KYBERNETIKA — VOLUME *33* (1997), NUMBER 5, PAGES 489 – 504

# ASYMPTOTIC BEHAVIOUR OF AN ESTIMATOR BASED ON RAO'S DIVERGENCE[1]

María Carmen Pardo

In this work the procedure of minimum divergence estimation based on Burbea and Rao [2] divergence is analyzed. Asymptotic behaviour for these estimators is given. A comparative study of Rao's estimator with other classical estimators is carried out by computer simulation.

## 1. INTRODUCTION

In this paper we consider a wide class of estimators which can be used when the data are discrete, either the underlying distribution is discrete or it is continuous but the observations are classificated into groups. The latter situation can occur either by experimental reasons or because the estimation problem without grouped data is not easy to resolve, see Fryer and Robertson [4]. For example, the maximum likelihood estimator for the five parameters of a mixture of two normal distributions based on non grouped data does not exist, cf. Kiefer and Wolfowitz [5]. An easy way to resolve this problem is to group the data and use a multinomial model.

Consider the probability densities $f_\theta(x)$ with respect to a $\sigma$-finite measure $\mu$ on the statistical space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, P_\theta)_{\theta\in\Theta\subseteq R^{M_0}}$ and the decomposition $\{A_1,\dots,A_M\}$ of $\mathcal{X}$. Then the formula $P_\theta(A_i) = q_i(\theta),\ i = 1,\dots,M$ defines a discrete statistical model. Let $X_1,\dots,X_n$ be a random sample drawn from the previous population and let $\hat{p}_i = \frac{n_i}{n}$ be the relative frequency of $A_i,\ i = 1,\dots,M$. If we are interested in estimating $\theta$, the most natural point estimator is the maximum likelihood estimator (MLE). The statistic $(N_1 = n_1,\dots,N_M = n_M)$ is obviously sufficient for the statistical model under consideration and multinomial, i. e.

$$P_\theta(N_1 = n_1,\dots,N_M = n_M) = \frac{n!}{n_1!\cdots n_M!}\, q_1(\theta)^{n_1}\cdots q_M(\theta)^{n_M}$$

so that

$$\log P_\theta(N_1 = n_1,\dots,N_M = n_M) = -n\, D^{\mathrm{KULLBACK}}(\hat{P},\, Q(\theta)) + o(n),$$

[1] This research has been supported by DGICYT grant PB 94–0308.

where $\hat{P} = (\hat{p}_1, \dots, \hat{p}_M)^t$, $Q(\theta) = (q_1(\theta), \dots, q_M(\theta))^t$ and $D^{\text{KULLBACK}}$ is the Kullback divergence (Kullback and Leibler [6]). Therefore, to estimate $\theta$ by the discrete model maximum likelihood estimator is equivalent to minimize on $\theta \in \Theta \subseteq R^{M_0}$ the Kullback divergence.

Since the Kullback divergence is not the unique divergence measure we can choose as an estimator the value $\tilde{\theta}$ which satisfies the condition

$$D(\hat{P}, Q(\tilde{\theta})) = \inf_{\theta \in \Theta \subseteq R^{M_0}} D(\hat{P}, Q(\theta)),$$

where $D$ is an arbitrary divergence measure. For example, Morales et al [7] considered the Csiszár divergence.

## 2. THE MINIMUM $R_\phi$–DIVERGENCE ESTIMATOR

Throughout this paper we consider the divergence of Burbea and Rao [2], i. e. we consider a continuous concave function $\phi(t) : (0, \infty) \to R$, and we put

$$\phi(0) = \lim_{t \downarrow 0} \phi(t) \in (-\infty, \infty].$$

The concavity of $\phi$ implies that the function $\delta_\phi : [0,1]^2 \to (-\infty, \infty]$ defined by

$$\delta_\phi(u,v) = \begin{cases} \phi\left(\dfrac{u+v}{2}\right) - \dfrac{\phi(u) + \phi(v)}{2} & \text{if} \quad (u,v) \neq (0,0) \\ 0 & \text{if} \quad (u,v) = (0,0) \end{cases}$$

is nonnegative. The corresponding distance

$$R_\phi(P, Q) = \sum_{i=1}^{M} \delta_\phi(p_i, q_i)$$

is the divergence of Burbea and Rao [2]. Some of the properties of this distance were studied by Pardo and Vajda [8].

Let $X_1, \dots, X_n$ be a random sample belonging to a population with an unknown parameter $\theta \in \Theta \subseteq R^{M_0}$, $M_0 < M - 1$, and let there exist a function $Q(\theta) = (q_1(\theta), \dots, q_M(\theta))^t$ that maps each $\theta = (\theta_1, \dots, \theta_{M_0})^t$ into a point in $\Delta_M = \{P = (p_1, \dots, p_M)^t | \sum_{i=1}^{M} p_i = 1,\, p_i \geq 0,\, i = 1, \dots, M\Big\}$. As $\theta$ ranges over the values of $\Theta$, $Q(\theta)$ ranges over a subset $T$ of $\Delta_M$. When we assume that a given model is 'correct', we just assume that there exists a value $\theta^0$ of $\theta$ such that $Q(\theta^0) = \pi$, where $\pi$ is the true value of the multinomial probability, i. e., $\pi \in T$.

**Definition 1.** Let us suppose that $n$ observations are drawn at random and with replacement from a population with the statistical space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P_\theta)_{\theta \in \Theta \subseteq R^{M_0}}$. Then the minimum $R_\phi$-divergence estimator of $\theta$ is $\hat{\theta}_\phi \in \overline{\Theta}$ satisfying the condition

$$R_\phi(\hat{P}, Q(\hat{\theta}_\phi)) = \inf_{\theta \in \Theta} R_\phi(\hat{P}, Q(\theta)),$$

where $\hat{P}$ is the relative frequency vector.

In other words, the minimum $R_\phi$-divergence estimator satisfies the condition $\hat{\theta}_\phi = \arg\inf_{\theta\in\Theta} R_\phi(\hat{P},\, Q(\theta))$.

**Example 1.** Suppose that $n$ independent and identically distributed Poisson variables with mean $\theta$ are observed, and let the observations be truncated at $x = 2$. Let $N_1$, $N_2$ and $N_3$ be the numbers of observations taking on the values 0, 1 and 2 or more, respectively. Then $X = (N_1, N_2, N_3)$ has the trinomial distribution $(n; q_1(\theta),\, q_2(\theta),\, q_3(\theta))$, where

$$\begin{aligned} q_1(\theta) &= P_\theta(X=0) = e^{-\theta} \\ q_2(\theta) &= P_\theta(X=1) = \theta\, e^{-\theta} \\ q_3(\theta) &= P_\theta(X \geq 2) = 1-(1+\theta)\, e^{-\theta}. \end{aligned}$$

If we consider the $R_\phi$-divergence for $\phi(x) = -x\ln x$ then the evaluation of $\hat{\theta}_1$ based on this divergence is equivalent to finding $\theta$ that minimizes the function

$$\begin{aligned} R(\hat{P},\, Q(\theta)) = {} & \frac{\hat{p}_1 \ln \hat{p}_1 + e^{-\theta}\ln e^{-\theta}}{2} + \frac{\hat{p}_2 \ln \hat{p}_2 + \theta\, e^{-\theta}\ln \theta\, e^{-\theta}}{2} \\ & + \frac{\hat{p}_3 \ln \hat{p}_3 + (1-(1+\theta)\, e^{-\theta})\ln(1-(1+\theta)\, e^{-\theta})}{2} \\ & - \left( \frac{\hat{p_1} + e^{-\theta}}{2} \ln \frac{\hat{p_1} + e^{-\theta}}{2} + \frac{\hat{p_2} + \theta\, e^{-\theta}}{2} \ln \frac{\hat{p_2} + \theta\, e^{-\theta}}{2} \right. \\ & \left. + \frac{\hat{p_3} + (1-(1+\theta)\, e^{-\theta})}{2} \ln \frac{\hat{p_3} + (1-(1+\theta)\, e^{-\theta})}{2} \right). \end{aligned}$$

For the observed frequency vector $\hat{P} = (0.2,\, 0.3,\, 0.5)^t$ we obtain $\hat{\theta}_1 = 1.661$. In this case

$$q_1(\hat{\theta}_1) = 0.19, \quad q_2(\hat{\theta}_1) = 0.31, \quad q_3(\hat{\theta}_1) = 0.5$$

and

$$R(\hat{P},\, Q(\hat{\theta}_1)) = 0.1734.$$

Geometrically, $\Delta_3$ is the triangle side ABC depicted in Figure 1, that we represent in the plane through the triangle of Figure 2.

As $\theta$ varies over $R^+ = [0, \infty)$, $Q(\theta) = \left(e^{-\theta},\, \theta\, e^{-\theta},\, 1-(1+\theta)\, e^{-\theta}\right)^t$, traces out an one-dimensional curve in $\Delta_3$. This curve is the subset $T$. When $\theta \to 0$, $Q(\theta) \to (1,\, 0,\, 0)^t$, and when $\theta \to \infty$, $Q(\theta) \to (0,\, 0,\, 1)^t$. Thus the boundary points of $\theta$ in this example correspond to the boundary points of $\Delta_3$. Figure 2 shows the relationships between $\Delta_3$, $T$, $\pi$ and $\hat{P}$ in this example. If the Poisson model is incorrect, then the true value of $\pi$ does not generally lie on the curve, although in principle it can. Because of the discreteness of the multinomial distribution, it often happens that $\hat{P}$ does not lie on $T$ (as is the case in the figure). The estimation method based on the minimum distance leads to a point in $T$ closest to $\hat{P}$ in the sense of the chosen distance.

**Fig. 1.**

**Fig. 2.**

## 3. PROPERTIES OF THE MINIMUM $R_\phi$–DIVERGENCE ESTIMATOR

Throughout the paper, we assume that the model is correct, so that $\pi = Q(\theta^0)$, and $M_0 < M - 1$. Furthermore, we restrict ourselves to unknown parameters $\theta^0$ satisfying the regularity conditions $1-6$ introduced by Birch [1]:

1. $\theta^0$ is an interior point of $\Theta$.

2. $\pi_i = q_i(\theta^0) > 0$ for $i = 1, \ldots, M$. Thus $\pi = (\pi_1, \ldots, \pi_M)^t$ is an interior point of $T$.

3. The mapping $Q : \Theta \to \Delta_M$ is totally differentiable at $\theta^0$ so that the partial derivatives of $q_i$ with respect to each $\theta_j$ exist at $\theta^0$ and $Q(\theta)$ has a linear approximation at $\theta^0$ given by

$$q_i(\theta) = q_i(\theta^0) + \sum_{j=1}^{M} (\theta_j - \theta_j^0) \, \frac{\partial q_i(\theta^0)}{\partial \theta_j} + o\left(\|\theta - \theta^0\|\right)$$

as $\theta \to \theta^0$.

4. The Jacobian matrix
$$\left(\frac{\partial Q(\theta)}{\partial \theta}\right)_{\theta=\theta^0} = \left(\frac{\partial q_i(\theta^0)}{\partial \theta_j}\right)_{\substack{i=1,\dots,M \\ j=1,\dots,M_0}}$$
is of full rank (i. e. of rank $M_0$).

5. The inverse mapping $Q^{-1} : T \to \Theta$ is continuous at $Q(\theta^0) = \pi$.

6. The mapping $Q : \Theta \to \Delta_M$ is continuous at every point $\theta \in \Theta$.

**Definition 2.** An estimator, $\hat{S}$, of $Q(\theta^0) = (q_1(\theta^0), \dots, q_M(\theta^0))^t$ is $c_n$-consistent if
$$c_n \left\| \hat{S} - Q(\theta^0) \right\| \le O_p(1).$$

For a sequence $\{Y_n\}_{n \in N}$ of random variables the relation
$$Y_n \le O_p(1)$$
means that
$$\lim_{c \to \infty} \liminf_{n \to \infty} P(|Y_n| < c) = 1,$$
i. e., either $Y_n$ is bounded in probability or $Y_n$ converges in probability to zero.

If $c_n \uparrow \infty$, then the $c_n$-consistency of an estimator is stronger than the consistency, i. e. every $c_n$-consistent estimator is consistent. It is also clear that if an estimator $\hat{S}_1$ is $c_n^1$-consistent and an estimator $\hat{S}_2$ is $c_n^2$-consistent then both $\hat{S}_1$ and $\hat{S}_2$ are $c_n$-consistent for $c_n = \min\{c_n^1, c_n^2\}$.

Let us introduce additional notations. We consider the linear differential operator
$$\frac{\mathrm{d}}{\mathrm{d}\theta} = \left(\frac{\partial}{\partial \theta_1}, \dots, \frac{\partial}{\partial \theta_{M_0}}\right),$$
and $M \times M_0$ Jacobian matrix $J(\theta) = (J_{jr}(\theta))$ where
$$J_{jr}(\theta) = \frac{\partial q_j(\theta)}{\partial \theta_r}.$$

Further, we define
$$q_j' = \frac{\mathrm{d} q_j(\theta)}{\mathrm{d}\theta} = (J_{j1}(\theta), \dots, J_{jM_0}(\theta))$$
and
$$A(\theta) = \operatorname{diag}\left(\sqrt{-\phi''(q_1(\theta))}, \dots, \sqrt{-\phi''(q_M(\theta))}\right) J(\theta).$$

To prove Theorem 1 we use the Implicit Function Theorem given below:

"Let $F = (F_1, \ldots, F_{M_0}) : R^{M+M_0} \to R^{M_0}$ be continuously differentiable in an open set $U \subset R^{M+M_0}$, containing the point $\left(x^* = (x_1^*, \ldots, x_M^*)^t;\, x_0 = (x_1^0, \ldots, x_{M_0}^0)^t\right)$ for which $F(x^{*t}, x_0^t) = 0$. Further, suppose that the matrix

$$\left(\frac{\partial F_i}{\partial x_j}\right)_{\substack{i=1,\ldots,M_0 \\ j=M+1,\ldots,M+M_0}}$$

is nonsingular at $(x^*, x_0)$. Then there exists a $M$-dimensional neighborhood $U_0$ of $x^*$ in $R^M$ and a unique, continuously differentiable function $g : U_0 \to R^{M_0}$ such that $g(x^*) = x_0$ and $F(x^t,\, g(x)^t) = 0 \;\; \forall\, x \in U_0$."

**Theorem 1.** Let $\phi : (0, \infty) \to R$ be a twice continuously differentiable concave function. Under the Birch regularity conditions and assuming that the function $Q : \Theta \to \Delta_M$ has continuous second partial derivatives in a neighbourhood of $\theta^0$, it holds

$$\begin{aligned}\hat{\theta}_\phi \;\; &= \;\; \theta^0 + \left(A(\theta^0)^t \,(A(\theta^0)\right)^{-1} A(\theta^0)^t \mathrm{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right)(\hat{P} - Q(\theta^0)) \\ &\qquad\qquad + o\left(\|\hat{P} - Q(\theta^0)\|\right)\end{aligned}$$

where $\hat{\theta}_\phi$ is unique in a neighbourhood of $\theta^0$.

*Proof.* Let $1^M$ be the interior of the unit $M$-dimensional cube. Let $U$ be a neighbourhood of $\theta^0$ on which $Q : \Theta \to \Delta_M$ has continuous second partial derivatives. Let

$$F = (F_1, \ldots, F_{M_0}) : 1^M \times U \to R_0^M$$

be defined by

$$F_j\left(p_1, \ldots, p_M; \theta_1, \ldots, \theta_{M_0}\right) = \frac{\partial R_\phi(P,\, Q(\theta))}{\partial \theta_j} \quad j = 1, \ldots, M_0.$$

It holds

$$F_j\left(\pi_1, \ldots, \pi_M; \theta_1^0, \ldots, \theta_{M_0}^0\right) = 0, \quad j = 1, \ldots, M_0,$$

due to

$$\frac{\partial R_\phi(P,\, Q(\theta))}{\partial \theta_j} = \frac{1}{2}\sum_{i=1}^{M}\left\{\phi'\left(\frac{p_i + q_i(\theta)}{2}\right) - \phi'(q_i(\theta))\right\}\frac{\partial q_i(\theta)}{\partial \theta_j}, \quad j = 1, \ldots, M_0.$$

Since

$$\begin{aligned}&\frac{\partial}{\partial \theta_r}\left(\frac{\partial R_\phi(P,\, Q(\theta))}{\partial \theta_j}\right) \\ &= \frac{1}{2}\sum_{i=1}^{M}\left\{\frac{\partial q_i(\theta)}{\partial \theta_r}\left(\frac{1}{2}\phi''\left(\frac{p_i + q_i(\theta)}{2}\right) - \phi''(q_i(\theta))\right)\frac{\partial q_i(\theta)}{\partial \theta_j}\right. \\ &\qquad \left. + \left(\phi'\left(\frac{p_i + q_i(\theta)}{2}\right) - \phi'(q_i(\theta))\right)\frac{\partial^2 q_i(\theta)}{\partial \theta_j \partial \theta_r}\right\},\end{aligned}$$

we have

$$\begin{aligned}\left(\frac{\partial F}{\partial \theta^0}\right) &= \left(\frac{\partial}{\partial \theta_r}\left(\frac{\partial R_\phi(\pi,\, Q(\theta^0))}{\partial \theta_j}\right)\right)_{\substack{i=1,\ldots,M_0\\ r=1,\ldots,M_0}}\\ &= \frac{1}{2}\sum_{i=1}^{M}\left\{\frac{\partial q_i(\theta^0)}{\partial \theta_r}\,\frac{\partial q_i(\theta^0)}{\partial \theta_j}\left(-\frac{1}{2}\right)\phi''(q_i(\theta^0))\right\}\\ &= \frac{1}{4}\,A(\theta^0)^t\,A(\theta^0).\end{aligned}$$

Taking into account that if $B$ is a $p \times q$ matrix and $C$ is a nonsingular matrix, then $\operatorname{rank}(BC) = \operatorname{rank}(B)$, and putting

$$B = \left(\frac{\partial q_i(\theta^0)}{\partial \theta_r}\right)^t_{\substack{i=1,\ldots,M\\ r=1,\ldots,M_0}} \qquad \text{and} \quad C = \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right)_{M\times M},$$

it follows that $A^t(\theta^0)$ and $A(\theta^0)$ have rank $M_0$. Also,

$$\operatorname{rank}(A^t(\theta^0)\,A(\theta^0)) = \operatorname{rank}(A(\theta^0)\,A^t(\theta^0)) = \operatorname{rank}(A(\theta^0)) = M_0.$$

Therefore, the matrix

$$\left(\frac{\partial F_j}{\partial \theta_r}\right)_{\substack{j=1,\ldots,M_0\\ r=1,\ldots,M_0}}$$

is nonsingular at $\theta^0$.

Applying the Implicit Function Theorem there exists an $M$-dimensional neighbourhood $U_0$ of $\pi = (\pi_1, \ldots, \pi_M)^t$ in $R^M$ and a unique, continuously differentiable function $\tilde{\theta} : U_0 \to R^{M_0}$ such that

$$F(P^t,\, \tilde{\theta}(P)^t) = 0 \qquad \forall\, P \in U_0$$

and

$$\tilde{\theta}(\pi) = \theta^0.$$

By the chain rule,

$$\frac{\partial F(P^t,\, \tilde{\theta}(P^t))}{\partial P} + \frac{\partial F(P^t,\, \tilde{\theta}(P^t))}{\partial \theta(P)}\,\frac{\partial \theta(P)}{\partial P} = 0$$

and, for $P = \pi$,

$$\frac{\partial F}{\partial \pi} + \frac{\partial F}{\partial \theta^0}\,\frac{\partial \theta^0}{\partial \pi} = 0.$$

Further, we know that

$$\frac{\partial F}{\partial \theta^0} = \frac{1}{4}\,A(\theta^0)^t\,A(\theta^0)$$

and

$$\frac{\partial F}{\partial \pi} = \frac{1}{4}\,J(\theta^0)^t\operatorname{diag}(\phi''(\pi)) = -\frac{1}{4}\,A(\theta^0)^t\operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right)$$

so that

$$\frac{\partial\theta^0}{\partial\pi} = \left(A(\theta^0)^t\, A(\theta^0)\right)^{-1}\, A(\theta^0)^t \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right).$$

The Taylor expansion of $\tilde{\theta}(P)$ around $\pi$ yields

$$\tilde{\theta}(P) = \tilde{\theta}(\pi) + \left(\frac{\partial\tilde{\theta}}{\partial P}\right)_{P=\pi} (P-\pi) + o\left(\|P-\pi\|\right).$$

For $\tilde{\theta}(\pi) = \theta^0$ we obtain from here

$$\tilde{\theta}(P) = \theta^0 + \left(A(\theta^0)^t\, A(\theta^0)\right)^{-1}\, A(\theta^0)^t \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right)(P-\pi) + o\left(\|P-\pi\|\right).$$

Therefore $\hat{P} \xrightarrow[n\to\infty]{\text{a. s.}} \pi$, so that $\hat{P} \in U_0$ and, consequently, $\tilde{\theta}(\hat{P})$ is the unique solution of equations

$$\frac{\partial R_\phi(\hat{P},\, \tilde{\theta}(\hat{P}))}{\partial\theta_j} = 0, \quad j = 1,\dots,M_0,$$

in the neighbourhood of $\pi$. Thus $\tilde{\theta}(\hat{P})$ is the minimum $R_\phi$-divergence estimator, $\hat{\theta}_\phi$, satisfying the relation

$$\begin{aligned}\hat{\theta}_\phi(\hat{P}) \quad &= \quad \theta^0 + \left(A(\theta^0)^t\, A(\theta^0)\right)^{-1}\, A(\theta^0)^t \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right)(\hat{P}-Q(\theta^0)) \\ &\qquad + o\left(\|\hat{P}-Q(\theta^0)\|\right).\end{aligned}$$

□

**Theorem 2.** Under the assumptions of Theorem 1 it holds:

a) $\sqrt{n}(\hat{\theta}_\phi - \theta^0) \approx N(0,\Sigma)$, where

$$\Sigma = B(\theta^0)\left(\operatorname{diag}(Q(\theta^0)) - Q(\theta^0)\, Q(\theta^0)^t\right) B(\theta^0)^t$$

and

$$B(\theta) = \left(A(\theta)^t\, A(\theta)\right)^{-1}\, A(\theta)^t \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta))}\right).$$

b) $Q(\hat{\theta}_\phi)$ is a $\sqrt{n}$-consistent estimator of $Q(\theta^0)$.

Proof.

a) Applying the Central Limit Theorem, we get

$$\sqrt{n}(\hat{P} - Q(\theta^0)) \xrightarrow[n\to\infty]{L} N\left(0,\, \Sigma_{Q(\theta^0)}\right),$$

where

$$\Sigma_{Q(\theta^0)} = \operatorname{diag}\left(Q(\theta^0)\right) - Q(\theta^0)\, Q(\theta^0)^t.$$

Consequently,

$$\sqrt{n}\, A(\theta^0)^t \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta))}\right)(\hat{P} - Q(\theta^0)) \xrightarrow[n\to\infty]{L} N(0, \Sigma_1)$$

where

$$\begin{aligned} \Sigma_1 &= A(\theta^0)^t \operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right)\left(\operatorname{diag}(Q(\theta^0)) - Q(\theta^0)\, Q(\theta^0)^t\right) \\ &\qquad \cdot\operatorname{diag}\left(\sqrt{-\phi''(Q(\theta^0))}\right) A(\theta^0). \end{aligned}$$

Hence the result follows from Theorem 1.

b) Consider the Taylor expansion of $q_j(\hat{\theta}_\phi)$ around $\theta^0$

$$q_j(\hat{\theta}_\phi) = q_j(\theta^0) + \sum_{s=1}^{M_0} \frac{\partial q_j(\theta^*)}{\partial \theta_s}(\hat{\theta}_s - \theta_s^0), \quad j = 1, \ldots, M,$$

or, equivalently,

$$Q(\hat{\theta}_\phi) - Q(\theta^0) = \left(\frac{\partial q_j(\theta)}{\partial \theta_s}\right)_{\substack{j=1,\ldots,M \\ s=1,\ldots,M_0}} (\hat{\theta}_\phi - \theta^0).$$

Since

$$\sqrt{n}\left(Q(\hat{\theta}_\phi) - Q(\theta^0)\right) \xrightarrow[n\to\infty]{L} N(0, \Sigma_Q)$$

for

$$\Sigma_Q = J(\theta^0)\, \Sigma\, J(\theta^0)^t$$

it holds

$$\sqrt{n}\left\|Q(\hat{\theta}_\phi) - Q(\theta^0)\right\| \le O_p(1). \qquad \square$$

## 4. MINIMUM $R_\phi$–DIVERGENCE FUNCTIONAL ROBUSTNESS

In this section we consider deviation of the discrete model, $Q(\theta) = (q_1(\theta), \ldots, q_M(\theta))^t$, given by the mixture

$$Q_\varepsilon(\theta) = (1-\varepsilon)\, Q(\theta) + \varepsilon P$$

for $\varepsilon > 0$, $\theta \in \Theta$ and $P \in \Delta_M$.

Let $\theta_\phi^\varepsilon(P)$ be the vector that minimizes the function

$$g_\varepsilon(P, \theta) = \sum_{i=1}^M \phi\left(\frac{p_i + q_i(\theta, \varepsilon)}{2}\right) - \frac{1}{2}\left\{\sum_{i=1}^M \phi(p_i) + \sum_{i=1}^M \phi(q_i(\theta, \varepsilon))\right\}$$

where $Q_\varepsilon(\theta) = (q_1(\theta, \varepsilon), \ldots, q_M(\theta, \varepsilon))$. To guarantee the robustness of $\theta_\phi(P)$, we have to verify that slight deviations of $Q(\theta)$ lead to slight deviations of $\theta_\phi^\varepsilon(P)$ or, analytically, that

$$\lim_{\varepsilon\to\infty} \theta_\phi^\varepsilon(P) = \theta_\phi(P).$$

The following theorem gives conditions that guarantee the functional robustness.

**Theorem 3.** Let the assumption of Theorem 1 be fulfilled. Then

$$\lim_{\varepsilon \to \infty} \theta_\phi^\varepsilon(P) = \theta_\phi(P).$$

Proof. Let $\{\varepsilon_n\}$ be an arbitrary sequence of positive numbers verifying $\varepsilon_n \underset{n\to\infty}{\longrightarrow} 0$. Since is $\phi$ continuous and $q_i(\theta, \varepsilon_n) \underset{\varepsilon_n \to 0}{\longrightarrow} q_i(\theta),\ i = 1, \dots, M$, we get that

$$g_{\varepsilon_n}(P, \theta) \underset{\varepsilon_n \to 0}{\longrightarrow} g(P, \theta) \quad \forall\, \theta \in \Theta.$$

Since $\Theta$ is compact the pointwise convergence implies the uniform convergence. Consequently,

$$\lim_{\varepsilon_n \to 0} \sup_{\theta \in \Theta} |g_{\varepsilon_n}(P, \theta) - g(P, \theta)| = 0$$

which implies that

$$\lim_{\varepsilon_n \to 0} \left| \inf_{\theta \in \Theta} g_{\varepsilon_n}(P, \theta) - \inf_{\theta \in \Theta} g(P, \theta) \right| = 0$$

or, equivalently,

$$\lim_{\varepsilon_n \to 0} \left| g_{\varepsilon_n}\left(P, \theta_\phi^{\varepsilon_n}(P)\right) - g(P, \theta_\phi(P)) \right| = 0.$$

So, we proved that

$$\lim_{\varepsilon_n \to 0} g_{\varepsilon_n}\left(P,\, \theta_\phi^{\varepsilon_n}(P)\right) = g(P,\, \theta_\phi(P)). \tag{1}$$

If $\lim\limits_{\varepsilon_n \to 0} \theta_\phi^{\varepsilon_n} \neq \theta_\phi(P)$ the compactness of $\Theta$ guarantees the existence of a subsequence

$$\left\{\theta_\phi^{\delta_n}(P)\right\} \subset \left\{\theta_\phi^{\varepsilon_n}(P)\right\}$$

such that

$$\lim_{\delta_n \to 0} \theta_\phi^{\delta_n}(P) = \theta_* \neq \theta_\phi(P).$$

But, by (1), $g(P,\, \theta_*) = g(P,\, \theta_\phi(P))$ for $\theta_* \neq \theta_\phi(P)$ which contradicts the assumed uniqueness of $\theta_\phi(P)$. The statement of theorem follows from here since the sequence $\{\varepsilon_n\}$ can be chosen arbitrarily. □

A more general way of studying the robustness is to assume that the true distribution $\pi \in \Delta_M$ satisfied the condition

$$\|\pi - Q(\theta)\| < \varepsilon \quad \text{for some } \theta \in \Theta$$

and to prove that if $\varepsilon$ is small, the value $\theta_\phi(\pi)$ is near to $\theta_\phi(Q(\theta)) = \theta$.

**Theorem 4.** Let the assumptions of Theorem 1 hold and let $\pi \in \Delta_M$. Then

$$\lim_{\|\pi - Q(\theta)\| \to 0} \theta_\phi(\pi) = \theta_\phi(Q(\theta)) = \theta.$$

Proof. The proof follows immediately since $\theta_\phi$ is a continuous function. □

## 5. NUMERICAL RESULTS

An important family of $R_\phi$-divergences introduced by Burbea and Rao [2] can be obtained by considering

$$\phi_\alpha(x) = \frac{1}{1-\alpha}\,(x^\alpha - x), \quad \alpha > 0,\ \alpha \neq 1$$

with $\phi_1(x) = \lim_{\alpha \to 1} \phi_\alpha(x) = -x \ln x$.

It follows from Theorem 1 that the minimum $R_\phi$-divergence estimator, $\hat{\theta}_\phi$, satisfies the asymptotic relation

$$\begin{aligned} \hat{\theta}_\phi &= \theta^0 + \left(A(\theta^0)^t\, A(\theta^0)\right)^{-1} A(\theta^0)^t\, \mathrm{diag}\left((Q(\theta^0))^{\frac{\alpha}{2}-1}\right)(\hat{P} - Q(\theta^0)) \\ &\quad + o\left(\|\hat{P} - Q(\theta^0)\|\right), \end{aligned}$$

where

$$A(\theta) = \mathrm{diag}\left(((Q(\theta^0))^{\frac{\alpha}{2}-1}\right) J(\theta).$$

Note that for $\phi = \phi_1$ we get

$$\begin{aligned} \hat{\theta}_{\phi_1} &= \theta^0 + \left(A(\theta^0)^t\, A(\theta^0)\right)^{-1} A(\theta^0)^t\, \mathrm{diag}\left((Q(\theta^0))^{-\frac{1}{2}}\right)(\hat{P} - Q(\theta^0)) \\ &\quad + o\left(\|\hat{P} - Q(\theta^0)\|\right), \end{aligned}$$

where

$$A(\theta) = \mathrm{diag}\left(((Q(\theta))^{-\frac{1}{2}}\right) J(\theta)$$

and

$$\sqrt{n}\left(\hat{\theta}_{\phi_1} - \theta^0\right)^t \xrightarrow[n\to\infty]{L} N\left(0,\, I(\theta^0)^{-1}\right),$$

where $I(\theta)$ is the Fisher information matrix.

As well known, the asymptotic efficiency of an estimator $\{\hat{\theta}_n\}$ satisfying the condition $\sqrt{n}(\hat{\theta}_n - \theta^0) \xrightarrow[n\to\infty]{L} N(0, \sigma^2_\theta)$ is defined by the ratio $\{I(\theta)\}^{-1}/\sigma^2_\theta$. Furthermore, if $\sigma^2_\theta = \{I(\theta)\}^{-1}$ the corresponding estimator $\hat{\theta}_n$ is asymptotically efficient. Such estimators are usually called BAN (Best Asymptotically Normal). So the estimator $\hat{\theta}_{\phi_1}$ is BAN.

In this section we consider the following two problems:

(1) To calculate the $\alpha$ value, $\alpha_{\min}$, that minimizes the mean quadratic error which is obtained when the two parameters of a Weibull population are estimated using the $R_{\phi_\alpha}$-divergence.

(2) To calculate the minimum $R_{\phi_\alpha}$-divergence estimators of the Weibull parameters for different $\alpha$. Moreover, to compare the obtained results with the maximum likelihood and the minimum Kolmogorov distance estimators.

**Definition 3.** The minimum $D_n$ estimator (minimum Kolmogorov distance estimator) for a distribution family $\{F_\theta(x),\, \theta \in \Theta\}$ is defined as the value $\hat{\theta} \in \Theta$ such that

$$D_n(\hat{\theta}) = \min\{D_n(\theta),\, \theta \in \Theta\}$$

for

$$D_n(\theta) = \sup_{x \in R} \{|F_n^*(x) - F_\theta(x)|\} = \max\left\{D_n^+(\theta),\, D_n^-(\theta)\right\}$$

where $F_n^*(x)$ is the empirical distribution function of the sample $x_1, \ldots, x_n$, so that

$$\begin{aligned} D_n^+(\theta) &= \sup_{x \in R} \{F_n^*(x) - F_\theta(x)\} = \max\left\{0, \max_{i=1,\ldots,n} \left\{\frac{i}{n} - F_\theta(x_{(i)})\right\}\right\} \\ D_n^-(\theta) &= \sup_{x \in R} \{F_\theta(x) - F_n^*(x)\} = \max\left\{0, \max_{i=1,\ldots,n} \left\{F_\theta(x_{(i)}) - \frac{i-1}{n}\right\}\right\} \end{aligned}$$

for the order statistics $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

**Definition 4.** We say that a random variable $X$ has a Weibull distribution $We(b,c)$ with parameters $(b,c)$, $b > 0$, $c > 0$, if the observation distribution function is

$$F_\theta(x) = 1 - \exp\left\{-\left(\frac{x}{b}\right)^c\right\}, \quad x \ge 0,$$

for $\theta = (b,c)$. Here $b$ is a scale parameter and $c$ is a shape parameter.

An algorithmic procedure for finding the optimum $\alpha$ in the sense of (1) is the following:

*Step 1:* We fix

(a) sample size ($n$),

(b) number of classes in the partition ($M$),

(c) number of simulated samples ($N$).
The values $a_0, \ldots, a_M$ are obtained by

$$\int_{a_{i-1}}^{a_i} f_\theta(x)\, \mathrm{d}x = 1/M, \quad i = 1, \ldots, M$$

for a partition $A_i = (a_{i-1}, a_i]$, $i = 1, \ldots, M$ of $\mathcal{X}$.

*Step 2:* The following function is minimized

$$mqe(\alpha) = \frac{\sum_{i=1}^{N}(\hat{\theta}_{1\alpha}^{i} - b)^2}{2N} + \frac{\sum_{i=1}^{N}(\hat{\theta}_{2\alpha}^{i} - c)^2}{2N}$$

where $\hat{\theta}_{1\alpha}^{i}$ is the minimum $R_\phi$-divergence estimator of $b$ and $\hat{\theta}_{2\alpha}^{i}$ of $c$ for the sample $i$. These values are calculated in the Step 3.

*Step 3:* Given $\alpha$ fixed, do for $i = 1$ until $N$

a) Generate a random sample of size $n$.

b) Calculate the classes relative frequency of the previous step.

c) Minimize on $\theta$ the function $R_\phi(\hat{P}, Q(\theta))$.

Go to Step 2.

Table 1 shows the $\alpha$ value, $\alpha_{\min}$, that minimizes the mean quadratic error committed when the parameters of $We(1,1)$ and $We(1,2)$ are estimated by the minimum $R_{\phi_\alpha}$-divergence for several sample sizes.

**Table 1.**

| | *We*(1,1) | *We*(1,2) |
|---|---|---|
| $n = 20$ | 1.76875 | 1.197607 |
| $n = 40$ | 1.1 | 0.613597 |
| $n = 60$ | 1.10158 | 1.205 |

Each $\alpha_{\min}$ has been evaluated from different initial points to check if the results are sensitive to them. The $\alpha$ with the least mean quadratic error has been chosen.

The general scheme for calculating the minimum $R_\phi$-divergence estimator is as follows:

*Step 1:* We fix

(a) sample size ($n$),

(b) number of classes in the partition ($M$),

(c) number of simulated samples ($N$).

The $a_0, \ldots, a_M$ values are obtained by

$$\int_{a_{i-1}}^{a_i} f_\theta(x)\,\mathrm{d}x = 1/M, \quad i = 1, \ldots, M$$

such that $A_i = (a_{i-1},\, a_i],\ i = 1, \ldots, M$, is a partition of $\mathcal{X}$.

*Step 2:* Given $\alpha$ fixed, do for $i = 1$ until $N$

(a) Generate a random sample size $n$.

(b) Calculate the classes relative frequency of the previous step.

(c) Minimize on $\theta$ the function $R_\phi(\hat{P},\, Q(\theta))$.

*Step 3:* Let $\hat{\theta}_\phi$ be the mean of the values obtained minimizing the function $R_\phi$ in step 2 (c) for all the samples and $mqe(\alpha)$ the same mean quadratic error of the estimated parameters as above.

Tables 2 and 3 show the maximum likelihood (MLE), the minimum $D_n$ ($D_nE$) and the minimum $R_\phi$-divergence ($R_\phi E$) estimators for Weibull population with parameters $b = 1,\ c = 1$ and $b = 1,\ c = 2$, respectively. These values have been calculated by computer simulation for 1 000 samples, classes number = 6 and sample sizes $n = 20$, 40 and 60. We move the shape parameter and fix the scale parameter because the estimates of $c$ are worse in general than the estimates of $b$. Therefore, it seems to be more interesting to observe the behaviour of estimates $\hat{c}$. In fact the estimates in Table 3 are worse than those in Table 2. The sums of the mean quadratic errors of the two parameters also appear in both tables.

**Table 2.**

| *We*(1,1) | | $n = 20$ | $n = 40$ | $n = 60$ |
|---|---|---|---|---|
| MLE | $\hat{b}$ | 0.998783 | 0.994317 | 0.994150 |
| | $\hat{c}$ | 1.06396 | 1.029655 | 1.019258 |
| | $mqe$ | 0.055893 | 0.025969 | 0.014047 |
| $D_nE$ | $\hat{b}$ | 0.984651 | 1.006958 | 0.978376 |
| | $\hat{c}$ | 1.565195 | 1.137134 | 1.185521 |
| | $mqe$ | 1.023289 | 1.121828 | 0.108134 |
| $R_{\phi_1}E$ | $\hat{b}$ | 1.008978 | 1.015399 | 0.983971 |
| | $\hat{c}$ | 1.386812 | 1.068477 | 1.117729 |
| | $mqe$ | 0.745879 | 0.112524 | 0.091662 |
| $R_{\phi_2}E$ | $\hat{b}$ | 1.002849 | 1.006966 | 0.979121 |
| | $\hat{c}$ | 1.414069 | 1.077251 | 1.133801 |
| | $mqe$ | 0.742312 | 0.098914 | 0.096844 |
| $R_{\phi_{\min}}E$ | $\hat{b}$ | 1.009632 | 1.006264 | 0.978674 |
| | $\hat{c}$ | 1.396201 | 1.051318 | 1.105390 |
| | $mqe$ | 0.734901 | 0.093519 | 0.083306 |

The mean quadratic error ($mqe$) generated by MLE based on the original Weibull values is smaller than the $R_\phi$-divergences for $n = 40$ and 60 and greater than that for $n = 20$. On the other hand, the $mqe$ $D_n$ is greater than that for the $R_\phi$-divergences in all cases although the minimum $D_n$ estimator is based on the original values and the $R_\phi$-divergence estimators classify the original values into classes.

So we can conclude from the obtained results that when the observations are classified into classes the $R_\phi$-divergences estimators are good.

The programs which calculate the minimum $R_\phi$-divergence and the minimum $D_n$ estimators need an initial estimates. These estimates have been calculated by the Dannenbring [3] method, i. e.:

$$\hat{b} = x_{([0.6321n]+1)}$$

and

$$\hat{c} = \frac{\ln(\log 2)}{\ln(x_M/\hat{b})}$$

where $x_M$ is the sample median.

**Table 3.**

| We(1,2) | | $n = 20$ | $n = 40$ | $n = 60$ |
|---|---|---|---|---|
| MLE | $\hat{b}$ | 0.992503 | 0.993879 | 0.994862 |
| | $\hat{c}$ | 2.127185 | 2.059309 | 2.038516 |
| | *mqe* | 0.093951 | 0.039142 | 0.023382 |
| $D_nE$ | $\hat{b}$ | 0.985581 | 0.997818 | 0.991716 |
| | $\hat{c}$ | 3.030805 | 2.214321 | 2.229546 |
| | *mqe* | 3.926376 | 0.385762 | 0.280197 |
| $R_{\phi_1}E$ | $\hat{b}$ | 0.992771 | 0.999766 | 0.991852 |
| | $\hat{c}$ | 2.699849 | 2.066902 | 2.135167 |
| | *mqe* | 2.675244 | 0.271025 | 0.208801 |
| $R_{\phi_2}E$ | $\hat{b}$ | 0.993954 | 0.997541 | 0.989239 |
| | $\hat{c}$ | 2.783239 | 2.145965 | 2.215762 |
| | *mqe* | 2.700610 | 0.326649 | 0.278683 |
| $R_{\phi_{\min}}E$ | $\hat{b}$ | 0.993631 | 0.997473 | 0.988435 |
| | $\hat{c}$ | 2.651145 | 2.019232 | 2.129087 |
| | *mqe* | 2.560270 | 0.231978 | 0.187459 |

## ACKNOWLEDGEMENT

I should like to thank I. Vajda and K. Sladký for their comments which have improved the paper.

(Received August 26, 1996.)

## REFERENCES

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

[2] J. Burbea and C. R. Rao: On the convexity of some divergence measures based on entropy functions. IEEE Trans. Inform. Theory *28* (1982), 489–495.

[3] D. G. Dannenbring: Procedures for estimating optimal solution values for large combinatorial problems. Management Sci. *23* (1977), 1273–1283.

[4] J. C. Fryer and C. A. Robertson: A comparison of some methods for estimating mixed normal distributions. Biometrika *59* (1972), 639–648.

[5] J. Kiefer and J. Wolfowitz: Consistency of the ML estimator in the presence of infinitely many incidental parameters. Ann. Math. Statist. *27* (1956), 887–906.

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

[7] D. Morales, L. Pardo and I. Vajda: Asymptotic divergence of estimates of discrete distributions. J. Statist. Plann. Inference *48* (1955), 347–369.

[8] M. C. Pardo and I. Vajda: About distances of discrete distributions satisfying the data processing theorem on information theory. IEEE Trans. Inform. Theory *43* (1997), 4, 1288–1293.

*Prof. Dr. María Carmen Pardo, Departamento de Estadística e I. O., Escuela Universitaria de Estadística, Universidad Complutense de Madrid, 28040 Madrid. Spain.*