mathrm{d}}{\mathrm{d}\tilde{\mu}}, \frac{\mathrm{d}}{\mathrm{d}\tilde{\sigma}} \right) M_{\alpha,\mu,\sigma}(Q,\tilde{\mu},\tilde{\sigma}) = P_{\tilde{\mu},\tilde{\sigma}} \cdot \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}} \right)^{\alpha} s_{\tilde{\mu},\tilde{\sigma}} - Q \cdot \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}} \right)^{\alpha} s_{\tilde{\mu},\tilde{\sigma}} \tag{67}$$

and the max$\underline{\mathrm{D}}_\alpha$-estimators as the argminima

$$(\tilde{\mu}_{\alpha,\mu,\sigma,n}, \tilde{\sigma}_{\alpha,\mu,\sigma,n}) = \mathrm{argmin}_{\tilde{\mu},\tilde{\sigma}} \left[\frac{1}{1-\alpha} P_{\tilde{\mu},\tilde{\sigma}} \cdot \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}} \right)^{\alpha} + \frac{1}{\alpha n} \sum_{i=1}^{n} \left(\frac{p_{\mu,\sigma}(X_i)}{p_{\tilde{\mu},\tilde{\sigma}}(X_i)} \right)^{\alpha} \right] \tag{68}$$

or, equivalently, as solutions of the equations

$$P_{\tilde{\mu},\tilde{\sigma}} \cdot \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}} \right)^{\alpha} s_{\tilde{\mu},\tilde{\sigma}} - \frac{1}{n} \sum_{i=1}^{n} \left(\frac{p_{\mu,\sigma}(X_i)}{p_{\tilde{\mu},\tilde{\sigma}}(X_i)} \right)^{\alpha} s_{\tilde{\mu},\tilde{\sigma}}(X_i) = 0. \tag{69}$$

By Theorem 3.1.1, the influence functions of these estimators at P_{μ_0,σ_0} are

$$\begin{aligned} & \mathrm{IF}\,(x; \tilde{T}_{\alpha,\mu,\sigma}, \mu_0, \sigma_0) \\ = \; & \boldsymbol{I}_{\mu,\sigma}(\mu_0,\sigma_0)^{-1} \left[\left(\frac{p_{\mu,\sigma}(x)}{p_{\mu_0,\sigma_0}(x)} \right)^{\alpha} s_{\mu_0,\sigma_0}(x) - P_{\mu_0,\sigma_0} \cdot \left(\frac{p_{\mu,\sigma}}{p_{\mu_0,\sigma_0}} \right)^{\alpha} s_{\mu_0,\sigma_0} \right] \end{aligned}$$

for

$$\boldsymbol{I}_{\mu,\sigma}(\mu_0,\sigma_0) = P_{\mu_0,\sigma_0} \cdot \left(\frac{p_{\mu,\sigma}}{p_{\mu_0,\sigma_0}} \right)^{\alpha} s^{\mathrm{t}}_{\mu_0,\sigma_0} s_{\mu_0,\sigma_0}. \tag{70}$$

Example 3.1.2. Power subdivergence estimators of location. Let in the frame of previous example $\mathcal{P} = \{P_\mu : \mu \in \mathbb{R}\}$ be the standard normal family with the location parameter μ and scale $\sigma = 1$. Then the function (63) takes on the form

$$M_{\alpha,\mu}(Q,\tilde{\mu}) = \frac{1}{1-\alpha} \left(\eta_{\alpha,\mu}(\mu,\tilde{\mu}) \right)^{\alpha-1} + \frac{1}{\alpha} Q \cdot \eta_{\alpha,\mu}(x,\tilde{\mu}) \tag{71}$$

for $\alpha > 0, \alpha \neq 1$ where

$$\eta_{\alpha,\mu}(x,\tilde{\mu}) = \exp\left\{ \alpha(\tilde{\mu}-\mu)(\tilde{\mu}+\mu-2x)/2 \right\}, \quad x \in \mathbb{R}.$$

The max$\underline{\mathrm{D}}_\alpha$-estimates $\tilde{\mu}_{\alpha,\mu,n}$ of location μ_0 with the divergence parameters $0 \leq \alpha < 1$ and escort parameters $\mu \in \mathbb{R}$ are the MLE's

$$\tilde{\mu}_{0,\mu,n} = \bar{\boldsymbol{X}}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{72}$$

if $\alpha = 0$. Otherwise they are the minimizers

$$\tilde{\mu}_{\alpha,\mu,n} = \text{argmin}_{\tilde{\mu}} M_{\alpha,\mu}(P_n, \tilde{\mu}) \tag{73}$$

or, equivalently, solutions of the equations

$$\Psi_{\alpha,\mu}(P_n, \tilde{\mu}) = 0$$

in they variable $\tilde{\mu} \in \mathbb{R}$ for

$$\Psi_{\alpha,\mu}(Q, \tilde{\mu}) = \frac{\mathrm{d}}{\mathrm{d}\tilde{\mu}} M_{\alpha,\mu}(Q, \tilde{\mu})$$

$$= Q \cdot (\tilde{\mu} - x)\eta_{\alpha,\mu}(x, \tilde{\mu}) - \alpha(\tilde{\mu} - \mu)\eta_{\alpha,\mu}^{\alpha-1}(\mu, \tilde{\mu}). \tag{74}$$

Let $\tilde{T}_{\alpha,\mu}(Q)$ be the solution of the equation $\Psi_{\alpha,\mu}(Q, \tilde{\mu}) = 0$ in the variable $\tilde{\mu} \in \mathbb{R}$ and let Q_{μ_0} denote the shift of the distribution Q by μ_0. Then

$$Q_{\mu_0} \cdot (\tilde{\mu} - x)\eta_{\alpha,\mu}(x, \tilde{\mu}) = Q \cdot (\tilde{\mu} - \mu_0 - x)\eta_{\alpha,\mu-\mu_0}(x, \tilde{\mu} - \mu_0))$$

so that $\tilde{T}_{\alpha,\mu}(Q_{\mu_0}) = \mu_0 + \tilde{T}_{\alpha,\mu-\mu_0}(Q)$. This means that the estimators (73) are Fisher consistent in the normal family $\mathcal{P}_\sigma = \{P_{\mu_0,\sigma} = N(\mu_0, \sigma^2) : \mu_0 \in \mathbb{R}\}$ with $\sigma > 0$ fixed if and only if the solution $\tilde{T}_{\alpha,\mu}(P_{0,\sigma})$ of the equation

$$P_{0,\sigma} \cdot (\tilde{\mu} - x)\eta_{\alpha,\mu}(x, \tilde{\mu}) - \alpha(\tilde{\mu} - \mu)\eta_{\alpha,\mu}^{\alpha-1}(\mu, \tilde{\mu}) = 0 \tag{75}$$

in the variable $\tilde{\mu}$ satisfies the condition

$$\tilde{T}_{\alpha,\mu}(P_{0,\sigma}) = 0 \quad \text{for all } \mu \in \mathbb{R}. \tag{76}$$

By evaluating the function $P_{0,\sigma} \cdot (\tilde{\mu} - x)\eta_{\alpha,\mu}(x, \tilde{\mu})$ of variables $\sigma, \mu, \tilde{\mu}$ and inserting it in (75), one can verify that (76) holds if and only if $\sigma = 1$. The "if" part follows from the Fisher consistency of $\tilde{T}_{\alpha,\mu}$ established in Theorem 3.2 which implies

$$\tilde{T}_{\alpha,\mu}(P_{0,1}) \equiv \tilde{T}_{\alpha,\mu}(P_0) = 0 \quad \text{for } P_{0,1} \equiv P_0 \in \mathcal{P} \text{ and all } \mu \in \mathbb{R}.$$

The "only if" assertion indicates a relatively easy loss of consistency of the max$\underline{\mathrm{D}}_\alpha$-estimators; this is due to the fact that the estimators are designed precisely making use of $\sigma = 1$, indicating that they are not free of the scale parameter.
The estimators $\tilde{\mu}_{\alpha,\bar{X}_n,n}$ with the adaptive MLE escort parameters $\bar{X}_n$ are Fisher consistent under all hypothetical models $P_{\mu,\sigma} = N(\mu, \sigma^2)$, $\sigma > 0$. More generally, the adaptive estimators

$$\tilde{\theta}_{\alpha,\tau_n,n} \quad \text{with the MLE escorts } \tau_n = \tilde{\theta}_{0,n} \text{ given by (44)} \tag{77}$$

are Fisher consistent under the hypothetical models P_{θ_0}.
Let us turn to the influence curves $\mathrm{IF}(x; T_{\alpha,\mu}, \mu_0), 0 < \alpha < 1$ at the data source P_{μ_0}.

Here $s_{\mu_0}^{\mathrm{t}}(x)s_{\mu_0}(x) = s_{\mu_0}^2(x) = (\mu_0 - x)^2$ so that, by (27) and (70),

$$\begin{aligned} I_{\alpha,\mu}(\mu_0) &= \boldsymbol{I}_{\alpha,\mu}(P_{\mu_0}) = P_{\mu_0} \cdot \left(\frac{p_\mu}{p_{\mu_0}}\right)^\alpha s_{\mu_0}^2 \\ &= \frac{1}{\sqrt{2\pi}} \int (\mu_0 - x)^2 \exp\left\{-\frac{\alpha(x-\mu)^2 + (1-\alpha)(x-\mu_0)^2}{2}\right\} \mathrm{d}x \qquad (78) \\ &= \left[1 + \alpha^2(\mu_0 - \mu)^2\right] \exp\left\{\frac{\alpha(\alpha-1)(\mu_0-\mu)^2}{2}\right\}. \end{aligned}$$

If we put

$$\begin{aligned} \psi_{\alpha,\mu}(x,\mu_0) &= \Psi_{\alpha,\mu}(\delta_x,\mu_0) \\ &= (\mu_0 - x)\eta_{\alpha,\mu}(x,\mu_0) - \alpha(\mu_0-\mu)\eta_{\alpha,\mu}^{\alpha-1}(\mu,\mu_0) \text{ (cf. (74))} \end{aligned}$$

then, by (70),

$$\begin{aligned} \mathrm{IF}(x; T_{\alpha,\mu}, \mu_0) &= -\frac{\psi_{\alpha,\mu}(x,\mu_0)}{I_{\alpha,\mu}(\mu_0)} \\ &= \frac{(x-\mu_0)e^{\alpha(\mu_0-\mu)(\mu_0+\mu-2x)/2} + \alpha(\mu_0-\mu)e^{\alpha(\alpha-1)(\mu_0-\mu)^2/2}}{[1+\alpha^2(\mu_0-\mu)^2]\, e^{\alpha(\alpha-1)(\mu_0-\mu)^2/2}}. \qquad (79) \end{aligned}$$

This formula remains valid also for $\alpha = 0$ because then it reduces to the well known influence function

$$\mathrm{IF}(x; MLE, \mu_0) = x - \mu_0$$

of the $MLE = T_{0,\mu}$ which is not depending on the escort parameter μ. We see that the influence curve (79) is unbounded for all $\mu, \mu_0 \in \mathbb{R}$ and $0 \le \alpha < 1$. For $0 < \alpha < 1$ and the escort parameters μ different from the true μ_0 the influence functions $\mathrm{IF}(x; T_{\alpha,\mu}, \mu_0)$ contain the constant terms $\mathrm{IF}(\mu_0; T_{\alpha,\mu}, \mu_0) \ne 0$ and, moreover, increase to infinity exponentially for $x \to \infty$ or $x \to -\infty$. Therefore $T_{\alpha,\mu}$ are strongly non-robust, as is the classical ML estimate in this case.

Example 3.1.3. Power subdivergence estimators of scale. Let in the frame of Example 3.1.1, $\mathcal{P} = \{P_\sigma : \sigma > 0\}$ be the standard normal family with the location parameter $\mu = 0$ and scale σ and let us consider the max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\sigma}_{\alpha,\sigma,n}$ of scale σ_0 with the divergence parameters $0 \le \alpha < 1$ and escort parameters $\sigma > 0$. For $\alpha = 0$ they reduce to the standard deviations

$$\tilde{\sigma}_{0,\sigma,n} = \left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{\boldsymbol{X}}_n\right)^2\right)^{1/2}$$

and otherwise they are of the form

$$\tilde{\sigma}_{\alpha,\sigma,n} = T_{\alpha,\sigma}(P_n) \quad \text{for} \quad T_{\alpha,\sigma}(Q) = \mathrm{argmin}_{\tilde{\sigma}} M_{\alpha,\sigma}(Q,\tilde{\sigma}),\ Q \in \mathcal{Q}$$

where

$$M_{\alpha,\sigma}(Q,\tilde{\sigma}) = \tilde{M}_{\alpha,\sigma}(Q,\tilde{\sigma}/\sigma)$$

for (cf. (63))

$$\tilde{M}_{\alpha,\sigma}(Q,s) = \frac{s^\alpha}{(1-\alpha)\sqrt{\alpha s^2+1-\alpha}} + \int \frac{s^\alpha}{\alpha}\exp\left\{\frac{\alpha x^2\left[s^{-2}-1\right]}{2\sigma^2}\right\}\mathrm{d}Q(x).$$

One can show that

$$\begin{aligned}\psi_{\alpha,\sigma}(x,\tilde{\sigma}) &= \frac{\mathrm{d}}{\mathrm{d}\tilde{\sigma}}M_{\alpha,\sigma}(\delta_x,\tilde{\sigma}) = \frac{1}{\sigma}\left(\frac{\mathrm{d}}{\mathrm{d}s}\tilde{M}_{\alpha,\sigma}(\delta_x,s)\right)_{s=\tilde{\sigma}/\sigma} \\ &= -\frac{1}{\sigma}\left[s^{\alpha-1}\left(\frac{\alpha\left(s^2-1\right)}{(\alpha s^2+1-\alpha)^{3/2}} + \left[\left(\frac{x}{\sigma s}\right)^2-1\right]e^{\alpha x^2\left[s^{-2}-1\right]/2\sigma^2}\right)\right]_{s=\tilde{\sigma}/\sigma} \\ &= -\left(\frac{\tilde{\sigma}}{\sigma}\right)^{\alpha-1}\left(\frac{\alpha\left(\tilde{\sigma}^2-\sigma^2\right)}{\left[\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2\right]^{3/2}} + \frac{1}{\sigma}\left[\left(\frac{x}{\tilde{\sigma}}\right)^2-1\right]e^{\alpha x^2\left[\tilde{\sigma}^{-2}-\sigma^{-2}\right]/2}\right).\end{aligned} \tag{80}$$

By differentiating this expression with respect to $\tilde{\sigma}$, we obtain the matrix

$$I_{\alpha,\sigma}(\tilde{\sigma}) := \boldsymbol{I}_{\alpha,\sigma}(P_{\tilde{\sigma}}) = \left(\frac{\tilde{\sigma}}{\sigma}\right)^{\alpha-1}\frac{2\sigma^4+\alpha^2(\tilde{\sigma}^2-\sigma^2)^2}{\tilde{\sigma}[\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2]^{5/2}}. \tag{81}$$

Hence, by Theorem 3.1.1, the influence function of max$\underline{\mathrm{D}}_\alpha$-estimators at the data generating distributions P_{σ_0} are for all $0<\alpha<1$

$$\begin{aligned}\mathrm{IF}(x;\tilde{T}_{\alpha,\sigma},\sigma_0) &= -\frac{\psi_{\alpha,\sigma}(x,\sigma_0)}{I_{\alpha,\sigma}(\sigma_0)} \\ &= \Delta_{\alpha,\sigma}(x;\sigma_0) + \frac{\alpha\sigma_0\left(\sigma_0^2-\sigma^2\right)\left[\alpha\sigma_0^2+(1-\alpha)\sigma^2\right]}{2\sigma^4+\alpha^2(\sigma_0^2-\sigma^2)^2},\end{aligned} \tag{82}$$

where

$$\Delta_{\alpha,\sigma}(x;\sigma_0) = \frac{\left[\alpha\sigma_0^2+(1-\alpha)\sigma^2\right]^{5/2}\left[\left(x/\sigma_0\right)^2-1\right]\exp\left\{\alpha x^2\left[\sigma_0^{-2}-\sigma^{-2}\right]/2\right\}}{\sigma\left[2\sigma^4+\alpha^2(\sigma_0^2-\sigma^2)^2\right]/\sigma_0}. \tag{83}$$

This formula remains valid also for $\alpha=0$ since in this case (82) reduces to the well known influence function

$$\mathrm{IF}(x;MLE,\sigma_0) = \frac{\sigma_0\left[\left(x/\sigma_0\right)^2-1\right]}{2}$$

obtained from the limit values

$$\psi_{0,\sigma}(x,\sigma_0) = -\left[\left(x/\sigma_0\right)^2-1\right]/\sigma_0 \quad \text{and} \quad I_{0,\sigma}(\tilde{\sigma}) = 2/\sigma_0^2$$

which do not depend on the escort parameter . We see from the formula (83) that the influence curve is unbounded for all $\sigma,\sigma_0>0$ and $\alpha\geq 0$. For $\alpha>0$ and $\sigma\neq\sigma_0$ we get $\mathrm{IF}(\sigma_0;\tilde{T}_{\alpha,\sigma},\sigma_0)\neq 0$. If moreover $\sigma<\sigma_0$ then $\mathrm{IF}(x;\tilde{T}_{\alpha,\sigma},\sigma_0)$ increases to infinity exponentially fast for $|x|\to\infty$. Thus $\tilde{T}_{\alpha,\sigma}$ with $\alpha>0$ and $\sigma\neq\sigma_0$ are strongly nonrobust.

Example 3.1.4. Power subdivergence estimator in Pareto family. It is hard to find simpler nontrivial examples of the max$\underline{\mathrm{D}}_\alpha$-estimators than the estimators of location (72), (73) from Example 2.1.2. Another relatively simple example is the family of max$\underline{\mathrm{D}}_\alpha$-estimators in the Pareto model with the family of measures $\mathcal{P} = \{P_\theta : \theta > 0\}$ defined on the interval $\mathcal{X} = (1, \infty)$ by the densities

$$p_\theta(x) = \frac{\theta}{x^{\theta+1}}. \tag{84}$$

with the mean values finite equal $\theta/(\theta - 1)$ in the domain $\theta > 1$ and variances finite and equal $\theta/[(\theta-2)(\theta-1)^2]$ in the domain $\theta > 2$. As before, the estimates $\tilde{\theta}_{\alpha,\theta,n}$ depend on the divergence parameters $\alpha \geq 0$ and escort parameters $\theta > 0$. By (50), for $\alpha = 0$ we get the MLE estimates

$$\tilde{\theta}_{0,\theta,n} = \mathrm{argmax}_{\tilde{\theta}}\ \Sigma_{i=1}^n \ln p_{\tilde{\theta}}(X_i) = \left(\frac{1}{n}\sum_{i=1}^n \ln X_i\right)^{-1}.$$

For $0 < \alpha < 1$ we can use the criterion function

$$M_{\alpha,\theta}(Q, \tilde{\theta}) = \frac{1}{1-\alpha} P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha + \frac{1}{\alpha} Q \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha, \quad Q \in \mathcal{Q} \tag{85}$$

of (53), or its derivative

$$\Psi_{\alpha,\theta}(Q, \tilde{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\tilde{\theta}} M_{\alpha,\theta}(Q, \tilde{\theta}) = P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} - Q \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} \tag{86}$$

given by (55), where in the present situation

$$P_{\tilde{\theta}} \cdot \left(\frac{p_\theta(x)}{p_{\tilde{\theta}}(x)}\right)^\alpha = \frac{\theta^\alpha \tilde{\theta}^{1-\alpha}}{\alpha\theta + (1-\alpha)\tilde{\theta}}, \quad \text{and} \quad s_\theta(x) = \frac{1}{\theta} - \ln x.$$

Substituting these expressions in (85), (86) we get the desired asymptotic characteristics of the max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta,n}$ obtained as argminima of the functions $M_{\alpha,\theta}(P_n, \tilde{\theta})$ or, equivalently, as solutions of the equations $\Psi_{\alpha,\theta}(P_n, \tilde{\theta}) = 0$ in the variable $\tilde{\theta}$. Further, by (22),

$$\psi_{\alpha,\theta}(x, \tilde{\theta}) = \Psi_{\alpha,\theta}(\delta_x, \tilde{\theta}) = P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} - \left(\frac{p_\theta(x)}{p_{\tilde{\theta}}(x)}\right)^\alpha s_{\tilde{\theta}}(x)$$

and using Theorem 3.1.1 one easily obtains the influence functions of the estimators $\tilde{\theta}_{\alpha,\theta,n}$ under consideration.

3.2. Power superdivergence estimators

In this subsection we deal with the min$\bar{\mathrm{D}}_\alpha$-estimators $\theta_{\alpha,n}$ with the power parameters $\alpha \geq 0$. For $\alpha = 0$ they coincide with the MLE's (50). Therefore we consider $\alpha > 0$ when these estimators are defined by (52) and (53). Restrict ourselves for simplicity to $0 < \alpha < 1$ and denote the function $\Psi_{\alpha,\theta}(Q,\tilde{\theta})$ from (55) in previous subsection temporarily by $\tilde{\Psi}_{\alpha,\theta}(Q,\tilde{\theta})$, i.e. let

$$\tilde{\Psi}_{\alpha,\theta}(Q,\tilde{\theta}) = P_{\tilde{\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}} - Q \cdot \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^\alpha s_{\tilde{\theta}}.$$

Further, let $\tilde{T}_{\alpha,\theta}(Q)$ be solution of the equation $\tilde{\Psi}_{\alpha,\theta}(Q,\tilde{\theta}) = 0$ in variable $\tilde{\theta}$, i.e.

$$\tilde{\Psi}_{\alpha,\theta}(Q,\tilde{T}_{\alpha,\theta}(Q)) = 0 \quad \text{for all } \theta \in \Theta. \tag{87}$$

Finally, let $M_{\alpha,\theta}(Q,\tilde{T}_{\alpha,\theta}(Q))$ be the function of variable $\theta \in \Theta$ obtained by inserting $\tilde{\theta} = \tilde{T}_{\alpha,\theta}(Q)$ in the function $M_{\alpha,\theta}(Q,\tilde{\theta})$ defined in (53). According to (52) and (53), the maximizers

$$T_\alpha(Q) = \text{argmax}_\theta \ M_{\alpha,\theta}(Q,\tilde{T}_{\alpha,\theta}(Q)) \tag{88}$$

generate the min$\bar{\mathrm{D}}_\alpha$-estimators $\theta_{\alpha,n}$ under consideration in the sense that $\theta_{\alpha,n} = T_\alpha(P_n)$. In the following theorem we consider the score function $s_\theta = \mathring{p}_\theta/p_\theta$ and we put for brevity $\tilde{\tau}_{\alpha,\theta} = \tilde{T}_{\alpha,\theta}(Q)$. The proof is in Broniatowski and Vajda [6].

Theorem 3.2.1. For all $0 < \alpha < 1$ the maximizers (88) solve the equations $\Psi_\alpha(Q,\theta) = 0$ in variable $\theta \in \Theta$ for the function

$$\Psi_\alpha(Q,\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} M_{\alpha,\theta}(Q,\tilde{\tau}_{\alpha,\theta}) = \frac{\alpha}{1-\alpha} P_{\tilde{\tau}_{\alpha,\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\tau}_{\alpha,\theta}}}\right)^\alpha s_\theta + Q \cdot \left(\frac{p_\theta}{p_{\tilde{\tau}_{\alpha,\theta}}}\right)^\alpha s_\theta. \tag{89}$$

Consequently the corresponding min$\bar{\mathrm{D}}_\alpha$-estimators $\theta_{\alpha,n} = T_\alpha(P_n)$ are solutions of the equations

$$\frac{\alpha}{1-\alpha} P_{\tilde{\tau}_{\alpha,\theta}} \cdot \left(\frac{p_\theta}{p_{\tilde{\tau}_{\alpha,\theta}}}\right)^\alpha s_\theta + \frac{1}{n}\sum_{i=1}^n \left(\frac{p_\theta(X_i)}{p_{\tilde{\tau}_{\alpha,\theta}}(X_i)}\right)^\alpha s_\theta(X_i) = 0. \tag{90}$$

Corollary 3.2.1. The influence functions $\mathrm{IF}(x;T_\alpha,\theta)$ of all min$\bar{\mathrm{D}}_\alpha$-estimators $\theta_{\alpha,n} = T_\alpha(P_n)$ with power parameters $0 < \alpha < 1$ at $P_\theta \in \mathcal{P}$ coincide with the influence function

$$\mathrm{IF}(x;T_0,\theta) = \boldsymbol{I}(\theta)^{-1} s_\theta(x) \qquad \text{(cf. (27) and (28))} \tag{91}$$

of the MLE $\theta_{0,n} = T_0(P_n)$.

P r o o f. By Theorem 3.2, the max$\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\theta}_{\alpha,\theta n} = \tilde{T}_{\alpha,\theta}(P_n)$ are Fisher consistent. Hence for $Q = P_{\theta_0}$ we get $\tilde{\tau}_{\alpha,\theta} := \tilde{T}_{\alpha,\theta}(P_{\theta_0}) = \theta_0$ in (89). Consequently it follows from (22) and (89) that the $\boldsymbol{\psi}$-functions

$$\boldsymbol{\psi}_\alpha(x,\tilde{\tau}_{\alpha,\theta}) \equiv \Psi_\alpha(\delta_x,\tilde{\tau}_{\alpha,\theta}) = \frac{\alpha}{1-\alpha} P_{\tilde{\tau}_{\alpha,\theta}} \cdot \left(\frac{p_{\theta_0}}{p_{\tilde{\tau}_{\alpha,\theta}}}\right)^\alpha s_{\theta_0} + \delta_x \cdot \left(\frac{p_{\theta_0}}{p_{\tilde{\tau}_{\alpha,\theta}}}\right)^\alpha s_{\theta_0}$$

of these estimators reduce for all $0 < \alpha < 1$ to the score function $s_{\theta_0}(x)$ which is the $\boldsymbol{\psi}$-function of MLE T_0. Similarly, we get from (27) and (24) for all $0 < \alpha < 1$ the matrix $\boldsymbol{I}(\theta_0) = P_{\theta_0} \cdot s_{\theta_0}^{\mathrm{t}} s_{\theta_0}$ corresponding to the MLE. Therefore the influence functions of all min$\bar{\mathrm{D}}_\alpha$-estimators under considerations reduce to the influence MLE function (91) which completes the proof. □

Formulas for the min$\bar{\mathrm{D}}_\alpha$-estimators of the normal location and/or scale are seen from the examples of Subsection 3.1.

4. DECOMPOSABLE PSEUDODISTANCES

The ϕ-divergences $D_\phi(P, Q)$, $\phi \in \boldsymbol{\Phi}$ 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* $\mathfrak{D}(P, Q)$ which may not satisfy the information processing property.

Definition 4.1. We say that $\mathfrak{D} : \mathcal{P} \otimes \mathcal{P}^+ \mapsto \mathbb{R}$ is a pseudodistance of probability measures $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $Q \in \mathcal{P}^+$if

$$\mathfrak{D}(P_\theta, P_{\tilde{\theta}}) \geq 0 \quad \text{for all } \theta, \tilde{\theta} \in \Theta \text{ with } \ \mathfrak{D}(P_\theta, P_{\tilde{\theta}}) = 0 \quad \text{iff} \quad \theta = \tilde{\theta}. \tag{92}$$

An additional restriction imposed in this section on pseudodistances $\mathfrak{D}(P, Q)$ will be the *decomposability*.

Definition 4.2. A pseudodistance $\mathfrak{D}$ on $\mathcal{P} \otimes \mathcal{P}^+$is a ***decomposable*** if there exist functionals $\mathfrak{D}^0 : \mathcal{P} \mapsto \mathbb{R}$, $\mathfrak{D}^1 : \mathcal{P}^+ \mapsto \mathbb{R}$ and measurable mappings

$$\rho_\theta : \mathcal{X} \mapsto \mathbb{R}, \qquad \theta \in \Theta \tag{93}$$

such that for all $\theta \in \Theta$ and $Q \in \mathcal{P}^+$ the expectations $Q \cdot \rho_\theta$ exist and

$$\mathfrak{D}(P_\theta, Q) = \mathfrak{D}^0(P_\theta) + \mathfrak{D}^1(Q) + Q \cdot \rho_\theta. \tag{94}$$

Definition 4.3. We say that a functional $T_{\mathfrak{D}} : \mathcal{Q} \mapsto \Theta$ for $\mathcal{Q} = \mathcal{P}^+ \cup \mathcal{P}_{\mathrm{emp}}$ defines a ***minimum pseudodistance estimator*** (briefly, min $\mathfrak{D}$-estimator)if $\mathfrak{D}(P_\theta, Q)$ is a decomposable pseudodistance on $\mathcal{P} \otimes \mathcal{P}^+$ and the parameters $T_{\mathfrak{D}}(Q) \in \Theta$ minimize $\mathfrak{D}^0(P_\theta) + Q \cdot \rho_\theta$ on Θ, in symbols

$$T_{\mathfrak{D}}(Q) = \mathrm{argmin}_\theta \left[\mathfrak{D}^0(P_\theta) + Q \cdot \rho_\theta\right] \qquad \text{for all } \ Q \in \mathcal{Q}. \tag{95}$$

In particular, for $Q = P_n \in \mathcal{P}_{\mathrm{emp}}$

$$\theta_{\mathfrak{D},n} := T_{\mathfrak{D}}(P_n) = \mathrm{argmin}_\theta \left[\mathfrak{D}^0(P_\theta) + \frac{1}{n}\sum_{i=1}^n \rho_\theta(X_i)\right] \quad \text{if } \ P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}. \tag{96}$$

Theorem 4.1. Every min $\mathfrak{D}$-estimator

$$\theta_{\mathfrak{D},n} = \mathrm{argmin}_\theta \left[\mathfrak{D}^0(P_\theta) + \frac{1}{n}\sum_{i=1}^n \rho_\theta(X_i)\right] \tag{97}$$

is Fisher consistent in the sense that

$$T_{\mathfrak{D}}(P_{\theta_0}) = \text{argmin}_\theta \mathfrak{D}(P_\theta, P_{\theta_0}) = \theta_0 \quad \text{for all } \theta_0 \in \Theta. \tag{98}$$

P r o o f. Consider arbitrary fixed $\theta_0 \in \Theta$. Then, by assumptions, $\mathfrak{D}^1(P_{\theta_0})$ is a finite constant. Therefore (95) together the definition of pseudodistance implies

$$\begin{aligned} T_{\mathfrak{D}}(P_{\theta_0}) &= \text{argmin}_\theta \left[\mathfrak{D}^0(P_\theta) + Q \cdot \rho_\theta\right] \\ &= \text{argmin}_\theta \left[\mathfrak{D}^0(P_\theta) + \mathfrak{D}^1(P_{\theta_0}) + Q \cdot \rho_\theta\right] \\ &= \text{argmin}_\theta \mathfrak{D}(P_\theta, P_{\theta_0}) = \theta_0. \end{aligned}$$

□

The decomposability of pseudodistance $\mathfrak{D}(P_\theta, Q)$ leads to the additive structure of the criterion

$$\mathfrak{D}(P_\theta, P_n) \sim \mathfrak{D}^0(P_\theta) + P_n \cdot \rho_\theta = \mathfrak{D}^0(P_\theta) + \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(X_i) \tag{99}$$

in the definition (97) of the min $\mathfrak{D}$-estimators which opens the possibility to apply the methods of the asymptotic theory of M-estimators (cf. Hampel et al. [7], van der Vaart and Wellner [20], van der Vaart [19] or Mieske and Liese [10]).
The general min $\mathfrak{D}$-estimators and their special classes studied in Subsections 4.1, 4.2 below were introduced in Vajda [18]. They contain as a subclass all the max$\underline{\mathrm{D}}_\phi$-estimators of Section 3. To see this suppose that the assumptions of Section 3 related to the estimators (101) hold and consider for arbitrary fixed $(\phi, \tau) \in \mathbf{\Phi} \otimes \Theta$ the well defined expressions

$$\mathfrak{D}^0_{\phi,\tau}(P_\theta) = -\ P_\tau \cdot \phi'\left(\frac{p_\tau}{p_\theta}\right), \qquad \rho_{\phi,\tau,\theta} = -\ \phi^\sharp\left(\frac{p_\tau}{p_\theta}\right)$$

and

$$\mathfrak{D}^1_{\phi,\tau}(Q) = -\ \inf_\theta \left[\mathfrak{D}^0_{\phi,\tau}(P_\theta) + Q \cdot \rho_{\phi,\tau,\theta}\right].$$

Theorem 4.2. The sum

$$\mathfrak{D}(P_\theta, Q) := \mathfrak{D}^0_{\phi,\tau}(P_\theta) + \mathfrak{D}^1_{\phi,\tau}(Q) + Q \cdot \rho_{\phi,\tau,\theta} \tag{100}$$

is a pseudodistance on $\mathcal{P}\otimes\mathcal{P}^+$ and the maximum subdivergence estimator

$$\theta_{\phi,\tau,n} = \text{argmax}_\theta \left[P_\tau \cdot \phi'\left(\frac{p_\tau}{p_\theta}\right) + \frac{1}{n}\sum_{i=1}^{n} \phi^\sharp\left(\frac{p_\tau(X_i)}{p_\theta(X_i)}\right)\right] \tag{101}$$

of Section 3 with the divergence parameter $\phi \in \mathbf{\Phi}$ and escort parameter $\tau \in \Theta$ is the min $\mathfrak{D}$-estimator for the decomposable pseudodistance (100).

P r o o f. Fix $(\phi, \tau) \in \mathbf{\Phi}\otimes\Theta$ and let the assumptions of Section 3 related to the estimators (101) hold. Then for any $\theta_0 \in \Theta$

$$\mathfrak{D}(P_{\theta_0}, Q) = \mathfrak{D}^0_{\phi,\tau}(P_{\theta_0}) + Q \cdot \rho_{\phi,\tau,\theta_0} - \inf_\theta \left[\mathfrak{D}^0_{\phi,\tau}(P_{\theta_0}) + Q \cdot \rho_{\phi,\tau,\theta_0}\right] \geq 0.$$

If $Q \in \mathcal{P}$ then, by (31) and (33),

$$\mathfrak{D}_{\phi,\tau}(P_{\theta_0}, Q) = \sup_\theta \left[P_{\theta_0} \cdot \phi' \left(\frac{p_\tau}{p_{\theta_0}} \right) + Q \cdot \phi^\sharp \left(\frac{p_\tau}{p_\theta} \right) \right] - P_{\theta_0} \cdot \phi' \left(\frac{p_\tau}{p_{\theta_0}} \right) + Q \cdot \phi^\sharp \left(\frac{p_\tau}{p_{\theta_0}} \right)$$

$$= D_\phi(P_{\theta_0}, Q) - \underline{\mathrm{D}}_{\phi,\tau}(P_{\theta_0}, Q).$$

By Theorem 3.1, this difference is zero if and only if $Q = P_{\theta_0}$ which proves that (100) is pseudodistance on $\mathcal{P} \otimes \mathcal{P}^+$. On the other hand, obviously, (101) satisfies

$$\theta_{\phi,\tau,n} = \operatorname{argmin}_\theta \left[\mathfrak{D}^0_{\phi,\tau}(P_\theta) + P_n \cdot \rho_{\phi,\tau,\theta} \right]$$

so that it is $\min \mathfrak{D}$-estimator for the pseudodistance (100) which completes the proof. □

The minimum superdivergence estimators $\theta_{\phi,n}$ of Section 3 (the $\min \bar{\mathrm{D}}_\phi$-estimators) minimize the suprema

$$\sup_\tau \mathfrak{D}(P_\theta, Q) \qquad \text{for } Q = P_n$$

of the decomposable pseudodistance (100). However, the suprema of decomposable pseudodistances are not in general decomposable pseudodistances. Therefore the standard theory of M-estimators is not applicable to this class of estimators. An exception is the MLE $\theta_{\phi_0,n}$ obtained for the logarithmic function ϕ_0 given in (7).

4.1. Power pseudodistance estimators

In this subsection we study a special class of pseudodistances $\mathfrak{D}_\psi(P_\theta, Q)$ defined on $\mathcal{P} \otimes \mathcal{P}^+$ by the integral formula

$$\mathfrak{D}_\psi(P_\theta, Q) = \int \psi(p_\theta, q)\, \mathrm{d}\lambda \quad \text{ for } p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda}, q = \frac{\mathrm{d}Q}{\mathrm{d}\lambda}, \tag{102}$$

where $\psi(s,t)$ are *reflexive* in the sense that they are nonnegative functions of arguments $s, t > 0$ with $\psi(s,t) = 0$ iff $s = t$. If a function ψ is reflexive and also *decomposable* in the sense

$$\psi(s,t) = \psi^0(s) + \psi^1(t) + \rho(s)\, t, \quad s, t \geq 0, \tag{103}$$

for some $\psi^0, \psi^1, \rho : (0, \infty) \to \mathbb{R}$ then the corresponding ψ-pseudodistance (102) is a *decomposable pseudodistance* satisfying

$$\mathfrak{D}_\psi(P_\theta, Q) = \mathfrak{D}^0_\psi(P_\theta) + \mathfrak{D}^1_\psi(Q) + Q \cdot \rho_\theta \quad \text{(cf. (94))} \tag{104}$$

for

$$\mathfrak{D}^0_\psi(P_\theta) = \int \psi^0(p_\theta)\, \mathrm{d}\lambda, \;\; \mathcal{D}^1_\psi(Q) = \int \psi^1(q)\, \mathrm{d}\lambda \;\text{ and }\; \rho_\theta = \rho(p_\theta). \tag{105}$$

Example 4.1.1. The ϕ-divergences $D_\phi(P_\theta, Q)$ are special ψ-pseudodistances (102) for the functions

$$\psi(s,t) = \phi(s/t)\, t - \phi'(1)(s-t), \quad s,t > 0 \tag{106}$$

since they are nonnegative and reflexive, and (106) implies $\mathfrak{D}_\psi(P_\theta, Q) = D_\phi(P_\theta, Q)$ for all $P \in \mathcal{P}, Q \in \mathcal{P}^+$ when $\phi \in \mathbf{\Phi}$ and ψ are related by (106). However, the functions (106) in general do not satisfy the decomposability condition (103) so that the ϕ-divergences are not in general decomposable pseudodistances. An exception is the logarithmic function $\phi = \phi_0$ defined in (7) for which the $\min \mathfrak{D}_{\phi_0}$-estimator is the MLE.

Example 4.1.2. **L_2-estimator** The quadratic function $\psi(s,t) = (s-t)^2$ is reflexive and also decomposable in the sense of (103). Thus it defines the decomposable pseudodistance

$$\mathfrak{D}_\psi(P_\theta, Q) = \int (p_\theta - q)^2 \,\mathrm{d}\lambda = \|p_\theta - q\|^2$$

on $\mathcal{P} \otimes \mathcal{P}^+$ for $\mathcal{P}^+ \subset L_2(\lambda)$. It is easy to verify that the decomposability in the sense of (104) holds for

$$\mathfrak{D}^0_\psi(P_\theta) = \int p_\theta^2 \,\mathrm{d}\lambda, \quad \mathcal{D}^1_\psi(Q) = \int q^2 \,\mathrm{d}\lambda_Q \quad \text{and} \quad \rho_\theta = -2p_\theta.$$

The corresponding $\min \mathfrak{D}_\psi$-estimator defined by (97) is in this case the L_2-estimator

$$\theta_n = \mathrm{argmin}_\theta \left[\int p_\theta^2 \,\mathrm{d}\lambda - \frac{2}{n} \sum_{i=1}^n p_\theta(X_i) \right] \tag{107}$$

which is known to be robust but not efficient (see e. g. Hampel et al. [7]).

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

$$\psi_\alpha(s,t) = \psi^0_\alpha(s) + \psi^1_\alpha(t) + \rho_\alpha(s)\, t \quad \text{for all } \alpha \geq 0 \quad \text{(cf. (103))} \tag{108}$$

with the limits at 0 satisfying for some constant $\varkappa$ all $s > 0$ the conditions

$$\psi^0_0(s) = \lim_{\alpha \downarrow 0} \psi^0_\alpha(s) = \varkappa\, s \quad \text{and} \quad \lim_{\alpha \downarrow 0} \rho_\alpha(s) = \rho_0(s) = -\ln s. \tag{109}$$

Then for all $\alpha \geq 0$ and $(P_\theta, Q) \in \mathcal{P} \otimes \mathcal{P}^+$ the family of ψ_α-pseudodistances

$$\mathfrak{D}_\alpha(P_\theta, Q) := \mathfrak{D}_{\psi_\alpha}(P_\theta, Q), \quad \alpha \geq 0 \tag{110}$$

satisfies the decomposability condition

$$\mathfrak{D}_\alpha(P_\theta, Q) = \mathfrak{D}^0_\alpha(Q) + \mathfrak{D}^1_\alpha(P_\theta) + Q \cdot \rho_{\alpha,\theta} \quad \text{(cf. (94))} \tag{111}$$

for

$$\mathfrak{D}^0_\alpha(P_\theta) = \int \psi^0_\alpha(p_\theta) \,\mathrm{d}\lambda, \quad \mathfrak{D}^1_\alpha(Q) = \int \psi^1_\alpha(q) \,\mathrm{d}\lambda \quad \text{and} \quad \rho_{\alpha,\theta} = \rho_\alpha(p_\theta). \tag{112}$$

In other words, the pseudodistances $\mathfrak{D}_\alpha(P_\theta, Q)$ defined by (110) are decomposable and define in accordance with (97) the family of min $\mathfrak{D}_\alpha$-estimators

$$\theta_{\alpha,n} = \arg\min_\theta \left[\mathfrak{D}^0_{\psi_\alpha}(P_\theta) + P_n \cdot \rho_{\alpha,\theta}\right] \tag{113}$$

$$= \operatorname{argmin}_\theta \left[\int \psi^0_\alpha(p_\theta)\,\mathrm{d}\lambda + \frac{1}{n}\sum_{i=1}^n \rho_\alpha(p_\theta(X_i))\right], \quad \alpha \geq 0. \tag{114}$$

Here (109) guarantees that this family contains as a special case for $\alpha = 0$ the efficient but non-robust MLE

$$\theta_{0,n} = \operatorname{argmin}_\theta \left[\text{const} - \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X_i)\right] \tag{115}$$

while for $\alpha > 0$ the $\theta_{\alpha,n}$'s are expected to be less efficient but more robust than $\theta_{0,n}$.

The rest of this subsection studies special family of decomposable pseudodistances $\mathfrak{D}_\alpha(P_\theta, Q)$. It is defined on $\mathcal{P} \otimes \mathcal{Q}$ in accordance with (110) and (102) by the functions

$$\psi_\alpha(s,t) = t^{1+\alpha}\left[\alpha\phi_{1+\alpha}\left(\frac{s}{t}\right) + (1-\alpha)\phi_\alpha\left(\frac{s}{t}\right)\right], \quad \alpha \geq 0 \tag{116}$$

of variables s, $t > 0$ where $\phi_{1+\alpha}$ and ϕ_α are the power functions defined by (6), (7). These functions satisfy (108), (109) as it is clarified by the next theorem. In this theorem and in the sequel we use for the function (116) the relations

$$\psi_\alpha(s,t) = \frac{s^{1+\alpha}}{1+\alpha} + t^{1+\alpha}\left(\frac{1}{\alpha} - \frac{1}{1+\alpha}\right) - \frac{ts^\alpha}{\alpha} \tag{117}$$

$$= \frac{s^{1+\alpha} - t^{1+\alpha}}{1+\alpha} + t\left(\frac{t^\alpha - 1}{\alpha} - \frac{s^\alpha - 1}{\alpha}\right) \tag{118}$$

when $\alpha > 0$ and

$$\psi_0(s,t) = s - t + t\ln t - t\ln s \tag{119}$$

$$= \lim_{\alpha\downarrow 0} \frac{s^{1+\alpha} - t^{1+\alpha}}{1+\alpha} + t\left(\frac{t^\alpha - 1}{\alpha} - \frac{s^\alpha - 1}{\alpha}\right) \tag{120}$$

when $\alpha = 0$.

Theorem 4.1.1. The power functions (116) are reflexive and decomposable in the sense of (108) with

$$\psi^0_\alpha(s) = \frac{s^{1+\alpha}}{1+\alpha}, \quad \psi^1_\alpha(t) = \begin{cases} t\left[\frac{t^\alpha - 1}{\alpha} - \frac{t^\alpha}{1+\alpha}\right] \\ t\ln t - t \end{cases} \quad \text{and} \quad \rho_\alpha(s) = \begin{cases} -\frac{s^\alpha - 1}{\alpha} & \text{if } \alpha > 0 \\ -\ln s & \text{if } \alpha = 0. \end{cases} \tag{121}$$

Moreover, this family is continuous in the parameter $\alpha \downarrow 0$ and satisfies (109) for $\varkappa = 1$.

P r o o f. Decomposition (108) for function $\psi_\alpha(s,t)$ of (116) into the components (121) is clear from (118) when $\alpha > 0$ and (119) when $\alpha = 0$. The continuity in the parameter $\alpha \downarrow 0$ and (109) for $\varkappa = 1$ follow from (120). We shall prove the nonnegativity and reflexivity. For arbitrary arguments s, $t > 0$ and fixed parameters a, $b > 0$ with the property $1/a + 1/b = 1$ it holds

$$st \le \frac{s^a}{a} + \frac{t^b}{b} \tag{122}$$

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

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

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

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

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. □

By (110), (102) and Theorem 4.1.1, the power functions (116) generate

$$\psi^0(p_\theta) = \frac{1}{1+\alpha}p_\theta^\alpha \quad \text{and} \quad \rho_\alpha(p_\theta) = \begin{cases} -\frac{1}{\alpha}p_\theta^\alpha & \text{if } \alpha > 0 \\ -\ln p_\theta & \text{if } \alpha = 0 \end{cases} \tag{123}$$

and define the family of decomposable pseudodistances

$$\begin{aligned} \mathfrak{D}_\alpha(P_\theta, Q) &= \textstyle\int \psi_\alpha(p_\theta, q)\,\mathrm{d}\lambda \\ &= \begin{cases} \frac{1}{1+\alpha}P_\theta \cdot p_\theta^\alpha + \frac{1}{\alpha(1+\alpha)}Q \cdot q^\alpha - \frac{1}{\alpha}Q \cdot p_\theta^\alpha & \text{if } \alpha > 0 \\ Q \cdot (\ln q - \ln p_\theta) & \text{if } \alpha = 0 \end{cases} \end{aligned} \tag{124}$$

in (114). Relation of this family to the family of power divergences $D_\alpha(P_\theta, Q)$ defined by (5) is rigorously established in the next theorem. It refers to the auxiliary family of functions

$$\varphi_\alpha(s,t) = t\left[\alpha\phi_{1+\alpha}\left(\frac{s}{t}\right) + (1-\alpha)\phi_\alpha\left(\frac{s}{t}\right)\right] \tag{125}$$

of arguments $s, t > 0$ parametrized by $\alpha \ge 0$.

Theorem 4.1.2. Decomposable pseudodistances (124) are for all $(P,Q) \in \mathcal{P}\otimes\mathcal{P}^+$ modifed power divergences $D_\alpha(P,Q)$ and $D_{1+\alpha}(P,Q)$ in the sense that the pseudodistance densities $\psi_\alpha(p,q)$ are weighted densities $\varphi_\alpha(p,q)$ of the mixed power divergences

$$\int \varphi_\alpha(p,q)\,\mathrm{d}\lambda_Q = \alpha\, D_{1+\alpha}(P,Q) + (1-\alpha)\, D_\alpha(P,Q) \tag{126}$$

with the power weights $w_\alpha(q) = q^\alpha$, i.e. $\psi_\alpha(p,q) = w_\alpha(q)\varphi_\alpha(p,q)$ on $(\mathcal{X}, \mathcal{A})$.

P r o o f. By (125),

$$\int \varphi_\alpha(p,q)\,\mathrm{d}\lambda = \alpha \int \phi_{1+\alpha}(p,q)\,\mathrm{d}\lambda + (1-\alpha)\int \phi_\alpha(p,q)\,\mathrm{d}\lambda$$

$$= \alpha\, D_{1+\alpha}(P,Q) + (1-\alpha)\, D_\alpha(P,Q). \tag{127}$$

By (116), $\psi_\alpha(s,t) = t^\alpha \varphi_\alpha(s,t)$ so that, by the first equality in (124),

$$\mathfrak{D}_\alpha(P_\theta,Q) = \int \psi_\alpha(p_\theta,q)\,\mathrm{d}\lambda = \int w_\alpha(q)\varphi_\alpha(p,q))\,\mathrm{d}\lambda.$$

This together with (127) implies the desired result. □

Due to Theorem 4.1.2, we call the pseudodistances $\mathfrak{D}_\alpha(P,Q)$ simply ***power pseudodistances*** of orders $\alpha \geq 0$. The next theorem guarantees finiteness and continuity of these divergences. It is restricted to the families $\mathcal{P}$ satisfying for some $\beta > 0$ the condition

$$p^\beta,\ q^\beta, \ln p \in \mathbb{L}_1(Q) \quad \text{for all } P\in\mathcal{P},\ Q\in\mathcal{P}^+. \tag{128}$$

Theorem 4.1.3. If (128) holds for some $\beta > 0$ then for all $0 \leq \alpha \leq \beta$, the modified power divergences are well defined by (124) and finite, satisfying for all $P\in\mathcal{P}$, $Q\in\mathcal{P}^+$ the continuity relation

$$\lim_{\alpha\downarrow 0}\mathfrak{D}_\alpha(P,Q) = \mathfrak{D}_0(P,Q). \tag{129}$$

P r o o f. By (118),

$$\mathfrak{D}_\alpha(P,Q) = \frac{1}{1+\alpha}\,(P\cdot p^\alpha - Q\cdot q^\alpha) + Q\cdot\left(\frac{q^\alpha-1}{\alpha} - \frac{p^\alpha - 1}{\alpha}\right).$$

By means of the indicator function **1** we can decompose

$$P\cdot p^\alpha = P\cdot(p^\alpha \mathbf{1}(p\leq 1)) + P\cdot(p^\alpha\mathbf{1}(p>1))$$

where

$$\lim_{\alpha\downarrow 0} P\cdot(p^\alpha\mathbf{1}(p\leq 1)) = P\cdot(\mathbf{1}(p\leq 1))$$

by the Lebesgue bounded convergence theorem for integrals and

$$\lim_{\alpha\downarrow 0} P\cdot(p^\alpha\mathbf{1}(p>1)) = P\cdot(\mathbf{1}(p>1))$$

by the monotone convergence theorem for integrals. Therefore

$$\lim_{\alpha\downarrow 0} P\cdot p^\alpha = P\cdot(\mathbf{1}(p\leq 1)) + P\cdot(\mathbf{1}(p>1)) = 1.$$

Similarly, $\lim_{\alpha\downarrow 0} Q\cdot q^\alpha = 1$. The convergences

$$\lim_{\alpha\downarrow 0} Q\cdot\frac{q^\alpha-1}{\alpha} = Q\cdot\ln q \quad\text{and}\quad \lim_{\alpha\downarrow 0} Q\cdot\frac{p^\alpha-1}{\alpha} = Q\cdot\ln p$$

follow from the monotone convergence as well, because for every fixed $t>0$

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

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

By (121) the expressions $\mathfrak{D}^0_{\psi_\alpha}(P_\theta)$ considered in(113), (114) are now given by

$$\mathfrak{D}^0_\alpha(P_\theta) = \frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathrm{d}\lambda \quad\text{for all } \alpha\geq 0.$$

Therefore the formulas (113), (114) and (123) lead to the ***power pseudodistance estimators*** (briefly, $\min\mathfrak{D}_\alpha$-estimators)

$$\theta_{\alpha,n} = \begin{cases} \operatorname{argmin}_\theta \left[\frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathrm{d}\lambda - \frac{1}{n\alpha}\sum_{i=1}^n p_\theta^\alpha(X_i)\right] & \text{if } \alpha>0, \\ \operatorname{argmax}_\theta \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X_i) & \text{if } \alpha=0. \end{cases} \tag{130}$$

Here the upper objective function can be replaced by

$$\begin{aligned} &\frac{1-\alpha}{\alpha} + \frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathrm{d}\lambda - \frac{1}{n\alpha}\sum_{i=1}^n p_\theta^\alpha(X_i) \\ &= \frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathrm{d}\lambda - \frac{1}{n}\sum_{i=1}^n \frac{p_\theta^\alpha(X_i)-1}{\alpha} - 1 \end{aligned}$$

which tends for $\alpha\downarrow 0$ to the lower criterion function. Therefore, if for a fixed n the minima of all functions in (130) are in a compact subset of Θ and the MLE $\theta_{n,0}$ is unique then

$$\lim_{\alpha\downarrow 0}\theta_{n,\alpha} = \theta_{n,0}. \tag{131}$$

Example 4.1.3. L_2-estimator revisited. By (130), the $\min\mathfrak{D}_\alpha$-estimator of order $\alpha=1$ is defined by

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

so that it is nothing but the L_2-estimator θ_n from Example 4.1.2. The family of estimators $\theta_{n,\alpha}$ from (130) smoothly connects this robust estimator with the efficient MLE $\theta_{n,0}$ when the parameter α decreases from 1 to 0.

Remark 4.1.1. The special class of the $\min \mathfrak{D}_\alpha$-estimators $\theta_{\alpha,n}$ given by (130) was proposed by Basu et al. [2] who confirmed their efficiency for $\alpha \approx 0$ and their intuitively expected robustness for $\alpha > 0$. These authors called $\theta_{\alpha,n}$ *minimum density power divergence estimators* without actual clarification of the relation of the "density power divergences" $\mathfrak{D}_\alpha(P,Q)$ to the standard power divergences $D_\alpha(P,Q)$ studied in Liese and Vajda [8] and Read and Cressie [11]. Theorem 3.1.2 which explains $\mathfrak{D}_\alpha(P,Q)$ as a convex mixture of modified power divergences $D_\alpha(P,Q)$ and $D_{1+\alpha}(P,Q)$ where the modification means weighting of the power divergence densities by the power q^α of the second probability density, is in this respect an interesting new result.

Remark 4.1.2. The formula (130) can be given the equivalent form

$$\theta_{\alpha,n} = \operatorname{argmax}_\theta \begin{cases} \frac{1}{n}\sum_{i=1}^n \frac{1}{\alpha}\left(p_\theta^\alpha(X_i) - 1\right) - \frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathrm{d}\lambda & \text{if } \alpha > 0, \\ \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X_i) - 1 & \text{if } \alpha = 0. \end{cases} \tag{132}$$

If the integral does not depend on θ then (132) is equivalent to

$$\theta_{\alpha,n} = \operatorname{argmax}_\theta \begin{cases} \frac{1}{n}\sum_{i=1}^n \frac{1}{\alpha}\left(p_\theta^\alpha(X_i) - 1\right) & \text{if } \alpha > 0, \\ \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X_i) & \text{if } \alpha = 0. \end{cases} \tag{133}$$

This subclass of general $\min \mathfrak{D}_\alpha$-estimators (132) was included in a wider family of generalized MLE's introduced and studied previously in Vajda ([14, 15]). However, the whole class (132) was not introduced there.

If the statistical model $\langle(\mathcal{X},\mathcal{A});\mathcal{P} = (P_\theta : \theta \in \Theta)\rangle$ is reparametrized by $\vartheta = \vartheta(\theta)$ then the new $\min \mathcal{D}_\alpha$-estimates ϑ_{α_n} are related to the original $\theta_{\alpha,n}$ by $\vartheta_{\alpha,n} = \vartheta(\theta_{\alpha,n})$. If the observations $x \in \mathcal{X}$ are replaced by $y = T(x)$ where $T : (\mathcal{X},\mathcal{A}) \mapsto (\mathcal{Y},\mathcal{B})$ is a measurable statistic with the inverse T^{-1} then the densities

$$\tilde{p}_\theta = \frac{\mathrm{d}\tilde{P}_\theta}{\mathrm{d}\tilde{\lambda}}$$

in the transformed model $\langle(\mathcal{Y},\mathcal{B});\tilde{\mathcal{P}} = (\tilde{P}_\theta = P_\theta T^{-1} : \theta \in \Theta)\rangle$ w.r.t. σ-finite dominating measure $\tilde{\lambda} = \lambda T^{-1}$ is related to the original densities p_θ by

$$\tilde{p}_\theta(y) = p_\theta(T^{-1}y)\,\mathcal{J}_T(y), \tag{134}$$

where $\mathcal{J}_T(y) = \mathrm{d}\lambda T^{-1}/\mathrm{d}\tilde{\lambda}$ is a generalized Jacobian of the statistic T. If $\mathcal{X}$, $\mathcal{Y}$ are Euclidean spaces, λ is the Lebesque measure and the inverse mapping $H = T^{-1}$ is differentiable then $\mathcal{J}_T(y)$ is the determinant

$$\mathcal{J}_T(y) = \left|\frac{\mathrm{d}}{\mathrm{d}y}H(y)\right|.$$

The $\min \mathcal{D}_\alpha$-estimators are in general not equivariant w.r.t. invertible transformations of observations T, unless $\alpha = 0$. The following theorem generalizes similar result of Section 3.4 in Basu et al. [2].

Theorem 4.1.4. The min $\mathfrak{D}_\alpha$-estimates $\tilde{\theta}_{\alpha,n}$ in the above considered transformed model coincide with the original min $\mathfrak{D}_\alpha$-estimates $\theta_{\alpha,n}$ if the Jacobian $\mathcal{J}_T$ of transformation is a nonzero constant on the transformed observation space $\mathcal{Y}$. Thus if $\mathcal{X},\mathcal{Y}$ are Euclidean spaces then the min $\mathfrak{D}_\alpha$-estimators are equivariant under linear statistics $Tx = ax + b$.

P r o o f. For $\alpha = 0$ the min $\mathfrak{D}_\alpha$-estimator is the MLE whose equivariance is well known. For $\alpha > 0$, by definition (130) and (134),

$$\begin{aligned}\tilde{\theta}_{\alpha,n} &= \operatorname{argmin}_\theta \left[\frac{1}{1+\alpha}\int_{\mathcal{Y}} \tilde{p}_\theta^{1+\alpha}\,\mathrm{d}\tilde{\lambda} - \frac{1}{n\alpha}\sum_{i=1}^n \tilde{p}_\theta^\alpha(TX_i)\right] \\ &= \operatorname{argmin}_\theta \left[\frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathcal{J}_T\,\mathrm{d}\lambda - \frac{1}{n\alpha}\sum_{i=1}^n p_\theta(X_i)\,\mathcal{J}_T(TX_i)\right].\end{aligned}$$

We see by comparison with (130) that $\tilde{\theta}_{\alpha,n} = \theta_{\alpha,n}$ if $\mathcal{J}_T$ is a nonzero constant on $\mathcal{Y}$. If $\alpha = 0$ then the estimator is MLE and its equivariance is well known. □

Next we derive the influence function of the min $\mathfrak{D}_\alpha$-estimators $\theta_{\alpha,n}$ of (130). Similarly as in (54), we use

$$s_\theta = \frac{\mathrm{d}}{\mathrm{d}\theta}\ln p_\theta \quad \text{and} \quad \mathring{s}_\theta = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} s_\theta.$$

It holds $\theta_{\alpha,n} = T_\alpha(P_n)$ where $T_\alpha(Q)$ for $Q \in \mathcal{Q}$ solves the equation $\Psi_\alpha(Q,\theta) \equiv Q \cdot \boldsymbol{\psi}(x,\theta) = 0$ for

$$\begin{aligned}\boldsymbol{\psi}_\alpha(x,\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{p_\theta^\alpha}{\alpha} - \frac{1}{1+\alpha}\int p_\theta^{1+\alpha}\,\mathrm{d}\lambda\right) \\ &= p_\theta^\alpha(x)\,s_\theta(x) - P_\theta \cdot p_\theta^\alpha s_\theta.\end{aligned} \tag{135}$$

Since

$$\mathring{\psi}_\alpha(x,\theta) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} \boldsymbol{\psi}_\alpha(x,\theta) = \Pi_{\alpha,\theta}(x) - P_\theta \cdot \left(\Pi_{\alpha,\theta} + p_\theta^\alpha s_\theta s_\theta^{\mathrm{t}}\right) \tag{136}$$

for

$$\Pi_{\alpha,\theta} = p_\theta^\alpha\,\left(\alpha s_\theta s_\theta^{\mathrm{t}} + \mathring{s}_\theta\right), \tag{137}$$

the matrix (24) is given for all $Q \in \mathcal{P}^+$ by the formula

$$\boldsymbol{I}_\alpha(Q) = Q \cdot \Pi_{\alpha,\tau_\alpha}(x) - P_{\tau_\alpha} \cdot \left(\Pi_{\alpha,\tau_\alpha} + p_{\tau_\alpha}^\alpha s_{\tau_\alpha} s_{\tau_\alpha}^{\mathrm{t}}\right) \text{ for } \tau_\alpha = T_\alpha(Q) \in \Theta \tag{138}$$

In particular,

$$\boldsymbol{I}_\alpha(\theta) \equiv \boldsymbol{I}_\alpha(P_\theta) = -P_\theta \cdot p_\theta^\alpha s_\theta s_\theta^{\mathrm{t}}. \tag{139}$$

By combining (135), (138) and (139) with Theorem 2.1 and Corollary 2.1, and taking into account the Fisher consistency in Theorem 4.1, we obtain the following extension of the influence function obtained in § 3.3 of Basu et al. [2] to arbitrary observation spaces $(\mathcal{X},\mathcal{A})$.

Theorem 4.1.5. If the influence function (21) at $Q \in \mathcal{P}^+$ or $P_\theta \in \mathcal{P}$ exists for some $\min \mathfrak{D}_\alpha$-estimator $\theta_{\alpha,n} = T_\alpha(P_n)$ then it is given by the formula

$$\mathrm{IF}(x; T_\alpha, Q) = -\boldsymbol{I}_\alpha(Q)^{-1}\left[p_{\tau_\alpha}^\alpha(x)\, s_{\tau_\alpha}(x) - P_{\tau_\alpha} \cdot p_{\tau_\alpha}^\alpha s_{\tau_\alpha}\right] \quad \text{for } \tau_\alpha = T_\alpha(Q) \tag{140}$$

or

$$\mathrm{IF}(x; T_\alpha, \theta) = -\boldsymbol{I}_\alpha(\theta)^{-1}\left[p_\theta^\alpha(x)\, s_\theta(x) - P_\theta \cdot p_\theta^\alpha s_\theta\right] \tag{141}$$

respectively.

4.2. Applications in the normal family

Consider the general normal family of Example 3.1.1. By (132), $\min \mathfrak{D}_\alpha$-estimator $\theta_{\alpha,n} = (\mu_{\alpha,n}, \sigma_{\alpha,n})$ is the MLE given by (62) when $\alpha = 0$. Since

$$\int p_\theta^{1+\alpha}\,\mathrm{d}x = \int \left(\frac{\exp\{-(x-\mu)^2/2\sigma^2\}}{(2\pi\sigma^2)^{1/2}}\right)^{1+\alpha} \mathrm{d}x = \frac{(1+\alpha)^{-1/2}}{(2\pi\sigma^2)^{\alpha/2}}, \tag{142}$$

we see from (132) that the $\min \mathfrak{D}_\alpha$-estimates are for $\alpha > 0$ given by

$$\begin{aligned}(\mu_{\alpha,n}, \sigma_{\alpha,n}) &= \mathrm{argmax}_{\mu,\sigma}\left[\frac{1}{\alpha n}\sum_{i=1}^n \frac{\exp\{-\alpha(X_i-\mu)^2/2\sigma^2\}}{(2\pi\sigma^2)^{\alpha/2}} - \frac{(1+\alpha)^{-3/2}}{(2\pi\sigma^2)^{\alpha/2}}\right] \\ &= \mathrm{argmax}_{\mu,\sigma}\frac{1}{n\sigma^\alpha}\sum_{i=1}^n\left(\exp\left\{-\alpha\frac{(X_i-\mu)^2}{2\sigma^2}\right\} - \frac{\alpha}{(1+\alpha)^{3/2}}\right).\end{aligned} \tag{143}$$

Notice that in practical applications, the trivial "solutions" $(\mu_{\alpha,n}, \sigma_{\alpha,n}) = (\max_i X_i, 0)$ can be avoided by restricting the maximization to the scales bounded avay from zero.

Example 4.2.1. Power pseudodistance estimators of location. Consider the normal family $\mathcal{P} = \{P_\mu : \mu \in \mathbb{R}\}$ of Example 3.1.2 where P_μ are given by the densities $p_\mu(x) = p(x-\mu)$ for the standard normal density $p(x)$. This family satisfies the condition of the formula (133) so that from (130) or (133) we obtain the $\min \mathfrak{D}_\alpha$-estimators $\mu_{\alpha,n} = T_\alpha(P_n)$ of location $\mu_0 \in \mathbb{R}$ in this family given by

$$\mu_{\alpha,n} = \mathrm{argmax}_\mu \begin{cases} \sum_{i=1}^n \exp\{-\alpha(X_i-\mu)^2/2\} & \text{if } \alpha > 0, \\ -\sum_{i=1}^n (X_i-\mu)^2 & \text{if } \alpha = 0. \end{cases} \tag{144}$$

Equivalently, they can be obtained by inserting $\sigma = 1$ in (143). If $\alpha = 0$ then $\mu_{\alpha,n}$ is the standard sample mean.

The estimators of location (144) were introduced and studied as part of larger class of estimators by Vajda [15, 16, 17]. He proved that if the observations are generated by $Q_{\mu_0} \in \mathcal{P}^+$ with density $q(x-\mu_0)$ for unimodal $q(x)$ symmetric about $x = 0$ then these estimators consistently estimate μ_0. For q differentiable with derivative q' he found the influence functions

$$\mathrm{IF}(x; T_\alpha, q) = \frac{x\exp\{-\alpha x^2/2\}}{\int x\exp\{-\alpha x^2/2\}\, q'(x)\,\mathrm{d}x} \quad \text{for } \alpha \geq 0. \tag{145}$$

This formula follows also from (139) and (140) where in this case

$$s_\mu(x) = x - \mu\Pi_{\alpha,\mu} = p_\mu^\alpha \left[\alpha\,(x-\mu)^2 - 1\right] \quad \text{and} \quad P_\mu \cdot p_\mu^\alpha s_\mu = 0. \tag{146}$$

Indeed, (146) implies $P_\mu \cdot p_\mu^\alpha s_\mu = 0$ and $p_0^\alpha(x)s_0(x) = x\exp\{-\alpha x^2/2\}.\,(2\pi)^{-\alpha/2}$ so that the numerator in (145) follows from (140). Using the identities

$$P_\mu \cdot \left(\Pi_{\alpha,\mu} + p_\mu^\alpha s_\mu^2\right) = \int p_\mu^{1+\alpha}\,\left[(1+\alpha)\,(x-\mu)^2 - 1\right]\,\mathrm{d}x = 0$$

and

$$\int x\,p_0(x)\,q'(x)\mathrm{d}x + \int \left[p_0(x) + xp_0'(x)\right]\,q(x)\,\mathrm{d}x = 0$$

we get from (146) and (138)

$$\boldsymbol{I}_\alpha(q) = (2\pi)^{-\alpha/2}\int x\exp\{-\alpha x^2/2\}\,q'(x)\,\mathrm{d}x$$

so that the denominator in (145) follows from (140).
The particular influence curve obtained in (145) for $\alpha = 1/5$ very closely and smoothly approximates the trapezoidal $\mathrm{IF}(x; 25A, q)$ of the estimator referred as the best under the name **Hampel's choice 25A** in the *Princeton Robustness Study* of Andrews et al. [1]. This study as well as the estimator of location $25A$ were influential and frequently cited in the first decades of robust statistics. The asymptotic normality

$$\sqrt{n}(\mu_{\alpha,n} - \mu_0) \longrightarrow N(0,\sigma_\alpha^2) \quad \text{for} \quad \sigma_\alpha^2 = \int \mathrm{IF}^2(x; T_\alpha, q)q(x)\mathrm{d}x$$

in the data generating model Q_{μ_0} was established in Vajda [15, 16, 17] too, and the simulations presented there demonstrated that the estimator $T_{1/5}$ overperformed the set of 6 robust estimators of location including those considered as the most prominent at that time.

Example 4.2.2. Power pseudodistance estimators of scale. Consider the normal family $\mathcal{P} = \{P_\sigma : \sigma > 0\}$ of Example 3.1.3 where P_σ are given by the densities $p_\sigma(x) = p(x/\sigma)/\sigma$ for the standard normal density $p(x)$. If $\alpha = 0$ then, by (132), the $\min \mathfrak{D}_\alpha$-estimator $\sigma_{\alpha,n} = T_\alpha(P_n)$ is the standard MLE of scale given in (62). Otherwise we get from (143) by inserting $\mu = 0$

$$\sigma_{\alpha,n} = \mathrm{argmax}_\sigma \frac{1}{\sigma^\alpha\; n}\sum_{i=1}^n \left[\exp\left\{-\frac{\alpha {X_i}^2}{2\sigma^2}\right\} - \frac{\alpha}{(1+\alpha)^{3/2}}\right], \quad \alpha > 0. \tag{147}$$

Taking into account here

$$\frac{1}{n}\sum_{i=1}^n \exp\left\{-\frac{\alpha {X_i}^2}{2\sigma^2}\right\} = \int \exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}\mathrm{d}P_n(x)$$

we find more general formula

$$T_\alpha(Q) = \operatorname{argmin}_\sigma M_\alpha(Q,\sigma) \quad \text{for} \quad Q \in \mathcal{P}^+,$$

where

$$M_\alpha(Q,\sigma) = \frac{1}{\sigma^\alpha}\int \exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\} \mathrm{d}Q(x) - \frac{\alpha}{(1+\alpha)^{3/2}}.$$

Referring to Broniatowski and Vajda [6] for details, it can be seen that for all $\sigma > 0$

$$\mathrm{IF}(x;T_\alpha,\sigma) = \frac{(1+\alpha)^{5/2}\sigma}{\alpha^2+2}\left[\left(\left(\frac{x}{\sigma}\right)^2 - 1\right)\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\} + \frac{\alpha}{(1+\alpha)^{3/2}}\right] \tag{148}$$

from which we conclude that the $\min \mathfrak{D}_\alpha$-estimators $\sigma_{\alpha,n} = T_\alpha(P_n)$ of normal scale are for all $\alpha > 0$ robust in the sense that their absolute sensitivity to the observations $x \in \mathbb{R}$ represented by

$$\sup_{x\in\mathbb{R}} |\mathrm{IF}(x;T_\alpha,\sigma)| = \max\{-\mathrm{IF}(0;T_\alpha,\sigma),\ \mathrm{IF}(\sigma_\alpha;T_\alpha,\sigma)\} \quad \text{for } \sigma_\alpha = \sigma\sqrt{\frac{2+\alpha}{\alpha}}$$

is bounded (cf. Hampel et al. [7]). However, they are not insensitive against extreme outliers because

$$\lim_{|x|\to\infty} \mathrm{IF}(x;T_\alpha,\sigma) = \mathrm{IF}(\sigma;T_\alpha,\sigma) = \frac{\alpha(1+\alpha)\sigma}{\alpha^2+2}. \tag{149}$$

4.3. Rényi pseudodistance estimators

In this subsection we propose for probability measures $P \in \mathcal{P}$ and $Q \in \mathcal{P}^+$ considered in the previous sections a family of pseudodistances $\mathfrak{R}_\alpha(P,Q)$ of a Rényi type of orders $\alpha \geq 0$ which are not of the integral type as $\mathfrak{D}_\psi(P,Q)$ of (102) or $\mathfrak{D}_\alpha(P,Q)$ of (124). Our proposal is based on the following theorem where

$$\mathfrak{R}^0_\alpha(P) = \frac{1}{1+\alpha}\ln(P\cdot p^\alpha) \quad \text{and} \quad \mathfrak{R}^1_\alpha(Q) = \frac{1}{\alpha(1+\alpha)}\ln(Q\cdot q^\alpha). \tag{150}$$

See Broniatowski, Toma and Vajda [5] for statistical properties of minimum Rényi pseudodistance estimators.

Theorem 4.3.1. Let the condition (128) hold for some $\beta > 0$. Then for all $0 < \alpha < \beta$

$$\mathfrak{R}_\alpha(P,Q) = \frac{1}{1+\alpha}\ln(P\cdot p^\alpha) + \frac{1}{\alpha(1+\alpha)}\ln(Q\cdot q^\alpha) - \frac{1}{\alpha}\ln(Q\cdot p^\alpha) \tag{151}$$

is a family of pseudodistances decomposable in the sense

$$\mathfrak{R}_\alpha(P,Q) = \mathfrak{R}^0_\alpha(P) + \mathfrak{R}^1_\alpha(Q) - \frac{1}{\alpha}\ln(Q\cdot p^\alpha) \tag{152}$$

for $\mathfrak{R}^0_\alpha(P), \mathfrak{R}^1_\alpha(Q)$ given by (150), and satisfying the limit relation

$$\mathfrak{R}_\alpha(P,Q) \to \mathfrak{R}_0(P,Q) := Q\ln q - Q\ln p \quad \text{for } \alpha \downarrow 0. \tag{153}$$

P r o o f. Under (128), the expressions $\ln(Q \cdot q^\alpha)$, $\ln(Q \cdot p^\alpha)$ and $Q \cdot \ln p$ appearing in (151) are finite so that the expressions $\mathfrak{R}_\alpha(P,Q)$ are well defined by (151). Taking $\alpha > 0$ and substituting

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

in the inequality (122), and integrating both sides by λ, 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. Since the expression (151) satisfies for $\alpha > 0$ the relation

$$\mathfrak{R}_\alpha(P,Q) = \frac{1}{\alpha}\left\{\ln\left[\left(\int p^{1+\alpha}\,\mathrm{d}\lambda\right)^{\alpha/(1+\alpha)}\left(\int q^{1+\alpha}\,\mathrm{d}\lambda\right)^{1/(1+\alpha)}\right] - \ln\int p^\alpha q\,\mathrm{d}\lambda\right\}, \tag{154}$$

we see that $\mathfrak{R}_\alpha(P,Q)$ is pseudodistance on the space $\mathcal{P} \otimes \mathcal{P}^+$. The decomposability in the sense of (152) on this space is obvious and the limit relation

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

can be proved in a similar manner as in the proof of Theorem 4.1.3 □

There is some similarity between the decomposable pseudodistances $\mathfrak{R}_\alpha(P,Q)$, $\alpha > 0$ of (151) and the Rényi divergences

$$R_\alpha(P,Q) = \frac{1}{\alpha-1}\ln\left(Q\cdot(p/q)^\alpha\right), \alpha > 0 \quad \text{(cf. Rényi [12]).}$$

Namely, rewriting the formula (154) into the form

$$\mathfrak{R}_\alpha(P,Q) = \frac{1}{\alpha+1}\ln\frac{Q\cdot\left(p^{1+\alpha}/q\right)}{Q\cdot p^\alpha} + \frac{1}{\alpha(\alpha+1)}\ln\frac{Q\cdot q^\alpha}{Q\cdot p^\alpha}$$

and replacing the ratios of expectations by the expectations of ratios, we get for $\alpha > 0$ the relation

$$\mathfrak{R}_\alpha(P,Q) = \frac{1}{\alpha+1}\ln(Q\cdot(p/q)) + \frac{1}{\alpha(\alpha+1)}\ln(Q\cdot(q/p)^\alpha) = \frac{1}{\alpha+1}R_{\alpha+1}(Q,P) \tag{155}$$

which can be extended to $\alpha = 0$ by taking on both sides the limits for $\alpha \downarrow 0$. Therefore the decomposable pseudodistances (151) are modified Rényi divergences and as such, they are called ***Rényi pseudodistances***.

Similarly as earlier in this section, we are interested in the estimators obtained by replacing the hypothetical distribution P_{θ_0} in the $\mathfrak{R}_\alpha$-pseudodistances $\mathfrak{R}_\alpha(P_\theta, P_{\theta_0})$ by the empirical distribution P_n. In other words, we are interested in the family of ***Rényi***

pseudodistance estimators of orders $0 \le \alpha \le \beta$ (in symbols, $\min \mathfrak{R}_\alpha$-estimators) defined as $\theta_{n,\alpha} = T_\alpha(P_n)$ for $T_\alpha(Q) \in \Theta$ with $Q \in \mathcal{Q} = \mathcal{P}^+ \cup \mathcal{P}_{\text{emp}}$ satisfying the condition

$$T_\alpha(Q) = \begin{cases} \arg\min_\theta \frac{1}{1+\alpha} \ln\left(P_\theta \cdot p_\theta^\alpha\right) - \frac{1}{\alpha} \ln(Q \cdot p_\theta^\alpha) & \text{if } 0 < \alpha \le \beta \\ \arg\min_\theta - \ln Q \cdot p_\theta & \text{if } \alpha = 0. \end{cases} \tag{156}$$

The upper formula is for

$$C_\theta(\alpha) = (P_\theta \cdot p_\theta^\alpha)^{\alpha/(1+\alpha)} \equiv \left(\int p_\theta^{1+\alpha} \mathrm{d}\lambda\right)^{\alpha/(1+\alpha)} \tag{157}$$

equivalent to

$$T_\alpha(Q) = \arg\max_\theta M_\alpha(Q,\theta) \quad \text{for} \quad M_\alpha(Q,\theta) = \frac{Q \cdot p_\theta^\alpha}{C_\theta(\alpha)}. \tag{158}$$

Alternatively, we can write

$$\theta_{n,\alpha} = \begin{cases} \arg\max_\theta C_\theta(\alpha)^{-1} \frac{1}{n} \sum_{i=1}^n p_\theta^\alpha(X_i) & \text{if } 0 < \alpha \le \beta \\ \arg\max_\theta \frac{1}{n} \sum_{i=1}^n \ln p_\theta(X_i) & \text{if } \alpha = 0. \end{cases} \tag{159}$$

For $\alpha \approx 0 \downarrow 0$ the approximations $C_\theta(\alpha) \approx 1$ and

$$\frac{1}{\alpha}\left(\frac{1}{n}\sum_{i=1}^n p_\theta^\alpha(X_i) - 1\right) = \frac{1}{n}\sum_{i=1}^n \frac{p_\theta^\alpha(X_i) - 1}{\alpha} \approx \frac{1}{n}\sum_{i=1}^n \ln p_\theta^\alpha(X_i)$$

indicate that the upper criterion function in (159) tends to the lower MLE criterion for $\alpha \downarrow 0$. If $C_\theta(\alpha)$ does not depend on θ then the $\min \mathfrak{R}_\alpha$-estimates reduce to the min $\mathfrak{D}_\alpha$-estimates considered in (133) of Remark 4.1.2, i.e.,

$$\theta_{\alpha,n} = \operatorname{argmax}_\theta \begin{cases} \frac{1}{n} \sum_{i=1}^n p_\theta^\alpha(X_i) & \text{if } 0 < \alpha < \beta, \\ \frac{1}{n} \sum_{i=1}^n \ln p_\theta(X_i) & \text{if } \alpha = 0. \end{cases} \tag{160}$$

If the extremal points of all functions in (159) are in a compact set of Θ then

$$\lim_{\alpha \downarrow 0} \theta_{n,\alpha} = \theta_{n,0}. \tag{161}$$

In the next theorem and its proof we use the auxiliary expressions

$$s_\theta = \frac{\mathrm{d}}{\mathrm{d}\theta} \ln p_\theta, \quad \mathring{s}_\theta = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} s_\theta \quad \text{(cf. (54))}$$

and

$$c_\theta(\alpha) = \frac{\int p_\theta^{1+\alpha} s_\theta \,\mathrm{d}\lambda}{\int p_\theta^{1+\alpha} \,\mathrm{d}\lambda}, \quad \mathring{c}_\theta(\alpha) = \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} c_\theta(\alpha) \quad \text{and} \quad \tau_\alpha = T_\alpha(Q).$$

Theorem 4.3.2. If the influence function (21) at $Q \in \mathcal{P}^+$ or $P_\theta \in \mathcal{P}$ exists for some $\min \mathfrak{R}_\alpha$-estimator $\theta_{\alpha,n} = T_\alpha(P_n)$ then it is given by the formula

$$\mathrm{IF}(x; T_\alpha, Q) = -\boldsymbol{I}_\alpha(Q)^{-1} \left[p_{\tau_\alpha}(x) \left(s_{\tau_\alpha}(x) - c_{\tau_\alpha}(\alpha)\right)\right] \tag{162}$$

or

$$\mathrm{IF}(x; T_\alpha, \theta) = -\boldsymbol{I}_\alpha(\theta)^{-1} \left[p_\theta(x) \left(s_\theta(x) - c_\theta(\alpha)\right)\right] \tag{163}$$

for the matrices

$$\boldsymbol{I}_\alpha(Q) = \int \left[\mathring{s}_{\tau_\alpha} - \mathring{c}_{\tau_\alpha}(\alpha) - \alpha p_{\tau_\alpha}^\alpha \left(s_{\tau_\alpha} - c_{\tau_\alpha}(\alpha)\right) \left(s_{\tau_\alpha} - c_{\tau_\alpha}(\alpha)\right)^{\mathrm{t}}\right] p_{\tau_\alpha}^\alpha \mathrm{d}Q \tag{164}$$

or

$$\boldsymbol{I}_\alpha(\theta) = \int \left[\mathring{s}_\theta - \mathring{c}_\theta(\alpha) - \alpha p_\theta^\alpha \left(s_\theta - c_\theta(\alpha)\right) \left(s_\theta - c_\theta(\alpha)\right)^{\mathrm{t}}\right] p_\theta^{1+\alpha} \mathrm{d}\lambda \tag{165}$$

respectively.

P r o o f. By (158), $T_\alpha(Q)$ for $Q \in \mathcal{Q}$ minimizes $Q \cdot (p_\theta^\alpha / C_\theta(\alpha))$, i. e. solves the equation $\Psi_\alpha(Q, \theta) \equiv Q \cdot \boldsymbol{\psi}(x, \theta) = 0$ for

$$\boldsymbol{\psi}_\alpha(x, \theta) \equiv \Psi_\alpha(\delta_x, \theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{p_\theta^\alpha}{C_\theta(\alpha)} = \frac{\alpha p_\theta^\alpha \left(s_\theta - c_\theta(\alpha)\right)}{C_\theta(\alpha)}. \tag{166}$$

Further,

$$C_\theta(\alpha) := \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} C_\theta(\alpha) = \alpha C_\theta(\alpha) c_\theta^{\mathrm{t}}(\alpha)$$

so that

$$\begin{aligned}
\mathring{\psi}_\alpha(x, \theta) &= \left(\frac{\mathrm{d}}{\mathrm{d}\theta}\right)^{\mathrm{t}} \boldsymbol{\psi}_\alpha(x, \theta) \\
&= \frac{C_\theta(\alpha) \left[\alpha^2 p_\theta^\alpha s_\theta^{\mathrm{t}} \left(s_\theta - c_\theta(\alpha)\right) + \alpha p_\theta^\alpha \left(\mathring{s}_\theta - \mathring{c}_\theta(\alpha)\right)\right] - \alpha p_\theta^\alpha \left(s_\theta - c_\theta(\alpha)\right) C_\theta(\alpha)}{C_\theta(\alpha)} \\
&= \frac{\alpha^2 p_\theta^\alpha s_\theta^{\mathrm{t}} \left(s_\theta - c_\theta(\alpha)\right) + \alpha p_\theta^\alpha \left(\mathring{s}_\theta - \mathring{c}_\theta(\alpha)\right) - \alpha^2 p_\theta^\alpha s_\theta^{\mathrm{t}} \left(s_\theta - c_\theta(\alpha)\right) c_\theta^{\mathrm{t}}(\alpha)}{C_\theta(\alpha)}.
\end{aligned}$$

Therefore the matrix (24) is given for all $Q \in \mathcal{P}^+$ by the formula (164) and (27) is given for $P_\theta \in \mathcal{P}$ by (165). The rest is clear from Theorems 2.1 and 4.1, and from Corollary 2.1. □

4.4. Applications in the normal family

Consider the general normal family of Example 3.1.1 for which the condition (128) is satisfied for all $\beta > 0$ and (142) implies

$$C_{\mu,\sigma}(\alpha) = C_\sigma(\alpha) = \left(\frac{(1+\alpha)^{-1/2}}{(2\pi\sigma^2)^{\alpha/2}}\right)^{\alpha/(1+\alpha)} = \frac{\sigma^{-\alpha^2/(1+\alpha)}}{c(\alpha)} \tag{167}$$

for all $\mu \in \mathbb{R}$ and the function

$$c(\alpha) = [(1+\alpha)(2\pi)^{\alpha}]^{\alpha/2(1+\alpha)}, \alpha > 0.$$

By (159), the $\min \mathfrak{R}_\alpha$-estimator $\theta_{\alpha,n} = (\mu_{\alpha,n}, \sigma_{\alpha,n})$ is the standard estimator of location and scale given by (62) if $\alpha = 0$. For $\alpha > 0$ we can use the relation

$$\frac{\sigma^{\alpha^2/(1+\alpha)}}{\sigma^\alpha} = \sigma^{-\alpha/(1+\alpha)}$$

to get from (159) and (167) the highly nonstandard estimator

$$(\mu_{\alpha,n}, \sigma_{\alpha,n}) = \operatorname{argmax}_{\mu,\sigma} \left[\frac{c_\alpha}{n\sigma^{\alpha/(1+\alpha)}} \sum_{i=1}^{n} \exp\left\{ -\alpha \frac{(X_i - \mu)^2}{2\sigma^2} \right\} \right] \tag{168}$$

which in general differs from the $\min \mathfrak{D}_\alpha$-estimator (143) as it will be seen in the submodel of scale below. Similarly as in the case of power pseudodistance estimator (143), the trivial "solutions" $(\mu_{\alpha,n}, \sigma_{\alpha,n}) = (\max_i X_i, 0)$ can be avoided in practical applications by restricting the maximization to the scales bouded avay from zero.
The next example of the submodel of location illustrates the situation where these two estimators coincide. Obviously, the constants $c_\alpha = c(\alpha)/(2\pi)^{\alpha/2}$ play no role in the maximization and can be replaced by 1.

Example 4.4.1. Rényi pseudodistance estimators of location. The normal family of location introduced in Example 3.1.2 satisfies the condition of the formula (133) so that from (130) or (133) we obtain the same $\min \mathfrak{R}_\alpha$-estimators $\mu_{\alpha,n}$ of location $\mu_0 \in \mathbb{R}$ as in (144). Thus to these estimators applies all what was seen in Example 4.2.1.

Example 4.4.2. Rényi pseudodistance estimators of scale. Consider the normal model of scale introduced in Example 3.1.3. If $\alpha = 0$ then, by (132), the $\min \mathfrak{R}_\alpha$-estimator $\sigma_{\alpha,n} = T_\alpha(P_n)$ is the standard MLE of scale given in (62). Otherwise by (168),

$$\sigma_{\alpha,n} = \operatorname{argmax}_{\sigma} \left[\frac{c_\alpha}{n\sigma^{\alpha/(1+\alpha)}} \sum_{i=1}^{n} \exp\left\{ -\alpha \frac{{X_i}^2}{2\sigma^2} \right\} \right], \quad \alpha > 0 \quad \text{(cf. (168))}. \tag{169}$$

It is easy to see e. g. by putting $n = 1$ and $\alpha X^2 = 2$ that these estimates differ from the $\mathfrak{D}_\alpha$-estimates of scale given in(147). Here (158) for the Dirac δ_x implies

$$M_\alpha(\delta_x, \sigma) = \frac{p_\sigma^\alpha(x)}{C_\sigma(\alpha)} = \frac{c_\alpha}{\sigma^{\alpha/(1+\alpha)}} \exp\left\{ -\frac{\alpha x^2}{2\sigma^2} \right\}$$

and by (20) and (22),

$$\psi_\alpha(x,\sigma) = \frac{\mathrm{d}}{\mathrm{d}\sigma} M_\alpha(\delta_x,\sigma) = c_\alpha \frac{\mathrm{d}}{\mathrm{d}\sigma}\left[\sigma^{-\alpha/(1+\alpha)} \exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}\right]$$

$$= \frac{c_\alpha}{\sigma^{\alpha/(1+\alpha)}}\left[\frac{\alpha x^2}{\sigma^3} - \frac{\alpha}{1+\alpha}\frac{1}{\sigma}\right]\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}$$

$$= \frac{\alpha c_\alpha}{\sigma^{1+\alpha/(1+\alpha)}}\left[\left(\frac{x}{\sigma}\right)^2 - \frac{1}{1+\alpha}\right]\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}. \tag{170}$$

This formula can be verified by checking the Fisher consistency known in general from Theorem 4.1. One can find that

$$\int\left[\left(\frac{x}{\sigma}\right)^2 - \frac{1}{1+\alpha}\right]\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}p_{\sigma_0}(x)\mathrm{d}x$$
$$= \frac{\sigma}{\sqrt{\sigma^2+\alpha\sigma_0^2}}\left[\left(\frac{\sigma_0^2}{\sigma^2+\alpha\sigma_0^2}\right)^2 - \frac{1}{1+\alpha}\right].$$

Since the right-hand side is zero if and only if $\sigma = \sigma_0$, the verification is positive. From (170) we evaluate after some effort the derivative

$$\mathring{\psi}_\alpha(x,\sigma) = \frac{\mathrm{d}}{\mathrm{d}\sigma}\psi_\alpha(x,\sigma) = \frac{\mathrm{d}}{\mathrm{d}\sigma}\frac{c_\alpha}{\sigma^{1+\alpha/(1+\alpha)}}\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}\left[\alpha\left(\frac{x}{\sigma}\right)^2 - \frac{\alpha}{1+\alpha}\right]$$

$$= \frac{\alpha c_\alpha}{\sigma^{2+\alpha/(1+\alpha)}}\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}\eta_\alpha\left(\frac{x}{\sigma}\right), \tag{171}$$

where

$$\eta_\alpha\left(\frac{x}{\sigma}\right) = \left[\alpha\left(\frac{x}{\sigma}\right)^4 - \frac{5\alpha+3}{1+\alpha}\left(\frac{x}{\sigma}\right)^2 + \frac{2\alpha+1}{(1+\alpha)^2}\right].$$

Thus, denoting for brevity

$$\tau_\alpha = T_\alpha(Q) \quad \text{for} \quad Q \in \mathcal{P}^+$$

we obtain from (170), (171) and Theorem 2.1 the influence functions of the $\min \mathfrak{D}_\alpha$-estimators $\sigma_{\alpha,n} = T_\alpha(P_n)$ at Q given for all $\alpha > 0$ by

$$\mathrm{IF}(x;T_\alpha,Q) = -\frac{\psi_\alpha(x,\tau_\alpha)}{\int \mathring{\psi}_\alpha(x,\tau_\alpha)\mathrm{d}Q}$$
$$= -\frac{\alpha}{\Upsilon_\alpha(Q)}\left[\left(\left(\frac{x}{\tau_\alpha}\right)^2 - \frac{1}{1+\alpha}\right)\exp\left\{-\frac{\alpha x^2}{2\tau_\alpha^2}\right\}\right], \tag{172}$$

where

$$\Upsilon_\alpha(Q) = \int \eta_\alpha\left(\frac{x}{\sigma}\right)\exp\left\{-\frac{\alpha x^2}{2\tau_\alpha^2}\right\}\mathrm{d}Q.$$

In the special case $Q = P_\sigma$ the Fisher consistency implies that $\tau_\alpha := T_\alpha(P_\sigma) = \sigma$. We use the relation

$$\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\} p_\sigma(x) = p_{\sigma_\alpha}(x)\frac{1}{\sqrt{1+\alpha}} \quad \text{for} \quad \sigma_\alpha = \frac{\sigma}{\sqrt{1+\alpha}}$$

to obtain

$$\begin{aligned}\Upsilon_\alpha(P_\sigma) &= \frac{1}{\sqrt{1+\alpha}}\int \eta_\alpha\left(\frac{x}{\sigma}\right)p_{\sigma_\alpha}(x)\mathrm{d}x \\ &= \frac{1}{(1+\alpha)^{1/2}}\left[\alpha\left(\frac{\sigma_\alpha}{\sigma}\right)^4 - \frac{5\alpha+3}{1+\alpha}\left(\frac{\sigma_\alpha}{\sigma}\right)^2 + \frac{2\alpha+1}{(1+\alpha)^2}\right] \\ &= \frac{1}{(1+\alpha)^{5/2}}\left[3\alpha - (5\alpha+3) + 2\alpha + 1\right] = -\frac{2}{(1+\alpha)^{5/2}}\end{aligned}$$

independently of $\sigma > 0$. Therefore at the normal location P_σ we get for all $\sigma > 0$ the influence functions

$$\mathrm{IF}(x; T_\alpha, P_\sigma) = \frac{(1+\alpha)^{5/2}\sigma}{2}\left[\left(\left(\frac{x}{\sigma}\right)^2 - \frac{1}{1+\alpha}\right)\exp\left\{-\frac{\alpha x^2}{2\sigma^2}\right\}\right]. \tag{173}$$

It is easy to verify that this is the influence function also in the MLE case $\alpha = 0$.

In conclusion we see that the $\min \mathfrak{R}_\alpha$-estimators $\sigma_{\alpha,n} = T_\alpha(P_n)$ of normal scale are for all $\alpha > 0$ robust in the sense that their influence functions are bounded. They are more robust against distant outliers than the corresponding $\min \mathfrak{D}_\alpha$-estimators studied in the Subsections 4.1 and 4.2 because their influence function is ***redescending***

$$\lim_{|x|\to\infty} \mathrm{IF}(x; T_\alpha, P_\sigma) = 0 \quad \text{(cf. (172))}. \tag{174}$$

ACKNOWLEDGEMENT

This research was supported by the grants GA ČR 102/07/1131 and Ministry of Education, Youth and Sports of the Czech Republic 1M 0572. The authors thank the PhD student Iva Frýdlová for careful reading and corrections of many previous versions of the first two sections. They thank also to the MSc student Radim Demut for simulations of the Rényi estimators in contaminated families. The very promising results obtained by him encouraged the theoretic research presented here.

The first author expresses his profound sorrow for the loss of an unvaluable colleague and friend.

(Received October 8, 2010)

REFERENCES

[1] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey: Robust Estimates of Location. Princeton University Press, Princeton N. J. 1972.

[2] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones: Robust and efficient estimation by minimizing a density power divergence. Biometrika *85* (1998), 3, 549–559.

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

[4] M. Broniatowski and A. Keziou: Parametric estimation and tests through divergences and the duality technique. J. Multivariate Anal. *100* (2009), 1, 16–31.

[5] M. Broniatowski, A. Toma, and I. Vajda: Decomposable pseudodistances and applications in statistical estimation. J. Statist. Plann. Inference. *142* (2012), 9, 2574-2585

[6] M. Broniatowski and I. Vajda: Several applications of divergence criteria in continuous families. arXiv:0911.0937v1, 2009.

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

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

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

[10] C. Miescke and F. Liese: Statistical Decision Theory. Springer, Berlin 2008.

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

[12] A. Rényi: On measures of entropy and information. In: Proc. 4th Berkeley Symp. on Probability and Statistics, Vol. 1, University of California Press, Berkeley 1961, pp. 547–561.

[13] A. Toma and M. Broniatowski: Minimum divergence estimators and tests: Robustness results. J. Multivariate Anal. *102* (2011), 1, 20–36.

[14] I. Vajda: Minimum divergence principle in statistical estimation. Statist. Decisions (1984), Suppl. Issue No. 1, 239–261.

[15] I. Vajda: Efficiency and robustness control via distorted maximum likelihood estimation. Kybernetika *22* (1986), 47–67.

[16] I. Vajda: Comparison of asymptotic variances for several estimators of location. Probl. Control Inform. Theory *18* (1989), 2, 79–89.

[17] I. Vajda: Estimators asymptotically minimax in wide sense. Biometr. J. *31* (1989), 7, 803–810.

[18] I. Vajda: Modifications od Divergence Criteria for Applications in Continuous Families. Research Report No. 2230, Institute of Information Theory and Automation, Prague 2008.

[19] A. W. van der Vaart: Asymptotic Statistics. Cambridge University Press, Cambridge 1998.

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

Michel Broniatowski, LSTA, Université Paris 6. France.
e-mail: michel.broniatowski@upmc.fr