KYBERNETIKA — VOLUME 20 (1984), NUMBER 3

MOTIVATION, EXISTENCE AND EQUIVARIANCE OF *D*-ESTIMATORS

IGOR VAJDA

This is the first in a series of papers on *D*-estimators to be published in Kybernetika. *D*-estimators are minimizing f-divergence or properly modified f-divergence between theoretical and empirical probability. Suitable specifications of convex functions f yield either new promising estimators, or well-known estimators such as the *MLE*, or *M*-estimators, or various minimum distance estimators, motivated so far quite diversely if motivated at all. The theory of *D*-estimators can be considered as an alternative to the loss-function-based theory in a systematic development of asymptotic as well as non-asymptotic properties of wide classes of estimators. The present paper is devoted to motivation and examples of *D*-estimators and to non-asymptotic aspects of the theory such as existence, measurability, continuity, invariance, and equivariance of *D*-estimators.

1. PRELIMINARIES

In this paper $\mathscr{X}$ denotes a Hausdorff topological space separable in the sense that it contains an at most countable dense subset. $\mathscr{B}$ denotes the Borel σ-algebra of subsets of $\mathscr{X}$. Thus if, in particular, $\mathscr{X}$ is a discrete space then it is finite or countable and $\mathscr{B} = \exp \mathscr{X}$. An identity mapping $\mathscr{X} \to \mathscr{X}$ is denoted by X, with additional indices when convenient.

$\mathscr{P}$ denotes the family of all probabilities on $(\mathscr{X}, \mathscr{B})$, $\mathscr{P}_e$ a subfamily of empirical probabilities P_n corresponding to the sample vectors $x = (x_1, \ldots, x_n) \in \mathscr{X}^n$ as follows

$$P_n(E) = \frac{1}{n} \sum_{i=1}^{n} 1_E(x_i) \quad \text{for} \quad E \in \mathscr{B} . \tag{1.1}$$

Obviously P_n is a mixture of empirical probabilities $1_{\{x_i\}}$ with a total mass concentrated at $x_i \in \mathscr{X}$ (i.e. $1_{\{x_i\}}(E) = 1_E(x_i)$, $E \in \mathscr{B}$) where the mixture is taken with a uniform weight distribution n^{-1}. For a normed $\mathscr{X}$ we define subfamilies $\mathscr{P}_\beta = \{P \in \mathscr{P} : \mathsf{E}_P \|X\|^\beta < \infty\}$, $\beta \geqq 0$.

By $\mathscr{P}_\Theta$ we denote parametric subfamilies $\{P_\theta : \theta \in \Theta\}$ of $\mathscr{P}$ where it is supposed

that $P_\theta \neq P_{\tilde\theta}$ for $\theta \neq \tilde\theta \in \Theta$ and that Θ is a Hausdorff topological space. Measurability on Θ is considered w.r.t. the Borel σ-algebra on Θ. The class $\exp \Theta$ is considered to be a topological space with open sets $\exp U$ for open $U \subset \Theta$. In this topology all "disjoint points" of the subspace $C(\Theta)$ of compact subsets of Θ can be separated by disjoint open neighbourhoods of these points. We restrict ourselves to parameter spaces satisfying an axiom of continuous choice: There exists a continuous mapping $\tau : C(\Theta) \to \Theta$ with $\tau(\Theta^*) \in \Theta^*$ for all $\Theta^* \in C(\Theta)$. For example, let Θ be a topological subspace of a finite product of ordered spaces with topologies induced by the respective orders (e.g. $\Theta \subset \mathbb{R}^m$) and let us consider a lexicographical order in Θ, i.e. the order defined first by the first-coordinate order, then by the second-coordinate order, etc. Then $\tau(\Theta^*) = \min \Theta^*$ is a continuous rule of choice on $C(\Theta)$.

Θ is said σ-compact if it can be covered by countably many compact sets $\Theta^{(s)} \subset \Theta$. For any σ-compact Θ we consider compact subsets

$$\Theta_j = \bigcup_{s=1}^{j} \Theta^{(s)} \tag{1.2}$$

tending to Θ as $j \to \infty$ in the usual set-theoretical sense.

Θ is said structural w.r.t. $\mathscr{X}$ if (i) Θ is a group with elements $\theta \in \Theta$ representing bijections $[\theta] : \mathscr{X} \to \mathscr{X}$ such that $[\theta]\,\mathscr{B} = \mathscr{B}$, (ii) $[\Theta] = \{[\theta] : \theta \in \Theta\}$ is a group with $[\theta] \,.\, [\tilde\theta] = [\theta]([\tilde\theta])$, (iii) the representation $\theta \to [\theta]$ is a homomorphism between groups Θ and $[\Theta]$, i.e. $[\theta\tilde\theta] = [\theta] \,.\, [\tilde\theta]$, and (iv) the mapping sending $(\theta, \tilde\theta) \in \Theta^2$ into $\theta^{-1}\tilde\theta \in \Theta$ is continuous. Here and in the sequel we consider product topologies on products of topological spaces. Since, by (iii), $[\theta]^{-1} = [\theta^{-1}]$, all bijections $[\theta]$ are $\mathscr{B}$-measurable. In what follows we exceptionally denote by $[\theta]$ $\mathscr{B}^n$-measurable mappings (statistics) defined by

$$[\theta](x) = ([\theta](x_1), \ldots, [\theta](x_n)), \quad x \in \mathscr{X}^n \tag{1.3}$$

as well. For structural Θ we restrict ourselves to parametric families $\mathscr{P}_\Theta = \{P_\theta = P[\theta]^{-1} : \theta \in \Theta\} \subset \mathscr{P}$ defined by parents $P \in \mathscr{P}$.

A well known example which we shall frequently refer to is a location-scale parameter space $\Theta = \mathbb{R} \times (0, \infty)$ which is structural w.r.t. the real line $\mathscr{X} = \mathbb{R}$ under bijections $[\theta](x) = [\mu, \sigma](x) = \mu + \sigma x$, $x \in \mathbb{R}$, and under the associative multiplication $\tilde\theta\theta = (\mu, \sigma)(\tilde\mu, \tilde\sigma) = (\mu + \sigma\tilde\mu, \sigma\tilde\sigma)$. The inverse element in $\mathbb{R} \times (0, \infty)$ is defined by $(\mu, \sigma)^{-1} = (-\mu/\sigma, 1/\sigma)$ and the unit element $e = (0, 1)$. The location or scale parameter spaces $\Theta = \mathbb{R}$ or $\Theta = (0, \infty)$ are obtained as subgroups $\mathbb{R} \times \{1\}$ or $\{0\} \times (0, \infty)$ of the group $\mathbb{R} \times (0, \infty)$ respectively.

A class $\mathscr{E} = \{E_x : x \in \mathscr{X}\}$ is said sufficient for $\mathscr{X}$ if $\mathscr{E} \subset \mathscr{B}$, $\mathscr{E}$ generates $\mathscr{B}$, and $P(E_x)$ are $\mathscr{B}$-measurable for all $P \in \mathscr{P}$. The intervals $(-\infty, x) : x \in \mathbb{R}$ are sufficient for $\mathbb{R}$ and their products for $\mathbb{R}^k$. Classes sufficient for product sample spaces of stochastic processes can be shown to exist as well. The $\mathscr{B}$-measurable functions $F(x) = P(E_x)$, $G(x) = Q(E_x)$, $F_\theta(x) = P_\theta(E_x)$, $F_n(x) = P_n(E_x)$, called simply distribution functions (d.f.'s), will be used throughout this paper as characteristics of the corres-

ponding probabilities P, Q, P_θ, P_n whenever a sufficient class $\mathscr{E}$ will be supposed on $\mathscr{X}$.

A family $\mathscr{P}_\Theta$ is said Θ-measurable if $P_\theta(E)$ are measurable functions of θ for all $E \in \mathscr{B}$. Analogically we define a Θ-continuous family $\mathscr{P}_\Theta$ as well.

By p, q, p_θ, p_n, w we denote Randon-Nikodym densities of P, Q, P_θ, P_n, W w.r.t. a dominating σ-finite measure λ on $(\mathscr{X}, \mathscr{B})$. Unless otherwise explicitly stated, on discrete $\mathscr{X}$ we consider the counting λ and on $\mathscr{X} = \mathbb{R}^k$ the Lebesgue λ. If $\mathscr{P}_\Theta$ dominated by a σ-finite λ is Θ-measurable then $p_\theta(x)$ is a measurable function of θ a.e. $[\lambda]$. This can be proved as follows. Since $\mathscr{X}$ is separable, there exists a net of finite or countable decompositions $\mathscr{D}^{(j)} : j = 1, 2, \ldots$ sufficient for $\mathscr{B}$ in the sense that the σ-algebras $\mathscr{B}^{(j)}$ are increasing and their union generates $\mathscr{B}$. Since λ is σ-finite, there exists at most countable decomposition $\mathscr{D} \subset \mathscr{B}$ of $\mathscr{X}$, $P^{(s)} \in \mathscr{P}$, and positive weights $w^{(s)}$ such that $\lambda = \sum w^{(s)} P^{(s)}$ and $P^{(s)}(E^{(s)}) = 1$ for some $E^{(s)} \in \mathscr{D}$. If $p_\theta^{(j,s)} = \mathsf{E}_{P^{(s)}}(p_\theta | \mathscr{B}^{(j)})$ on $E^{(s)}$ and 0 elsewhere then, by a theorem of Lévy (e.g. Theorem 2.8 in [12]), $p_\theta^{(j,s)} \to$ $\to p_\theta$ a.s. $[P^{(s)}]$ as $j \to \infty$. Define $p_\theta^{(j)}$ equal $p_\theta^{(j,s)}$ on $E^{(s)}$ for every s. Since the Θ-measurability of $\mathscr{P}_\Theta$ implies that all $p_\theta^{(j,s)}$ are measurable on Θ a.e. $[\lambda]$, $p_\theta^{(j)}$ is measurable on Θ a.e. $[\lambda]$. Since further

$$p_\theta^{(j)} \to p_\theta \quad \text{a.e.} \quad [\lambda] \quad \text{as} \quad j \to \infty\,, \tag{1.4}$$

p_θ is measurable on Θ a.e. $[\lambda]$ too. □

An estimator of a parameter from Θ is defined as a mapping T from a non-empty subset $\mathscr{P}(T) \subset \mathscr{P}$ into Θ. An estimator T is said *well-defined* if $\mathscr{P}_e \subset \mathscr{P}(T)$ and if $T(P_n)$ as a function of $x \in \mathscr{X}^n$ (cf. (1.1)) is $\mathscr{B}^n$-measurable. While the first condition is purely practical, the second one is unavoidable for any probabilistic theory of estimation since it enables to interpret $T(P_n)$ as a random variable defined on $(\mathscr{X}^n, \mathscr{B}^n, P^n)$, $P \in \mathscr{P}$. To facilitate the requirement of $\mathscr{B}^n$-measurability we accept in this paper the following convention.

Convention 1.1. If some criterion defines an estimate $T(Q)$ as a point from a non-empty set $\mathbb{T}(Q) \subset \Theta$, we assume $T(Q) = \tau(\mathbb{T}(Q))$ for an arbitrary fixed extension of the above considered mapping $\tau : C(\Theta) \to \Theta$ to the whole domain $\exp \Theta - \emptyset$. Moreover, on structural Θ we restrict ourselves to the rules of continuous choice homogeneous in the following sense

$$\tau(\theta\Theta^*) = \theta\,\tau(\Theta^*) \quad \text{for all} \quad \theta \in \Theta,\ \Theta^* \in C(\Theta)\,. \tag{1.5}$$

Notice that if the group multiplication by an arbitrary constant $\theta \in \Theta$ preserves the above considered lexicographical order on Θ (this takes place e.g. in the location and scale case $\Theta = \mathbb{R} \times (0, \infty) \subset \mathbb{R}^2$) then $\tau(\Theta^*) = \min \Theta^*$ is homogeneous in the sense of (1.5).

2. f-DIVERGENCE AND ITS MODIFICATIONS

Hereafter we denote by f a real valued function continuous and convex on $(0, \infty)$ which is strictly convex at 1 and $f(1) = 0$. As proved on pp. 76–77 in [23], the limits

$$f(0) = \lim_{u \downarrow 0} f(u), \quad 0f(\infty) = \lim_{u \uparrow \infty} \frac{f(u)}{u}$$

under these assumptions exist in the extended real line and their sum

$$\|f\| = f(0) + 0f(\infty) \tag{2.1}$$

is well-defined and positive. We say that f is semibounded if $\|f\| < \infty$.

In accordance with Csiszár [5, 6] we define an f-divergence of probabilities $P, Q \in \mathscr{P}$ by*)

$$D_f(P, Q) = \mathsf{E}_\lambda q f(p/q) = \mathsf{E}_Q f(p/q) \quad \text{for some} \quad \lambda \gg P, Q \tag{2.2}$$

where $0f(0/0) = 0$ and $0f(u/0) = u\,0f(\infty)$ for $u > 0$. By [6] the f-divergence is well-defined by (2.2) and it is independent of λ. We next list some properties of f-divergences for references later.

Lemma 2.1. $0 \leqq D_f(P, Q) \leqq \|f\|$ where the left equality holds iff $P = Q$ and the right equality holds if (for $\|f\| < \infty$ iff) $P \perp Q$.

Proof. The left inequality including the sufficient condition $P = Q$ for equality has been proved by Csiszár [5], the right inequality including the sufficient condition $P \perp Q$ for equality has been proved by Vajda [21]. Both conditions have also been shown necessary there for f strictly convex on $(0, \infty)$. For f under consideration the necessity of $P = Q$ easily follows from Lemma 1.1 in [6]. As to the necessity of $P \perp Q$, we refer to the proof given on pp. 89–90 in [23]. □

Example 2.1. The function $f(u) = u \ln u$ yields *I-divergence*

$$I(P, Q) = \mathsf{E}_Q(p/q) \ln (p/q) = \mathsf{E}_P \ln (p/q), \quad \|f\| = \infty \quad \text{(cf. [10])}. \tag{2.3}$$

The functions $f(u) = \operatorname{sign}(1 - \alpha)(1 - u^\alpha)/\alpha$, $\alpha \in [0, 1) \cup (1, \infty)$ with a limit $f(u) = -\ln u$ at $\alpha = 0$ yield *D^α-divergences*

$$D^\alpha(P, Q) = \begin{cases} \mathsf{E}_Q(-\ln (p/q)) = I(Q, P), & \|f\| = \infty, \quad \alpha = 0 \\ \alpha^{-1}(1 - \mathsf{E}_\lambda p^\alpha q^{1-\alpha}), & \|f\| = \alpha^{-1}, \quad \alpha \in (0, 1) \\ \alpha^{-1}(\mathsf{E}_\lambda p^\alpha q^{1-\alpha} - 1), & \|f\| = \infty, \quad \alpha \in (1, \infty). \end{cases} \tag{2.4}$$

*) The f-divergence has first been introduced by Cziszár in [5] and then, independently, by Ali and Slivey in J. Roy. Statist. Soc. Ser. B *28* (1966), 131–142. Perez in Kybernetika *3* (1967), 1–21, found the first statistical application of this concept – an upper bound for the Bayes risk increase caused by a reduction of σ-algebras figuring in a statistical decision problem. His Lemma 2.1 on p. 9 concerning a linearly constrained f-divergence minimum may be helpful in evaluating standard D-estimates defined in (3.1) below provided $\mathscr{P}_\theta$ satisfies the corresponding linear constraint.

The function $f(u) = |1 - u^\alpha|^{1/\alpha}$ for $\alpha \in (0, 1]$ and $f(u) = |1 - u|^\alpha$ for $\alpha \in (1, \infty)$ yield *χ^α-divergences*

$$(2.5) \quad \chi^\alpha(P, Q) = \begin{cases} \mathsf{E}_\lambda |p^\alpha - q^\alpha|^{1/\alpha}, & \|f\| = 2, \quad \alpha \in (0, 1] \quad \text{(cf. [3])} \\ \mathsf{E}_\lambda |p - q|^\alpha q^{1-\alpha}, & \|f\| = \infty, \quad \alpha \in (1, \infty) \quad \text{(cf. [22])}. \end{cases}$$

The I-divergence, the total variation χ^1, the χ^2-divergence, and the *Hellinger distance* $\chi^{1/2} = D^{1/2}$ are well-known in statistics.

The next lemma has been proved by Csiszár [5].

Lemma 2.2. If $P^{\mathscr{A}}$, $Q^{\mathscr{A}}$ are restrictions of P, Q on a sub-σ-algebra $\mathscr{A} \subset \mathscr{B}$, then $D_f(P^{\mathscr{A}}, Q^{\mathscr{A}}) \leqq D_f(P, Q)$ where the equality holds if $\mathscr{A}$ is sufficient for the family $\{P, Q\} \subset \mathscr{P}$.

Corollary 2.1. If $\mathscr{A} = \{\emptyset, E, \mathscr{X} - E, \mathscr{X}\} \subset \mathscr{B}$ then $D_f(P^{\mathscr{A}}, Q^{\mathscr{A}}) = d_f(P(E), Q(E)) \leqq D_f(P, Q)$, where

$$d_f(u, v) = v f\left(\frac{u}{v}\right) + (1 - v) f\left(\frac{1 - u}{1 - v}\right) \quad \text{for} \quad (u, v) \in [0, 1]^2$$

with the same conventions as those considered in (2.2).

The next three results can be found in Vajda [21].

Lemma 2.3. $D_f(P, Q) = \sup_{\mathscr{A}} D_f(P^{\mathscr{A}}, Q^{\mathscr{A}})$ provided the supremum extends over all algebras $\mathscr{A}$ generated by finite $\mathscr{B}$-measurable decompositions of $\mathscr{X}$.

Lemma 2.4. If $\{\mathscr{B}^{(j)} : j = 1, 2, \ldots\}$ is a sequence of sub-σ-algebras of $\mathscr{B}$ with $\mathscr{B}^{(j)} \subset \mathscr{B}^{(j+1)}$ and with $\mathscr{B}^{(1)} \cup \mathscr{B}^{(2)} \cup$ generating $\mathscr{B}$, and if $P^{(j)}$, $Q^{(j)}$ are restrictions of P, Q on $\mathscr{B}^{(j)}$, then $D_f(P^{(j)}, Q^{(j)})$ is a non-decreasing sequence with a limit $D_f(P, Q)$ as $j \to \infty$.

Lemma 2.5. The functions $f^*(u)$ or $f_*(u)$, equal $+\infty$ on $(2, \infty)$ and equal $u f((2 - u)/u)$ or $(2 - u) f(u/(2 - u))$ respectively on $[0, 2]$, are satisfying all conditions imposed on f in this paper and $D_f(P, Q) = D_{f^*}(Q, W) = D_{f_*}(P, W)$ for $W = (P + Q)/2 \in \mathscr{P}$.

Lemma 2.6. If $f^c(u) = u f(1/u) + c(u - 1)$, $c \in \mathbb{R}$, then f^c satisfies all conditions imposed on f in this paper and the f^c-divergence is conjugated to an f-divergence in the sense $D_{f^c}(P, Q) = D_f(Q, P)$ for every $P, Q \in \mathscr{P}$. Consequently, an f-divergence is symmetric if $f = f^c$ for some $c \in \mathbb{R}$.

Proof. For $c = 0$ this statement follows from (1.13) in [6]. Its extension to $c \neq 0$ is clear from (2.2). □

Lemma 2.7. If $\mathscr{P}_\Theta, Q_\Theta$ are Θ-measurable and W is a probability on Θ then $\mathsf{E}_W P_\theta$, $\mathsf{E}_W Q_\theta \in \mathscr{P}$ and $\mathsf{E}_W D_f(P_\theta, Q_\theta) \geqq D_f(\mathsf{E}_W P_\theta, \mathsf{E}_W Q_\theta)$.

Proof. It is easy to verify that $\Phi(u, v) = u f(u/v)$ is convex in the domain $(u, v) \in \in [0. \infty)^2 \subset \mathbb{R}^2$. The inequality of Lemma 2.7 then follows from Jensen's inequality. The rest is clear. □

In the next lemma $L_{\mathscr{F}}$ denotes a lower envelope of a class $\mathscr{F}$ of functions defined on $[0, 1]$. The lower envelope is defined on $[0, 1]$ by the condition that it is a convex function satisfying

$$L_{\mathscr{F}}(u) \leqq \inf_{f \in \mathscr{F}} f(u) \quad \text{for} \quad u \in [0, 1] \tag{2.6}$$

and dominating any other convex function satisfying (2.6).

Lemma 2.8. It holds

$$L_{\{f_0, f_1\}}\left(\frac{\chi^1(P, Q)}{2}\right) \leqq D_f(P, Q) \leqq \max\{f(0), 0 f(\infty)\} \cdot \chi^1(P, Q)$$

where $f_0(u) = (1 + u) f((1 - u)/(1 + u))$, $f_1(u) = (1 - u) f((1 + u)/(1 - u))$ for $u \in [0, 1]$. If $f_0 = f_1$ then $f(0) = 0 f(\infty) = \|f\|/2$ and the lower as well as the upper bound is attainable in the domain $\mathscr{P}$.

Proof. It is easy to verify that both f_i are convex on $[0, 1]$ so that

$$f_i(u) \leqq (1 - u) f_i(0) + u f_i(1) = \begin{cases} 2 f(0)\, u & \text{if} \quad i = 0 \\ 2\, 0 f(\infty)\, u & \text{if} \quad i = 1 . \end{cases}$$

Hence

$$f_i(u) \leqq 2 \max\{f(0), 0 f(\infty)\}\, u \quad \text{for} \quad u \in [0, 1] . \tag{2.7}$$

Further, by Lemma 2.5 and (2.2), (2.5),

$$\chi^1(P, Q) = 2\, \mathsf{E}_W |p_* - 1| , \quad D_f(P, Q) = \mathsf{E}_W f_*(p_*) , \quad p_* = \frac{\mathrm{d}P}{\mathrm{d}W} \in [0, 2]$$

where W, f_* are defined in Lemma 2.5. Define on $(\mathscr{X}, \mathscr{B}, W)$ r.v.'s

$$U = |p_* - 1| , \quad V = f_0(U)\, 1_{\{p_* < 1\}} + f_1(U)\, 1_{\{p_* \geqq 1\}} .$$

It holds

$$\mathsf{E}_W U = \frac{\chi^1(P, Q)}{2} , \quad \mathsf{E}_W V = D_f(P, Q) \quad \text{(cf. Lemma 2.5)} \tag{2.8}$$

and

$$0 \leqq U \leqq 1 , \quad L_{\{f_0, f_1\}}(U) \leqq V \leqq 2 \max\{f(0), 0 f(\infty)\}\, U$$

(cf. the definition of $L_{\{f_0, f_1\}}$ and (2.7)). Since the set of all points $(u, v) \in \mathbb{R}^2$ such that

$$0 \leqq u \leqq 1 , \quad L_{\{f_0, f_1\}}(u) \leqq v \leqq 2 \max\{f(0), 0 f(\infty)\}\, u$$

is convex in $\mathbb{R}^2$, the point

$$\mathsf{E}_W(U, V) = (\mathsf{E}_W U, \mathsf{E}_W V) = (\chi^1(P, Q)/2, D_f(P, Q)) \in \mathbb{R}^2 \quad \text{(cf. (2.8))}$$

belongs to this set too which means that the desired inequalities hold. As to the attainability of bounds represented by these inequalities, if $f_0 = f_1$ then $L_{(f_0,f_1)} = f_0, f_0(1) = 2.f(0) = 2.0f(\infty) > 0$, and $V = f_0(U)$ on $(\mathcal{X}, \mathcal{B}, W)$. Since f_0 is convex with $f_0(0) = 0$, any point (u_0, v_0) of the convex subset of R^2 defined by

$$0 \leqq u \leqq 1\,, \quad f_0(u) \leqq v \leqq f_0(1)\,u$$

is a convex combination of finitely many points $(u, f_0(u))$. Therefore $(u_0, v_0) = \mathsf{E}_{W_0}(U, V)$ for some $W_0 \in \mathcal{P}$, i.e.

$$(u_0, v_0) = \left(\frac{\chi^1(P_0, Q_0)}{2},\ D_f(P_0, Q_0)\right)$$

for some $P_0, Q_0 \in \mathcal{P}$. This proves that if $f_0 = f_1$, then the bounds are attainable. □

Example 2.2. For $f(u) = |1 - u^\alpha|^{1/\alpha}$ and $\alpha \in (0, 1)$ we get from Lemma 2.8 attainable bounds

$$(2.9)\quad \left[\left(1 + \frac{\chi^1(P, Q)}{2}\right)^\alpha - \left(1 - \frac{\chi^1(P, Q)}{2}\right)^\alpha\right]^{1/\alpha} \leqq \chi^\alpha(P, Q) \leqq \chi^1(P, Q) \quad \text{(cf. (2.5))}\,.$$

The lower bound (2.9) is for $\alpha = \frac{1}{2}$ sharper than a bound used for the Hellinger distance $\chi^{1/2}(P, Q)$ by Le Cam [11]. Analogically the functions $f(u) = (1 - u^\alpha)/\alpha + (u - 1)/2$ for $\alpha \in (0, 1)$ yield attainable bounds

$$(2.10)\quad \left[1 - \left(1 + \frac{\chi^1(P, Q)}{2}\right)^{\max(\alpha, 1-\alpha)} \left(1 - \frac{\chi^1(P, Q)}{2}\right)^{\min(\alpha, 1-\alpha)}\right] \leqq$$

$$\leqq \alpha\, D^\alpha(P, Q) \leqq \frac{\chi^1(P, Q)}{2}\,.$$

These bounds follow, however, from inequality (23) in [20] too.

Lemma 2.8 together with the next complementary result proved in [21] are useful for comparisons of topologies in $\mathcal{P}$ which f-divergences give rise to.

Lemma 2.9. If

$$L_f(u) = \inf_{\chi^1(P,Q) = 2u} D_f(P, Q)\,, \quad U_f(v) = \sup_{\chi^1(P,Q) = 2v} D_f(P, Q)$$

then $L_f(u)$ is a convex increasing function on $[0, 1]$ with $L_f(0) = 0$. If $\|f\| = \infty$ then $U_f(v) = \infty$ unless $v = 0$ when it is zero.

Remark 2.1. An old idea expressed already on p. 224 of [21] is to use $D_f(P, Q)$ as a global measure of differences between probabilities $P(E)$, $Q(E)$ on $\mathcal{B}$ in a minimum distance estimation. There is however an obstacle to this project. Namely, $D_f(P, P_n) = \|f\|$ on Θ whenever $\mathcal{P}_e \perp \mathcal{P}_\Theta$ so that $T(P_n) = \Theta$ for all minimum divergence estimators T. At the same time the undesirable singularity $\mathcal{P}_e \perp \mathcal{P}_\Theta$ takes place

whenever $\mathscr{P}_\Theta$ contains nonatomic probabilities, e.g. in as frequent problems as estimation of location and scale or estimation of parameters of stochastic processes.

This difficulty becomes quite comprehensible from the point of view of interpretation of theoretical probability in Kolmogorov's theory. According to this interpretation, the numbers $P_\theta(E)$ approximate the observed proportions $P_n(E)$ on a subclass $\mathscr{E} \subset \mathscr{B}$ of "observable" events. Thus, for any global measure $D(P, Q)$ of differences between $P(E)$, $Q(E)$ on the class $\mathscr{E}$ of observable events E, P_θ is the better model of a reality represented by P_n, the smaller is the quantity $D(P_\theta, P_n)$. (Estimators minimizing such a quantity are considered *well-motivated* in this paper).

If the global measure D is specified as an f-divergence D_f, then $D(P, Q)$ measures differences between $P(E)$, $Q(E)$ on $\mathscr{B}$, i.e. $\mathscr{E} = \mathscr{B}$ and all events from $\mathscr{B}$ are considered observable. When $\mathscr{B}$ permits families $\mathscr{P}_\Theta$ singular with the whole family $\mathscr{P}_e$, then this specification obviously leads to a contradiction with what is considered "observable" and, consequently, "simple" (e.g. the support S_Θ of $\mathscr{P}_\Theta$ as a complement of an infinite set dense in $\mathscr{X}$ is hardly observable). The specification $D(P, Q) = D_f(P, Q)$ is thus justified only if $\mathscr{B}$ is simple in the sense that the underlying topological space $\mathscr{X}$ is discrete. In all other cases (including, of course, the previous one) Lemma 2.3 suggests to replace the above considered specification by $D(P, Q) = D_f(P^{\mathscr{A}}, Q^{\mathscr{A}})$, for a fixed algebra $\mathscr{A}$ generated by a class $\mathscr{E} \subset \mathscr{B}$ of sufficiently simple "observable" events.

The most elementary case is $\mathscr{E} = \{E\}$ for a fixed event $E \in \mathscr{B}$. In this case we get from Corollary 2.1 and Lemma 2.1

$$D_f(P_\theta^{\mathscr{A}}, P_n^{\mathscr{A}}) = d_f(P_\theta(E), P_n(E)) < \|f\| \quad \text{a.s.} \quad [P_\theta^n], \quad n = 1, 2, \ldots$$

even if $\mathscr{P}_e \perp \mathscr{P}_\Theta$ on $\mathscr{B}$. This strict inequality remains to be preserved even by a mean divergence

$$\mathsf{E}_{W_\theta} d_f(P_\theta(E_x), P_n(E_x)) = \mathsf{E}_{W_\theta} d_f(F_\theta, F_n) \quad \text{for} \quad W_\theta \equiv P_\theta$$

where the mean is taken over particular f-divergences $d_f(P_\theta(E_x), P_n(E_x))$ yielded by events $E = E_x$ from a class $\mathscr{E} = \{E_x : x \in \mathscr{X}\}$. If $\mathscr{E}$ is sufficient for $\mathscr{X}$, then the mean divergence exists (cf. Sec. 1) and it represents similarities between theoretical model P_θ and a reality P_n more objectively than any of the particular f-divergences. This motivates the following definition.

Let there is a class $\mathscr{E} = \{E_x : x \in \mathscr{X}\}$ sufficient for $\mathscr{X}$. A *weak f-divergence* of probabilities $P, Q \in \mathscr{P}$ is defined by

$$W D_f(P, Q) = \mathsf{E}_W d_f(F, G) \quad \text{(cf. Section 1 and Corollary 2.1)} \tag{2.11}$$

where W is a measure on $(\mathscr{X}, \mathscr{B})$ called weight, possibly depending on P, Q. It is convenient to consider weights factorized as follows

$$W(E) = \int_E \varphi(F(x), G(x)) \, \mathrm{d}\tilde{W} \quad \text{(in symbols } W \triangleq \varphi\tilde{W}\text{)}, \tag{2.12}$$

where $\tilde{W}$ is a factorweight and $\varphi(u, v)$, $(u, v) \in [0, 1]^2$, a measurable factorfunction.

Example 2.3. If $\mathscr{X} = \mathbb{R}$, $E_x = (-\infty, x)$ for $x \in \mathbb{R}$ and the total weight $W(\mathbb{R}) = 1$ is concentrated at the x which maximizes $|F(x) - G(x)|$, then $W\chi^1(P, Q) = \sup|F(x) - G(x)|$ is the Kolmogorov-Smirnov distance. The factorization $\varphi(u, v) = v(1 - v)$ or $v(1 - v)/u(1 - u)$ yields a generalized Cramér-von Mises or Anderson-Darling distance

$$W\chi^2(P, Q) = \mathsf{E}_W(F - G)^2 \quad \text{or} \quad W\chi^2(P, Q) = \mathsf{E}_W \frac{F(1 - F)}{(F - G)^2}$$

respectively (usual versions are yield by $W = P$). Extensions of both distances to $\mathscr{X} = \mathbb{R}^k$ is straightforward. Extensions to infinite product spaces are possible too.

Notice that replacing q in $\mathsf{E}_Q f(p/q)$ of (2.2) by another density $w = \mathrm{d}W/\mathrm{d}\lambda$, one can avoid the difficulties mentioned in Remark 2.1 too. This motivates the following definition.

A *directed f-divergence* of probabilities $P, Q \in \mathscr{P}$ with a directing measure W on $(\mathscr{X}, \mathscr{B})$ possibly depending on P and Q is defined by

$$D_f(P, Q\,|W) = \mathsf{E}_\lambda q\, f(p/w) = \mathsf{E}_Q f(p/w) \quad \lambda \gg P, Q, W, \tag{2.13}$$

where conventions $0 f(0/0) = 0$ and (for $f(0) = \infty$) $0 f(0/w) = w \lim_{\varepsilon \downarrow 0} \varepsilon f(\varepsilon)$ are added to those considered in (2.2).

Example 2.3. The directing density $w(x) = p(x)\,|x - \mathsf{E}_P X|^{-1}$ on $\mathscr{X} = \mathbb{R}$ yields the directed D^α-divergences

$$D^\alpha(P, Q \mid W) = \begin{cases} \alpha^{-1}(1 - \mathsf{E}_Q|X - \mathsf{E}_P X|^\alpha) & \alpha \in (0, 1) \\ \alpha^{-1}(\mathsf{E}_Q|X - \mathsf{E}_P X|^\alpha - 1) & \alpha \in (1, \infty) \end{cases}$$

for all $P, Q \in \mathscr{P}_1$.

3. STANDARD D-ESTIMATORS

Let $\mathscr{P}_\Theta$ be a family of probabilities on a sample space $(\mathscr{X}, \mathscr{B})$. A mapping $T : \mathscr{P}(T) \to \Theta$ defined by the criterion (for the definition of $D_f(P_\theta, Q)$ see (2.2))

$$T(Q) \text{ minimizes } D_Q(\theta) = D_f(P_\theta, Q) \text{ on } \Theta \tag{3.1}$$

is called *standard D-estimator* with projection family $\mathscr{P}_\Theta$ and symbolically denoted $T \triangleq \mathscr{P}_\Theta/D_f$.

Suppose that $\{x\} \in \mathscr{B}$ for all $x \in \mathscr{X}$. By (2.2)

$$D_{P_n}(\theta) = \|f\| \qquad \text{if} \quad P_\theta(S_n) = 0\,,$$

$$D_{P_n}(\theta) = \sum_{S_n} p_n f(p_\theta/p_n) + 0 f(\infty)\, P_\theta(\mathscr{X} - S_n) \quad \text{if} \quad P_\theta(S_n) \in (0, 1)\,,$$

$$D_{P_n}(\theta) = \sum_{S_n} p_n f(p_\theta/p_n) \qquad \text{if} \quad P_\theta(S_n) = 1\,,$$

where, here and in the sequel, S_n denotes the support of P_n and where $p_\theta(x) = P_\theta(\{x\})$, $p_n(x) = P_n(\{x\})$ for $x \in S_n$. We see from here that if $\mathscr{B}$ is "complicated" in the sense that $\mathscr{X}$ is uncountable and $\{x\} \in \mathscr{B}$, $x \in \mathscr{X}$, then a non-triviality condition $P_\theta(S_n) > 0$ for some $\theta \in \Theta$ cannot be satisfied by many $P_n \in \mathscr{P}_e$. If moreover $\mathscr{B}$ admits families $\mathscr{P}_\Theta$ continuous in the sense $P_\theta(\{x\}) = 0$ for all $x \in \mathscr{X}$ and $\theta \in \Theta$, then this condition is satisfied for none $P_n \in \mathscr{P}_e$. Indeed, it is necessary to restrict the standard D-estimators to projection families $\mathscr{P}_\Theta$ supported by discrete $\mathscr{X}$ as proposed in Remark 2.1 above.

The following theorem and its corollary have been proved in [25].

Theorem 3.1. Let $\mathscr{P}_\Theta$ be a Θ-continuous family on a discrete $\mathscr{X}$. If for each sample x either $D_{P_n}(\theta)$ is constant on Θ, or $\Theta(x) = \{\theta \in \Theta : D_{P_n}(\theta) < \infty\}$ is compact, or $\Theta(x)$ is σ-compact and

$$D_{P_n}(\theta) < \inf_{\{\theta_j\}} \liminf_{j \to \infty} D_{P_n}(\theta_j)$$

for some $\theta \in \Theta(x)$ and $\theta_j \in \Theta(x) - \Theta_j$ (cf. (1.2)), then $T \triangleq \mathscr{P}_\Theta | D_f$ is well defined.

Corollary 3.1. Let $\mathscr{P}_\Theta$ be a Θ-continuous family on a discrete $\mathscr{X} \subset R$, let S_θ be supports of $P_\theta \in \mathscr{P}_\Theta$, let $\|f\| < \infty$, $\Theta \subset \mathbb{R}$, and let $\Theta \cap [-j, j]$ be compact for all naturals $j > j_0$. If there exists $j_1 > j_0$ such that for all $j > j_1$ either $\Theta - [-j, j] = \emptyset$, or $\lim_{j \to \infty} p_{\theta_j}(x) = 0$ for all $x \in \mathscr{X}$ and $\theta_j \in \Theta - [-j, j]$, or $[-s_j, s_j] \cap S_{\theta_j} = \emptyset$ for all $\theta_j \in \Theta - [-j, j]$ and some sequence $s_j \uparrow \infty$, then $T \triangleq \mathscr{P}_\Theta | D_f$ is well-defined. □

If $\bigcap_\Theta S_\theta$ is infinite then $0 f(\infty) < \infty$ is necessary for the non-triviality of $T \triangleq \mathscr{P}_\Theta | D_f$ even under restriction to discrete $\mathscr{X}$'s. Analogically, if S_n exceeds $\bigcup_\Theta S_\theta$ and $f(0) = \infty$ then $T(P_n) = \Theta$ so that $f(0) < \infty$ is necessary for robustness of standard estimators $T \triangleq \mathscr{P}_\Theta | D_f$ w.r.t. the support modifying contaminations $\mathscr{Q}_\Theta = (1 - \varepsilon)\mathscr{P}_\Theta + \varepsilon \mathscr{H}_\Theta$. Only the standard D_f-estimators with $\|f\| < \infty$ are simultaneously non-trivial and robust in the stated sense. This explains the semiboundedness considered in Corollary 3.1.

Example 3.1. The usual projection families $\mathscr{P}_\Theta$ such as binomial, Poisson, uniform etc. are satisfying the conditions of Corollary 3.1 so that all $T \triangleq \mathscr{P}_\Theta | D_f$ with $\|f\| < \infty$ are well-defined. If $\Theta = [0, 1]$ and

$$P_\theta(\{1\}) = \tfrac{1}{2}[1 + \theta(1_{\Theta_0}(\theta) - 1_{\Theta_1}(\theta))], \quad P_\theta(\{0\}) = 1 - P_\theta(\{1\}),$$

where Θ_0, Θ_1 denote the sets of rational and irrational numbers, then we get a Θ-discontinuous family $\mathscr{P}_\Theta$. It is easy to see that for all $T \triangleq \mathscr{P}_\Theta | D_f$ with this $\mathscr{P}_\Theta$ it holds $T(P_n) \neq \emptyset$ iff the sample mean $\bar{x}$ belongs to $[\frac{1}{2}, 1]$ (in which case $T(P_n) = \{2\bar{x} - 1\}$ so that $T(P_n) = 2\bar{x} - 1$), i.e. none $T \triangleq \mathscr{P}_\Theta | D_f$ is well-defined. □

Example 3.2. Let us consider $T \triangleq \mathscr{P}_\Theta / D^0$ (cf. (2.4)) with $f(u) = -\ln u$, $\|f\| = f(0) = \infty$, i.e. without the robustness property considered above. By (3.1), $T(Q)$ maximizes $\mathsf{E}_Q \ln p_\theta$ whenever the entropy $H(Q) = -\mathsf{E}_Q \ln q$ is finite. If we denote by $\hat{T}$ the *MLE* with projection family $\mathscr{P}_\Theta$ under consideration then obviously $\tilde{T}(Q) = T(Q)$ for all $Q \in \mathscr{P}$ with $H(Q) < \infty$. In particular, $\tilde{T}(P_n) = \hat{T}(P_n)$ for all $P_n \in \mathscr{P}_e$. Thus for finite $\mathscr{X}$ the standard D^0-estimator is the well-known *MLE*.

The first estimators ever labelled as "minimum distance estimators" have been the standard D_f-estimators with $\mathscr{X} = \{0, 1, \ldots, k\}$ and $f(u) = (1-u)^2$ (Cramér [7] and Neyman [15]), $f(u) = (1-u)^2/u$, $f(u) = u \ln u$ (Rao [17, 18]). The general f has been introduced by Vajda [23]. Efficiency of estimators from this class has been studied by Vošvrda [26] under the same conditions on $\mathscr{X}$ and $\mathscr{P}_\Theta$ as considered by Rao. According to [26] all $T \triangleq \mathscr{P}_\Theta / D_f$ with $f''(1) \neq 0$ are efficient and if, moreover, $2f''(1) + f'''(1) = 0$ then also efficient in a second order sense introduced by Rao [17].

4. WEAK *D*-ESTIMATORS

Let $\mathscr{P}_\Theta$ be a family of probabilities and $\mathscr{W}_\Theta$ a family of weights on a sample space $(\mathscr{X}, \mathscr{B})$ and let a class $\mathscr{E} = \{E_x : x \in \mathscr{X}\}$ be sufficient for $\mathscr{X}$. A mapping $T : \mathscr{P}(T) \to \Theta$ defined by the criterion (for the definition of $W_\theta D_f(F_\theta, G)$ see (2.11))

$$T(Q) \text{ minimizes } D_Q(\theta) = W_\theta D_f(F_\theta, G) \quad \text{on} \quad \Theta \tag{4.1}$$

is called *weak D-estimator* with projection family $\mathscr{P}_\Theta$ and family of weights $\mathscr{W}_\Theta$ and symbolically denoted $T \triangleq \mathscr{P}_\Theta / \mathscr{W}_\Theta D_f$. If Θ is structural then we write simply $P / W D_f$ for parents P, W of $\mathscr{P}_\Theta$, $\mathscr{W}_\Theta$, with W replaced by $\varphi \tilde{W}$ whenever convenient (cf. (2.12)).

The next theorem and its corollary have been proved in [25] (the expression $1_E(x)$ denotes $1_E(x_1) + \ldots + 1_E(x_n)$; connected means that $\emptyset$, Θ are the only open and at the same time closed subsets of Θ).

Theorem 4.1. Let Θ be a structural connected space with $P[\theta](E_x) = P(E_{[\theta](x)})$ on $\Theta \times \mathscr{X}$ and with a continuous mapping sending (θ, x) into $[\theta](x)$. Let $\Phi(u, v) = d_f(u, v)\, \varphi(u, v)$ be bounded on $[0, 1]^2$ and $W(\{x : 1_{E_x}(x) \neq 1_{E_x}(\tilde{x})\}) \to 0$ as $\tilde{x} \to x$ in $\mathscr{X}^n$. Then the compactness of Θ implies the continuity of $T \triangleq P / \varphi \tilde{W} D_f$ on $\mathscr{X}^n$ and the σ-compactness of Θ together with the conditions

$$D_{P_n}(\theta(x)) < \inf_{\{\theta_j\}} \liminf_{j \to \infty} D_{P_n}(\theta_j)$$

for some $\theta(x) \in \Theta$, $x \in \mathscr{X}_n \subset \mathscr{X}^n$ and $\theta_j \in \Theta - \Theta_j$ (cf. (1.2)), and

$$\tilde{W}(\{x : 1_{E_x}([\theta](x)) \neq 1_{E_x}([\theta](\tilde{x}))\}) \to 0$$

uniformly on Θ as $\tilde{x} \to x$ in $\mathscr{X}^n$, imply the continuity of T on $\mathscr{X}_n$.

Corollary 4.1. The estimates of location $T \triangleq P/\varphi\tilde{W}\chi^\alpha$, $\alpha \in [1, \infty)$ with factor-functions $\varphi(u, v) = [v(1-v)]^{\alpha-1}/[v^{\alpha-1} + (1-v)^{-\alpha 1}]$ and factorweights $\tilde{W} \ll \lambda$ satisfying condition

$$\tilde{W}((-\infty, F^{-1}(\tfrac{1}{2}))\, \tilde{W}((F^{-1}(\tfrac{1}{2}), \infty)) \neq 0 \quad \text{for} \quad F(x) = P((-\infty, x)),$$

are well-defined in the sense that they are continuous on $\mathscr{X}^n$ for all n under consideration.

Note that if the first condition of Theorem 4.1 concerning $\tilde{W}$ holds and $[\theta] : \mathscr{X} \to \mathscr{X}$ is continuous uniformly for all $\theta \in \Theta$, then the second condition concerning $\tilde{W}$ holds too. It is also to be noted that for all weak D-estimators of location and scale all conditions of Theorem 4.1 concerning Θ hold.

Example 4.1. It can be shown that if, in addition to what has been supposed concerning φ and $\tilde{W}$ in Corollary 4.1, $\sup |x|\, \tilde{w}(x) < \infty$, then all weak χ^α-estimators of location and scale $(M, S) \triangleq P/\varphi\tilde{W}\chi^\alpha$, $\alpha \in [1, \infty)$, are well-defined with estimates $(M(P_n), S(P_n))$ continuous on $\mathscr{X}_n = \mathbb{R}^n - H_n$, where $H_n = \{x \in \mathbb{R}^n : x_1 = \ldots = x_n\}$. For x form the hyperplane H_n, $(\mathbb{M}, \mathbb{S})(P_n) = \{T(1_{\{0\}})\} \times (0, \infty) \subset \mathbb{R} \times (0, \infty)$, where T is the corresponding weak χ^α-estimator of location considered in Corollary 4.1 and $1_{\{0\}} \in \mathscr{P}_e$ has been defined in Section 1.

The method of minimum distance estimation, first outlined by Cramér and Neyman (see Sec. 3) and by Wolfowitz [27], has been revived as an alternative to *MLE*'s after the point of view of robustness has been introduced into statistics by Huber [8]. A weak χ^1-estimator of location has been studied by Rao et al. [19] and later by Parr and Schucany [16] together with some weak χ^2-estimators. Weak χ^2-estimators of location have been more systematically studied by Millar [13] and Boos [4]. Some weak χ^1-estimators with discrete projection families have been studied even earlier (see Mood et al. [14]). The concepts of weak f-divergence and weak D-estimator have been in a general form introduced by Vajda [23, 25] (cf. also [24]).

5. DIRECTED D-ESTIMATORS

Let $\mathscr{P}_\Theta$ be a family of probabilities and $\mathscr{W}_\Theta$ a family of measures on a sample space $(\mathscr{X}, \mathscr{B})$. A mapping $T : \mathscr{P}(T) \to \Theta$ defined by the criterion (for the definition of $D_f(P_\theta, Q \mid W)$ see (2.13))

$$T(Q) \text{ minimizes } D_Q(\theta) = D_f(P_\theta, Q \mid W) \text{ on } \Theta \tag{5.1}$$

is called *directed D-estimator* with projection family $\mathscr{P}_\Theta$ and family of directing measures $\mathscr{W}_\Theta$ and symbolically denoted $T \triangleq \mathscr{P}_\Theta/\mathscr{W}_\Theta/D_f$. Since we shall mainly consider one-measure classes $\mathscr{W}_\Theta = \{W\}$, we shall usually write $\mathscr{P}_\Theta/W/D_f$ or simply $P/W/D_f$ for a parent P of $\mathscr{P}_\Theta$ when Θ is structural.

Remark 5.1. $D_f(P_\theta, Q \mid W)$ as a formal generalization of the f-divergence cannot

in general be considered as a quantity the minimization of which yields well-motivated estimators. Indeed, one can meet within this class estimators with quite curious properties. This is in particular true for many M-estimators of location with loss functions $M(x) = f(|x|)$ yielded by the directing parent density $w(x) = p(x)\,|x|^{-1}$ (any convex M symmetric about 0 is in this class). If we restrict to projection parents P with $\mathsf{E}_P X = 0$ then the directed D^α-estimate $T(Q)$ minimizes $\mathsf{E}_Q|X - \theta|^\alpha$ if $\alpha \in \in (1, \infty)$ (cf. Example 2.3) while it maximizes (!) $\mathsf{E}_Q|X - \theta|^\alpha$ for $\alpha \in (0, 1)$. It is therefore very important to be able to characterize directing measures and functions f yielding well-motivated directed D-estimators. This is the aim of the next our considerations.

Let $\Delta = \{\mathscr{D}^{(j)} : j = 1, 2, \ldots\}$ be a net of countable decompositions of $\mathscr{X}$ sufficient for $\mathscr{B}$ in the sense specified in Section 1 and let $\mathscr{B}^{(j)}$ be a sub-σ-algebra of $\mathscr{B}$ generated by $\mathscr{D}^{(j)}$.

Suppose that a σ-finite W dominates $\mathscr{P}_\Theta$ and let $W^{(j)}$, $\mathscr{P}_\Theta^{(j)}$ be restrictions of W, $\mathscr{P}_\Theta$ on $\mathscr{B}^{(j)}$ and $\widetilde{\mathscr{P}}_\Theta^{(j)}$ extensions of $\mathscr{P}_\Theta^{(j)}$ back to $\mathscr{B}$ with Randon-Nikodym densities $\mathrm{d}\tilde{P}_\theta^{(j)}/\mathrm{d}W$ identical with $p_\theta^{(j)} = \mathrm{d}P_\theta^{(j)}/\mathrm{d}W^{(j)}$. In accordance with Remark 2.1, a directed D-estimator $T \triangleq \mathscr{P}_\Theta/W_\Theta/D_f$ is said well-motivated if for some net Δ the corresponding standard D-estimators $T^{(j)} \triangleq \mathscr{P}_\Theta^{(j)}/D_f$ tend to T in the sense that

$$\lim_{j\to\infty} \mathbb{T}_{\Theta^*}^{(j)}(P_n) = \mathbb{T}_{\Theta^*}(P_n) \tag{5.2}$$

for all compact $\Theta^* \subset \Theta$ and all $x \in \mathscr{X}^n$ with mutually distinct coordinates $x_1, \ldots, x_n$, where $\mathbb{T}_{\Theta^*}$ denotes the points of minima of $D_f(P_\theta, Q)$ on $\Theta^* \subset \Theta$.

A measure W on $(\mathscr{X}, \mathscr{B})$ is said *equiuniform* if there exists a net Δ of the above described properties such that, for every j, W is constant on all sets from $\mathscr{D}^{(j)}$. For example, the Lebesque measure λ or any $W \equiv \lambda$ is equiuniform on $\mathscr{X} = \mathbb{R}^k$ (on a more general locally compact group $\mathscr{X}$ one can take a Haar measure W).

Theorem 5.1. All directed D-estimators $T \triangleq \mathscr{P}_\Theta/W/D^\alpha$, $\alpha \in (0, 1)$, with $\mathscr{P}_\Theta$ dominated by an equiuniform σ-finite W are well motivated.

Proof. (I) Fix $\alpha \in (0, 1)$ and $x \in \mathscr{X}^n$ with mutually distinct coordinates, put $f(u) = = (1 - u^\alpha)/\alpha$ (cf. (2.4)), and denote by D_{ij} the disjoint events of $\mathscr{D}^{(j)}$ containing coordinates x_i, $i = 1, \ldots, n$. By (2.2)

$$D_f(\widetilde{\mathscr{P}}_\theta^{(j)}, P_n) = \sum_{i=1}^n \frac{1}{nW(D_{ij})} \int_{D_{ij}} f\left(p_\theta^{(j)} \Big/ \frac{1}{nW(D_{ij})}\right) \mathrm{d}W + P_\theta(\mathscr{X} - \bigcup_{i=1}^n D_{ij})\, 0 f(\infty)\,.$$

Taking into account

$$0 f(\infty) = 0\,, \quad f(uv) = v^\alpha f(u) + f(v) \quad \text{for} \quad u, v \geqq 0\,, \tag{5.3}$$

together with the assumption $W(D_{ij}) = W(D_{1j})$ we get

$$D_f(\widetilde{\mathscr{P}}_\theta^{(j)}, P_n) = \sum_{i=1}^n \frac{(nW(D_{ij}))^\alpha}{nW(D_{ij})} \int_{D_{ij}} f(p_\theta^{(j)})\, \mathrm{d}W + \sum_{i=1}^n \frac{f(nW(D_{ij}))}{nW(D_{ij})} =$$

$$= (n\,W(D_{ij}))^2 \sum_{i=1}^{n} \frac{1}{nW(D_{ij})} \int_{D_{ij}} f(p_\theta^{(j)})\,\mathrm{d}W + \frac{f(nW(D_{1j}))}{W(D_{1j})}.$$

It follows from here that $T^{(j)}(P_n)$ is a set of parameters minimizing

$$D_{P_n}^{(j)}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{W(D_{ij})} \int_{D_{ij}} f(p_\theta^{(j)})\,\mathrm{d}W.$$

(II) Since W is σ-finite, it is easy to see that $p_\theta^{(j)}$ defined in Section 1 coincides a.e. $[W]$ with $p_\theta^{(j)}$ defined here provided $\lambda = W$. Hence it follows from (1.4) and from the continuity of f that $f(p_\theta^{(j)}) \to f(p_\theta)$ a.e. $[W]$. Since $p_\theta^{(j)}(x) = p_\theta^{(j)}(x_i)$ on D_{ij}, the just established result yields

$$\frac{1}{W(D_{ij})} \int_{D_{ij}} f(p_\theta^{(j)})\,\mathrm{d}W = f(p_\theta^{(j)}(x_i)) \to f(p_\theta(x_i))$$

for all $x_i \in \mathscr{X}$ except a set of W-measure zero. Therefore

$$D_{P_n}^{(j)} \to D_{P_n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} f(p_\theta(x_i)) = \mathsf{E}_{P_n} f(p_\theta) \quad \text{a.e.} \quad [W^n]$$

and, consequently,

$$D_{P_n}^{(j)} \to \mathsf{E}_{P_n} f(p_\theta) \quad \text{a.s.} \quad [P_\Theta^n].$$

Combining this convergence with the result proved in (I) and taking into account the compactness of Θ^* we see that (5.2) holds. □

Note that a particular variant of this theorem with $\mathscr{X} = \mathbb{R}$ and $W = \lambda$ on $\mathbb{R}$ has first been proved in [24].

In an attempt to extend this proof to other functions f we find the following two properties indispensable: (i) $\|f\| < \infty$, (ii) a functional equation $f(uv) = \Phi(v)f_1(u) + f_2(v)$ is required to be satisfied in the domain $u, v > 0$ by some Φ, f_1, f_2 (cf. (5.3)). It is easy to see that (ii) holds only if $f_1(u) = f_2(u) = f(u)$ for $u > 0$. Further, by Aczél [1], all continuous solutions f, Φ of the equation $f(u, v) = \Phi(v)\,f(u) + f(v)$ with $f(1) = 0$ are of the form $\Phi(v) = v^\alpha$, $\alpha \in R$, $f(u) = c(\alpha)\,(1 - u^\alpha)$ for $\alpha \neq 0$ and $f(u) = c \ln u$ for $\alpha = 0$. Therefore the only functions admissible in the proof are those considered in Theorem 5.1 ($f(u)$ for $\alpha \geqq 1$ does not satisfy assumptions of Section 1; the limit values $\alpha = 0$, $\alpha = 1$ are analyzed separately below). This conclusion is still not a proof that the only well-motivated directed D-estimators are the D^α-estimators but it provides certain evidence in favour of such a conjecture. In any case this problem deserves a deeper attention.

Hereafter we denote the well-motivated D-estimators $T \triangleq \mathscr{P}_\Theta | W | D^\alpha$ with $\mathscr{P}_\Theta \ll W$ briefly by $T^\alpha \triangleq \mathscr{P}_\Theta // W$. If Θ is structural and P is a parent of $\mathscr{P}_\Theta$ then we shall write simply $T^\alpha \triangleq P // W$ instead of $T^\alpha \triangleq \mathscr{P}_\Theta // W$.

Since the functions $D_{P_n}(\theta) = \mathsf{E}_{P_n}(-\ln p_\theta)$ or $D_{P_n}(\theta) = \mathsf{E}_{P_n}(1 - p_\theta)$ are the limits of the function $\alpha^{-1}\,\mathsf{E}_{P_n}(1 - p^\alpha)$ as $\alpha \uparrow 0$ or $\alpha \uparrow 1$ respectively, it holds for the sets

of parameters $\mathbb{T}^0(P_n)$, $\mathbb{T}^1(P_n)$ minimizing these two functions

$$\lim_{\alpha \downarrow 0} \mathbb{T}^{\alpha}_{\Theta^*}(P_n) = \mathbb{T}^0_{\Theta^*}(P_n), \quad \lim_{\alpha \uparrow 1} \mathbb{T}^{\alpha}_{\Theta^*}(P_n) = \mathbb{T}^1_{\Theta^*}(P_n)$$

for every compact Θ^* and every $P_n \in \mathscr{P}_e$. Hence, if W is equiuniform then the limits $T^0 \triangleq \mathscr{P}_\Theta/\!/W$, $T^1 \triangleq \mathscr{P}_\Theta/\!/W$ of the well-motivated estimators $T^\alpha \triangleq \mathscr{P}_\Theta/\!/W$, $\alpha \in (0, 1)$, are well-motivated too.

Thus in what follows all estimators $T^\alpha \triangleq \mathscr{P}_\Theta/\!/W$, $\alpha \in [0, 1]$, with equiuniform W will be considered well-motivated. Remind that, by (2.4) and (5.1), T^α is a mapping $\mathscr{P}(T) \to \Theta$ defined by

$$T^\alpha(Q) \text{ maximizes } D_Q(\theta) = \begin{cases} \mathsf{E}_Q p_\theta^\alpha & \text{if } \alpha \in (0, 1] \\ \mathsf{E}_\theta \ln p_\theta & \text{if } \alpha = 0 \end{cases} \text{ on } \Theta \text{ where } p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}W}.$$

Theorem 5.2. Let us consider estimators $T^\alpha \triangleq \mathscr{P}_\Theta/\!/W$, $\alpha \in [0, 1]$, with projection densities $p_\theta(x)$ continuous on $\Theta \times \mathscr{X}$. (a) If Θ is compact, then the estimates $T^\alpha(P_n)$ are continuous on $\mathscr{X}^n$. (b) If Θ is σ-compact, if $p_\theta(x)$ are continuous on $\mathscr{X}$ uniformly for all $\theta \in \Theta$, and if $\lim_{j \to \infty} p_{\theta_j}(x) = 0$ for all $x \in \mathscr{X}$ and $\theta_j \in \Theta - \Theta_j$ (cf. (1.2)), then T^α is well-defined and the estimates $T^\alpha(P_n)$ are continuous on $S^n_\Theta \cup \operatorname{int}(\mathscr{X}^n - S^n_\Theta)$, where $S_\Theta = \bigcup_\Theta \{x \in \mathscr{X} : p_\theta(x) > 0\}$ is an open support of $\mathscr{P}_\Theta$ and int denotes the interior.

Proof. (a) Since all functions

$$g(u) = \begin{cases} u^\alpha & \alpha \in (0, 1] \\ \ln u & \alpha = 0 \end{cases}$$

are continuous on $(0, \infty)$, the functions

$$D_{P_n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(p_\theta(x_i))$$

are continuous on $\Theta \times \mathscr{X}^n$. Therefore, if Θ is compact, then $\mathbb{T}^\alpha(P_n)$ is non-empty and $\mathbb{T}^\alpha : \mathscr{X}^n \to C(\Theta)$ is continuous. Consequently (cf. Convention 1.1), $T^\alpha(P_n) = \tau(\mathbb{T}^\alpha(P_n))$ is continuous on $\mathscr{X}^n$.

(b) If $x \in \mathscr{X}^n - S^n_\Theta$, then $D_{P_n}(\theta)$ is constant on Θ (either 0 or $-\infty$) so that $\mathbb{T}^\alpha(P_n) = \Theta$ for all $\alpha \in [0, 1]$. Hence $T^\alpha(P_n) = \tau(\Theta)$ is constant on $\mathscr{X}^n - S^n_\Theta$. If $x \in S^n_\Theta$ then

$$D_{P_n}(\theta) > \begin{cases} 0 & \text{for } \alpha \in (0, 1] \\ -\infty & \text{for } \alpha = 0. \end{cases}$$

Since by assumptions

$$\lim_{j \to \infty} D_{P_n}(\theta_j) = \begin{cases} 0 & \text{for } \alpha \in (0, 1] \\ -\infty & \text{for } \alpha = 0 \end{cases} \quad \text{for all } \theta_j \in \Theta - \Theta_j,$$

$\mathbb{T}^\alpha(P_n) \subset \Theta_n \subset \Theta_j$ for all sufficiently large j. Thus the continuity of $D_{P_n}(\theta)$ in part (a) of the proof implies that $\mathbb{T}^\alpha(P_n)$ is non-empty compact. Analogically as in the proof

of Theorem 4.1 in [25], the uniform continuity of $p_\theta(x)$ on $\mathscr{X}$ w.r.t. $\theta \in \Theta$ yields the continuity of $\mathbb{T}^\alpha : \mathscr{X}^n \to C(\Theta)$ so that the continuity of $T^\alpha(P_n) = \tau(\mathbb{T}^\alpha(P_n))$ on S_Θ^n follows from Convention 1.1. □

Corollary 5.1. Let $p = \mathrm{d}P/\mathrm{d}\lambda$ be continuous on $\mathbb{R}$ with a support $S \subset \mathbb{R}$. All estimators of location and scale $T^\alpha = (M^\alpha, S^\alpha) \triangleq P/\!/\lambda$, $\alpha \in [0, 1]$, with projection subgroups $\Theta_\varepsilon = \mathbb{R} \times [\varepsilon, \varepsilon^{-1}]$ of the group $\mathbb{R} \times (0, \infty)$, $\varepsilon \in (0, 1)$, are well-defined and the estimates $T^\alpha(P_n)$ are continuous on $S^n \cup (\mathrm{int}\,(\mathbb{R}^n - S^n) \subset R^n$.

Example 5.1. A general *MLE* is defined by a projection family $\mathscr{P}_\Theta$ dominated by some (directing) σ-finite measure W on a sample space $(\mathscr{X}, \mathscr{B})$. It follows from (5.4) that this *MLE* coincides with $T^0 \triangleq \mathscr{P}_\Theta/\!/W$. The T^0 is the only member of the family $T^\alpha \triangleq \mathscr{P}_\Theta/\!/W$, $\alpha \in [0, 1]$, independent of W in the sense that, if $\tilde{T}^0 \triangleq \mathscr{P}_\Theta/\!/\tilde{W}$, $w, \tilde{w} = \mathrm{d}W, \mathrm{d}\tilde{W}/\mathrm{d}\lambda$ for some $\lambda \gg W, \tilde{W}$, and $\mathsf{E}_Q \ln (w/\tilde{w}) < \infty$, then $Q \in \mathscr{P}(T^0) \cap \cap \mathscr{P}(\tilde{T}^0)$ and $\mathbb{T}^0(Q) = \mathbb{T}^0(Q)$ (thus, in particular, $\mathbb{T}^0(P_n) = \mathbb{T}^0(P_n)$ on $\mathscr{X}^n$ a.s. $[\mathscr{P}_\Theta^n]$). The dependence of the rest of this family on W vanishes when Θ becomes structural since there is usually a unique directing W satisfying the equivariance conditions of Theorem 6.3 below. This W is usually the Lebesque or Haar measure producing at the same time well-motivated variants of T^α, $\alpha \neq 0$, (cf. Theorem 5.1).

Example 5.2. It is well known that the location and scale estimator $T^0 = (M^0, S^0) \triangleq$ $\triangleq No\,(0,1)/\!/\lambda$ is the sample mean-sample deviation

$$(5.5) \qquad (M^0(Q), S^0(Q)) = (\mathsf{E}_Q X, [\mathsf{E}_Q(X - \mathsf{E}_Q X)^2]^{1/2}) \quad \text{for} \quad Q \in \mathscr{P}(T^0) = \mathscr{P}_2 .$$

It is also well known that the estimator of location $T^0 \triangleq P/\!/\lambda$ with doubly exponential P is the sample median

$$(5.6) \qquad T^0(Q) = G^{-1}(\tfrac{1}{2}) \quad \text{for} \quad Q \in \mathscr{P}(T) = \mathscr{P} .$$

The estimator of location $T^1 \triangleq No(0, 1)/\!/\lambda$ is the "mean likelihood" estimator of [2] while T^1 with projection parent density

$$(5.7) \qquad p(x) = \tfrac{3}{2}\, 1_{(-1/2,\, 1/2)}(x)\,(1 - 4x^2)$$

is the "skipped mean" of Huber [9].

Example 5.3. $T^\alpha \triangleq No(0, 1)/\!/\lambda$ with $\alpha \in (0{\cdot}1, 0{\cdot}3)$ are highly recommended estimators of location. They are good from the point of view of both efficiency and robustness because their sensitivity curves quite closely approximate the curves of estimators A 17 – A 25 and *AMT* which emerged as most promising robust estimators of location from the extensive experimental study [2].

The estimators $T^\alpha \triangleq \mathscr{P}_\Theta/\!/\lambda$, $\alpha \in [0, 1]$, with projection families $\mathscr{P}_\Theta$ on $\mathbb{R}$ have first been introduced in [23] and first motivated in the sense of Theorem 5.1 in [24]. The general $T^\alpha \triangleq \mathscr{P}_\Theta/\!/W$ we have introduced in [25] but Theorem 5.1 is first proved here.

6. INVARIANCE AND EQUIVARIANCE OF D-ESTIMATORS

We shall establish an invariance of D-estimators $T: \mathscr{P}(T) \to \Theta$ with projection families $\mathscr{P}_\theta$ defined by a minimization specified in (3.1) or (4.1) or (5.1) and by the Convention 1.1.

Theorem 6.1. Let t be a mapping from Θ into a parameter space $\tilde{\Theta}$ and let $\tilde{T}$: $: \mathscr{P}(\tilde{T}) \to \Theta$ be a D-estimator of $\tilde{\theta} = t(\theta)$ defined by the criterion

$$\tilde{T}(Q) \text{ minimizes } \tilde{D}_Q(\tilde{\theta}) = \inf_{\theta \in t^{-1}(\tilde{\theta})} D_Q(\theta) \quad \text{on} \quad \tilde{\Theta}\,.$$

Then $\mathscr{P}(\tilde{T}) \subset \mathscr{P}(T)$ and $\tilde{T}(Q) = t(T(Q))$ for $Q \in \mathscr{P}(T)$.

Proof. The equivariance of *MLE*'s T with projection $\mathscr{P}_\theta \ll \lambda$ and its proof given by Zehna [28] remain unchanged if the domain $\mathscr{P}_e$ of these estimators is extended to $\mathscr{P}(T) \subset \mathscr{P}$ provided the functions $D_{P_n}(\theta) = \mathsf{E}_{P_n}(-\ln p_\theta)$ minimized by $T(P_n)$, $P_n \in \mathscr{P}_e$, in [28] are replaced by $D_Q(\theta) = \mathsf{E}_Q(-\ln p_\theta)$ minimized by $T(Q)$, $Q \in \mathscr{P}(T)$. Since this modified proof employs no specific properties of functions $\mathsf{E}_Q(-\ln p_\theta)$ it can be applied to arbitrary functions $D_Q(\theta)$, in particular to those figuring in (3.1), (4.1), or (5.1) respectively. We avoid reproduction of details here. □

Corollary 6.1. D-estimators are *invariant* w.r.t. 1-1 reparametrizations $t: \Theta \to \tilde{\Theta}$ of projection families $\mathscr{P}_\theta$ in the sense that if T is a D-estimator with projection family $\tilde{\mathscr{P}}_{\tilde{\Theta}} = \{\tilde{P}_{\tilde{\theta}} = P_{t^{-1}(\tilde{\theta})} \in \mathscr{P}_\theta : \theta \in \tilde{\Theta}\}$ then $\mathscr{P}(\tilde{T}) = \mathscr{P}(T)$ and $\tilde{T}(Q) = t(T(Q))$ for any $Q \in \mathscr{P}(T)$.

In the rest of this section we suppose that Θ is structural on $\mathscr{X}$. An estimator $T: \mathscr{P}(T) \to \Theta$ of a structural parameter is said *equivariant* if $\mathscr{P}(T)[\theta] = \{Q[\theta] : Q \in \mathscr{P}(T)\} \subset \mathscr{P}(T)$ and (for the notation $[\theta]$, θ^{-1} see Sec. 1)

$$T(Q[\theta]) = \theta^{-1}\, T(Q) \quad \text{for all} \quad Q \in \mathscr{P}(T)\,, \quad \theta \in \Theta\,. \tag{6.1}$$

For estimators of location or scale M or S, (6.1) takes on the following form

$$M(Q[\mu, \sigma]) = \frac{M(Q) - \mu}{\sigma} \quad \text{or} \quad S(Q[\mu, \sigma]) = \frac{S(Q)}{\sigma} \tag{6.2}$$

respectively (cf. the identity $[\mu, \sigma]^{-1}(M(Q), S(Q)) = ((M(Q) - \mu)/\sigma,\ S(Q)/\sigma)$ following from the definition of $[\mu, \sigma]^{-1} = [(\mu, \sigma)^{-1}]$ in Section 1). If (6.2) holds, M or S are said *location-scale equivariant*. M is location equivariant of S is *scale equivariant* if (6.2) holds with $\sigma = 1$ or $\mu = 0$ respectively.

Theorem 6.2. If a family $\mathscr{E} = \{E_x : x \in \mathscr{X}\}$ sufficient for $\mathscr{X}$ satisfies the condition $[\theta](E_x) = E_{[\theta](x)}$ for all $x \in \mathscr{X}$, $\theta \in \Theta$, then all weak D-estimators $T \triangleq P/\varphi \tilde{W} D_f$ of $\theta \in \Theta$ are equivariant.

Proof. Analogically as in part (I) of the proof of Theorem 4.1 in [25], the assump-

tion of Theorem 6.2 implies

$$D_Q(\tilde{\theta}) = \mathsf{E}_W\, d_f(F, G[\tilde{\theta}])\, \varphi(F, G[\tilde{\theta}]) \quad \text{for all} \quad \tilde{\theta} \in \Theta$$

where D_Q is the function defined in (4.1) and $G[\tilde{\theta}]$ is a d.f. of $Q[\tilde{\theta}] \in \mathscr{P}$. Therefore, for any fixed $\theta \in \Theta$,

$$D_{Q[\theta]}(\tilde{\theta}) = D_Q(\theta\tilde{\theta}) \quad \text{for all} \quad \tilde{\theta} \in \Theta\,.$$

We see from here that $D_Q(\tilde{\theta})$ attains its minimum on Θ at some $\tilde{\theta}_*$ iff $D_{Q[\theta]}(\tilde{\theta})$ attains its minimum on Θ at $\tilde{\theta}^{-1}\tilde{\theta}_*$ which implies $\mathbb{T}(Q[\theta] = \theta^{-1}\mathbb{T}(Q)$. Since this is true for any $\theta \in \Theta$, it is obvious that $Q \in \mathscr{P}(T)$ iff $Q[\theta] \in \mathscr{P}(T)$ for all $\theta \in \Theta$ and, moreover, by (1.5)

$$T(Q[\theta] = \tau(\mathbb{T}(Q[\theta])) = \tau(\theta^{-1}\mathbb{T}(Q)) = \theta^{-1}\tau(\mathbb{T}(Q)) = \theta^{-1}\, T(Q)$$

for all $\theta \in \Theta$. □

Theorem 6.3. If $W[\theta] \ll W$, if the Jacobians $J(\theta) = \mathrm{d}W[\theta]/\mathrm{d}W$ are constant on $\mathscr{X}$ for all $\theta \in \Theta$, and if $J(\theta\tilde{\theta}) = J(\theta)\, J(\tilde{\theta})$ for all $\theta, \tilde{\theta} \in \Theta$, then all directed D^α-estimators $T^\alpha \triangleq P/\!/W$, $\alpha \in [0, 1]$, of $\theta \in \Theta$ are equivariant.

Proof. It is easy to see that

$$p_\theta(x) = J(\theta^{-1})\, p([\theta]^{-1}(x)) \quad \text{for} \quad p_\theta = \frac{\mathrm{d}P[\theta]^{-1}}{\mathrm{d}W}, \quad p = \frac{\mathrm{d}P}{\mathrm{d}W}. \tag{6.3}$$

Thus it holds for D_Q figuring in (5.4) and for g defined in part (a) of the proof of Theorem 5.2

$$D_Q(\tilde{\theta}) = \mathsf{E}_Q\, g(J(\tilde{\theta}^{-1})\, p([\tilde{\theta}]^{-1}(x)) = \mathsf{E}_{Q[\tilde{\theta}]}\, g(J(\tilde{\theta}^{-1})\, p(x)) \quad \text{for all} \quad \tilde{\theta} \in \Theta\,.$$

Therefore the multiplicativity of J yields for any fixed $\theta \in \Theta$

$$D_{Q[\theta]}(\tilde{\theta}) = \mathsf{E}_{Q[\theta\tilde{\theta}]}\, g(J(\tilde{\theta}^{-1})\, p(x)) = \mathsf{E}_{Q[\theta\tilde{\theta}]}\, g(J(\theta^{-1})^{-1}\, J((\theta\tilde{\theta})^{-1}\, p(x))\,.$$

Since obviously $J(\theta^{-1})^{-1} = c \in (0, \infty)$ on $\mathscr{X}$ for the fixed $\theta \in \Theta$, it follows

$$D_{Q[\theta]}(\tilde{\theta}) = \begin{cases} g(c)\, D_Q(\theta\tilde{\theta}) & \text{if} \quad \alpha \in (0, 1] \\ g(c) + D_Q(\theta\tilde{\theta}) & \text{if} \quad \alpha = 0 \end{cases} \quad \text{where} \quad g(c) \in \begin{cases} (0, \infty) \\ \mathbb{R} \end{cases}.$$

We see from here that $D_Q(\tilde{\theta})$ attains its maximum on Θ at some $\tilde{\theta}_*$ iff $D_{Q[\theta]}(\tilde{\theta})$ attains its maximum on Θ at $\theta^{-1}\tilde{\theta}_*$ which implies $\mathbb{T}^\alpha(Q[\theta]) = \theta^{-1}\, \mathbb{T}^\alpha(Q)$. The rest is the same as in the proof of Theorem 6.2.

Corollary 6.2. Let us consider the parameter of location and scale $\theta = (\mu, \sigma) \in \mathbb{R} \times$ $\times\, (0, \infty)$. (a) The assumption of Theorem 6.2 holds for intervals $E_x = (-\infty, x)$ on $\mathscr{X} = \mathbb{R}$ so that all weak D-estimators of locations and/or scale $T \triangleq P_I/\varphi\tilde{W}\, D_f$ are location and/or scale equivariant respectively. (b) The assumptions of Theorem 6.3 hold iff W is the Lebesgue measure (or a constant multiple of it) so that all directed D^α-estimators of location and/or scale $T^\alpha \triangleq P/\!/\lambda$, $\alpha \in [0, 1]$, are location and/or scale equivariant respectively.

It is easy to verify from the definition that if M is a location equivariant estimator of location with $\mathscr{P}(M)[0, \sigma] \subset \mathscr{P}(M)$ for all σ (S is scale equivariant estimator of scale with $\mathscr{P}(S)[\mu, 1) \subset \mathscr{P}(S)$ for all μ) and T is a location-scale equivariant estimator of scale with $\mathscr{P}(T) \subset \mathscr{P}(M)$ (of location with $\mathscr{P}(T) \subset \mathscr{P}(S)$), then a *factorized variant* M_T of M (S_T of S) defined by

$$M_T(Q) = M(Q[0, T(Q)])\, T(Q) \quad (S_T(Q) = S(Q[T(Q), 1])) \tag{6.4}$$

is location-scale equivariant estimator of location with $\mathscr{P}(M_T) = \mathscr{P}(M)$ (of scale with $\mathscr{P}(S_T) = \mathscr{P}(S)$).

(Received September 9, 1983.)

REFERENCES

[1] J. Aczél: Lectures on Functional Equations and Their Applications. Academic Press, New York 1966.

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

[3] D. E. Boekee: The D_f-information of order s. In: Trans. 8th Prague Conf. on Inform. Theory, etc., Vol. C, Academia, Prague 1979, 55—68.

[4] D. D. Boos: Minimum distance estimators for location and goodness of fit. J. Amer. Statist. Assoc. 76 (1981), 663—670.

[5] I. Csiszár: Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publ. Math. Inst. Hungar. Acad. Sci. Ser. A *8* (1963), 85—108.

[6] I. Csiszár: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. *2*, (1967) 209—318.

[7] H. Cramér: Mathematical Methods of Statistics. Princeton Univ. Press, Princeton, N. J. 1946.

[8] P. I. Huber: Robust estimation of a location parameter. Ann. Math. Statist. *35* (1964), 73—101.

[9] P. J. Huber: Robust statistics: a review. Ann. Math. Statist. *43* (1972), 1041—1067.

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

[11] L. Le Cam: On the information contained in additional observations. Ann. Statist. *2* (1974), 630—649.

[12] R. Š. Lipcer and A. N. Širjaev: Statistics of Random Processes (in Russian). Nauka, Moscow 1974.

[13] P. W. Millar: Robust estimation via minimum distance methods. Z. Wahrsch. verw. Gebiete *55* (1981), 73—89.

[14] A. M. Mood, F. A. Graybill and D. C. Boes: Introduction to the Theory of Statistics. McGraw-Hill, New York 1963.

[15] J. Neyman: Contributions to the theory of χ^2-test. In: Proc. 1st Berkeley Symp. on Math. Statist., etc., Univ. of Calif. Press, Berkeley 1949, 239—273.

[16] W. C. Parr and W. R. Schucany: Minimum distance and robust estimation. J. Amer. Statist. Assoc. *75* (1980), 616—624.

[17] C. R. Rao: Asymptotic efficiency and limiting information. In: Proc. 4th Berkeley Symp. on Math. Statist., etc., Vol. 1, Univ. of Calif. Press, Berkeley 1961, 531—546.

[18] C. R. Rao: Criteria of estimation in large samples. Sankhya *25* (1963), 189—206.

[19] P. V. Rao et al.: Estimation of shift and center of symmetry based on Kolmogorov-Smirnov statistic. Ann. Statist. *3* (1975), 862–873.
[20] I. Vajda: Limit theorems for total variation of Cartesian product measures. Studia Sci. Math. Hungar. *6* (1971), 317–333.
[21] I. Vajda: On the f-divergence and singularity of probability measures. Period. Math. Hungar. *2* (1972), 223–234.
[22] I. Vajda: χ^α-divergence and generalized Fisher information. In: Trans. 6th Prague Conf. on Inform. Theory, etc., Academia, Prague 1973, 873–886.
[23] I. Vajda: Theory of Information and Statistical Decision (in Slovak), Alfa, Bratislava 1981.
[24] I. Vajda: A new general approach to minimum distance estimation In: Trans. 9th Prague Conf. on Inform. Theory, etc., Vol. C, Academia, Prague 1983.
[25] I. Vajda: Minimum divergence principle in statistical estimation. Statistics and Decisions (submitted).
[26] M. Vošvrda: On second order efficiency of minimum divergence estimators. In: Trans. 9th Prague Conf. on Inform. Theory, etc., Vol. C, Academia, Prague 1983.
[27] J. Wolfowitz: The minimum distance method. Ann. Math. Statist. *28* (1957), 75–88.
[28] P. W. Zehna: Invariance of maximum likelihood estimation. Ann. Math. Statist. *37* (1966), 755.

Ing. Igor Vajda, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Prague 8. Czechoslovakia.