Ann. Inst. Statist. Math.
Vol. 53, No. 2, 277–288 (2001)

MINIMUM DIVERGENCE ESTIMATORS BASED ON GROUPED DATA*

M. MENÉNDEZ[1], D. MORALES[2], L. PARDO[3] AND I. VAJDA[4]

[1] *Department of Applied Mathematics, Technical University of Madrid, 28040 Madrid, Spain*
[2] *Operations Research Center, Miguel Hernández University of Elche, 03202 Elche, Spain*
[3] *Department of Statistics & O. R., Complutense University of Madrid, 28040 Madrid, Spain*
[4] *Institute of Information Theory, Academy of Sciences of the Czech Republic, CZ-18208 Prague, Czech Republic*

(Received September 8, 1997; revised July 9, 1999)

Abstract. The paper considers statistical models with real-valued observations i.i.d. by $F(x, \theta_0)$ from a family of distribution functions $(F(x, \theta); \theta \in \Theta)$, $\Theta \subset R^s$, $s \geq 1$. For random quantizations defined by sample quantiles $(F_n^{-1}(\lambda_1), \dots, F_n^{-1}(\lambda_{m-1}))$ of arbitrary fixed orders $0 < \lambda_1 < \cdots < \lambda_{m-1} < 1$, there are studied estimators $\theta_{\phi,n}$ of θ_0 which minimize ϕ-divergences of the theoretical and empirical probabilities. Under an appropriate regularity, all these estimators are shown to be as efficient (first order, in the sense of Rao) as the MLE in the model quantified nonrandomly by $(F^{-1}(\lambda_1, \theta_0), \dots, F^{-1}(\lambda_{m-1}, \theta_0))$. Moreover, the Fisher information matrix $I_m(\theta_0, \lambda)$ of the latter model with the equidistant orders $\boldsymbol{\lambda} = (\lambda_j = j/m : 1 \leq j \leq m-1)$ arbitrarily closely approximates the Fisher information $\mathcal{J}(\theta_0)$ of the original model when m is appropriately large. Thus the random binning by a large number of quantiles of equidistant orders leads to appropriate estimates of the above considered type.

Key words and phrases: Minimum divergence estimators, random quantization, asymptotic normality, efficiency, Fisher information, optimization.

1. Introduction and basic concepts

This paper deals with the minimum distance point estimation in the case where the initial information about data and hypothetical parametrized models is reduced by partitioning the observation space, and the distance is measured by the divergence of reduced hypothetical and empirical distributions. Partitioning is sometimes practical because it reduces the numerical complexity of estimation. Often data are themselves 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 unreduced data and models, for example the minimum Pearson divergence estimator has in this sense been employed by Neyman (1949), or the maximum likelihood estimator (MLE) is obtained by minimizing the information divergence of Kullback.

The MLE is known to be efficient in regular models but is also known to be nonrobust. The main reason for introducing ϕ-divergences different from that of Kullback into the point estimation is the efficiency and, at the same time, robustness of many

*This work was supported by the Sabbatical Program of Complutense University of Madrid, by the GACR grant 102/99/1137, by the DGI grant BFM 2000–0800 and by GV grant 99-159-1-01.

ϕ-divergence estimators, see Lindsay (1994).

In this paper we consider arbitrary parametrized models $(F(x,\theta) : \theta \in \Theta)$ with parameter spaces $\Theta \subset R^s$, $s \geq 1$, unknown true values $\theta_0 \in \Theta$, and random observations $X_1, \ldots, X_n$ i.i.d. by $F(x, \theta_0)$, $x \in R$. By quantization we mean a partition of the observation space R into m intervals (bins) specified by a vector

$$(1.1) \qquad \boldsymbol{y} = (y_1, \ldots, y_{m-1}), \qquad y_0 = -\infty < y_1 < \cdots < y_{m-1} < \infty = y_m.$$

If $F(x,\theta_0)$ is absolutely continuous at all $x \in \{y_1, \ldots, y_m\}$ then the exact specification of the bins at their ends is irrelevant. The binning leads to the theoretical and empirical probability distributions

$$(1.2) \qquad p(\boldsymbol{y}, \theta) = (p_j(\boldsymbol{y}, \theta) = F(y_j, \theta) - F(y_{j-1}, \theta) : 1 \leq j \leq m)$$

and

$$(1.3) \qquad p_n(\boldsymbol{y}) = (p_{nj}(\boldsymbol{y}) = F_n(y_j) - F_n(y_{j-1}) : 1 \leq j \leq m),$$

where $F_n(x)$, $x \in R$, is the empirical distribution function. By minimizing the ϕ-divergence $D_\phi(p(\boldsymbol{y}, \boldsymbol{\theta}); p_n(\boldsymbol{y}))$ of the discrete distributions (1.2) and (1.3) over the parameter space Θ we obtain a *minimum ϕ-divergence estimator* $\theta_n(\boldsymbol{y}, \phi)$. More precisely, we define this estimator as a sequence of Θ-valued measurable functions of the sample

$$\theta_n(\boldsymbol{y}, \phi) = \theta_n(\boldsymbol{y}, \phi, X_1, \ldots, X_n), \qquad n = 1, 2, \ldots,$$

with parameters $\boldsymbol{y}$ and ϕ, satisfying the asymptotic relation

$$P\{D_\phi(p(\boldsymbol{y}, \boldsymbol{\theta}_n(y, \phi)); p_n(\boldsymbol{y})) \neq \inf_{\Theta} D(p(\boldsymbol{y}, \boldsymbol{\theta}); p_n(\boldsymbol{y}))\} = o(1).$$

Note that the *ϕ-divergence* of arbitrary probability m-vectors p and q is defined by the formula

$$(1.4) \qquad D_\phi(p; q) = \sum_{j=1}^{m} q_j \phi\left(\frac{p_j}{q_j}\right), \qquad \phi \in \Phi,$$

where Φ is the class of all convex functions $\phi(t)$, $t > 0$, equal to 0 at $t = 1$. For every $\phi \in \Phi$ differentiable at $t = 1$

$$(1.5) \qquad \phi(t) \sim \phi(t) - \phi'(t)(t-1),$$

where the right hand side belongs to Φ and the equivalence means that the two functions define the same divergence (1.4).

Hereafter Φ stands for the subclass of convex functions twice continuously differentiable in the neighborhood of $t = 1$ with $\phi(1) = 0$, $\phi''(1) \neq 0$. Obviously, we can assume without loss of generality that $\phi'(1) = 0$ and $\phi''(1) = 1$ for every $\phi \in \Phi$.

Example 1.1. The nonnegative functions

$$\phi_a(t) = \frac{t^{(a+1)/2} - \frac{1}{2}(a+1)(t-1) - 1}{\frac{(|a|-1)}{2}} \sim \frac{t^{(a+1)/2} - 1}{\frac{(|a|-1)}{2}}$$

(cf. the equivalence relation $\sim$ in (1.5)) with limits

$$\phi_1(t) = t\ln t - t + 1 \sim t\ln t$$

and

$$\phi_{-1}(t) = -\ln t + t - 1 \sim -\ln t$$

have continuous and positive second derivatives

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

They define a class of *modified power divergences*

$$(1.6) \qquad D_a(p,q) = \frac{2}{|a|-1}\left(\sum_{j=1}^m \sqrt{p_j^{1+a} q_j^{1-a}} - 1\right) \qquad \text{for all} \quad a \neq -1,\ a \neq 1,$$

with the well known Kullback and reversed Kullback divergences

$$(1.7) \qquad D_1(p;q) = \sum_{j=1}^m q_j \ln \frac{p_j}{q_j} \qquad \text{and} \qquad D_{-1}(p;q) = D_1(q;p)$$

as the limits for $a \to 1$ and $a \to -1$. (The *skew symmetry* $D_{-a}(p;q) = D_a(q;p)$ for remaining $a \in R$ is clear from (1.6)). Well known are also the Pearson divergence

$$(1.8) \qquad D_3(p;q) = \sum_{j=1}^m \frac{p_j^2}{q_j} - 1 = \sum_{j=1}^m \frac{(p_j - q_j)^2}{q_j},$$

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

$$D_0(p;q) = 2\left(1 - \sum_{j=1}^m \sqrt{p_j q_j}\right) = \sum_{j=1}^m (\sqrt{p_j} - \sqrt{q_j})^2.$$

The original power divergences of Cressie and Read (1984) are 1-1 transforms of (1.6),

$$(1.9) \qquad I_\lambda(p,q) = \frac{4D_{2\lambda+1}(p;q)}{|2\lambda+1|+1}, \qquad \lambda \in R.$$

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 similar symmetry about 0 in the family (1.6). For example, it may not be easy to recognize at 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 (1.6) and (1.9) can be obtained as 1-1 transforms of the α-divergences $R_\alpha(p;q)$, $\alpha > 0$, see Liese and Vajda (1987). E.g.,

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

It is easy to see that $\hat{\theta}_n(\boldsymbol{y}, \phi_1)$ is the MLE in the discrete model (1.2). Birch (1964) formulated conditions on an arbitrary discrete model $p(\theta) = (p_j(\theta) : 1 \leq j \leq m)$, $\theta \in \Theta$, under which the MLE in this model is efficient (first order in the sense of Rao (1961, 1973) briefly *efficient in the sense of Rao*). Morales *et al.* (1995) proved that these conditions are sufficient for the Rao efficiency of all minimum ϕ-divergence estimators, $\phi \in \Phi$, in this model.

In the next section we present the Birch conditions for the model (1.2) and evaluate the Fisher information $I_m(\theta_0, \boldsymbol{y})$ in this model. If the Fisher information $\mathcal{J}(\theta_0)$ in the original model $(F(x, \theta) : \theta \in \Theta)$ exists then $\mathcal{J}(\theta_0) - I_m(\theta_0, \boldsymbol{y})$ is positive semidefinite (cf. Vajda (1973), in particular $\operatorname{tr} \mathcal{J}(\theta_0) \geq \operatorname{tr} I_m(\theta_0, \boldsymbol{y})$). In typical situations this inequality is strict, i.e. the estimators $\hat{\theta}_n(\boldsymbol{y}, \phi)$, $\phi \in \Phi$, are not efficient in the original model. The maximization of $\operatorname{tr} I_m(\theta_0, \boldsymbol{y})$ leads to

$$\boldsymbol{y}_{opt} = \arg\max_{\boldsymbol{y}} \operatorname{tr} I_m(\theta_0, \boldsymbol{y}) \tag{1.10}$$

which depends on the unknown θ_0 (cf. Ferentinos and Papaioannou (1979) and Tsairidis *et al.* (1998)). Moreover, it is not clear whether for any $\varepsilon > 0$ there exists m such that

$$\sup_{\theta \in \Theta} (\operatorname{tr} \mathcal{J}(\theta) - \operatorname{tr} I_m(\theta, \boldsymbol{y})) < \varepsilon. \tag{1.11}$$

To avoid these difficulties, we consider in the next section the partitions

$$\boldsymbol{y}_0 = (y_{0j} = F^{-1}(\lambda_j, \theta_0) : 1 \leq j \leq m), \qquad \lambda_0 = 0 < \lambda_1 < \cdots < \lambda_{m-1} < 1 = \lambda_m, \tag{1.12}$$

where $F^{-1}(\lambda, \theta) = \inf\{x \in R : F(x, \theta) \geq \lambda\}$ is the generalized inverse (quantile of the order $0 < \lambda < 1$). If $F(x, \theta_0)$ is increasing on R then all partitions (1.1) are in the class (1.12). If $F(x, \theta_0)$ is constant in an interval (x_1, x_2) then the partition of this interval has no influence on $I_m(\theta_0, \boldsymbol{y})$. One can deduce from here that $\boldsymbol{y}_{opt}$ in (1.10) is achieved in the class (1.12). Hence for

$$\tilde{I}_m(\theta_0, \boldsymbol{\lambda}) = I_m(\theta_0, \boldsymbol{y}_0), \qquad \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m-1}), \tag{1.13}$$

with $\boldsymbol{y}_0$ given by (1.12), the optimization (1.10) is equivalent to the evaluation of

$$\boldsymbol{\lambda}_{opt} = \arg\max_{\boldsymbol{\lambda}} \operatorname{tr} \tilde{I}_m(\theta_0, \boldsymbol{\lambda}). \tag{1.14}$$

For some models, e.g. for the models of location with $\Theta = R$, $\boldsymbol{\lambda}_{opt}$ is independent of $\theta_0 \in \Theta$, and for each $\varepsilon > 0$ there exists m such that

$$\sup_{\theta \in \Theta} (\operatorname{tr} \mathcal{J}(\theta) - \operatorname{tr} \tilde{I}_m(\theta, \boldsymbol{\lambda}_{unif})) < \varepsilon, \tag{1.15}$$

where

$$\boldsymbol{\lambda}_{unif} = (\lambda_{unif,j} = j/m : 1 \leq j \leq m - 1) \tag{1.16}$$

leads to the uniform empirical distribution

$$p_n(\boldsymbol{y}_n) = (1/m, \ldots, 1/m), \tag{1.17}$$

i.e. to the partitions $\boldsymbol{y}_n = (F_n^{-1}(j/m) : 1 \leq j \leq m)$ of the observation space into statistically equivalent blocks (see Devroye *et al.* (1996)).

Of course, partitions (1.12) depend on the unknown quantiles $F^{-1}(\lambda_j, \theta_0)$, but these can be replaced by the empirical quantiles $F_n^{-1}(\lambda_j)$ of the same orders which almost surely tend to the theoretical ones for arbitrary $\theta_0 \in \Theta$.

Therefore, in the next section we consider the random partitons

$$\boldsymbol{y}_n = (y_{nj} = F_n^{-1}(\lambda_j) : 1 \leq j \leq m) \tag{1.18}$$

for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m-1})$ assumed in (1.12), and the corresponding estimators

$$\tilde{\theta}_n(\boldsymbol{\lambda}, \phi) = \hat{\theta}_n(\boldsymbol{y}_n, \phi), \qquad \phi \in \Phi. \tag{1.19}$$

These estimators are defined by the asymptotic condition

$$P\{D_\phi(p(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)), q(\boldsymbol{\lambda})) \neq \inf_\Theta D_\phi(p(\boldsymbol{y}_n, \theta), q(\boldsymbol{\lambda}))\} = o(1) \text{ as } n \to \infty \tag{1.20}$$

for $\boldsymbol{y}_n$ given by (1.18) and

$$q(\boldsymbol{\lambda}) = q = (q_j = \lambda_j - \lambda_{j-1} : 1 \leq j \leq m) \qquad \text{(cf. (1.12))}. \tag{1.21}$$

We formulate conditions for the Rao efficiency of all these estimators in the discrete model $p(\boldsymbol{y}_0, \theta), \theta \in \Theta$, with $\boldsymbol{y}_0$ given by (1.12).

By (1.13), under the mentioned conditions (1.14) defines an optimal binning and (1.15) guarantees an ε-efficiency of the bins defined by empirical quantiles of the equidistant orders (1.16). The optimization (1.14) and relation (1.15) for $\boldsymbol{\lambda}_{unif}$ are investigated in the last section.

2. The Rao efficiency

Let us formulate in a slightly stronger form the conditions of Birch (1964) and Morales *et al.* (1995) for the Rao efficiency of estimators $\hat{\theta}_n(\boldsymbol{y}_0, \phi)$, $\phi \in \Phi$, in the model $(p(\boldsymbol{y}_0, \theta) : \theta \in \Theta)$ where $\boldsymbol{y}_0$ is an arbitrary vector satisfying (1.1).

(B1) True θ_0 is in the interior of Θ and $p(\boldsymbol{y}_0, \theta_0)$ has all coordinates positive.

(B2) Gradient $\Gamma(\boldsymbol{y}_0, \theta) = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_s) F(\boldsymbol{y}_0, \theta)^t$ exists and is continuous at every point θ from the neighborhood of θ_0.

Under (B1), (B2) also the gradient $G(\boldsymbol{y}_0, \theta) = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_s) p(\boldsymbol{y}_0, \theta)^t$ and the matrix $A(\boldsymbol{y}, \theta) = \operatorname{diag} p(\boldsymbol{y}_0, \theta_0)^{-1/2} G(\boldsymbol{y}, \theta)$ exist and are continuous at every point θ from the neighborhood of θ_0. Note that for any k-vector p and mapping $\psi : R \mapsto R$, $\operatorname{diag} \psi(p)$ in this paper denotes the diagonal $(k \times k)$-matrix with entries $\psi(p_1), \ldots, \psi(p_k)$ at the diagonal.

(B3) Matrix $A(\boldsymbol{y}_0, \theta_0)$ is of rank s and $s < m$.

(B4) Mapping $\theta \mapsto F(\boldsymbol{y}_0, \theta)$ is 1-1 on Θ.

We are interested in similar conditions for the Rao efficiency of estimators $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi)$, $\phi \in \Phi$. Note that such conditions were previously formulated for the estimator $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi_3)$ but, as we shall see, they are not sufficient even for the consistency of this estimator. In the conditions that follow we consider $\boldsymbol{y}_0$ given by (1.12), and we need the identity

$$p(\boldsymbol{y}_0, \theta_0) = q \tag{2.1}$$

valid for this $\boldsymbol{y}_0$ and q given by (1.21). Obviously, these conditions imply (B1)–(B4) for $\boldsymbol{y}_0$ under consideration.

(A1) True θ_0 is in the interior of Θ and all coordinates of q are positive.

(A2) In the neighborhood of $(\boldsymbol{y}_0;\theta_0)$, $F(\boldsymbol{y},\theta)$ is continuous and the gradient $\Gamma(\boldsymbol{y},\theta) = (\partial/\partial\theta_1,\dots,\partial/\partial\theta_s)F(\boldsymbol{y},\theta)^t$ 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 neighborhood of $(\boldsymbol{y}_0;\theta_0)$, and has all coordinates positive, similarly as $p(\boldsymbol{y}_0,\theta_0) = q$. Thus, in particular, we can consider in this neighborhood the $(m \times s)$-matrix functions

$$(2.2) \qquad G(\boldsymbol{y},\theta) = (\partial/\partial\theta_1,\dots,\partial/\partial\theta_s)p(\boldsymbol{y},\theta)^t \quad \text{and} \quad A(\boldsymbol{y},\theta) = \operatorname{diag} q^{-1/2}G(\boldsymbol{y},\theta),$$

with $G = G(\boldsymbol{y}_0,\theta_0)$ and $A = A(\boldsymbol{y}_0,\theta_0)$.

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

$$(2.3) \qquad I = A^tA$$

is under (A3) positive definite. Due to the continuity assumed in (A2), also $I(\boldsymbol{y},\theta) = A(\boldsymbol{y},\theta)^tA(\boldsymbol{y},\theta)$ is positive definite in the neighborhood of $(\boldsymbol{y}_0;\theta_0)$. Obviously, (2.3) is the Fisher information matrix of the reduced statistical model $(p(\boldsymbol{y}_0,\theta) : \theta \in \Theta)$ at the point θ_0.

The continuity of $F(\boldsymbol{y},\theta)$ which follows from (A2) implies in particular that, for all θ from the neighborhood of θ_0, the functions $x \mapsto F(x,\theta)$ are continuous in the neighborhood of y_{0j}, $1 \le j \le m-1$. At $\theta = \theta_0$ we assume more.

(A4) $F(x,\theta_0)$ is increasing in the neighborhood of every y_{0j}, $1 \le j \le m-1$.
This assumption implies that $F(\boldsymbol{y},\theta_0)$ is invertible in the neighborhood 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 $\boldsymbol{y}_n$

$$(2.4) \qquad \|F(\boldsymbol{y}_n,\theta_0) - \boldsymbol{\lambda}\| = o(1) \Rightarrow \|\boldsymbol{y}_n - \boldsymbol{y}_0\| = o(1).$$

In the sequel we need the inequality

$$(2.5) \qquad \frac{1}{2}\|\gamma\| \le \|\Gamma\| \le m\|\gamma\|$$

which follows for all vectors $\gamma = (\gamma_1,\dots,\gamma_m)$ with the sum of coordinates equal zero, and for $\Gamma = (\Gamma_j \equiv \gamma_1 + \dots + \gamma_j : 1 \le j \le m)$, from the obvious relations

$$\gamma_j^2 \le 2(\Gamma_{j-1}^2 + \Gamma_j^2) \quad \text{and} \quad \Gamma_j^2 \le j\|\gamma\|^2, \quad 1 \le j \le m, \text{ where } \Gamma_0 = 0.$$

Using (2.5) one obtains for any θ_1, $\theta_2 \in \Theta$ and $\boldsymbol{y}_1$, $\boldsymbol{y}_2$ satisfying (1.1),

$$(2.6) \ \frac{1}{2}\|p(\boldsymbol{y}_1,\theta_1) - p(\boldsymbol{y}_2,\theta_2)\| \le \|F(\boldsymbol{y}_1,\theta_1) - F(\boldsymbol{y}_2,\theta_2)\| \le m\|p(\boldsymbol{y}_1,\theta_1) - p(\boldsymbol{y}_2,\theta_2)\|,$$

where

$$F(\boldsymbol{y},\theta) = (F(y_1,\theta),\dots,F(y_{m-1},\theta)), \qquad \theta \in \Theta.$$

We also need the asymptotic formula

$$(2.7) \qquad \|F_n(\boldsymbol{y}_0) + F(\boldsymbol{y}_n,\theta_0) - 2\boldsymbol{\lambda}\| = o_p(n^{-1/2})$$

proved in Theorem 1 of Bofinger (1973) under the assumption that $F(x,\theta_0)$ is continuous and increasing in the neighborhood of y_{0j}, $1 \le j \le m-1$. Using (2.5) one obtains from (2.7) the following useful relation

$$(2.8) \qquad \|p_n(\boldsymbol{y}_0) + p(\boldsymbol{y}_n,\theta_0) - 2q\| = o_p(n^{-1/2}).$$

LEMMA 2.1. *If* (A1)–(A4) *hold then*

$$\|\boldsymbol{y}_n - \boldsymbol{y}_0\| = o_p(1) \tag{2.9}$$

and

$$n^{1/2}(p(\boldsymbol{y}_n, \theta_0) - q) \xrightarrow{w} N(0, \operatorname{diag} q - q^t q). \tag{2.10}$$

PROOF. As stated above, (A2) implies (2.7) and (2.8). Using the inequality $|\|a\| - \|b\|| \leq \|a - b\|$ valid for all vectors a, b, one obtains from (2.7)

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

and from (2.8)

$$\|p_n(\boldsymbol{y}_0) - q\| = \|p(\boldsymbol{y}_n, \theta_0) - q\| + o_p(n^{-1/2}). \tag{2.11}$$

Since $n^{1/2}(F_n(\boldsymbol{y}_0) - \boldsymbol{\lambda}) \to^w N(0, \boldsymbol{\lambda}^t(\mathbf{1} - \boldsymbol{\lambda}))$, (2.9) follows from the first relation using (2.4). Further, since $p_n(\boldsymbol{y}_0) - q = n^{-1}(Z_n - nq)$ where Z_n is multinomially distributed random vector with parameters n and q, it holds

$$n^{1/2}(p_n(\boldsymbol{y}_0) - q) \xrightarrow{w} N(0, \operatorname{diag} q - q^t q). \tag{2.12}$$

Relations (2.11) and (2.12) imply (2.10).

LEMMA 2.2. *If $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi)$ is consistent and* (A1)–(A4) *hold then $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi)$ is efficient in the model $(p(\boldsymbol{y}_0, \theta) : \theta \in \Theta)$ in the sense*

$$\tilde{\theta}_n(\boldsymbol{\lambda}, \phi) = \theta_0 + (p_n(\boldsymbol{y}_0) - q) \operatorname{diag} q^{-1/2} A I^{-1} + o_p(n^{-1/2}) \tag{2.13}$$

and asymptotically normal in the sense

$$\sqrt{n}(\tilde{\theta}_n(\boldsymbol{\lambda}, \phi) - \theta_0) \xrightarrow{w} N(0, I^{-1}), \tag{2.14}$$

where A is the matrix figuring in (A3) *and I is the Fisher information matrix defined by* (2.3).

PROOF. By assumptions about Φ, let $\phi(1) = \phi'(1) = 0$ and let us introduce auxiliary function

$$v(\boldsymbol{y}, \theta) = \left(q_j^{1/2} \phi' \left(\frac{p_j(\boldsymbol{y}, \theta)}{q_j} \right) : 1 \leq j \leq m \right)$$

of vector variables $(\boldsymbol{y}; \theta)$ from the neighborhood of $(\boldsymbol{y}_0; \theta_0)$. It follows from (1.20) that

$$P\{v(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)) A(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)) \neq 0\} = o(1)$$

where $A(\boldsymbol{y}, \theta)$ is defined in (A2). If we apply the Taylor formula to $v(\boldsymbol{y}_n, \theta_0) - v(\boldsymbol{y}_0, \theta_0)$ and $v(\boldsymbol{y}_n, \theta) - v(\boldsymbol{y}_n, \theta_0)$ and use the fact that $\phi'(1) = 0$ implies $v(\boldsymbol{y}_0, \theta_0) = 0$, then we get the desired result from Lemma 2.1. For details we refer to Menéndez *et al.* (1998).

It remains to formulate an appropriate consistency condition for the estimators $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi)$, $\phi \in \Phi$. To this end is needed an identifiability condition for true θ_0 similar to (B4) in the model $(p(\boldsymbol{y}_0, \theta) : \theta \in \Theta)$. Bofinger (1973) 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 in other places of that paper. In Menéndez *et al.* (1998) we presented an example which demonstrates that (B4) is under (A1)–(A4) not sufficient for the consistency. For the consistency is also needed the following extension of Proposition 9.49 in Vajda (1989), established in Menéndez *et al.* (1998).

LEMMA 2.3. *Let $p_n = (p_{n0}, \dots, p_{nm})$ be a sequence of random probability vectors. If for a fixed probability $(m+1)$-vector q with all coordinates positive, and for $\phi \in \Phi$,*

$$D_\phi(p_n; q) = o_p(1)$$

then $\|p_n - q\|^2$ tends stochastically to zero with at least the same rate as $D_\phi(p_n; q)$ or, more precisely,

$$\|p_n - q\|^2 \le \frac{2}{\phi''(1)} D_\phi(p_n; q) + o_p(D_\phi(p_n; q)).$$

Now we can formulate the consistency condition.

(A5) For all $\boldsymbol{y}$ from the neighborhood of $\boldsymbol{y}_0$, the mappings $\theta \mapsto F(\boldsymbol{y}, \theta)$ are 1-1 on Θ.

LEMMA 2.4. *If* (A1)–(A5) *hold then all estimators $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi)$, $\phi \in \Phi$, are consistent.*

PROOF. See the Appendix.

The results of Lemmas 2.2–2.4 can be summarized as follows.

THEOREM 2.1. *If* (A1)–(A5) *hold then all estimators $\tilde{\theta}_n(\boldsymbol{\lambda}, \phi)$, $\phi \in \Phi$, are efficient in the sense of* (2.13) *and asymptotically normal in the sense of* (2.14).

3. Optimum partitions and efficiency

In Section 2 we have shown that the random quantization (1.18) leads to the same efficiency of minimum disparity estimators as the quantization (1.12). This efficiency is in some sense characterized by the Fisher information figuring in (A3) and denoted in Section 1 by $I_m(\theta_0, \boldsymbol{\lambda})$. In this section we suppose for simplicity that the parameter θ is real, from an open interval $\Theta \subset R$. Then, by (2.2) and (2.3),

$$I_m(\theta_0, \boldsymbol{\lambda}) = \sum_{j=1}^{m} \frac{\dot{\pi}_j(\theta_0)^2}{\pi_j(\theta_0)}, \tag{3.1}$$

where are used the alternative symbols

$$\pi(\theta) = (\pi_1(\theta), \dots, \pi_m(\theta)) \overset{\triangle}{=} p(\boldsymbol{y}_0, \theta), \qquad \dot{\pi}(\theta) \overset{\triangle}{=} \frac{d\pi(\theta)^t}{d\theta}, \qquad \theta \in \Theta, \tag{3.2}$$

for $p(\boldsymbol{y}_0, \theta)$ given by (1.2) and (1.12) and for the gradient $G(\boldsymbol{y}_0, \theta)$ given in (A2).

Since the partitions are specified by vectors $\boldsymbol{\lambda}$, an *optimum partition* $\boldsymbol{\lambda}_{opt}$ can be defined by the condition

$$I_m(\boldsymbol{\lambda}_{opt}) = \max_{\boldsymbol{\lambda}} I_m(\theta_0, \boldsymbol{\lambda}).$$

By (2.1), $\pi(\theta_0) = (\lambda_j - \lambda_{j-1} : 1 \le j \le m)$, and by (3.2)

$$\dot{\pi}(\theta) = (s(\theta, \lambda_j) - s(\theta, \lambda_{j-1}) : 1 \le j \le m), \qquad \text{where}$$

$$s(\theta, \tau) = \frac{dF(F^{-1}(\tau, \theta_0), \theta)}{d\theta}, \qquad 0 \le \tau \le 1. \tag{3.3}$$

Thus

$$I_m(\theta_0, \boldsymbol{\lambda}) = \sum_{j=1}^{m} \frac{[s(\theta_0, \lambda_j) - s(\theta_0, \lambda_{j-1})]^2}{\lambda_j - \lambda_{j-1}} \tag{3.4}$$

and a necessary condition for (1.14) is the stationarity

$$\left(\frac{\partial}{\partial \lambda_1}, \dots, \frac{\partial}{\partial \lambda_{m-1}}\right) I_m(\theta_0, \boldsymbol{\lambda}) = 0. \tag{3.5}$$

If these equations have only one solution in the domain $0 < \lambda_1 < \cdots < \lambda_{m-1} < 1$, and the function $I_m(\theta_0, \boldsymbol{\lambda})$ is in this domain concave, then (3.5) is necessary and sufficient for $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{opt}$.

Maximization of functions of the type (3.4) has been studied by Cheng (1975), Nagahata (1985), Pötzelberger and Felsenstein (1993) and Tsairidis *et al.* (1998). These authors established under mild restrictions on the basic model $(F(x, \theta) : \theta \in \Theta)$ the existence of a solution of (3.5) which, under additional reasonable restrictions, is the desired $\boldsymbol{\lambda}_{opt}$. Unfortunately, as can be expected, $\boldsymbol{\lambda}_{opt}$ in general depends on the parameter θ_0, with the exception of models invariant in an apropriate sense, such as e.g. the location models. It follows from the results presented by these authors that the uniform $\boldsymbol{\lambda}_{unif}$ defined by (1.16) need not in general be $\boldsymbol{\lambda}_{opt}$. However, as follows from the numerical results presented in these papers, and also from our own numerical studies, if m is not too small then in the most common statistical models the absolute as well as relative inefficiencies of the quantization (1.18) using $\boldsymbol{\lambda}_{unif}$,

$$I_m(\theta_0, \boldsymbol{\lambda}_{opt}) - I_m(\theta_0, \boldsymbol{\lambda}_{unif}) \qquad \text{and} \qquad \frac{I_m(\theta_0, \boldsymbol{\lambda}_{opt}) - I_m(\theta_0, \boldsymbol{\lambda}_{unif})}{I_m(\theta_0, \boldsymbol{\lambda}_{opt})},$$

are close to zero. A typical situation is illustrated by Table 1 presenting the situation in three common location families $(F(x - \theta) : \theta \in R)$.

An obvious advantage of the empirical quantization using $\boldsymbol{\lambda}_{unif}$ is that it requires neither the knowledge of the true parameter θ_0 nor the knowledge of the basic model $(F(x, \theta) : \theta \in \Theta)$ itself. Nevertheless, it guarantees the efficiency characterized by the Fisher information $I_m(\theta_0, \boldsymbol{\lambda}_{unif})$ in any underlying reduced model satisfying (A1)–(A5). This is a practical gain which certainly compensates the small relative inefficiency $I_m(\theta_0, \boldsymbol{\lambda}_{opt}) - I_m(\theta_0, \boldsymbol{\lambda}_{unif})$. To see how this argument practically works, look at the following simple example.

Example 3.1. Consider the estimation of location, and assume that the model is normal. If we employ an estimator $\tilde{\theta}_n(\boldsymbol{\lambda}_{unif}, \phi)$, $\phi \in \Phi$, with $m = 16$, then it follows from the last column of Table 1 that the relative loss

$$\rho_m = 100 \cdot [I_m(\boldsymbol{\lambda}_{opt}) - I_m(\boldsymbol{\lambda}_{unif})] / I_m(\boldsymbol{\lambda}_{opt})$$

Table 1. Fisher information $I_m(\theta_0, \boldsymbol{\lambda}_{unif}) = I_m(\boldsymbol{\lambda}_{unif})$ for given location families. For the logistic and doubly exponential families $I_m(\theta_0, \boldsymbol{\lambda}_{opt}) = I_m(\boldsymbol{\lambda}_{opt})$ coincides with $I_m(\boldsymbol{\lambda}_{unif})$. For the normal family the values of $I_m(\boldsymbol{\lambda}_{opt})$ are in parentheses.

Family	$m=2$	$m=3$	$m=4$	$m=5$	$m=8$	$m=16$
Logistic $F(x) = \frac{1}{1+e^{-x}}$	0.2500	0.2963	0.3125	0.3200	0.3281	0.3320
Normal	(0.6366)	(0.8098)	(0.8825)	(0.9201)	(0.9665)	(0.9905)
$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2}$	0.6366	0.7932	0.8606	0.8970	0.9450	0.9778
Doubly exponential $f(x) = \frac{1}{2} e^{-\lvert x \rvert}$	1	$\frac{2}{3}$	1	$\frac{4}{5}$	1	1

Table 2. Relative asymptotic inefficiencies $\eta_m = 100 \cdot (\mathcal{J} - I_m(\boldsymbol{\lambda}_{unif}))/\mathcal{J}$ (in %) of the ϕ–divergence estimators using $\boldsymbol{\lambda}_{unif}$ in the location families of Table 1. For the location families $\mathcal{J} = \mathcal{J}(\theta)$ is constant for all $\theta \in R$.

Family	$\mathcal{J} = \int \frac{\dot{f}^2}{f} dx$	$m=2$	$m=3$	$m=4$	$m=5$	$m=8$	$m=16$
Logistic	1/3	25	11.1	6.3	4	1.6	0.4
Normal	1	36.3	20.6	14	8.1	4.7	2.1
Doubly exponential	1	0	33.33	0	20	0	0

of asymptotic accuracy against $\tilde{\theta}_n(\boldsymbol{\lambda}_{opt}, \phi)$ is $1.27/0.9905 = 1.28\%$. In this case the model is invariant, so that $\boldsymbol{\lambda}_{opt}$ does not depend on the true θ_0, but it strongly depends on the assumption that the true model is normal. If the true model is logistic or doubly exponential, then $\boldsymbol{\lambda}_{opt} = \boldsymbol{\lambda}_{unif}$. Therefore the relative loss ρ_m of $\tilde{\theta}_n(\boldsymbol{\lambda}_{unif}, \phi)$ in these models will be zero. Replacing $\boldsymbol{\lambda}_{unif}$ by $\boldsymbol{\lambda}_{opt}$ for normal, we raise this loss to nonzero levels. Hence the use of $\boldsymbol{\lambda}_{unif}$ guarantees a robustness of all estimators under consideration in the class of location models with ρ_m small enough.

The problem of efficiency mentioned in the title of this section can be naturally interpreted as the evaluation of absolute or relative asymptotic inefficiency

$$\mathcal{J}(\theta_0) - I_m(\theta_0, \boldsymbol{\lambda}) \quad \text{or} \quad \frac{\mathcal{J}(\theta_0) - I_m(\theta_0, \boldsymbol{\lambda})}{\mathcal{J}(\theta_0)}$$

for estimators studied in this paper, where in this case we mean the inefficiency with respect to what is achievable in the basic continuous model $(F(x,\theta) : \theta \in \Theta)$.

Our regularity conditions (A1)–(A5), guaranteeing the existence of informations $I_m(\theta_0, \boldsymbol{\lambda})$, 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 $I_m(\theta_0, \boldsymbol{\lambda})$ in the reduced models given by (3.1) or (3.4).

We shall consider the conditions for existence of Fisher informations $\mathcal{J}(\theta_0)$, $\theta_0 \in \Theta$, introduced in Vajda (1973) (condition $C_{\in}$ on p. 280 ibid.), namely that the derivatives $\dot{f}(x,\theta) = df(x,\theta)/d\theta$ of densities $f(x,\theta) = dF(x,\theta)/dx$ exist at θ_0 for almost all x, and for some $\varepsilon > 0$ (possibly depending on θ_0)

$$(3.6) \qquad \int \sup_{|\theta - \theta_0| < \varepsilon} \left(\frac{f(x,\theta) - f(x,\theta_0)}{(\theta - \theta_0) f(x,\theta_0)} \right)^2 f(x,\theta_0) dx < \infty.$$

Under this condition

$$\mathcal{J}(\theta_0) = \int \frac{\dot{f}(x,\theta)^2}{f(x,\theta)} dx < \infty. \tag{3.7}$$

For example, the doubly exponential model of location considered in Table 1 does not satisfy the standard regularity assumptions of the asymptotic statistics but satisfies (3.6) and, by (3.7), $\mathcal{J}(\theta_0) = 1$ for every location $\theta_0 \in R$.

By Theorem 3 in Vajda (1973), if $\mathcal{J}(\theta_0)$ is finite then $\mathcal{J}(\theta_0) - I_m(\theta_0, \boldsymbol{\lambda}) \geq 0$ for every $\boldsymbol{\lambda}$ under consideration. If all densities $\{f(x,\theta) : \theta \in \Theta\}$ have a common support then Theorem 4 ibid. implies $\mathcal{J}(\theta_0) = I_m(\theta_0, \boldsymbol{\lambda}_{unif}) + o(1)$ asymptotically for $m = r^k$, any integer $r > 1$, and $k \to \infty$.

The values of $\mathcal{J} = \mathcal{J}(\theta_0)$ and the relative inefficiencies of the ϕ-divergence estimators using $\boldsymbol{\lambda}_{unif}$ in the location models of Table 1 can be seen in Table 2.

Acknowledgements

We thank the referees for their helpful remarks and a stimulation to include the new Section 3.

Appendix

Proof of Lemma 2.4. Obviously,

$$0 \leq D_\phi(p(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)); q) \leq D_\phi(p(\boldsymbol{y}_n, \theta_0); q).$$

Using the Taylor expansion of $\phi(t)$ around $t = 1$ one obtains from (1.4) and (2.10) that $D_\phi(p(\boldsymbol{y}_n, \theta_0); q) = O_p(n^{-1})$. Consequently also

$$D_\phi(p(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)); q) = O_p(n^{-1}).$$

This together with the Lemma 2.3 implies $\|p(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)) - q\| = O_p(n^{-1/2})$. We shall use only the weaker relation

$$\|p(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)) - q\| = o_p(1). \tag{A.1}$$

Further, (A5) implies that there exists an open neighborhood U of θ_0 such that for all $\boldsymbol{y}$ from a closed ball V centered at $\boldsymbol{y}_0$ and $\varepsilon = \varepsilon(\boldsymbol{y})$ possibly depending on $\boldsymbol{y}$,

$$\|F(\boldsymbol{y}, \theta) - \boldsymbol{\lambda}\| < \varepsilon \Rightarrow \theta \in U. \tag{A.2}$$

However, due to the compactness of V,

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

Moreover, the neighborhoods V and U can be chosen so 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 neighborhood of $\boldsymbol{\lambda} = F(\boldsymbol{y}_0, \theta_0)$. Finally, by (A.1) and (2.12), $\|p(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)) - p(\boldsymbol{y}_n, \theta_0)\| = o_p(1)$. By the right-hand inequality in (2.6), this implies

$$\|F(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi)) - F(\boldsymbol{y}_0, \theta_0)\| = o_p(1).$$

Consistency follows from this relation and (A.2) by using the identities

$$\varphi(F(\boldsymbol{y}_n, \tilde{\theta}_n(\boldsymbol{\lambda}, \phi))) = \tilde{\theta}_n(\boldsymbol{\lambda}, \phi), \qquad \varphi(F(\boldsymbol{y}_0, \theta_0)) = \theta_0$$

and the continuity of φ.

REFERENCES

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

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

Cheng, R. C. H. (1975). A unified approach to choosing optimum quantiles for the ABLE's, *J. Amer. Statist. Assoc.*, **70**, 155–159.

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

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

Ferentinos, K. and Papaioannou, T. (1979). Loss of information due to groupings, *Transactions of the Eighth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, 87–94, Prague Academia.

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

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

Menéndez, M. L., Morales, D. and Pardo, L. (1997). Maximum entropy principle and statistical inference on condensed ordered data, *Statist. Probab. Lett.*, **34**, 85–93.

Menéndez, M. L., Morales, D., Pardo, L. and Vajda, I. (1998). Two approaches to grouping of data and related disparity statistics, *Comm. Statist. Theory Methods*, **27**(3), 609–633.

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

Nagahata, H. (1985). Optimal spacing for grouped observations from the information view-point, *Mathematica Japonica*, **30**, 277–282.

Neyman, J. (1949). Contribution to the theory of the χ^2 test, *Proceeding of the First Berkeley Symposium on Mathematical Statistics and Probability*, 239–273. Berkeley University Press, Berkeley, California.

Pötzelberger, K. and Felsenstein, K. (1993). On the Fisher information of discretized data, *J. Statist. Comput. Simulation*, **46**, 125–144.

Rao, C. R. (1961). Asymptotic efficiency and limiting information, *Proc. Fourth Berkeley Symp. on Math. Statist. Prob.*, Vol. 1, 531–545, Berkeley University Press, Berkeley, California.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York.

Tsairidis, Ch., Zografos, K. and Ferentinos, T. (1998). Fisher's information matrix and divergence for finite optimal partitions of the sample space, *Comm. Statist. Theory Methods.*, **26**(9), 2271–2289.

Vajda, I. (1973). χ^2–divergence and generalized Fisher information, *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, 223–234, Prague Academia.

Vajda, I. (1989). *Theory of Statistical Inference and Information*, Kluwer, Boston.