Applications of Mathematics

María Luisa Menéndez; Domingo Morales; Leandro Pardo; Igor Vajda
Minimum disparity estimators for discrete and continuous models

Applications of Mathematics, Vol. 46 (2001), No. 6, 439–466

Persistent URL: http://dml.cz/dmlcz/134477

MINIMUM DISPARITY ESTIMATORS FOR DISCRETE AND CONTINUOUS MODELS*

M. Menéndez, Madrid, D. Morales, Elche, L. Pardo, Madrid and I. Vajda, Praha

(Received June 1, 1999)

Abstract. Disparities of discrete distributions are introduced as a natural and useful extension of the information-theoretic divergences. The minimum disparity point estimators are studied in regular discrete models with i.i.d. observations and their asymptotic efficiency of the first order, in the sense of Rao, is proved. These estimators are applied to continuous models with i.i.d. observations when the observation space is quantized by fixed points, or at random, by the sample quantiles of fixed orders. It is shown that the random quantization leads to estimators which are robust in the sense of Lindsay [9], and which can achieve the efficiency in the underlying continuous models provided these are regular enough.

Keywords: divergence, disparity, minimum disparity estimators, robustness, asymptotic efficiency

MSC 2000: 62B10, 62E20

1. Introduction

This paper deals with the minimum distance point estimation with i.i.d. observations in the case when model is discrete, or when the initial information about the data and hypothetical parametrized models is reduced by partitioning the observation space. The distance is in both cases measured by a disparity (divergence) between hypothetical and empirical distributions. Partitioning is sometimes practical because it reduces the numerical complexity of estimation. Often data themselves

* This work was supported by grants A1075101 of the Academy of Sciences of the Czech Republic, DGES PB 96–0635 and GV99-159-1-01.

are grouped into classes satisfying various easily verifiable criteria, e.g. in the econometry and sociometry. Partitioning also allows one to use distances not applicable to the unreduced data and models. For example, the minimum Pearson divergence estimator was in this sense used by Neyman [13], or the maximum likelihood estimator (MLE) was obtained by minimizing the information divergence of Kullback in Rao [15].

The MLE is known to be efficient in regular models but is also known to be nonrobust. The efficiency as well as the nonrobustness are resulting from specific properties of the logarithmic function used in the definition of the information divergence. Replacing the logarithmic function by other functions $\varphi(t)$ with appropriate properties in the neighbourhood of $t = 1$, one obtains estimators which are in the discrete models efficient (first order, in the sense of Rao [15], [16]) and robust in the sense of Lindsay [9].

Lindsay paid special attention to the family of functions

$$\varphi_b(t) = \left(\frac{t-1}{1-b+b\sqrt{t}}\right)^2, \quad 0 \leqslant b \leqslant 1, \tag{1.1}$$

leading for $b = 0$ and $b = 1$ to the Pearson and Neyman divergences ([22], [13], see also [17]), and for $0 < b < 1$ to the so-called φ_b-disparities. In this paper we consider the classes $\Phi_{\rm div}$ and $\Phi_{\rm disp} \supset \Phi_{\rm div}$ of functions $\varphi(t)$ which define φ-divergences and φ-disparities leading in the discrete parametrized models to consistent and efficient φ-estimators of the true parameter. In this respect we extend the results of Morales, Pardo and Vajda [12] restricted to the minimum φ-divergence estimators in discrete parametrized models. We also apply the minimum φ-disparity estimators to the continuous models and study their efficiency in these models. In this respect we go considerably beyond the framework of the paper of Lindsay [9] and other papers in this area (see [1], [2], [14]).

Minimum φ-divergence and φ-disparity estimators (briefly, φ-estimators) are in Sections 3 and 4 applied to the discrete models $p(\theta) = (p_j(\theta)\colon\ 1 \leqslant j \leqslant m)$, $\theta \in \Theta$, in particular to the models

$$p(\boldsymbol{y},\theta) = \big(p_j(\boldsymbol{y},\theta) = F(y_j,\theta) - F(y_{j-1},\theta)\colon\ 1 \leqslant j \leqslant m\big), \quad \theta \in \Theta, \tag{1.2}$$

obtained from the continuous models $(F(x,\theta)\colon\ x \in R)$, $\theta \in \Theta$, by partitioning the observation space R by given points

$$\boldsymbol{y} = (y_1,\dots,y_{m-1}), \quad y_0 = -\infty < y_1 < \dots < y_{m-1} < \infty = y_m. \tag{1.3}$$

This is the *deterministic partition* of R specified by the vector $\boldsymbol{y}$. In Section 5 attention is paid to the *random partitions* of R by the components of the random vector

$$\boldsymbol{y}_n = (y_{nj} = F_n^{-1}(\lambda_j)\colon\ 1 \leqslant j \leqslant m), \tag{1.4}$$

where $F_n(x)$, $x \in R$, is the empirical distribution function and

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m-1}), \quad \lambda_0 = 0 < \lambda_1 < \ldots < \lambda_{m-1} < 1 = \lambda_m. \tag{1.5}$$

This partition is specified by the vector $\boldsymbol{\lambda}$ and the observed i.i.d. data.

The results of the present paper concerning the asymptotic properties of φ-disparity estimators in the models of all three types mentioned extend similar results in the above mentioned papers, and also in [11] and other papers cited there, dealing with various particular φ-divergence estimators. Moreover, we show that both the deterministic and random partitions lead to estimators which are ε-efficient in the original model $(F(x,\theta)\colon\ \theta \in \Theta)$ for any given $\varepsilon > 0$, and that the estimators with random partitions can achieve this efficiency without any a priori information about the true parameter. It is important that, to this end, one can use any of the φ-disparity estimators, including all those which are robust in the sense of Lindsay [9]. This means that, in fact, we introduce an infinite class of estimation procedures which are robust and, from the practical point of view, efficient for all sufficiently regular continuous models with i.i.d. observations.

Applicability of the present results in testing composite hypotheses about continuous statistical results can be seen from [10].

2. φ-DIVERGENCES AND φ-DISPARITIES

By a stochastic vector we mean a vector with nonnegative components the sum of which is 1. By a stochastic m-vector we mean a stochastic vector with m components for $m > 1$.

The φ-*divergence* of arbitrary stochastic m-vectors p and q is defined by the formula

$$D_\varphi(p;q) = \sum_{j=1}^{m} q_j \varphi\left(\frac{p_j}{q_j}\right), \tag{2.1}$$

where φ is from the class Φ_{div} of all convex functions $\varphi(t)$, $t > 0$, equal to 0 at $t = 1$. For every $\varphi \in \Phi_{\mathrm{div}}$ differentiable at $t = 1$ we have

$$\varphi(t) \sim \varphi(t) - \varphi'(1)(t-1), \tag{2.2}$$

where the right hand side belongs to Φ_{div} and the equivalence $\sim$ means that the two functions define the same divergence (2.1).

Hereafter Φ_{div} stands for the subclass of the convex functions twice continuously differentiable in the neighbourhood of $t = 1$ with $\varphi(1) = 0$, $\varphi''(1) > 0$. Obviously, we can assume without loss of generality that $\varphi'(1) = 0$ and $\varphi''(1) = 1$ for every $\varphi \in \Phi_{\text{div}}$.

Note that the concept of φ-divergence was introduced by Csiszár [23] and Ali and Silvey [24], and that in (2.1) it is assumed

$$0\varphi\left(\frac{0}{0}\right) = 0, \quad q\varphi\left(\frac{0}{q}\right) = \varphi(0) \triangleq \lim_{t\to 0}\varphi(t) \quad \text{and} \quad 0\varphi\left(\frac{p}{0}\right) = p\lim_{t\to\infty}\frac{\varphi(t)}{t}.$$

For the properties of φ-divergences we refer to Liese and Vajda [8] or Vajda [20].

E x a m p l e 2.1. The nonnegative functions

$$\varphi_a(t) = \frac{t^{(a+1)/2} - \frac{1}{2}(a+1)(t-1) - 1}{(|a|-1)/2} \sim \frac{t^{(a+1)/2} - 1}{(|a|-1)/2}, \quad a \neq 1,$$

(cf. the equivalence relation $\sim$ in (2.2)) with limits

$$\varphi_1(t) = t\ln t - t + 1 \sim t\ln t \quad \text{and} \quad \varphi_{-1}(t) = -\ln t + t - 1 \sim -\ln t$$

have continuous and positive second derivatives

$$\varphi_a''(t) = \frac{|a|+1}{2} t^{(a-3)/2}, \quad a \in R.$$

They define a class of *modified power divergences*

$$D_a(p;q) = \frac{2}{|a|-1}\left(\sum_{j=1}^{m}\sqrt{p_j^{1+a}q_j^{1-a}} - 1\right) \quad \text{for all } a \neq -1,\ a \neq 1, \tag{2.3}$$

with the well known *information divergence* and *reversed information divergence* of Kullback

$$D_1(p;q) = \sum_{j=1}^{m} p_j \ln\frac{p_j}{q_j} \quad \text{and} \quad D_{-1}(p;q) = D_1(q;p) \tag{2.4}$$

obtained from φ_1 and φ_{-1}, or as the limits of $D_a(p;q)$ for $a \to 1$ and $a \to -1$. (The *skew symmetry* $D_{-a}(p;q) = D_a(q;p)$ for the remaining $a \in R$ is clear from (2.3).) Well known is also the *Pearson divergence*

$$D_3(p;q) = \sum_{j=1}^{m}\frac{p_j^2}{q_j} - 1 = \sum_{j=1}^{m}\frac{(p_j - q_j)^2}{q_j}, \tag{2.5}$$

the reversed Pearson divergence (*Neyman divergence*) $D_{-3}(p;q)$, and the *Hellinger divergence* (squared Hellinger distance)

$$D_0(p;q) = 2\Bigl(1 - \sum_{j=1}^{m} \sqrt{p_j q_j}\Bigr) = \sum_{j=1}^{m} \bigl(\sqrt{p_j} - \sqrt{q_j}\bigr)^2.$$

The original power divergences of Cressie and Read [6] are 1-1 transforms of (2.3),

$$I_\lambda(p,q) = \frac{4D_{2\lambda+1}(p;q)}{|2\lambda+1|+1}, \quad \lambda \in R. \tag{2.6}$$

These divergences do not provide exactly the squared Hellinger distance (at $\lambda = -1/2$ they are proportional, with the factor 4). Also the skew symmetry about $\lambda = -1/2$ in this family seems to be less practical than that about 0 in the family (2.3). For example, it may be not easy to recognize at the first sight that $I_{-0.357}(p;q)$ means the same as $I_{-0.643}(q;p)$, while for $D_{-0.357}(p;q)$ and $D_{0.357}(q;p)$ this is easy. Note that both families (2.3) and (2.6) are one-one transforms of the α-divergences $R_\alpha(p;q)$, $\alpha > 0$, of Rényi [25]. E.g.,

$$\frac{\lambda(\lambda+1)}{2} I_\lambda(p;q) = \begin{cases} \exp\{\lambda R_{\lambda+1}(p;q)\} - 1 & \text{for } \lambda > -1, \\ \exp\{-(\lambda+1)R_{-\lambda}(q;p)\} - 1 & \text{for } \lambda \leqslant -1. \end{cases}$$

The φ-*disparity* is defined by the same formula (2.1) as the φ-divergence, but for much wider class of functions $\varphi \in \Phi_{\text{disp}}$. The class Φ_{disp} is assumed to contain all functions $\varphi(t)$, $t > 0$, twice continuously differentiable in the neighbourhood of $t = 1$, with $\varphi(1) = 0$, $\varphi''(1) > 0$, such that $\varphi_*(t) = \varphi(t) - \varphi'(1)(t-1)$ is monotone on the intervals (0,1) and $(1,\infty)$. Then, the function $\varphi_*(t)$ is nonincreasing on (0,1), nondecreasing on $(1,\infty)$ and strictly convex near $t = 1$. Especially, $\varphi_*(t)$ is unimodal with the mode (the minimum) at $t = 1$.

It is obvious that any $\varphi, \psi \in \Phi$ define the same disparity (2.1) if there exists $a \in R$ satisfying for all $t > 0$ the relation $\varphi(t) - \psi(t) = a(t-1)$, in symbols $\varphi \sim \psi$. Since any $\varphi \in \Phi_{\text{disp}}$ satisfies the relation (2.2), we can assume without loss of generality $\varphi'(1) = 0$ and the unimodality of all $\varphi \in \Phi_{\text{disp}}$.

Φ_{disp} contains the class Φ_{div} of convex functions introduced above. Convex as well as nonconvex functions $\varphi \in \Phi_{\text{disp}}$ can be generated by the composition formula

$$\varphi(t) = \psi(\varphi(t)), \quad t > 0, \tag{2.7}$$

under the conditions presented in the next two theorems.

Theorem 2.1. *If $\varphi(t)$, $t > 0$, is monotone and twice continuously differentiable in the neighbourhood of $t = 1$ with $\varphi(1) = 0$, $\varphi'(1) \neq 0$, and $\psi(y)$, $y \in R$, is* *monotone on the intervals* $(-\infty, 0]$ *and* $[0, \infty)$, *twice continuously differentiable in the neighbourhood of* $y = 0$ *with* $\psi(0) = \psi'(0) = 0$ *and* $\psi''(0) > 0$, *then the composite function* (2.7) *belongs to* Φ_{disp}.

P r o o f. Under the assumptions we obtain in the neighbourhood of 1

$$\varphi'(t) = \psi'(\varphi(t))\varphi'(t), \tag{2.8}$$

$$\varphi''(t) = \psi''(\varphi(t))(\varphi'(t))^2 + \psi'(\varphi(t))\varphi''(t). \tag{2.9}$$

Hence

$$\varphi(1) = \psi(\varphi(1)) = 0 \quad \text{and} \quad \varphi''(1) = \psi''(0)(\varphi'(1))^2 > 0.$$

If $\varphi'(t) \geqslant 0$ for all $t > 0$ then $\varphi(t)$ is nonnegative and nondecreasing for $t \geqslant 1$, so that $\psi(\varphi(t))$ is nondecreasing on $[1, \infty)$ and $\varphi(t)$ is nonpositive for $0 < t \leqslant 1$. Hence (2.7) is nonincreasing on $(0, 1]$. □

Theorem 2.2. *If* $\varphi \in \Phi_{\mathrm{disp}}$, $\varphi'(1) = 0$, *and* $\psi(y)$, $y \geqslant 0$, *is nondecreasing, twice continuously differentiable in the neighbourhood of* $y = 0$ *(from the right at* $y = 0$*) with* $\psi'_+(0) > 0$, *then* $\varphi(t) = \psi(\varphi(t))$, $t > 0$, *belongs to* Φ_{disp}.

P r o o f. Under the stated assumptions, $\varphi(t)$ is nonincreasing on $(0, 1]$ and nondecreasing on $[1, \infty)$ with $\varphi(1) = 0$. This implies the desired monotonicity of $\psi(\varphi(t))$ and $\psi(\varphi(1)) = 0$. The differentiability in the neighbourhood of $t = 1$ follows from (2.8), (2.9) if $t \neq 1$. If $t = 1$ then it follows by taking limits for $t \downarrow 1$ and $t \uparrow 1$ on both sides of (2.8), (2.9). □

It is easy to see that $\psi(y) = 1 - e^{-y^2}$ satisfies the assumptions of Theorem 2.1 and $\psi(y) = 1 - e^{-y}$ the assumptions of Theorem 2.2. In both cases the application of ψ to $\varphi(p_j/q_j)$ behind the sum in (2.1) reduces the influence of extremal deviations of q_j from p_j. Next an example follows where $\psi(y)$ enhances the influence of such deviations.

E x a m p l e 2.2. It is easy to verify that the functions

$$\varphi_b(t) = \frac{t-1}{1-b+b\sqrt{t}}, \quad 0 \leqslant b \leqslant 1,$$

are convex and twice continuously differentiable in the domain $t > 0$ with $\varphi_b'(1) \neq 0$, i.e. they belong to Φ_{div} and satisfy the assumptions of Theorem 2.1. Since $\psi(y) = y^2$ satisfies the assumptions of Theorem 2.1 as well, it follows that the functions

$$\varphi_b(t) = \psi(\varphi_b(t)) = \left(\frac{t-1}{1-b+b\sqrt{t}}\right)^2, \quad 0 \leqslant b \leqslant 1,$$

belong to $\Phi_{\rm disp}$. Somewhat tedious calculations show that $\varphi_b'(1) = 0$ and that the second derivatives are positive everywhere with $\varphi_b''(1) = 2$. The corresponding φ_b-disparities are (cf. p. 1101 in [9])

$$\mathcal{D}_b(p;q) = \sum_{j=0}^{m} \left(\frac{p_j - q_j}{(1-b)\sqrt{q_j} + b\sqrt{p_j}} \right)^2, \quad 0 \leqslant b \leqslant 1. \tag{2.10}$$

Here $\mathcal{D}_0(p;q)$ is the Pearson divergence $D_3(p;q)$ from Example 2.1 and $\mathcal{D}_1(p;q)$ is the Neyman divergence $D_{-3}(p;q)$. Thus the disparities (2.10) are *blends* of the Pearson and Neyman divergences. The symmetric blend $\mathcal{D}_{1/2}(p;q)$ is four times the Hellinger divergence $D_0(p;q)$ from Example 2.1.

Note that the replacement of the square function in Example 2.2 by $\psi(y) = 1 - e^{-y^2}$ leads to functions $\varphi_b(t)$ which still belong to $\Phi_{\rm disp}$ but are not convex. Indeed, the functions $\varphi(t)$ convex in the domain $t > 0$ with $\varphi'(1) = 0$ and $\varphi''(1) > 0$ cannot be bounded. In general, for convex φ and nondecreasing ψ the composition (2.7) is Schur-convex (see [26]). The disparities (2.1) with Schur-convex functions φ possess some, but not all, the nice properties of φ-divergences.

Some properties of φ-divergences hold for all φ-disparities. For example, the inequality

$$D_\varphi(p;q) \geqslant 0,$$

with the sign of equality if and only if $p = q$, remains true for every $\varphi \in \Phi_{\rm disp}$.

Conventions. Recall that by a stochastic vector we mean in this paper a vector with nonnegative coordinates the sum of which is one. If the components of this vector are random, but nonnegative and summing up to one, then we speak about a random stochastic vector. Unless otherwise explicitly stated, m is in the sequel assumed to be fixed and all convergences and asymptotic relations including $o(1)$ and $o_p(1)$ are considered for $n \to \infty$. If a sequence of random variables Z_n satisfies the asymptotic relation $Z_n = o_p(1)$, i.e. if $P(|Z_n| > \varepsilon) = o(1)$ for any $\varepsilon > 0$, then we say that Z_n tends stochastically to zero.

Theorem 2.3. *Let $p_n = (p_{n1}, \ldots, p_{nm})$ be a sequence of random stochastic vectors. If*

$$D_\varphi(p_n;q) = o_p(1) \tag{2.11}$$

for a fixed stochastic m-vector q with all coordinates positive and for $\varphi \in \Phi$, then $\|p_n - q\|^2$ tends stochastically to zero with at least the same rate as $D_\varphi(p_n;q)$ or, more precisely,

$$\|p_n - q\|^2 \leqslant \frac{2}{\varphi''(1)} D_\varphi(p_n;q) + o_p(D_\varphi(p_n;q)). \tag{2.12}$$

P r o o f. (I) First we prove that (2.11) implies

$$\|p_n - q\| = o_p(1). \tag{2.13}$$

Let $U \subset (0, \infty)$ be an open neighbourhood of 1 where φ is twice continuously differentiable and where the infimum

$$\alpha = \inf_{t \in U} \varphi''(t)$$

is positive. We can assume without loss of generality $\varphi'(1) = 0$ so that φ is unimodal with the unique mode at $t = 1$. This together with $\alpha > 0$ implies

$$\beta = \inf_{t \notin U} \varphi(t) > 0.$$

Define random variables

$$Z_n = \sum_{j=0}^{m} I\left\{\frac{p_{nj}}{q_j} \notin U\right\} \quad \text{and} \quad W_n = \prod_{j=0}^{m} I\left\{\frac{p_{nj}}{q_j} \in U\right\},$$

where $I\{\cdot\}$ denotes the indicator of an event. If $\varphi \in \Phi$ then the Taylor theorem implies for all t from the neighbourhood of 1

$$\varphi(t) = \frac{1}{2}\varphi''(t^*)(t-1)^2, \tag{2.14}$$

where t^* is between 1 and t. By (2.1), (2.5) and (2.14),

$$D_\varphi(p_n; q) \geqslant Z_n \beta \min_{0 \leqslant j \leqslant m} q_j + W_n \alpha D_3(p_n; q).$$

Hence (2.11) implies $Z_n = o_p(1)$ and, since $W_n \neq 1$ takes place only if $Z_n \geqslant 1$, also $W_n = 1 + o_p(1)$. Using this one obtains that (2.11) implies $D_3(p_n; q) = o_p(1)$ which, by virtue of the inequality

$$D_3(p_n; q) \geqslant \|p_n - q\|^2, \tag{2.15}$$

implies (2.13).

(II) It follows from (2.1) and (2.14) that under (2.13) we have

$$D_\varphi(p_n; q) = \frac{\varphi''(1)}{2} D_3(p_n; q) + \|p_n - q\|^2 o_p(1),$$

and (2.12) follows from here and from (2.15). □

3. Discrete models

Consider a theoretical discrete model $p(\theta) = ((p_1(\theta), \dots, p_m(\theta)),\ \theta \in \Theta)$, $\Theta \subset R^s$, and independent random data $(x_i\colon\ 1 \leqslant i \leqslant n)$ taking on values $1, \dots, m$ with probabilities $p_1(\theta_0), \dots, p_m(\theta_0)$. Here $\theta_0 \in \Theta$ is a true parameter. Let p_n be the vector of relative frequencies of $1, \dots, m$ defining an empirical discrete model,

$$p_n = \left(p_{nj} = \frac{1}{n} \sum_{i=1}^{n} I_{\{j\}}(x_i)\colon\ 1 \leqslant j \leqslant m \right). \tag{3.1}$$

Further, consider $\varphi \in \Phi_{\mathrm{disp}}$ and the disparity $D_\varphi(p_n; p(\theta))$ for $\varphi \in \Phi_{\mathrm{disp}}$—a disparity between the empirical and theoretical models.

If a Θ-valued statistics $\theta_{\varphi,n} = \theta_{\varphi,n}(x_1, \dots, x_n)$ asymptotically minimizes the φ-disparity $D_\varphi(p_n; p(\theta))$ in the sense that

$$P\Big\{ D_\varphi(p_n; p(\theta_{\varphi,n})) \neq \inf_{\theta \in \Theta} D_\varphi(p_n; p(\theta)) \Big\} = o(1), \tag{3.2}$$

then it is called a *φ-disparity estimator* of θ_0 (briefly, φ-estimator).

Let $\widehat{\theta}_n$ be the MLE in the discrete model under consideration. As is easy to see, then (3.2) with $\varphi(t) = t \ln t$ holds for $\theta_{\varphi,n}$ replaced by $\widehat{\theta}_n$. Thus $\widehat{\theta}_n$ is the minimum information divergence estimator $\theta_{\varphi_1,n}$ (cf. the family $\{\varphi_a\colon\ a \in R\} \subset \Phi_{\mathrm{div}}$ introduced in Example 2.1). Birch [3] formulated regularity assumptions for the model under which the MLE is for this model efficient in the sense made precise in Theorem 3.1 below (it is the so-called first order efficiency in the sense of Rao [15], considered throughout the paper unless otherwise explicitly stated).

Special φ-disparity estimators can be found in many papers and books, e.g. in [7], [5], [13], [21], [15], [18], [20], [9], [1], [2] and [14]. By using the relative deviations

$$\delta_{nj}(\theta) = \frac{p_{nj} - p_j(\theta)}{p_j(\theta)}, \quad 1 \leqslant j \leqslant m,$$

of data from the models (the *residuals* of Lindsay [9]), one can transform the criterion function $D_\varphi(p_n; p(\theta))$ from (3.2) into the form

$$M_n(\theta) = \sum_{j=1}^{m} p_{nj}\, \varrho(\delta_{nj}(\theta)), \tag{3.3}$$

where

$$\varrho(t) = \frac{\varphi(1+t)}{t+1}, \quad t > -1, \qquad \text{with } \varrho(0) = 0$$

is continuous and (when assuming that φ is convex with $\varphi'(1) = 0$) decreasing on $(-1, 0]$ and increasing on $[0, \infty)$. This displays certain affinity with the M-estimators of mathematical statistics. The estimators which minimize the criterion function (3.3) have been studied by Lindsay [9]. He argued that these estimators are *robust* if the function φ is bounded.

Next we introduce the regularity assumptions of Birch [3] in a slightly modified form, more convenient for the purpose of the present paper.

(R1) The true θ_0 is in the interior of Θ and all coordinates of $p(\theta_0)$ are positive.

(R2) The gradient matrix $G(\theta) = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_s)p(\theta)^t$ exists in the neighbourhood of $\theta = \theta_0$ and is continuous there.

Under (R1) and (R2) also the matrix function $A(\theta) = \operatorname{diag} p(\theta_0)^{-1/2}G(\theta)$ exists in the neighbourhood of $\theta = \theta_0$ and is continuous there. Note that for any k-vector p and mapping $\psi\colon\ R \mapsto R$, $\operatorname{diag}\psi(p)$ in this paper denotes the diagonal $(k \times k)$ matrix with the entries $\psi(p_1), \ldots, \psi(p_k)$ at the diagonal.

(R3) The matrix $A(\theta_0)$ is of rank s and $s < m$.

(R4) The mapping $\theta \mapsto p(\theta)$ is one-one on Θ.

Under (R1)–(R3) the matrix

$$I(\theta) = A(\theta)^t A(\theta) = \left(\sum_{j=1}^{m} \frac{1}{p_j(\theta)} \frac{\partial p_j(\theta)}{\partial\theta_k} \frac{\partial p_j(\theta)}{\partial\theta_\ell}\right)_{k,\ell=1}^{s} \tag{3.4}$$

is well defined, continuous and positive definite in the neighbourhood of $\theta = \theta_0$. It is the *Fisher information matrix* of the model $(p(\theta)\colon\ \theta \in \Theta)$.

Theorem 3.1. *If the model $(p(\theta)\colon\ \theta \in \Theta)$ under consideration and the true parameter θ_0 fulfil* (R1)–(R4) *then the estimators $\theta_{\varphi,n}$, $\varphi \in \Phi_{\text{disp}}$, defined by* (3.2) *are efficient in the sense that they satisfy the asymptotic relation*

$$\theta_{\varphi,n} = \theta_0 + (p_n - p(\theta_0)) \operatorname{diag} p(\theta_0)^{-1/2} A(\theta_0) I(\theta_0)^{-1} + o_p(n^{-1/2}), \tag{3.5}$$

and they are also asymptotically normal in the sense that

$$\sqrt{n}(\theta_{\varphi,n} - \theta_0) \xrightarrow{w} N\bigl(0, I(\theta_0)^{-1}\bigr). \tag{3.6}$$

Proof. For φ from the subclass $\Phi_{\text{div}} \subset \Phi_{\text{disp}}$ of divergence-generating functions the relations (3.5) and (3.6) were proved in Theorem 3 of [12]. By inspecting the arguments used there one can see that the reference to Proposition 9.49 in [20], which is sufficient for $\varphi \in \Phi_{\text{div}}$, can in the case $\varphi \in \Phi_{\text{disp}}$ be replaced by the reference to the present Theorem 2.3. The remaining steps are based on the local properties of $\varphi(u) \in \Phi_{\text{div}}$ in the neighbourhood of $u = 1$, which are assumed for $\varphi \in \Phi_{\text{disp}}$ as well.

The results of the previous section are easily extended to the models $(p(\boldsymbol{y},\theta)\colon \theta \in \Theta)$ defined by (1.2) with $\Theta \subset R^s$. Let the observations $(x_i\colon 1 \leqslant i \leqslant n)$ be i.i.d. on R with the common distribution function $F(x,\theta_0)$ assumed in (1.2) and let

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[x_i,\infty)}(x), \quad x \in R$$

be the empirical distribution function on R. Define an empirical probability vector

$$p_n(\boldsymbol{y}) = (p_{nj}(\boldsymbol{y}) = F_n(y_j) - F_n(y_{j-1})\colon 1 \leqslant j \leqslant m) \tag{4.1}$$

for $\boldsymbol{y}$ given by (1.3) and consider $D_\varphi(p_n(\boldsymbol{y}); p(\boldsymbol{y},\theta))$ for $\varphi \in \Phi_{\mathrm{disp}}$ and $\theta \in \Theta$.

Obviously, $p_n(\boldsymbol{y})$ can formally be defined by (3.1) with x_i replaced by the new random variables

$$\widetilde{x}_i = \widetilde{x}_i(\boldsymbol{y}, x_i) = \sum_{j=1}^{m} j I_{[y_{j-1},y_j)}(x_i)$$

taking on values from the sample space $\{1,\dots,m\}$ and indicating the random events $y_{j-1} \leqslant x_i < y_j$. Obviously, $(\widetilde{x}_1,\dots,\widetilde{x}_m)$ are sufficient statistics of the samples $(x_1,\dots,x_n)$ for the quantized continuous models $(p(\boldsymbol{y},\theta)\colon \theta\in\Theta)$.

A Θ-valued statistics $\theta_{\varphi,n} = \theta_{\varphi,n}(x_1,\dots,x_n)$ satisfying the condition

$$P\Big\{D_\varphi(p_n(\boldsymbol{y}); p(\boldsymbol{y},\theta_{\varphi,n})) \neq \inf_{\theta\in\Theta} D_\varphi(p_n(\boldsymbol{y}); p(\boldsymbol{y},\theta))\Big\} = o(1) \tag{4.2}$$

is called a φ-*disparity estimator* of θ_0 (briefly, φ-estimator). Since the random function $D_\varphi(p_n(\boldsymbol{y}); p(\boldsymbol{y},\theta))$ of the variable $\theta\in\Theta$ figuring in (4.1) depends only on the sample statistic $(\widetilde{x}_1,\dots,\widetilde{x}_m)$, the estimator $\theta_{\varphi,n}$ is in fact a function of this statistics as well, i.e. $\theta_{\varphi,n} = \theta_{\varphi,n}(\widetilde{x}_1,\dots,\widetilde{x}_m)$.

The quantized continuous model $(p(\boldsymbol{y},\theta)\colon \theta\in\Theta)$ under consideration is a particular case of the discrete model $(p(\theta)\colon \theta\in\Theta)$, $\Theta\subset R^s$, of Section 3. For this model the Birch regularity conditions (R1)–(R4) reduce with help of the notation

$$F(\boldsymbol{y},\theta) = (F(y_1,\theta),\dots,F(y_{m-1},\theta)) \tag{4.3}$$

as follows.

(B1) The true θ_0 is in the interior of Θ and the partition vector $\boldsymbol{y}$ of (1.3) satisfies the condition $F(y_j,\theta_0) \neq F(y_{j-1},\theta_0)$ for all $1\leqslant j\leqslant m$.

(B2) The gradient matrix $\Gamma(\boldsymbol{y},\theta) = (\partial/\partial\theta_1,\dots,\partial/\partial\theta_s)F(\boldsymbol{y},\theta)^t$ exists in the neighbourhood of θ_0 and is continuous there.

Under (B1), (B2) also the matrix functions $G(\boldsymbol{y},\theta) = (\partial/\partial\theta_1,\dots,\partial/\partial\theta_s)p(\boldsymbol{y},\theta)^t$ and $A(\boldsymbol{y},\theta) = \operatorname{diag} q^{-1/2} G(\boldsymbol{y},\theta)$ exist and are continuous in the neighbourhood of $\theta = \theta_0$.

(B3) The matrix $A(\boldsymbol{y},\theta_0)$ is of rank s and $s < m$.

(B4) The mapping $\theta \mapsto F(\boldsymbol{y},\theta)$ is one-one on Θ.

Obviously, (B1)–(B3) imply (R1)–(R3) for $p(\theta) = p(\boldsymbol{y},\theta)$, $\theta \in \Theta$. The validity of the implication (B4) $\Rightarrow$ (R4) for this particular model will be proved with help of the following lemma.

Lemma 4.1. *Let $\Delta = (\Delta_j\colon\ 1 \leqslant j \leqslant m)$ be the vector of the partial sums*

$$\Delta_j = \sum_{k=1}^{j} \delta_k$$

of coordinates of any vector $\delta = (\delta_j\colon\ 1 \leqslant j \leqslant m) \in R^m$. Then

$$\frac{\|\delta\|}{2} \leqslant \|\Delta\| \leqslant m\|\delta\|.$$

P r o o f. Define $\Delta_0 = 0$ and consider any $1 \leqslant j \leqslant m$. Since

$$(\delta_j + 2\Delta_{j-1})^2 = \delta_j^2 + 4\delta_j\Delta_{j-1} + 4\Delta_{j-1}^2 = 2[(\delta_j + \Delta_{j-1})^2 + \Delta_{j-1}^2] - \delta_j^2$$

is nonnegative, we have

$$\delta_j^2 \leqslant 2\Delta_j^2 + \Delta_{j-1}^2.$$

Therefore

$$\sum_{j=1}^{m} \delta_j^2 \leqslant 4\sum_{j=1}^{m} \Delta_j^2$$

so that $\|\delta\| \leqslant 2\|\Delta\|$. Further, by the Jensen inequality,

$$\left(\frac{1}{j}\sum_{k=1}^{j} \delta_k\right)^2 \leqslant \frac{1}{j}\sum_{k=1}^{j} \delta_k^2 \leqslant \frac{1}{j}\|\delta\|^2,$$

so that $\Delta_j^2 \leqslant j\|\delta\|^2$. Therefore

$$\|\Delta\| \leqslant \sqrt{\frac{m(m+1)}{2}}\|\delta\| \leqslant m\|\delta\|,$$

which completes the proof. $\square$

Corollary 4.1. *For arbitrary vectors* $\boldsymbol{y}, \widetilde{\boldsymbol{y}}$ *of the type* (1.3) *and arbitrary* θ_0, $\theta \in \Theta$,

$$\frac{\|p(\boldsymbol{y},\theta_0) - p(\widetilde{\boldsymbol{y}},\theta)\|}{2} \leqslant \|F(\boldsymbol{y},\theta_0) - F(\widetilde{\boldsymbol{y}},\theta)\| \leqslant m\|p(\boldsymbol{y},\theta_0) - p(\widetilde{\boldsymbol{y}},\theta)\|.$$

P r o o f. Clear from Lemma 4.1 if we put

$$\delta_j = p(\boldsymbol{y},\theta_0) - p_j(\widetilde{\boldsymbol{y}},\theta)$$

so that

$$\Delta_j = F(y_j,\theta_0) - F(\widetilde{y}_j,\theta).$$

□

The implication (B4) $\Rightarrow$ (R4) now follows from Corollary 4.1 used for $\widetilde{\boldsymbol{y}} = \boldsymbol{y}$. Indeed, we see that, for any $\boldsymbol{y}$ considered in (1.3) and $\theta_1, \theta_2 \in \Theta$,

$$F(\boldsymbol{y},\theta_1) - F(\boldsymbol{y},\theta_2) \neq 0 \quad \text{if and only if } p(\boldsymbol{y},\theta_1) - p(\boldsymbol{y},\theta_2) \neq 0.$$

Since (B1)–(B4) imply (R1)–(R4) for the model $p(\theta) = p(\boldsymbol{y},\theta)$, the following result can be deduced from Theorem 3.1. In this result we use the Fisher information matrix

$$I(\boldsymbol{y},\theta) = A(\boldsymbol{y},\theta)^t A(\boldsymbol{y},\theta) = \left(\sum_{j=1}^{m} \frac{1}{p_j(\boldsymbol{y},\theta)} \frac{\partial p_j(\boldsymbol{y},\theta)}{\partial \theta_k} \frac{\partial p_j(\boldsymbol{y},\theta)}{\partial \theta_\ell} \right)_{k=1,\ell=1}^{m} \tag{4.4}$$

of the model $(p(\boldsymbol{y},\theta)\colon\ \theta \in \Theta)$, which is under (B1)–(B3) well defined and positive definite in the neighbourhood of $\theta = \theta_0$.

Theorem 4.1. *If the quantized continuous model* $(p(\boldsymbol{y},\theta)\colon\ \theta \in \Theta)$ *and the true parameter* θ_0 *fulfil* (B1)–(B4) *then the estimators* $\theta_{\varphi,n}$, $\varphi \in \Phi_{\text{disp}}$, *defined by* (4.2) *are efficient in the sense that*

$$\theta_{\varphi,n} = \theta_0 + (p_n(\boldsymbol{y}) - p(\boldsymbol{y},\theta_0)) \operatorname{diag} p(\boldsymbol{y},\theta_0)^{-1/2} A(\boldsymbol{y},\theta_0) I(\boldsymbol{y},\theta_0)^{-1} + o_p\big(n^{-1/2}\big) \tag{4.5}$$

and asymptotically normal in the sense that

$$\sqrt{n}(\theta_{\varphi,n} - \theta_0) \xrightarrow{w} N(0, I(\boldsymbol{y},\theta_0)^{-1}). \tag{4.6}$$

The estimation method characterized by Theorem 4.1. is applicable to an arbitrary continuous model

$$\mathcal{P} = (F(x,\theta)\colon\ \theta \in \Theta), \quad \Theta \subset R^s. \tag{4.7}$$

It consists in the quantization of the observation space by $\boldsymbol{y}$ from (1.3) and a subsequent application of the φ-disparity estimator. In the rest of the section we study the efficiency of this method in the model (4.7). As is well known, if this model satisfies certain regularity assumptions in the neighbourhood of $\theta = \theta_0$ then a positive definite Fisher information $s \times s$ matrix $\mathcal{J}(\theta)$ exists in this neighbourhood and the efficient estimators of θ_0 in this model achieve the asymptotic covariance matrix $\mathcal{J}(\theta_0)^{-1}$. Thus the trace

$$\operatorname{tr} \mathcal{J}(\theta_0)^{-1} = \frac{1}{\operatorname{tr} \mathcal{J}(\theta_0)}$$

characterizes the least asymptotic variance achievable in a reasonably wide class of possible estimators of θ_0 in the model (4.7). Theorem 4.1 shows that, under an appropriate regularity, all φ-disparities estimators of θ_0 in the quantized version (1.2) of the model (4.7) achieve the asymptotic variance $\operatorname{tr} I(\boldsymbol{y}, \theta_0)^{-1}$. Therefore the quantity

$$\mathcal{S}(\boldsymbol{y}, \theta) = \operatorname{tr}(\mathcal{J}(\theta) - I(\boldsymbol{y}, \theta))$$

can serve as a *measure of subefficiency* at $\theta \in \Theta$ of the estimation method studied in this section.

In the sequel we explicitly denote the dependence of the information (4.4) on the partition size m by writing $I_m(\boldsymbol{y}, \theta)$ and $\mathcal{S}_m(\boldsymbol{y}, \theta)$ instead of $I(\boldsymbol{y}, \theta)$ and $\mathcal{S}(\boldsymbol{y}, \theta)$. For simplicity we restrict ourselves to the univariate parameters θ, i.e. we assume that $\Theta \subset R$ and, consequently,

$$(4.8) \qquad I_m(\boldsymbol{y}, \theta) = \sum_{j=1}^{m} \frac{\dot{p}_j(\boldsymbol{y}, \theta)^2}{p_j(\boldsymbol{y}, \theta)} \quad \text{and} \quad \mathcal{S}_m(\boldsymbol{y}, \theta) = \mathcal{J}(\theta) - I_m(\boldsymbol{y}, \theta),$$

where $\dot{p}_j(\boldsymbol{y}, \theta) = \mathrm{d}p_j(\boldsymbol{y}, \theta)/\mathrm{d}\theta$.

Our regularity conditions (B1)–(B4), guaranteeing the existence of informations $I_m(\boldsymbol{y}, \theta_0)$ for vectors $\boldsymbol{y}$ of (1.3), do not imply the existence of the information $\mathcal{J}(\theta_0)$. The first question is, therefore, when the informations $\mathcal{J}(\theta)$, $\theta \in \Theta$, exist and whether $\mathcal{J}(\theta_0)$ is always greater than the information $I_m(\boldsymbol{y}, \theta_0)$ in the quantized models (1.2).

We shall consider conditions for the existence of the Fisher information $\mathcal{J}(\theta_0)$, $\theta_0 \in \Theta$, in the model (4.7) introduced in [19] (condition $\mathcal{C}_2$ on p. 280 ibid.), namely that the densities $f(x, \theta) = \mathrm{d}F(x, \theta)/\mathrm{d}x$ and their derivatives $\dot{f}(x, \theta) = \mathrm{d}f(x, \theta)/\mathrm{d}\theta$ exist at θ_0 for almost all x, and for some $\varepsilon > 0$ (possibly depending on θ_0)

$$(4.9) \qquad \int \sup_{|\theta - \theta_0| < \varepsilon} \left(\frac{f(x, \theta) - f(x, \theta_0)}{(\theta - \theta_0) f(x, \theta_0)} \right)^2 f(x, \theta_0) \, \mathrm{d}x < \infty.$$

Under this condition

$$\mathcal{J}(\theta_0) = \int \frac{\dot{f}(x,\theta_0)^2}{f(x,\theta_0)}\,\mathrm{d}x < \infty \tag{4.10}$$

is finite—it is the *Fisher information in the continuous model* (4.7).

E x a m p l e 4.1. Let (4.7) be the location family with $F(x,\theta)$ defined by the densities

$$f(x,\theta) = f_0(x-\theta) \quad \text{for} \quad x,\theta \in R.$$

If

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is the standard normal density then $\dot{f}(x,\theta)$ is continuous in the variables $x,\theta \in R$ and (4.10) implies that $\mathcal{J}(\theta_0) = 1$ for all $\theta_0 \in \Theta$. Similarly, for the logistic model of location with

$$f_0(x) = \frac{e^x}{(1+e^x)^2}$$

we obtain the continuity of $\dot{f}(x,\theta)$ and $\mathcal{J}(\theta_0) = 1$ for all $\theta_0 \in \Theta$. On the other hand, the doubly exponential model of location with

$$f_0(x) = \frac{1}{2} e^{-|x|}$$

does not satisfy the standard continuous differentiability assumption but satisfies (4.9) and $\dot{f}(x,\theta)$ exists for all $x \neq \theta_0$. Thus, by (4.10), $\mathcal{J}(\theta_0) = 1$ for every location $\theta_0 \in R$.

By Theorem 3 in [19], if $\mathcal{J}(\theta_0)$ is finite then

$$\mathcal{J}(\theta_0) - I_m(\boldsymbol{y},\theta_0) \geqslant 0$$

for all models (4.7) satisfying the conditions considered above and all $\boldsymbol{y}$ considered in (1.3).

In what follows we are interested in special $\boldsymbol{y} = \boldsymbol{y}_0 = \boldsymbol{y}_0(\theta_0)$ with the coordinates y_{0j} defined as the quantiles of the sample distribution $F(x,\theta_0)$ of orders j/m, $1 \leqslant j \leqslant m-1$, i.e. with

$$y_{0j} = F^{-1}(j/m,\theta_0) \quad \text{for} \quad F^{-1}(\tau,\theta) = \inf\{x\colon\ F(x,\theta) \geqslant \tau\}. \tag{4.11}$$

Obviously, $\boldsymbol{y}_0$ is one of the vectors (1.3) which is uniquely specified by θ_0 and the partition size $m > 1$. If all densities $\{f(x, \theta)\colon \theta \in \Theta\}$ are positive on a common interval of support then Theorem 4 in [19] implies that the relation

$$\mathcal{S}_m(\boldsymbol{y}_0, \theta_0) = \mathcal{J}(\theta_0) - I_m(\boldsymbol{y}_0, \theta_0) = o(1)$$

holds asymptotically for $m = r^k$, any integer $r > 1$, and $k \to \infty$.

Now we present heuristic argument leading to a more universal and precise asymptotic formula for the subefficiency at $\boldsymbol{y}_0$, namely

$$\mathcal{S}_m(\boldsymbol{y}_0, \theta_0) = O\left(\frac{1}{m^2}\right) \quad \text{as} \quad m \to \infty. \tag{4.12}$$

The argument is valid under appropriate regularity of the model (4.7) assumed in addition to (B1)–(B4).

In order to obtain (4.12), let us suppose that the densities $f(x, \theta)$ and their derivatives $\dot{f}(x, \theta)$ considered in (4.10) exist and that $f(x, \theta)$ is positive and $\dot{f}(x, \theta)$ continuous in the variable $x \in R$ for every $\theta \in \Theta$. For fixed θ_0, $\theta \in \Theta$ let us introduce the function

$$\varphi(\tau) = F(F^{-1}(\tau, \theta_0), \theta), \quad \tau \in (0, 1),$$

with the derivative

$$\varphi'(\tau) = \frac{f(F^{-1}(\tau, \theta_0), \theta)}{f(F^{-1}(\tau, \theta_0), \theta_0)}, \tag{4.13}$$

and the functions

$$\psi_\tau(x) = \varphi(\tau + x) - \varphi(\tau - x)$$

of the variable x from the neighbourhood of 0. If $\varphi'''(\tau) = \mathrm{d}^3\varphi(\tau)/\mathrm{d}\tau^3$ exists and is continuous on $(0, 1)$ then, by the Taylor theorem,

$$\psi_\tau(x) = 2\varphi'(\tau)x + \frac{\varphi'''(\tau_*)}{3}x^3 \quad \text{for} \quad \tau - x < \tau_* < \tau + x. \tag{4.14}$$

For $1 \leqslant j \leqslant m$ put

$$\tau_j = \frac{2j - 1}{2m}.$$

Since

$$\frac{j}{m} = \tau_j + \frac{1}{2m} \quad \text{and} \quad \frac{j-1}{m} = \tau_j - \frac{1}{2m},$$

we have

$$p_j(\boldsymbol{y}_0, \theta) = \varphi\left(\frac{j}{m}\right) - \varphi\left(\frac{j-1}{m}\right) = \psi_{\tau_j}\left(\frac{1}{2m}\right).$$

By (4.14), this implies

$$\begin{aligned}
p_j(\boldsymbol{y}_0,\theta) &= \varphi'(\tau_j)\,\frac{1}{m} + \frac{\varphi'''(\tau_{*j})}{24}\,\frac{1}{m^3}\\
&= \frac{f\big(F^{-1}(\tau_j,\theta_0),\theta\big)}{f\big(F^{-1}(\tau_j,\theta_0),\theta_0\big)}\,\frac{1}{m} + \frac{\varphi'''(\tau_{*j})}{24}\,\frac{1}{m^3} \qquad \text{(cf. 4.13).}
\end{aligned}$$

It follows from here that, under an additional regularity of the function $\varphi'''(\tau)$,

$$\dot{p}_j(\boldsymbol{y}_0,\theta_0) = \frac{\dot{f}\big(F^{-1}(\tau_j,\theta_0),\theta_0\big)}{f\big(F^{-1}(\tau_j,\theta_0),\theta_0\big)}\,\frac{1}{m} + O\Big(\frac{1}{m^3}\Big) \qquad \text{as} \quad m\to\infty.$$

By (4.11), $p_j(\boldsymbol{y}_0,\theta_0) = 1/m$ so that (4.8) implies for densities $f(x,\theta)$ with appropriate properties

$$\begin{aligned}
I_m(\boldsymbol{y}_0,\theta_0) &= m\sum_{j=1}^{m} \dot{p}_j(\boldsymbol{y}_0,\theta_0)^2\\
&= \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\dot{f}\big(F^{-1}(\tau_j,\theta_0),\theta_0\big)}{f\big(F^{-1}(\tau_j,\theta_0),\theta_0\big)}\right)^2 + O\Big(\frac{1}{m^2}\Big)\\
&= \int_0^1\left(\frac{\dot{f}\big(F^{-1}(\tau,\theta_0),\theta_0\big)}{f\big(F^{-1}(\tau,\theta_0),\theta_0\big)}\right)^2 \mathrm{d}\tau + O\Big(\frac{1}{m^2}\Big) \qquad \text{as} \quad m\to\infty.
\end{aligned}$$

Finally, the substitution $x = F^{-1}(\tau,\theta_0)$ transforms the last integral into the Fisher information (4.10). Thus under suitable regularity (4.12) holds.

On the basis of what has been said above, we can make several important *conclusions*. First of all, the optimal partition of any given size m is independent of the disparity function φ used to estimate the true parameter θ_0, and it is defined by the condition

$$\boldsymbol{y}_{\text{opt}} = \operatorname{argmax}_{\boldsymbol{y}} I_m(\boldsymbol{y},\theta_0), \tag{4.15}$$

where the maximization extends over all vectors $\boldsymbol{y}$ of (1.3) satisfying, together with θ_0, the conditions (B1)–(B4). Further, under the weak regularity of the basic continuous model (4.7) guaranteeing the existence of the Fisher information (4.10), the suboptimality $\mathcal{S}_m(\boldsymbol{y}_{\text{opt}},\theta_0)$ of the optimal partition is finite and tends to zero for $m\to\infty$. Under the stronger regularity of the model (4.7) guaranteeing the validity of (4.12), the suboptimality $\mathcal{S}_m(\boldsymbol{y}_{\text{opt}},\theta_0)$ tends to zero at least as fast as $1/m^2$ for $m\to\infty$. Finally, the number

$$\eta_m(\boldsymbol{y},\theta_0) = \frac{\mathcal{S}_m(\boldsymbol{y},\theta_0)}{\mathcal{J}(\theta_0)}\cdot 100 \tag{4.16}$$

characterizes in % the *relative asymptotic inefficiency* of all minimum φ-disparity estimators of θ_0 using the partition $\boldsymbol{y}$ in the continuous model (4.7). The number $\eta_m(\boldsymbol{y}_{\mathrm{opt}}, \theta_0)$ characterizes the relative asymptotic inefficiency of the estimation method of the present section in the continuous model (4.7). Unfortunately, this inefficiency is rarely practically achievable because the partition $\boldsymbol{y}_{\mathrm{opt}}$ usually depends on the true θ_0 which is a priori unknown. However, as we shall see in the next section, the inefficiency $\eta_m(\boldsymbol{y}_0, \theta_0)$ for $\boldsymbol{y}_0$ given by (4.11) is practically achievable for all $\theta_0 \in \Theta$ and all models satisfying regularity assumptions slightly stronger than (B1)–(B4). Under (4.12) this inefficiency is negligible for large m, tending to zero with the rate $1/m^2$ for $m \to \infty$.

Example 4.2. Table 4.1 presents for selected m the values of $I_m(\boldsymbol{y}_0, \theta_0)$, $\mathcal{S}_m(\boldsymbol{y}_0, \theta_0)$, $\eta_m(\boldsymbol{y}_0, \theta_0)$ and $I_m(\boldsymbol{y}_{\mathrm{opt}}, \theta_0)$, $\mathcal{S}_m(\boldsymbol{y}_{\mathrm{opt}}, \theta_0)$, $\eta_m(\boldsymbol{y}_{\mathrm{opt}}, \theta_0)$ in the normal and logistic models of location studied in Example 4.1. These values, as well as the corresponding Fisher informations

$$\mathcal{J}(\theta_0) = 1 \quad \text{and} \quad \mathcal{J}(\theta_0) = 1/3$$

do not depend on the location $\theta_0 \in R$. The partition vectors $\boldsymbol{y}_0$ and $\boldsymbol{y}_{\mathrm{opt}}$ corresponding to $\theta_0 = 0$ are presented in Table 4.1, too. Since their coordinates are symmetric about 0, we show only the nonnegative ones.

5. Continuous models. Random partitions

In this section we consider the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m-1})$ introduced in (1.5), the vector $\boldsymbol{y}_n$ of sample quantiles defined by (1.4), further the vector $\boldsymbol{y}_0 = (y_{01}, \ldots, y_{0m-1})$ of theoretical quantiles

(5.1) $$y_{0j} = F^{-1}(\lambda_j, \theta_0), \quad 1 \leqslant j \leqslant m-1,$$

and the stochastic vector $q = (q_1, \ldots, q_m)$ one-one related to $\boldsymbol{\lambda}$ by the formula

(5.2) $$q_j = \lambda_j - \lambda_{j-1}, \quad 1 \leqslant j \leqslant m.$$

We see that $\boldsymbol{y}_0$ of the present section is more general than that of Section 4. It coincides with $\boldsymbol{y}_0$ defined by (4.11) if the distribution q is uniform, i.e. if the quantile orders are equidistant.

By $\widetilde{\theta}_{\varphi,n} = \widetilde{\theta}_{\varphi,n}(x_1, \ldots, x_n)$ we denote the Θ-valued statistics satisfying the asymptotic relation

(5.3) $$P\bigl(D_\varphi(p(\boldsymbol{y}_n, \widetilde{\theta}_{\varphi,n}); q) \neq \inf_{\theta \in \Theta} D_\varphi(p(\boldsymbol{y}_n, \theta); q)\bigr) = o(1).$$

Family	$I.E.C.$	$m=5$	$m=10$	$m=15$	$m=20$
Normal	$\boldsymbol{y}_0$	0.280 0.842	0.000 0.320 0.640 0.961 1.281	0.115 0.346 0.577 0.808 1.039 1.270 1.501	0.000 0.183 0.365 0.548 0.731 0.913 1.097 1.279 1.462 1.645
	$I_m(\boldsymbol{y}_0, \theta_0)$	0.897	0.959	0.976	0.984
	$\mathcal{S}_m(\boldsymbol{y}_0, \theta_0)$	0.103	0.041	0.024	0.016
	$\eta_m(\boldsymbol{y}_0, \theta_0)$	**10.298**	**4.063**	**2.377**	**1.631**
	$\boldsymbol{y}_{\text{opt}}$	0.382 1.244	0.000 0.405 0.834 1.325 1.968	0.137 0.414 0.703 1.013 1.360 1.776 1.344	0.000 0.208 0.420 0.637 0.866 1.111 1.381 1.690 2.068 2.593
	$I_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$	0.920	0.977	0.989	0.994
	$\mathcal{S}_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$	0.080	0.023	0.011	0.006
	$\eta_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$	**7.994**	**2.294**	**1.074**	**0.621**
Logistic	$\boldsymbol{y}_0 \equiv \boldsymbol{y}_{\text{opt}}$	0.405 1.386	0.000 0.405 0.847 1.386 2.197	0.133 0.405 0.693 1.012 1.386 1.872 2.639	0.000 0.201 0.405 0.619 0.847 1.099 1.386 1.735 2.197 2.944
	$I_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$	0.320	0.330	0.332	0.332
	$\mathcal{S}_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$	0.013	0.003	0.001	0.001
	$\eta_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$	**4.000**	**1.000**	**0.444**	**0.250**

Table 4.1. Information and efficiency characteristics ($I.E.C.$) considered in Example 4.2. All values are rounded off to three decimals. The relative inefficiencies, given in percents, are printed in bold.

It is the φ-estimator of θ_0 in the model (4.6) with the observation space quantized by the sample quantiles (1.4).

If the sample space R is quantized by the sample quantiles $\boldsymbol{y}_n$ from (1.4) then the observation space is R^{m-1} and the true distribution functions G_θ in this space depend on the size of the sample $(x_i\colon\ 1 \leqslant i \leqslant n)$ in the original model (4.7). Namely, for the true θ_0 we have

$$G_{\theta_0}(\boldsymbol{y}) = G_{n,\theta_0}(\boldsymbol{y}) = P(\boldsymbol{y}_n < \boldsymbol{y}), \quad \boldsymbol{y} \in R^{m-1}.$$

This leads to quite complicated expressions for the families $\mathcal{G} = (G_{n,\theta}\colon\ \theta \in \Theta)$. Fortunately, it will be sufficient to use the asymptotic formula

$$\|F_n(\boldsymbol{y}_0) + F(\boldsymbol{y}_n, \theta_0) - 2\boldsymbol{\lambda}\| = o_p(n^{-1/2}) \tag{5.4}$$

proved in Theorem 1 of [4] under the assumption that the function $x \mapsto F(x,\theta_0)$ is continuous and increasing in the neighbourhood of $x = y_{0j}$ for every $1 \leqslant j \leqslant m-1$. Employing Corollary 4.1, one obtains from (5.4) a useful relation

$$\|p_n(\boldsymbol{y}_0) + p(\boldsymbol{y}_n,\theta_0) - 2q\| = o_p(n^{-1/2}). \tag{5.5}$$

Assumptions (A1)–(A5) that follow are analogues of (B1)–(B4) for the reduced models $(p(\boldsymbol{y},\theta)\colon\ \theta \in \Theta)$ of Section 4 with $\boldsymbol{y}$ from the neighbourhood of $\boldsymbol{y}_0$ given by (5.1). In particular, as is easy to see, these assumptions imply (B1)–(B4) for the model $(p(\boldsymbol{y}_0,\theta)\colon\ \theta \in \Theta)$.

(A1) $\equiv$ (B1) for $\boldsymbol{y} = \boldsymbol{y}_0$.

(A2) In the neighbourhood of $(\boldsymbol{y}_0;\theta_0)$, $F(\boldsymbol{y};\theta)$ is continuous, and also the gradient matrix $\Gamma(\boldsymbol{y};\theta) = (\partial/\partial\theta_1,\dots,\partial/\partial\theta_s)F(\boldsymbol{y},\theta)^t$ exists and is continuous.

Under (A2) also the function $p(\boldsymbol{y},\theta)$ is continuous and continuously differentiable in θ at all points $(\boldsymbol{y};\theta)$ from the neighbourhood of $(\boldsymbol{y}_0;\theta_0)$. Under (A1) it has all coordinates in this neighbourhood positive, due to the similar property of $p(\boldsymbol{y}_0,\theta_0) = q$ assumed in (A1). Thus, in particular, we can consider in this neighbourhood the $(m \times s)$ matrix functions

$$G(\boldsymbol{y},\theta) = (\partial/\partial\theta_1,\dots,\partial/\partial\theta_s)p(\boldsymbol{y},\theta)^t \quad \text{and} \quad A(\boldsymbol{y},\theta) = \operatorname{diag} q^{-1/2} G(\boldsymbol{y},\theta). \tag{5.6}$$

(A3) The matrix $A = A(\boldsymbol{y}_0,\theta_0)$ is of the rank s and $s < m$.

The $(s \times s)$ matrix

$$I = A^t A \tag{5.7}$$

is under (A3) positive definite. Due to the continuity assumed in (A2), also $I(\boldsymbol{y},\theta) = A(\boldsymbol{y},\theta)^t A(\boldsymbol{y},\theta)$ is positive definite in the neighbourhood of $(\boldsymbol{y}_0;\theta_0)$. Obviously, (5.7) is the Fisher information matrix of the reduced statistical model $(p(\boldsymbol{y}_0,\theta)\colon\ \theta \in \Theta)$ at the point θ_0.

The continuity of $F(\boldsymbol{y},\theta)$ assumed in (A2) implies in particular that, for all θ from the neighbourhood of θ_0, the functions $x \longmapsto F(x,\theta)$ are continuous in the neighbourhood of y_{0j}, $1 \leqslant j \leqslant m-1$. At $\theta = \theta_0$ we assume an additional property of $F(x,\theta)$.

(A4) $F(x,\theta_0)$ is increasing in the neighbourhood of $x = y_{0j}$ for every $1 \leqslant j \leqslant m-1$.

This assumption implies that $F(\boldsymbol{y},\theta_0)$ is invertible in the neighbourhood of $\boldsymbol{y} = \boldsymbol{y}_0$. Combining this with the monotonicity of $F(x,\theta_0)$ in the variable $x \in R$, one obtains for any sequence $\boldsymbol{y}_n$ the implication

$$\|F(\boldsymbol{y}_n,\theta_0) - \boldsymbol{\lambda}\| = o(1) \implies \|\boldsymbol{y}_n - \boldsymbol{y}_0\| = o(1). \tag{5.8}$$

Lemma 5.1. *If* (A1)–(A4) *hold then*

$$\|\boldsymbol{y}_n - \boldsymbol{y}_0\| = o_p(1) \tag{5.9}$$

and

$$n^{1/2}(p(\boldsymbol{y}_n, \theta_0) - q) \xrightarrow{w} N(0, \operatorname{diag} q - q^t q). \tag{5.10}$$

P r o o f. As stated above, (A1)–(A4) imply (5.4) and (5.5). Using the inequality $\big|\|a\| - \|b\|\big| \leqslant \|a - b\|$ valid for all vectors a, b, one obtains from (5.4)

$$\|F_n(\boldsymbol{y}_0) - \boldsymbol{\lambda}\| = \|F(\boldsymbol{y}_n, \theta_0) - \boldsymbol{\lambda}\| + o_p(n^{-1/2}) \tag{5.11}$$

and from (5.5)

$$\|p_n(\boldsymbol{y}_0) - q\| = \|p(\boldsymbol{y}_n, \theta_0) - q\| + o_p(n^{-1/2}). \tag{5.12}$$

Since $n^{1/2}(F_n(\boldsymbol{y}_0) - \boldsymbol{\lambda}) \xrightarrow{w} N(0, \boldsymbol{\lambda}^t(\mathbf{1} - \boldsymbol{\lambda}))$, (5.9) follows from (5.11) and (5.8). Further, since $p_n(\boldsymbol{y}_0) - q = n^{-1}(Z_n - nq)$ where Z_n is the multinomially distributed random vector with parameters n and q, we conclude

$$n^{1/2}(p_n(\boldsymbol{y}_0) - q) \xrightarrow{w} N(0, \operatorname{diag} q - q^t q). \tag{5.13}$$

Relations (5.12) and (5.13) imply (5.10). $\square$

The following result has been proved for $\varphi = \varphi_3$ from Example 2.1, i.e. for the minimum Pearson divergence estimator $\widetilde{\theta}_{\varphi_3,n}$, by Bofinger [4].

Lemma 5.2. *If* $\widetilde{\theta}_{\varphi,n}$ *is consistent and* (A1)–(A4) *hold then* $\widetilde{\theta}_{\varphi,n}$ *is efficient in the model* $(p(\boldsymbol{y}_0, \theta)\colon\ \theta \in \Theta)$ *in the sense*

$$\widetilde{\theta}_{\varphi,n} = \theta_0 + (p_n(\boldsymbol{y}_0) - q) \operatorname{diag} q^{-1/2} A I^{-1} + o_p(n^{-1/2}) \tag{5.14}$$

and asymptotically normal in the sense

$$\sqrt{n}(\widetilde{\theta}_{\varphi,n} - \theta_0) \xrightarrow{w} N(0, I^{-1}), \tag{5.15}$$

where A *is the matrix figuring in* (A3) *and* I *is the Fisher information matrix defined by* (5.7).

P r o o f. By the assumptions concerning Φ_{disp}, let $\varphi(1) = \varphi'(1) = 0$ and let us introduce an auxiliary function

$$v(\boldsymbol{y}, \theta) = \left(q_j^{1/2} \varphi' \Big(\frac{p_j(\boldsymbol{y}, \theta)}{q_j} \Big) \colon\ 1 \leqslant j \leqslant m \right)$$

of the vector variables $(\boldsymbol{y};\theta)$ from the neighbourhood of $(\boldsymbol{y}_0;\theta_0)$. It follows from (5.3) that

$$P\{v(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n})A(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n})\neq 0\}=o(1)$$

where $A(\boldsymbol{y},\theta)$ is defined in (5.6). If we apply the Taylor formula to

$$v(\boldsymbol{y}_n,\theta_0)-v(\boldsymbol{y}_0,\theta_0)\quad\text{ and }\quad v(\boldsymbol{y}_n,\theta)-v(\boldsymbol{y}_n,\theta_0)$$

and use the fact that $\varphi'(1)=0$ implies $v(\boldsymbol{y}_0,\theta_0)=0$, then we get the equation

$$v(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n})A(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n})=(p(\boldsymbol{y}_n,\theta_0)-q)B_n+(\widetilde{\theta}_{\varphi,n}-\theta_0)C_n$$

where

$$B_n=\operatorname{diag}\Big(\xi_{nj}^{-1/2}\varphi''\Big(\frac{\xi_{nj}}{q_j}\Big)\colon\ 1\leqslant j\leqslant m\Big)A(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n})$$

and

$$C_n=A(\boldsymbol{y}_n,\tau_n)^t\operatorname{diag}\Big(\varphi''\Big(\frac{p_j(\boldsymbol{y}_n,\tau_n)}{q_j}\Big)\colon\ 1\leqslant j\leqslant m\Big)A(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n})$$

for $\xi_n=(\xi_{n1},\dots,\xi_{nm})$ "between" q and $p(\boldsymbol{y}_n,\theta_0)$ and $\tau_n=(\tau_{n1},\dots,\tau_{ns})$ "between" θ_0 and $\widetilde{\theta}_{\varphi,n}$. Obviously,

$$B_n\xrightarrow{p}\varphi''(1)\operatorname{diag}q^{-1/2}A\quad\text{ and }\quad C_n\xrightarrow{p}\varphi''(1)A^tA. \tag{5.16}$$

If

$$(p(\boldsymbol{y}_n,\theta_0)-q)B_n+(\widetilde{\theta}_{\varphi,n}-\theta_0)C_n=0 \tag{5.17}$$

then the relation $\|p(\boldsymbol{y}_n,\theta_0)-q\|=O_p(n^{-1/2})$ obtained from (5.10) and the regularity of A^tA following from (A3) imply $\|\widetilde{\theta}_{\varphi,n}-\theta_0\|=O_p(n^{-1/2})$. Therefore (5.14) follows from (5.16) and (5.17).

Relation (5.17) follows from (5.13) and (5.14), or from (5.10) and (5.17), by using the formula

$$A^t\operatorname{diag}q^{-1/2}(\operatorname{diag}q-q^tq)\operatorname{diag}q^{-1/2}A=A^tA$$

deducible from the fact that $q\operatorname{diag}q^{-1/2}A=q\operatorname{diag}q^{-1}G(\boldsymbol{y}_0,\theta_0)$ is the zero s-vector of sums of the columns of the gradient $G(\boldsymbol{y}_0,\theta_0)$. □

It remains to formulate an appropriate consistency condition for the estimators $\widetilde{\theta}_{\varphi,n}$, $\varphi\in\Phi$. To this end we need an identifiability condition for the true θ_0 in the model under consideration, similar to (B4) in the model $(p(\boldsymbol{y}_0,\theta)\colon\ \theta\in\Theta)$. Bofinger [4] in Theorem 2 formulated an identifiability condition denoted there by (i), which is equivalent to (B4). Note that (A1)–(A4) are equivalent to the remaining conditions (ii)–(iv) of the mentioned theorem, and to the conditions formulated at other places of that paper. The next example demonstrates that (B4) is under (A1)–(A4) not sufficient for consistency.

E x a m p l e 5.1. Let $\Theta = (0,1)$ and let for all $(x;\theta) \in [0,\infty) \times (0,1) - S$ and for the square $S = [0, 0.6) \times (0.4, 1)$

$$F(x,\theta) = \min\{\theta x, 1\}.$$

If $(x;\theta) \in S$ then we put

$$F(x,\theta) = \frac{0.4}{0.6}(1-\theta)x + (\theta - 0.4)\Big(\frac{x}{0.6}\Big)^{(\theta-0.4)/(1-\theta)}.$$

The function $F(x,\theta)$ is continuous on $[0,\infty) \times (0,1)$, and also coordinatewise linear on $[0,1] \times (0,1) - S$, as seen from Fig. 5.1. Let $\theta_0 = 0.2$ and $m = 2$, so that $\boldsymbol{\lambda} = \lambda$ is a scalar from $(0,1)$ and similarly $\boldsymbol{y}_0 = y_0 = F^{-1}(\lambda, \theta_0)$ is a scalar from $(0,1)$. If $\lambda = 0.12$ then $y_0 = F^{-1}(\lambda, \theta_0) = 0.6$. In the neighbourhood $(0,1) \times (0, 0.4]$ of $(y_0; \theta_0) = (0.6; 0.2)$, $F(x,\theta)$ is linear and increasing in both variables x and θ. Also $F(0.6, \theta)$ is linear and increasing in θ on the whole parameter space (0,1). Hence (A1)–(A4) as well as (B4) hold. On the other hand, the sample quantiles of order $\lambda = 0.12$,

$$y_n = F_n^{-1}(0.12),$$

tend in probability to $y_0 = 0.6$. Since $F(x,\theta)$ is for every $x \in (0, 0.6]$ continuous in $\theta \in (0.4, 1)$, it takes on all values between $F(x, 0.4) = 0.4x$ and $F(x,1) = 0$ (cf. Fig. 5.1). Since the solution of the equation $0.4x = 0.12$ is $x = 0.3$, we obtain that

$$y_n \in (0.3, 0.6) \implies F(y_n, \widetilde{\theta}_n) = 0.12 \quad \text{for some } \widetilde{\theta}_n \in (0.4, 1).$$

Finally, in this example we have $p(y_n, \theta) = (F(y_n,\theta), 1 - F(y_n,\theta))$ and $q = (0.12, 0.88)$. Consequently, (2.5) implies for the Pearson divergence the formula

$$D_3(p(y_n,\theta); q) = \Big(\frac{1}{0.12} + \frac{1}{0.88}\Big)(F(y_n,\theta) - 0.12)^2.$$

Therefore if $y_n \in (0.3, 0.6)$ then $D_3(p(y_n, \widetilde{\theta}_n); q) = 0$ for some $\widetilde{\theta}_n \in (0.4, 1)$. But the sample quantiles y_n take on values in $(0.3, 0.6)$ with probability π_n tending to 1/2 for $n \to \infty$. This means that the minimum Pearson divergence estimator $\widetilde{\theta}_n$ attains values outside the neighbourhood $(0, 0.4]$ of $\theta_0 = 0.2$ with probability at least 1/2 when $n \to \infty$.

The insufficiency of the assumptions (A1)–(A4), (B4) of Bofinger [4] for the consistency of minimum Pearson divergence estimators $\widetilde{\theta}_{\varphi_3,n}$ demonstrated in Example 5.1 motivates the following condition.

(A5) For each $\boldsymbol{y}$ from the neighbourhood of $\boldsymbol{y}_0$, $\theta \mapsto F(\boldsymbol{y}, \theta)$ is a one-one mapping on Θ.

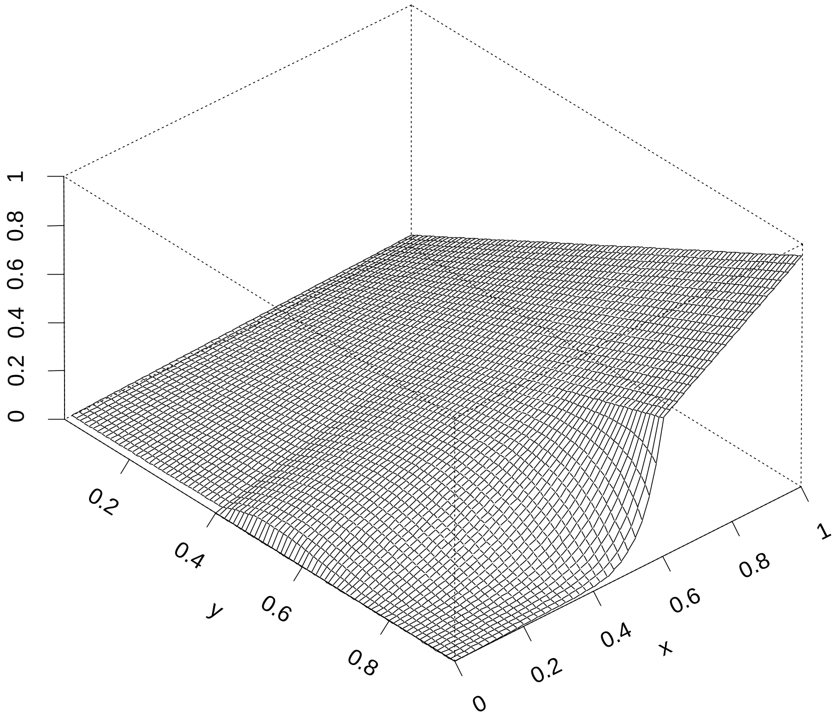

Figure 5.1. $F(x,\theta)$ for $x \in (0,1)$ and $\theta \equiv y \in (0,1)$.

We shall see that (A1)–(A4), (B4) are sufficient for the consistency of all disparity estimators $\widetilde{\theta}_{\varphi,n}$ if the continuity of $F(\boldsymbol{y},\theta)$, assumed in (A2), is replaced by the following stronger property.

(A6) The system of functions $\{F(\boldsymbol{y},\theta)\colon\ \theta \in \Theta\}$ is equicontinuous at $\boldsymbol{y} = \boldsymbol{y}_0$.

Lemma 5.3. *If* (A1)–(A5) *hold then all estimators* $\widetilde{\theta}_{\varphi,n}$, $\varphi \in \Phi$, *are consistent. This statement remains true with* (A1)–(A5) *replaced by* (A1)–(A4), (B4) *and* (A6).

Proof. (I) Obviously,

$$0 \leqslant D_\varphi(p(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n});q) \leqslant D_\varphi(p(\boldsymbol{y}_n,\theta_0);q).$$

Using the Taylor expansion of $\varphi(t)$ around $t=1$ one obtains from (2.1) and (5.10) that $D_\varphi(p(\boldsymbol{y}_n,\theta_0);q) = O_p(n^{-1})$. Consequently also

$$D_\varphi(p(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n});q) = O_p(n^{-1}).$$

This together with Theorem 2.3 implies $\|p(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n}) - q\| = O_p(n^{-1/2})$. We will use only the weaker relation

$$\|p(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n}) - q\| = o_p(1). \tag{5.18}$$

Let us assume that (A6) holds, i.e.

$$\sup_{\theta\in\Theta} \|F(\boldsymbol{y},\theta) - F(\boldsymbol{y}_0,\theta)\| \longrightarrow 0 \quad \text{as} \quad \boldsymbol{y} \to \boldsymbol{y}_0. \tag{5.19}$$

Then (5.9) implies

$$\sup_{\theta\in\Theta} \|F(\boldsymbol{y}_n,\theta) - F(\boldsymbol{y}_0,\theta)\| = o_p(1). \tag{5.20}$$

It follows from here and from Corollary 4.1 with $\widetilde{\boldsymbol{y}} = \boldsymbol{y}_n$ and $\theta = \theta_0$ that

$$\sup_{\theta\in\Theta} \|p(\boldsymbol{y}_n,\theta) - p(\boldsymbol{y}_0,\theta)\| = o_p(1). \tag{5.21}$$

By the triangle inequality,

$$\|p(\boldsymbol{y}_0,\widetilde{\theta}_{\varphi,n}) - q\| \leqslant \|p(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n}) - p(\boldsymbol{y}_0,\widetilde{\theta}_{\varphi,n})\| + \|p(\boldsymbol{y}_n,\widetilde{\theta}_{\varphi,n}) - q\|.$$

Therefore (5.21) and (5.18) imply

$$\|p(\boldsymbol{y}_0,\widetilde{\theta}_{\varphi,n}) - p(\boldsymbol{y}_0,\theta_0)\| = \|p(\boldsymbol{y}_0,\widetilde{\theta}_n) - q\| = o_p(1)$$

and, by virtue of the right-hand inequality in Corollary 4.1 with $\boldsymbol{y} = \widetilde{\boldsymbol{y}} = \boldsymbol{y}_0$ and $\theta = \widetilde{\theta}_{\varphi,n}$,

$$\|F(\boldsymbol{y}_0,\widetilde{\theta}_{\varphi,n}) - F(\boldsymbol{y}_0,\theta_0)\| = o_p(1). \tag{5.22}$$

The continuity of the mapping $\theta \mapsto F(\boldsymbol{y}_0,\theta)$ in an open neighbourhood U of θ_0 and the one-one property assumed in (B4) imply that the inverse mapping $\boldsymbol{\tau} \mapsto \Psi(\boldsymbol{\tau})$, defined for all $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_{m-1})$ from the domain $D = \{F(\boldsymbol{y}_0,\theta)\colon\ \theta \in \Theta\} \subset R^m$, is continuous on the subdomain $S = \{F(\boldsymbol{y}_0,\theta)\colon\ \theta \in U\}$. (B4) also implies that the image $F(\boldsymbol{y}_0,\theta)$ of $\theta \notin U$ is not in S. Since S is open and contains $\boldsymbol{\lambda} = F(\boldsymbol{y}_0,\theta_0)$, there exists $\varepsilon > 0$ such that

$$\|F(\boldsymbol{y}_0,\theta) - \boldsymbol{\lambda}\| < \varepsilon \implies \theta \in U. \tag{5.23}$$

Thus (5.22) implies that the probability of $\widetilde{\theta}_n \notin U$ tends to zero. Since

$$\widetilde{\theta}_n \in U \implies \Psi(F(\boldsymbol{y}_0,\widetilde{\theta}_{\varphi,n})) = \widetilde{\theta}_{\varphi,n}$$

and $\Psi(F(\boldsymbol{y}_0,\theta_0)) = \theta_0$, the desired result

$$\|\widetilde{\theta}_{\varphi,n} - \theta_0\| = o_p(1) \tag{5.24}$$

follows from (5.22) by the continuity of $\Psi(\boldsymbol{\tau})$.

(II) If instead of (B4) and (A6) one assumes (A5) then the proof is simpler. One obtains the same implication as in (5.23) with $\boldsymbol{y}_0$ replaced by $\boldsymbol{y}$ from a closed ball $V \subset \Theta$ centered at $\boldsymbol{y}_0$, but with $\varepsilon = \varepsilon(\boldsymbol{y})$ possibly depending on $\boldsymbol{y}$. However, due to the compactness of V,

$$\inf_{\boldsymbol{y} \in V} \varepsilon(\boldsymbol{y}) > 0.$$

By (A5), the neighbourhoods V and U can be chosen such that the mapping $F(\boldsymbol{y}, \theta)$ is invertible on $V \times U$, with the inverse $\varphi(\boldsymbol{\tau})$ defined and continuous for $\boldsymbol{\tau}$ from the neighbourhood of $\boldsymbol{\lambda} = F(\boldsymbol{y}_0, \theta_0)$. Finally, (5.18) means that $\|p(\boldsymbol{y}_n, \widetilde{\theta}_{\varphi,n}) - p(\boldsymbol{y}_0, \theta_0)\| = o_p(1)$. By the right-hand inequality in Corollary 4.1 this implies

$$\|F(\boldsymbol{y}_n, \widetilde{\theta}_{\varphi,n}) - F(\boldsymbol{y}_0, \theta_0)\| = o_p(1).$$

(5.4) follows from this relation in the same way as it followed above from (5.22), by using the identities

$$\varphi(F(\boldsymbol{y}_n, \widetilde{\theta}_{\varphi,n})) = \widetilde{\theta}_{\varphi,n}, \quad \varphi(F(\boldsymbol{y}_0, \theta_0)) = \theta_0$$

and the continuity of φ. □

Condition (A5) does not seem to be weaker than the conjunction of (B4) and (A6), and vice versa. If we consider for simplicity only (A5) then the results of the previous three lemmas can be summarized as follows.

Theorem 5.1. *If* (A1)–(A5) *hold then all estimators* $\widetilde{\theta}_{\varphi,n}$, $\varphi \in \Phi$, *are efficient in the sense of* (5.14) *and asymptotically normal in the sense of* (5.15).

We see from Theorem 5.1 that if the continuous model (4.7) satisfies for the true $\theta_0 \in \Theta$ and for $\boldsymbol{y}_0$ given by (4.11) the regularity assumptions (A1)–(A5) then all minimum disparity estimators $\widetilde{\theta}_{\varphi,n}$, $\varphi \in \Phi_{\text{disp}}$ achieve under the random partition by the sample quantiles $\boldsymbol{y}_n$ of the equidistant orders j/m, $1 \leqslant j \leqslant m-1$, the same asymptotic variances as the estimators $\theta_{\varphi,n}$, $\varphi \in \Phi_{\text{disp}}$ under the deterministic partitions $\boldsymbol{y}_0$ of (4.11). Therefore in this case $\mathcal{S}_m(\boldsymbol{y}_0, \theta_0)$ from Section 4 (defined by (4.8) if $\Theta \subset R$) characterizes the subefficiency of the estimation procedure of the present section in the model (4.7). Or, equivalently, $\eta_m(\boldsymbol{y}_0, \theta_0)$ defined by (4.16) characterizes in % the asymptotic inefficiency of the present procedure in the original model (4.7). As was indicated in Example 4.2, $\eta_m(\boldsymbol{y}_0, \theta_0)$ is for larger m only slightly below the minimal asymptotic inefficiency $\eta_m(\boldsymbol{y}_{\text{opt}}, \theta_0)$ achievable by the partitions of size m. If the models (4.7) satisfy an additional regularity condition, then the inefficiency $\eta_m(\boldsymbol{y}_0, \theta_0)$ tends to zero for $m \to \infty$ with the rate $1/m^2$. Thus the efficient estimation in sufficiently regular models (4.7) can be practically achieved by any of the estimators $\widetilde{\theta}_{\varphi,n}$, $\varphi \in \Phi_{\mathrm{disp}}$, and by taking a partition of size m large enough. This estimation is at the same time robust in the sense of Lindsay [9] whenever the used disparity function φ is bounded.

References

[1] *A. Basu, S. Sarkar*: Minimum disparity estimation in the errors-in-variables model. Statist. Probab. Lett. *20* (1994), 69–73.

[2] *A. Basu, S. Sarkar*: The trade-off between robustness and efficiency and the effect of model smoothing. J. Statist. Comput. Simulation *50* (1994), 173–185.

[3] *M. W. Birch*: A new proof of the Pearson-Fisher theorem. Ann. Math. Statist. *35* (1964), 817–824.

[4] *E. Bofinger*: Goodness-of-fit using sample quantiles. J. Roy. Statist. Soc. Ser. B *35* (1973), 277–284.

[5] *H. Cramér*: Mathematical Methods of Statistics. Princeton University Press, Princeton, 1946.

[6] *N. A. C. Cressie, R. C. Read*: Multinomial goodness-of-fit tests. J. Roy. Statist. Soc. Ser. B *46* (1984), 440–464.

[7] *R. A. Fisher*: Statistical Methods for Research Workers (8th edition). London, 1941.

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

[9] *B. G. Lindsay*: Efficiency versus robutness: The case for minimum Hellinger distance and other methods. Ann. Statist. *22* (1994), 1081–1114.

[10] *M. L. Menéndez, D. Morales, L. Pardo and I. Vajda*: Two approaches to grouping of data and related disparity statistics. Comm. Statist. Theory Methods *27* (1998), 609–633.

[11] *M. L. Menéndez, D. Morales, L. Pardo and I. Vajda*: Minimum divergence estimators based on grouped data. Ann. Inst. Statist. Math. *53* (2001), 277–288.

[12] *D. Morales, L. Pardo and I. Vajda*: Asymptotic divergence of estimates of discrete distributions. J. Statist. Plann. Inference *48* (1995), 347–369.

[13] *J. Neyman*: Contribution to the theory of the χ^2 test. In: Proc. Berkeley Symp. Math. Statist. Probab., Berkeley, CA. Berkeley University Press, Berkeley, 1949, pp. 239–273.

[14] *Ch. Park, A. Basu and S. Basu*: Robust minimum distance inference based on combined distances. Comm. Statist. Simulation Comput. *24* (1995), 653–673.

[15] *C. R. Rao*: Asymptotic efficiency and limiting information. In: Proc. 4th Berkeley Symp. Math. Stat. Probab., Berkeley, CA. Berkeley University Press, Berkeley, 1961, pp. 531–545.

[16] *C. R. Rao*: Linear Statistical Inference and its Applications (2nd edition). Wiley, New York, 1973.

[17] *R. C. Read, N. A. C. Cressie*: Goodness-of-fit Statistics for Discrete Multivariate Data. Springer-Verlag, New York, 1988.

[18] *C. A. Robertson*: On minimum discrepancy estimators. Sankhyä Ser. A *34* (1972), 133–144.

[19] *I. Vajda*: χ^2-divergence and generalized Fisher information. In: Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes. Academia, Prague, 1973, pp. 223–234.

[20] *I. Vajda*: Theory of Statistical Inference and Information. Kluwer Academic Publishers, Boston, 1989.

[21] *B. L. van der Waarden*: Mathematische Statistik. Springer-Verlag, Berlin, 1957.

[22] *K. Pearson*: On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine *50* (1990), 157–172.

[23] *I. Csiszár*: Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, Series A *8* (1963), 85–108.

[24] *M. S. Ali, S. D. Silvey*: A general class of coefficients of divergence of one distribution from another. J. Roy. Statist. Soc. Ser. B *28* (1966), 131–140.

[25] *A. Rényi*: On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Probability Theory and Mathematical Statistics, Vol. 1. University of California Press, Berkeley, 1961, pp. 531–546.

[26] *A. W. Marshall, I. Olkin*: Inequalities: Theory of Majorization and its Applications. Academic Press, New York, 1979.

Authors' addresses: *M. Menéndez*, Department of Applied Mathematics, Technical University of Madrid, Juan de Herrera 4, 28040 Madrid, Spain, e-mail: `mmenende@aq.upm.es`; *D. Morales*, Operations Research Center, Miguel Hernández University of Elche, Elche, Spain, e-mail: `d.morales@umh.es`; *L. Pardo*, Department of Statistics & O. R., Complutense University of Madrid, Madrid, Spain, e-mail: `Leandro_Pardo@mat.ucm.es`; *I. Vajda*, Institute of Information Theory, Academy of Sciences of the Czech Republic, Prague, Czech Republic, e-mail: `vajda@utia.cas.cz`.