*The Annals of Statistics*
1997, Vol. 25, No. 2, 553–576

# TRIMMED $k$-MEANS: AN ATTEMPT TO ROBUSTIFY QUANTIZERS[1]

BY J. A. CUESTA-ALBERTOS, A. GORDALIZA AND C. MATRÁN

*Universidad de Cantabria, Universidad de Valladolid and Universidad de Valladolid*

A class of procedures based on "impartial trimming" (self-determined by the data) is introduced with the aim of robustifying $k$-means, hence the associated clustering analysis. We include a detailed study of optimal regions, showing that only nonpathological regions can arise from impartial trimming procedures. The asymptotic results provided in the paper focus on strong consistency of the suggested methods under widely general conditions. A section is devoted to exploring the performance of the procedure to detect anomalous data in simulated data sets.

**1. Introduction.** The development and study of methods to detect clusters is a very important goal in data analysis [see, e.g., Hartigan (1975) and Kaufman and Rousseeuw (1990)]. Closely connected, but from the populational point of view, in statistics or in information theory, the quantization of a random variable is a well-known problem widely studied in the literature [see, e.g., the special issue of IEEE (1982) devoted to this topic]. Particular attention has been paid to $k$-mean clustering procedures [see, e.g., Hartigan (1975, 1978), Pollard (1981, 1982), Sverdrup-Thygeson (1981), Cambanis and Gerr (1983), Cuesta-Albertos and Matrán (1988), Arcones and Giné (1992) and Serinko and Babu (1992)] based on the minimization of the expected value of a "penalty function" $\Phi$ of the distance to $k$-sets (sets of $k$ points), through the problem:

> Given an $\Re^p$-valued random vector $X$, find the $k$-set $M = \{m_1, m_2, \dots, m_k\}$ in $\Re^p$ that minimizes $V_\Phi(M) = \int \Phi(\inf_{i=1,\dots,k}\|X - m_i\|)\, dP$.

Principal points [see, e.g., Tarpey, Li and Flury (1995)] is another recent meaning of this concept in the population framework.

The motivation for this work lies in the fact that, although this formulation is similar to that of obtaining $k$ joint location $M$-estimators, robustness properties behave very differently and quantizers based on typically robust methods can be highly unsatisfactory. For instance, although the median of a random variable may be considered a very robust centralization measure, the

Received January 1994; revised April 1996.
[1]Research partially supported by DGICYT PB91-0306-C02-01 and 02.
*AMS* 1991 *subject classifications*. Primary 62H30, 60F15; secondary 62F35.
*Key words and phrases.* $k$-means, trimmed $k$-means, clustering methods, consistency, robustness.

selection of two "joint" medians through that formulation is very unstable: *the introduction of one, even very improbable, sufficiently remote value implies the selection of such a value as one of the medians*!

This difficulty shows the necessity of designing new clustering procedures with emphasis on robustness properties. Among the available standard techniques in robust estimation, those based on removing part of the data (trimming procedures) present a good performance, often being an obligatory benchmark to compare new estimators. However, the arbitrariness in the selection of zones to remove data is a serious drawback of such procedures.

Gordaliza (1991a) introduced a class of best approximants based on the idea of "impartial trimming." As in the case of the least trimmed squares estimator of Rousseeuw [see, e.g., Rousseeuw and Leroy (1987)], the trimmings depend only on the joint structure of the data and not on arbitrarily selected directions or zones for removing data. Therefore, they are especially suitable in the multivariate case [see also Gordaliza (1991b)].

The main aims of this paper are to suggest a natural extension of Gordaliza's procedure to obtain robustified $k$-means and to provide some mathematical analysis of the method. Consistency properties have a high priority in our study.

The methodology of "impartial trimming," as a way to obtain a trimmed set (at a given level $\alpha$) with the lowest possible variation (penalized by $\Phi$), leads us to formulate the procedures of interest as follows.

Let $\alpha \in (0,1)$, $k$ a natural number and $\Phi$ a penalty function be given. For every set $A$ such that $P(A) \geq 1 - \alpha$ and every $k$-set $M = \{m_1, m_2, \dots, m_k\}$ in $\Re^p$, let us consider the variation about $M$ given $A$:

$$V_\Phi^A(M) := \frac{1}{P(A)} \int_A \Phi\Big( \inf_{i=1,\dots,k} \|X - m_i\|\Big)\, dP.$$

$V_\Phi^A(M)$ measures how well the set $M$ represents the probability mass of $P$ living on $A$ and our job is to choose the best representation to the "more adequate" set containing a given amount of probability mass. This is done by minimizing $V_\Phi^A(M)$ on $A$ and $M$ in the following way:

1. obtain the $k$-variation given $A$, $V_{k,\Phi}^A$, by minimizing in $M$:

$$V_{k,\Phi}^A := \inf_{\substack{M \subset \Re^p \\ \# M = k}} V_\Phi^A(M);$$

2. obtain the trimmed $k$-variation, $V_{k,\Phi,\alpha}$, by minimizing in $A$:

$$V_{k,\Phi,\alpha} := V_{k,\Phi,\alpha}(X) := V_{k,\Phi,\alpha}(P_X) := \inf_{\substack{A \in \beta^p \\ P(A) \geq 1-\alpha}} V_{k,\Phi}^A.$$

We wish to obtain a trimmed set $A_0$, if it exists, and a $k$-set $M_0 = \{m_1^0, m_2^0, \dots, m_k^0\}$, if it exists, through the condition

$$V_\Phi^{A_0}(M_0) = V_{k,\Phi,\alpha}.$$

"Impartially $\alpha$-trimmed $k$-$\Phi$-mean" seems a suitable name for the quantizer $M_0$ just introduced. However, the shorter "trimmed $k$-mean" will be used.

We use the approach in Gordaliza (1991a) and employ "trimming functions." These are a more tractable tool than trimmed sets. The tuning of the technical background necessary for our purposes is made in Section 2.

We prove in Corollary 3.2 that the best trimming function essentially coincides with the indicator function of a nonpathological set: the union of $k$ balls with the same radius. In fact, Section 3 is mainly devoted to analyzing the existence and characterization of the trimmed $k$-means and the associated clusters as well as showing the strong consistency of the method.

An important question remains: what about the applicability of our results in the practical setting? This is analyzed in Section 4 with hopeful results. Our analysis is carried out by applying our methodology to a bivariate data set randomly generated from a mixture of three normal distributions, contaminated both by outlayers and inlayers. In this framework we consider some illustrative situations to discuss the scope of the method.

A main difficulty arises from the nonexistence of a deterministic (nonexhaustive) optimal algorithm to handle the problem. However, a simulated annealing based algorithm suitably performed the procedure in an efficient way, leading to quickly recognized anomalous data and a clusterized data set.

Finally, most of the proofs are given in the Appendix.

**2. Notation and preliminary results.** In this paper $(\Omega, \sigma, P)$ is a probability space and $X$ is an $\Re^p$-valued random vector defined in $(\Omega, \sigma, P)$, with probability law $P_X$ in the $\sigma$-algebra $\mathscr{B}^p$ of all Borel sets in $\Re^p$.

The "penalty function" under consideration, $\Phi: \Re^+ \to \Re^+$, is assumed to be continuous, nondecreasing and such that $\Phi(0) = 0$ and $\Phi(x) < \Phi(\infty)$ for all $x$.

For a set $B \subset \Re^p$, $\overline{B}$ denotes its closure and $B^c$ its complementary set. We denote by $d(x, y)$ the usual distance on $\Re^p$. For $m \in \Re^p$ and $r \geq 0$, $B(m, r)$ denotes the open ball with radius $r$ centered at $m$. Moreover, for $x \in \Re^p$ and $C$, $D \subset \Re^p$, we denote

$$d(x, C) = \inf_{y \in C} d(x, y)$$

and

$$d(C, D) = \sup\left\{ \sup_{x \in C} d(x, D), \sup_{y \in D} d(y, C) \right\}$$

(note that the last expression coincides with the Hausdorff distance between bounded closed sets in $\Re^p$, although in this paper we only use it to obtain distances between sets of $k$ elements).

For $\alpha \in (0, 1)$, $\tau_\alpha[\equiv \tau_\alpha(X)]$ denotes the nonempty set of trimming functions for $X$ of level $\alpha$, that is,

$$\tau_\alpha = \left\{ \tau: \Re^p \to [0, 1], \text{measurable and} \int \tau(X)\, dP = 1 - \alpha \right\}$$

and, $\tau_{\alpha-}$ denotes the set of trimming functions for level $\beta \le \alpha$, that is,

$$\tau_{\alpha-} = \left\{ \tau\colon \Re^p \to [0,1], \text{ measurable and } \int \tau(X)\, dP \ge 1-\alpha \right\} = \bigcup_{\beta \le \alpha} \tau_\beta.$$

Note that the functions in $\tau_\alpha$ (resp. $\tau_{\alpha-}$) are a natural generalization of the indicator functions of sets which have probability $\alpha$ (resp. at least $\alpha$) obtained by introducing the possibility of partial participation of the points in the trimmings.

Now the problem stated in Section 1 can be generalized in a natural way: let $\alpha \in (0,1)$, $k$ a natural number and $\Phi$ a penalty function be given, and, for every $\tau \in \tau_{\alpha-}$ and every $k$-set $M = \{m_1, m_2, \dots, m_k\}$ in $\Re^p$, let us consider the variation about $M$ given $\tau$:

$$V_\Phi^\tau(M) := \frac{1}{\int \tau(X)\, dP} \int \tau(X)\Phi(d(X,M))\, dP.$$

Then:

1. obtain the $k$-variation given $\tau$, $B_{k,\Phi}^\tau$, by minimizing in $M$:

$$V_{k,\Phi}^\tau := \inf_{\substack{M \subset \Re^p\\ \#M = k}} V_\Phi^\tau(M);$$

2. obtain the trimmed $k$-variation, $V_{k,\Phi,\alpha}$, by minimizing in $\tau \in \tau_{\alpha-}$:

$$V_{k,\Phi,\alpha} := V_{k,\Phi,\alpha}(X) := V_{k,\Phi,\alpha}(P_X) := \inf_{\tau \in \tau_{\alpha-}} V_{k,\Phi}^\tau.$$

We wish to obtain a trimming function $\tau_0$, if it exists, and a $k$-set $M_0 = \{m_1^0, m_2^0, \dots, m_k^0\}$, if it exists, through the condition

$$V_\Phi^{\tau_0}(M_0) = V_{k,\Phi,\alpha}. \tag{1}$$

Obviously, $I_B \in \tau_{\alpha-}$ for every set $B \in \mathscr{B}^p$ with $P_X(B) \ge 1-\alpha$. Therefore, the approximation obtained through trimming functions is better than the one obtained through trimmed sets.

Also note that $V_{k,\Phi,\alpha}(P_X) < \infty$ for every $k$, $\Phi$, $X$ and $\alpha > 0$; in fact, by taking a ball $B = B(0,r)$ such that $P_X(B) \ge 1-\alpha$, we have

$$\begin{aligned} V_{k,\Phi,\alpha}(X) &\le \frac{1}{P_X(B)} \int I_B(X)\Phi(d(X,0))\, dP \\ &\le \Phi(r) < \infty. \end{aligned} \tag{2}$$

The following simple results provide the bases for our subsequent work. Their proofs are related to those given in Gordaliza (1991a) and will be omitted.

LEMMA 2.1. *Let $M = \{m_1, \dots, m_k\} \subset \Re^p$ a $k$-set and $\beta \in (0,1)$. Let us denote the (generalized) ball centered at $M$ by*

$$B(M,r) = \bigcup_{i=1}^k B(m_i, r) \quad \textit{for all } r \ge 0,$$

*and let*

$$r_\beta(M) = \inf\left\{r \geq 0: P_X(B(M,r)) \leq 1-\beta \leq P_X\left(\overline{B}(M,r)\right)\right\}$$

*and*

$$\tau_{M,\beta} = \left\{\tau \in \tau_\beta: I_{B(M,r_\beta(M))} \leq \tau \leq I_{\overline{B}(M,r_\beta(M))}, a.e.\ P_X\right\},$$

*then, for all* $\tau \in \tau_{M,\beta}$ *we have*:

(a) $\int \tau(X)\Phi(d(X,M))\, dP \leq \int \tau'(X)\Phi(d(X,M))\, dP$ *for all* $\tau' \in \tau_\beta$;

(b) *If* $\Phi$ *is strictly increasing, then the inequality in* (a) *is strict if and only if* $\tau' \in \tau_\beta - \tau_{M,\beta}$.

From Lemma 2.1, the $\beta$-trimmed variation about $M$:

$$V_{\Phi,\beta}(M) := \frac{1}{1-\beta}\int \tau(X)\Phi(d(X,M))\, dP$$

is the same for every function $\tau$ in $\tau_{M,\beta}$. Therefore, unless necessary, no explicit reference to any particular choice in $\tau_{M,\beta}$ will be made and the same notation $\tau_{M,\beta}$ will be used for any function in $\tau_{M,\beta}$.

LEMMA 2.2. *With the same notation as in Lemma* 2.1, *if* $\beta \leq \alpha$, *then*:

(a) $V_{\Phi,\alpha}(M) \leq V_{\Phi,\beta}(M)$;

(b) *if* $\Phi$ *is strictly increasing, then the equality holds in* (a) *if and only if* $r_\alpha(M) = r_\beta(M)$ *and* $P_X[B(M, r_\alpha(M))] = 0$.

PROPOSITION 2.3. *With the same notation as in Lemmas* 2.1 *and* 2.2,

$$V_{k,\Phi,\alpha} = \inf_{\substack{M \subset \Re^p \\ \#M = k}} V_{\Phi,\alpha}(M).$$

The notation introduced in the previous lemmas will be maintained throughout the paper.

REMARK 2.1. After Lemma 2.1 we know that the $\beta$-trimmed variation about $M$ is minimized by taking any trimming function in $\tau_{M,\beta}$, that is, essentially an indicator function of a ball centered at $M$.

REMARK 2.2. After Lemma 2.2 we know that in order to minimize the $\alpha$-trimmed variation about $M$, it is strictly better to trim the exact quantity $\alpha$, except in the case where all the probability mass of $\overline{B}(M, r_\alpha(M))$ is concentrated on the boundary.

REMARK 2.3. After Proposition 2.3 the problem stated in (1) can be restated as follows: select a $k$-set $M_0 = \{m_1^0, \dots, m_k^0\} \subset \Re^p$ such that

$$V_{\Phi,\alpha}(M_0) = V_{k,\Phi,\alpha}. \tag{3}$$

**3. Existence and consistency of trimmed *k*-means.** The existence of $k$-means is shown in the Appendix. There we prove the following theorem.

THEOREM 3.1 (Existence of trimmed $k$-means). *Let $X$ be an $\Re^p$-valued random vector. Let $\alpha \in (0,1)$, $k \in \mathcal{N}$ and let $\Phi\colon \Re^+ \to \Re^+$ be a continuous, nondecreasing function such that $\Phi(0) = 0$ and $\Phi(x) < \Phi(\infty)$ for all $x$. Then there exists a trimmed $k$-mean of $X$.*

Once the existence of $k$-means is established, Lemma 2.1 provides an important relationship between trimmed $k$-means and the best trimming functions, which we state next.

COROLLARY 3.2. *Under the hypotheses of Theorem 3.1, if $\Phi$ is strictly increasing and $\tau_0$ and $M_0$ are a solution of (1), then*

$$I_{B(M_0, r_\alpha(M_0))} \le \tau_0 \le I_{\overline{B}(M_0, r_\alpha(M_0))}, \quad P_X\text{-}a.e.$$

*Moreover, if $P_X$ is absolutely continuous with respect to the Lebesgue measure on $\Re^p$, then*

$$I_{B(M_0, r_\alpha(M_0))} = \tau_0, \quad P_X\text{-}a.e.$$

REMARK 3.1. Consider a trimmed $k$-mean of $X$, $M_0 = \{m_1^0, \dots, m_k^0\}$, with associated optimal trimming function $\tau_0$ and optimal radius $r_0$, that is,

$$I_{B(M_0, r_0)} \le \tau_0 \le I_{\overline{B}(M_0, r_0)},$$

where $\overline{B}(M_0, r_0)$ is the optimal set except at most by part of the boundary. Note that every trimmed $k$-mean, $M_0$, induces a partition of $\overline{B}(M_0, r_0)$ into $k$ clusters in the following way: the cluster $A_i$ consists of all points $x \in \Re^p$ which are closer to $m_i^0$ than to the remaining $k-1$ points in $M_0$. The points in the boundary between the clusters could be assigned in any way because, obviously, the trimmed $k$-variation remains unchanged.

The set $M_0$ also induces a partition of the trimmed $k$-variation of $X$ into the variations corresponding to each cluster:

$$\begin{aligned} V_{k,\Phi,\alpha}(X) &= \frac{1}{1-\alpha} \int \tau_0(X) \Phi\big(d(X, M_0)\big)\, dP \\ &= \frac{1}{1-\alpha} \sum_{i=1}^{k} \int_{A_i} \tau_0(X) \Phi\big(d\big(X, m_i^0\big)\big)\, dP. \end{aligned}$$

Moreover, for every $i = 1, \dots, k$, $m_i^0$ has to be a $\Phi$-mean of the corresponding cluster $A_i$, or, more precisely, a $\Phi$-mean of $X$ given $A_i$; that is, $m_i^0$ is a solution of

$$\inf_{m \in \Re^p} \int_{A_i} \tau_0(X) \Phi(d(X, m))\, dP.$$

On the contrary, we could diminish the variation in some clusters by replacing $m_i^0$, $i = 1, \dots, k$, by $\Phi$-means of the corresponding clusters, and then $M_0$ would not be a trimmed $k$-mean of $X$. Thus we have proved not only that the trimmed $k$-mean induces a partition of $\overline{B}(M_0, r_0)$ into $k$ clusters but also that the partition determines the trimmed $k$-mean. We summarize this result in the following proposition, which relates the trimmed $k$-mean to a joint set of $\Phi$-means.

PROPOSITION 3.3. *With the same notation as above, $m_i^0$ is a $\Phi$-mean of $X$ given the cluster $A_i$, $i = 1, \dots, k$.*

As a consequence, uniqueness of the trimmed $k$-mean depends not only on the uniqueness of the optimal trimming set, but also on the uniqueness of the $\Phi$-mean given each cluster (consider, e.g., the median as the particular case where $\Phi$ is the identity). This kind of difficulty can be avoided by imposing restrictions on the penalty functions. For instance, we have proved in Cuesta-Albertos, Gordaliza and Matrán (1995) that if $\Phi$ is a continuously differentiable, strictly convex function, then there is no probability mass at the boundary between the clusters.

We have even proved in that paper that, under the same hypotheses, the mass on the external boundary of the clusters cannot be placed in an arbitrary way, because the optimal $B(M_0, r_0)$ necessarily satisfies one of the following:

1. The boundary does not lie at all in the optimal trimming set, that is, $P_X[B(M_0, r_0)] = 1 - \alpha$.
2. All the boundary lies in the optimal trimming set, that is, $P_X[\overline{B}(M_0, r_0)] = 1 - \alpha$.
3. There exists $x_0 \in Bd(B(M_0, r_0))$ such that all the probability mass of the boundary is concentrated at $x_0$, that is, $P_X[Bd(b(M_0, r_0))] = P_X[\{x_0\}]$.

In Cuesta-Albertos, Gordaliza and Matrán (1995), we also provide examples of the necessity of some kind of condition on $\Phi$ to get such conclusions.

The main result related to the consistency of the trimmed $k$-means is based on a previous, more general result of continuity of trimmed $k$-means and trimmed $k$-variations as well as on the Skorohod representation theorem. The latter allows us to represent the convergence of the empirical measures in terms of an almost sure convergent sequence and then to apply the continuity result. This scheme is similar to that used in Cuesta-Albertos and Matrán (1988) to establish the SLLN for $k$-means. However, some difficulties arise from the presence of trimmings, because the trimming functions are discontinuous on the boundary of the corresponding balls so that some care is needed with the convergences. The continuity of the probability distribution of the limit random vector will be imposed in order to guarantee the results.

In what follows, $\{X_n\}_n$ is a sequence of $\Re^p$-valued random vectors defined on $(\Omega, \sigma, P)$ and $M_n = \{m_1^n, \dots, m_k^n\}$, $n = 0, 1, 2, \dots,$ is a trimmed $k$-mean of $X_n$ with associated optimal trimming function $\tau_n$ and optimal radius $r_n$. Moreover, $V_n[= V_{k,\Phi,\alpha}(X_n)]$, $n = 0, 1, 2, \dots,$ denotes the trimmed $k$-variation of $X_n$.

THEOREM 3.4. *With the same notation as above, assume that*:

(a) $X_n \to X_0$, *P-a.e.*;
(b) $P_{X_0}$ *is continuous*;
(c) $M_0 = \{m_1^0, \dots, m_k^0\}$ *is the unique trimmed k-mean of* $X_0$.

*Then*

$$M_n \to M_0 \text{ (in the Hausdorff distance)} \quad \text{as } n \to \infty$$

*and*

$$V_n \to V_0 \quad \text{as } n \to \infty.$$

COROLLARY 3.5. *If we assume that every hypothesis in Theorem* 3.4 *is satisfied and* (a) *is replaced by*:

(a*) $X_n \to X_0$ *in distribution.*

*Then*

$$M_n \to M_0 \text{ (in the Hausdorff distance)} \quad \text{as } n \to \infty$$

*and*

$$V_n \to V_0 \quad \text{as } n \to \infty.$$

PROOF. By applying the a.s. Skorohod representation theorem, there exists a sequence $\{Y_n\}_n$ of $\Re^p$-valued random vectors such that $P_{Y_0} = P_X$, $P_{Y_n} = P_{X_n}$ and $Y_n \to Y_0$ a.s. Hence, the result follows by applying Theorem 3.4 to the sequence $\{Y_n\}_n$. □

Now we obtain the consistency of trimmed $k$-means as a simple consequence of Corollary 3.5

THEOREM 3.6 (Consistency of trimmed $k$-means). *Let* $\{X_n\}_n$ *be a sequence of independent, identically distributed random vectors with distribution* $P_X$ *and let* $\{P_n^\omega\}$ *be the sequence of empirical probability measures* (*i.e.*, $P_n^\omega(A) = n^{-1}\sum_{1 \le i \le n} I_A[X_i(\omega)]$). *Let us assume that* $P_X$ *is continuous and that there exists a unique trimmed k-mean for* $P_X$, $M_0$. *if* $\{M_n^\omega\}_n$ *is a sequence of empirical trimmed k-means, then*:

(a) $d(M_n^\omega, M_0) \to 0$, *P-a.s.*;
(b) $V_{k,\Phi,\alpha}(P_n^\omega) \to V_{k,\Phi,\alpha}(P_X)$, *P-a.s.*

PROOF. Let $A := \{\omega \in \Omega$ such that $P_n^\omega \mapsto_d P_X\}$. It is well known that $P(A) = 1$, so the result follows from Corollary 3.5. □

**4. Application.** The objective of this section is to show the ability of the procedure, on the one hand, to detect anomalous data and rightly assign data to clusters and, on the other hand, to estimate the mean of clusters in the presence of anomalous observations. For simplicity, we consider the quadratic loss, and we will be concerned with the $\alpha$-trimmed $k$-mean (in fact, we always consider $k = 3$) for different sizes of $\alpha$. The general scheme will be the following.

First, we will randomly generate a set of points which are divided into three clusters and we will add a proportion $\beta$ of anomalous points. This set will be denoted by $A$. Then we will choose $\alpha \in (0,1)$. According to our procedure, we will delete a proportion $\alpha$ of points in $A$ and we will divide the remaining points into three groups in order to minimize the within-group variance.

As stated, our analysis of the method is focused in two directions. First, in the spirit of cluster analysis, we make a sensitivity study by exploring different departures of what could be called the ideal model. Here our interest relies on the capacity of the method to detect the anomalous data and to divide the remaining ones in the original clusters.

Note that, from Corollary 3.2, it follows that relatively nearby clusters could be badly detected if they are nonspherical or even if their shapes are too different because, according to that corollary, every cluster obtained with our method is spherical and all of them have the same radius. Therefore, the ideal model would be one consisting of spherical, not too close, clusters. Moreover, in that ideal model, the anomalous observations should appear clearly separated from the nonanomalous ones.

Here we have chosen a sample of situations in which each requirement to have an ideal model is violated. Moreover, all of them have in common that many of the anomalous data are not so anomalous because the only restriction we have imposed on them is that they are not allowed to be in the 75% level confidence ellipsoids of the distributions generating the points in the clusters.

More precisely, in every situation we have simulated three bivariate normal distributions, $N_1$, $N_2$ and $N_3$, with means at $(0,0)$, $(0,10)$ and $(6,0)$. These means were chosen to avoid harmonizing effects which could appear if we place the means on the vertices of an equilateral triangle. The anomalous data were randomly generated from $N_4$, a bivariate normal centered at $(2, 10/3)$ (the mean of the means above) with a dispersion large enough to produce both inner and outer contaminations. The points from $N_4$ lying in the 75% level confidence ellipsoids of $N_1$, $N_2$, or $N_3$ were replaced by other ones not belonging to that area. With this selection of contaminating data, we wanted to produce an inner additional (small) cluster, zones of uncertainty and masking to render difficult a right classification, and even a clear bias due to the greater proportion of anomalous data in the middle area. We fixed the size of every cluster and the number of anomalous observations separately.

Therefore, in every situation, the model is specified by $(n_1, n_2, n_3, n_4, \Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4)$, where $n_i$ is the sample size from the distribution $N_i$, and $\Sigma_i$ is the covariance matrix of the distribution $N_i$, $I = 1,2,3,4$. In order to improve the final display, we have chosen moderate sample sizes. However, every time we have chosen $n_4 = 40$ and $\Sigma_4 = 20\text{Id}$, where Id denotes the identify matrix. Thus, only the value of $n_1$, $n_2$, $n_3$, $\Sigma_1$, $\Sigma_2$ and $\Sigma_3$ need to be specified.

The capability of the method, under reasonable deviations from homogeneity and sphericity of the right clusters, is shown in Figures 1–4. The figures

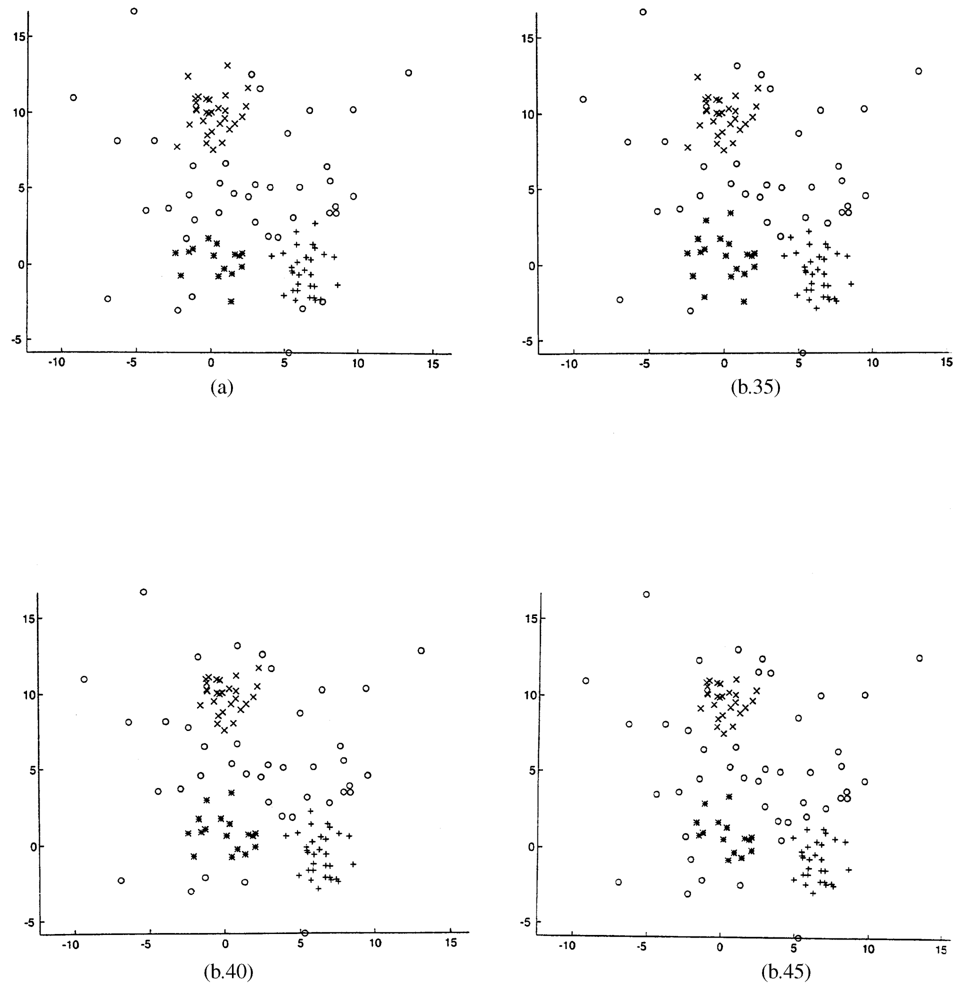
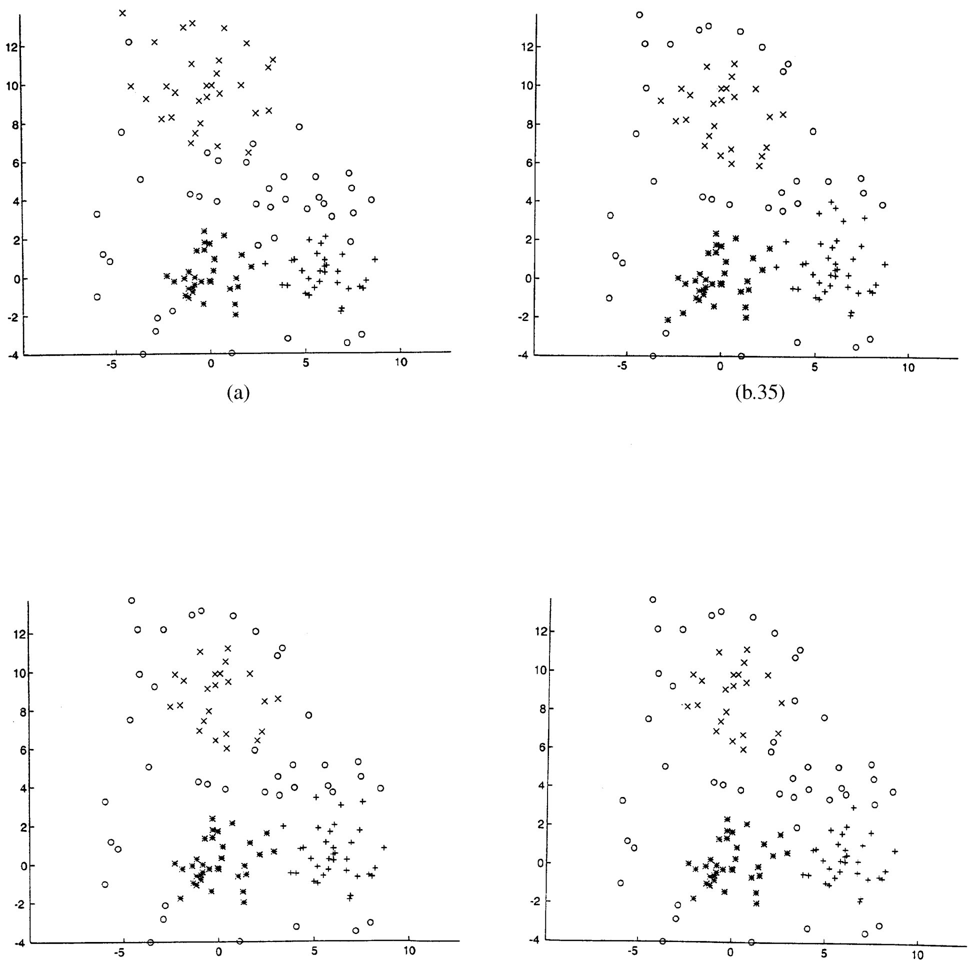



FIG. 1.

labeled (a) correspond to the initial set of points, while the figures labeled (b.35), (b.40) and (b.45) show the results of our method for the different trimming sizes which were chosen around the number of anomalous observations as 35, 40 and 45. In this figures the symbol ○ denotes an anomalous observation in the figures labeled (a) or a trimmed observation in the figures labeled (b). The symbols $+$, $\times$ and $*$ denote the initial clusters in the figures labeled (a) or the clusters suggested by our method in the figures labeled (b).

The closest situation to the ideal model is shown in Figure 1. Here the model is given by

$$n_1 = 15, \qquad n_2 = n_3 = 30, \qquad \Sigma_1 = \Sigma_2 = \Sigma_3 = 1.5\mathrm{Id}.$$



(a) (b.35) (b.40) (b.45)

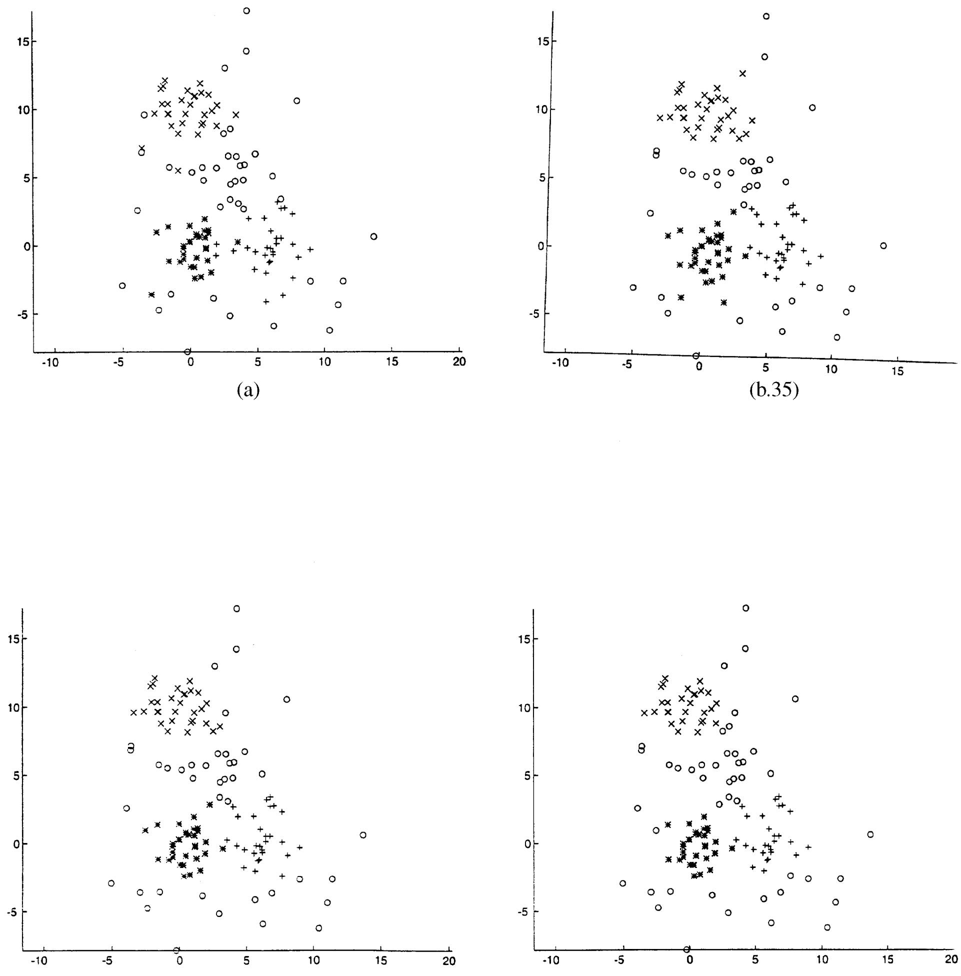
FIG. 2.

Therefore, every cluster is spheric and some separation appears between clusters, but the situation is also conflicting because the size of one of the clusters is a half of those of the other ones and it amounts to less than a half of the contamination.

In the remaining cases we have fixed the values for $n_1 = n_2 = n_3 = 30$ and we have varied the covariance matrices. So in Figure 2 we have increased the dispersion of $N_2$ by taking $\Sigma_2 = 4\text{Id}$ while $\Sigma_1 = \Sigma_3 = 1.5\text{Id}$.

In the data in Figure 3 we have increased the dispersion of the distributions $N_1$ and $N_3$ to get the associated clusters in touch. We have also increased the dispersion of $N_2$ (with respect to the values in Figure 1). In this case we have chosen

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = 2\text{Id}.$$



(a) (b.35) (b.40) (b.45)

FIG. 3.

Perhaps the most difficult situation considered is that shown in Figure 4. Here we have introduced a nonspheric cluster by taking $\Sigma_2 = \Sigma_3 = 2\mathrm{Id}$ and

$$\Sigma_1 = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

The results are summarized in Table 1, where we show, for every case and every trimming size, the number of rightly trimmed data and the number of mistakes (where we include the incorrectly trimmed data, the data which were incorrectly assigned to a cluster and those anomalous observations which were not trimmed). Note that the number of incorrectly trimmed data is necessarily grater than or equal to 5 when trimming 45 points because

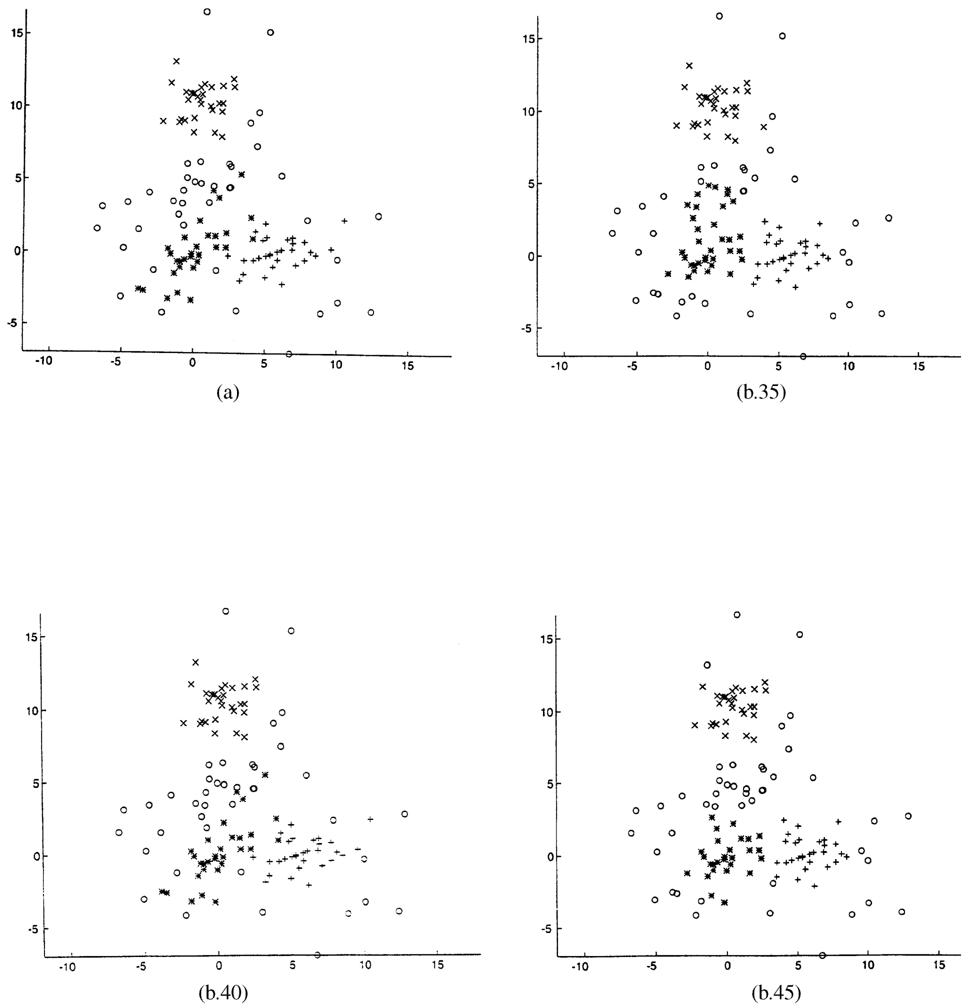



FIG. 4.

TABLE 1

| Cases | Trimming size = 35 | | Trimming size = 40 | | Trimming size = 45 | |
|---|---|---|---|---|---|---|
| | Rightly trimmed | Mistakes | Rightly trimmed | Mistakes | Rightly trimmed | Mistakes |
| Case 1 | 33 | 9 | 35 | 10 | 35 | 15 |
| Case 2 | 30 | 19 | 34 | 16 | 37 | 15 |
| Case 3 | 26 | 23 | 30 | 21 | 33 | 20 |
| Case 4 | 27 | 24 | 32 | 19 | 35 | 18 |

there are only 40 anomalous observations. Analogously, the number of nontrimmed anomalous observations is at least 5 when trimming 35 points.

We want to remark that the procedure was quite successful in spite of the fact that, in every case, we have a really high proportion of contamination, which introduces enough noise as to make the original clusters badly identifiable just by eye.

In a different way we have also analyzed the behavior of the method for the estimation of the means of the distributions $N_i$, $i = 1, 2, 3$.

For this task we have generated 25 data sets obtained from the same model as above, but taking $n_i = 50$, $i = 1, 2, 3, 4$, $\Sigma_1 = \Sigma_2 = \Sigma_3 = 1.5\mathrm{Id}$ and $\Sigma_4 = 20\mathrm{Id}$. However, for obvious reasons, now only those points from $N_4$ not included on the 90% level confidence ellipsoids of $N_i$, $i = 1, 2, 3$, were considered as anomalous and included in the whole sample. We successively obtained, for each set, the sample impartial trimmed 3-mean by using trimming sizes in the range 40 to 100 points.

We have also computed, to be used as elements of comparison, the estimates consisting of the 3-mean when computed, respectively, from a set without anomalous observations and with 10 and 50 anomalous observations. That is, here no trimming is allowed, so that we have just divided the data, in each case, into three groups, by minimizing the within-group variance and then we have estimated the mean of every cluster as the sample mean of the points included in that cluster. This job has been carried out for the same 25 data sets which we have employed with our method.

This general comparison process can be summarized as follows:

1. We have randomly generated a set of 150 points by taking $n_1 = n_2 = n_3 = 50$ and $n_4 = 0$ and we have computed its 3-mean without trimming.
2. We have added $n_4 = 10$ anomalous points from $N_4$ and we have computed the 3-mean without trimming of this data set.
3. We have included 40 additional data points from $N_4$ (thus $n_i = 50$, $i = 1, 2, 3, 4$) and we have computed the 3-mean of those points in the following cases:
   a. without trimming;
   b. with trimming sizes equal to 40, 50, 60, 70, 80, 90 and 100.
4. We have repeated the previous steps for the 25 data sets.

The results are summarized in Table 2, in which we show as "mean vector" the mean of the values that we have obtained for each data set in previous steps. "Distance" is the Euclidean distance between the six-dimensional mean vector and the theoretical mean vector $((0, 0), (0, 10), (6, 0))$. We have finally sorted the different estimates according to the values of those distances.

These data show some bias in the estimator, which is more apparent for low levels of trimming. To explain the nature of this bias, let us suppose that there exists a high density of probability mass "on a side" of trimmed cluster (at a given level), given by the ball $B(x_0, r)$. To fix the ideas, let us assume that $x_0 = (0, 0)$, that the zone of high density is approximately placed on the

TABLE 2

| Case | Vector of $k$-means: Cluster 1 | Cluster 2 | Cluster 3 | Distance | Order |
|---|---|---|---|---|---|
| No anomalous | $(-0.003, 0.051)$ | $(0.011, 9.974)$ | $(5.965, -0.030)$ | 0.074 | 1 |
| 10 anomalous | $(-0.03, 0.11)$ | $(0.05, 9.82)$ | $(6.03, 0.14)$ | 0.262 | 8 |
| 50 anomalous | $(-0.22, 0.10)$ | $(0.37, 9.44)$ | $(6.24, 0.64)$ | 0.988 | 10 |
| Trimming = 40 | $(0.07, 0.16)$ | $(0.06, 9.83)$ | $(5.88, 0.11)$ | 0.299 | 9 |
| Trimming = 50 | $(0.08, 0.11)$ | $(0.02, 9.91)$ | $(5.86, 0.01)$ | 0.216 | 7 |
| Trimming = 60 | $(0.04, 0.08)$ | $(-0.02, 9.94)$ | $(5.89, 0.01)$ | 0.156 | 5 |
| Trimming = 70 | $(-0.010, 0.069)$ | $(-0.044, 9.950)$ | $(5.898, 0.009)$ | 0.141 | 2 |
| Trimming = 80 | $(-0.024, 0.067)$ | $(-0.005, 9.952)$ | $(5.873, 0.005)$ | 0.154 | 4 |
| Trimming = 90 | $(-0.027, 0.087)$ | $(-0.038, 9.933)$ | $(5.868, 0.004)$ | 0.178 | 6 |
| Trimming = 100 | $(-0.014, 0.080)$ | $(-0.047, 9.947)$ | $(5.902, 0.012)$ | 0.146 | 3 |

point $(r - \varepsilon, 0)$ and that a greater trimming level is required. Then the procedure tries to maintain the zone of high density of probability also in the new trimmed cluster and searches for a less "inhabited" zone for trimming. The center of the new ball corresponding to this trimmed cluster will be moved from the old $(0, 0)$, producing the bias.

In the examples of our simulations this notably happens, due to the kind of contamination under consideration and to the relative proximity between two clusters, when the trimming size does not suffice for trimming to a greater extent than that corresponding to the 90% level confidence spheres. This happens, in mean, around the trimming corresponding to 65 points. The bias is less dramatic as the trimming level increases.

The main difficulty in accomplishing our goal was the nonexistence of a deterministic optimal (nonexhaustive) algorithm to choose the optimal trimming set. Moreover, as is widely recognized, optimal algorithms do not exist for the $k$-means problem even without trimming. However, the employment of a random algorithm along the lines of the so-called "simulated annealing" procedures, in the Matlab setting, has shown a quick and suitable behavior with different data sets to handle both problems.

**5. Conclusions.** From a general point of view, the behavior of the procedure seems hopeful because the objectives were successfully reached. The procedure is orthogonally equivariant and, as shown in the simulations in Section 4, its robustness against contamination seems to be high when the probability is supported by a set of relatively well-shaped spherical clusters.

However, let us emphasize that there we used the right number of clusters in the analyses, but let us consider the following example. Let us assume that $\Phi(t) = t^2$ and that we try to compute the 1/3-trimmed 2-mean of the set in $\Re$:

$$A = \{-3, -2, -1, 1, 2, 3, 20, 23, 26\}.$$

That is, we are allowed to delete up to three points in $A$, and then to split the remaining points into two groups and to compute the within-group variance. The goal is to minimize this quantity.

It is obvious that the optimal points to trim are 20, 23 and 26 and that the associated 2-mean is $\{-2, 2\}$. Now let us contaminate $A$ by replacing the point $-3$ by $-100$. Then it happens that the points to trim are again 20, 23 and 26, but now the associated 2-mean is $\{-100, 0.6\}$. Thus the trimming procedure, when applied to $A$, has a breakdown point which is less than or equal to $1/9$. Moreover, it is clear that, by modifying the set $A$, we would have that for every $\alpha$, $k$ and $\Phi$ there exists a probability $P$ such that the breakdown point of the $\alpha$-trimmed $k$-mean of $P$ is as close to 0 as desired.

A careful look at this example shows that the cause of this behavior of the 2-mean procedure when applied to the uniform probability on $A$ is that, if we are looking for just two clusters, then the set $A$ already contains $1/3$ of anomalous points (independently of when those points constitute or do not constitute a new cluster). Therefore, it seems that for probabilities $Q$, supported by a set which is divided into exactly $k$ clusters with respective probabilities $q_1, q_2, \dots, q_k$, the breakdown point of the $\alpha$-trimmed $k$-mean of $Q$ is, at most, $\inf\{\alpha, q_1, \dots, q_k\}$ independently of $\Phi$. Of course, this fact is more apparent when we use other methods like the "two joint medians" example in the Introduction, which is clearly unstable in every circumstance.

In other words, the breakdown point generally depends, of course, on the procedure, but it also depends heavily on the data, in the sense that the same procedure can be highly stable with reasonable clusterized data when considering the right number of clusters, but it can also be very unstable in other cases. This is in some way natural and clearly related to the traditional key problem in cluster analysis: how to choose the number of clusters to look for?

As an added conclusion, from our point of view, the analysis of robustness of the cluster analysis procedures needs some fit of the available theory to analyze problems like the previous one.

An extreme case which naturally arises from our study is that of the trimmed $k$-means associated with the $L_\infty$-criterion, the so-called *trimmed $k$-nets*. This case is quite different from the one treated here and we have studied it in a separate paper [Cuesta-Albertos, Gordaliza and Matrán (1996)].

To give a hint to the difference between $k$-nets and $k$-means may be enough to say that the strong consistency of trimmed $k$-nets does not generally hold if the level of trimming in the sampling remains constant. In fact, in order to get consistency, we need suitable sizes of trimming that vary with the size of the sample.

## APPENDIX

We begin with two results of a different scope. The first one shows the continuity of the trimmed variation $V_{\Phi,\alpha}(M)$ with respect to $M$, and the second one shows the natural fact that the trimmed variation is strictly improved by increasing the number of clusters. Both results are needed in the proof of the existence of trimmed $k$-means.

PROPOSITION A.1. *Let* $M_n = \{m_1^n, \dots, m_k^n\}$, $n = 0, 1, 2, \dots,$ *be a sequence of* $k$*-sets in* $\Re^p$ *satisfying*

$$M_n \to M_0 \text{ in the Hausdorff distance} \quad \text{as } n \to \infty.$$

*Then we have*

$$V_{\Phi,\alpha}(M_n) \to V_{\Phi,\alpha}(M_0) \quad \text{as } n \to \infty.$$

PROOF. Let us set $r_n = r_\alpha(M_n)$ and $\tau_n = \tau_{M_n,\alpha}$, $n = 0, 1, \dots$ . It is easy to see that $\lim_{n\to\infty} r_n = r_0$. Let us denote $D_n = |V_{\Phi,\alpha}(M_n) - V_{\Phi,\alpha}(M_0)|$. Then

$$\begin{aligned}
(1-\alpha)D_n &= \left| \int \tau_n(X)\Phi(d(X, M_n))\, dP - \int \tau_0(X)\Phi(d(X, M_0))\, dP \right| \\
&\le \left| \int \tau_n(X)(\Phi(d(X, M_n)) - \Phi(d(X, M_0)))\, dP \right| \\
&\quad + \left| \int (\tau_n(X) - \tau_0(X))\Phi(d(X, M_0))\, dP \right| \\
&=: D_n^{(1)} + D_n^{(2)}.
\end{aligned}$$

Notice now that $d(X, M_n) - d(X, M_0) \to 0$ as $n \to \infty$ and that $\Phi$ is uniformly continuous on every compact set, so that we have

$$\begin{aligned}
D_n^{(1)} &\le \int \tau_n(X)|\Phi(d(X, M_n)) - \Phi(d(X, M_0))|\, dP \\
&\le (1-\alpha)\left( \sup_{x\in \overline{B}(M_n, r_n)} |\Phi(d(x, M_n)) - \Phi(d(x, M_0))| \right) \\
&\to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

In order to prove that also $D_n^{(2)} \to 0$ as $n \to \infty$, let us denote

$$E_n := \{x \in \Re^p : \tau_n(x) > \tau_0(x)\}$$

and

$$F_n := \{x \in \Re^p : \tau_n(x) < \tau_0(x)\}.$$

We have

$$\begin{aligned}
0 &= \int (\tau_n(x) - \tau_0(x))\, dP_X \\
&= \int_{E_n} (\tau_n(x) - \tau_0(x))\, dP_X + \int_{F_n} (\tau_n(x) - \tau_0(x))\, dP_X
\end{aligned} \tag{4}$$

and therefore

$$\int_{E_n} (\tau_n(x) - \tau_0(x))\, dP_x = \int_{F_n} (\tau_0(x) - \tau_n(x))\, dP_x.$$

Moreover, for every $x \in E_n$,

$$\begin{aligned}
\Phi(d(x, M_0)) &\le \Phi(d(x, M_n) + d(M_n, M_0)) \\
&\le \Phi(r_n + d(M_n, M_0))
\end{aligned} \tag{5}$$

and, for every $x \in F_n$,

$$\Phi(d(x, M_0)) \geq \Phi(d(x, M_n) - d(M_n, M_0)) \geq \Phi(r_n - d(M_n, M_0)), \tag{6}$$

because $E_n \subset B^c(M_0, r_0) \cap \overline{B}(M_n, r_n)$ and $F_n \subset \overline{B}(M_0, r_0) \cap B^c(M_n, r_n)$. On the other hand, by the definition of $\tau_0$,

$$\int (\tau_n(X) - \tau(X))\Phi(d(X, M_0))\, dP \geq 0,$$

so that we have

$$\begin{aligned}
D_n^{(2)} &= \int (\tau_n(X) - \tau_0(X))\Phi(d(X, M_0))\, dP \\
&= \int_{E_n} (\tau_n(X) - \tau_0(X))\Phi(d(X, M_0))\, dP \\
&\quad - \int_{F_n} (\tau_0(X) - \tau_n(X))\Phi(d(X, M_0))\, dP \\
&\leq \Phi(r_n + d(M_n, M_0)) \int_{E_n} (\tau_n(X) - \tau_0(X))\, dP \\
&\quad - \Phi(r_n - d(M_n, M_0)) \int_{F_n} (\tau_0(X) - \tau_n(X))\, dP \\
&\leq \Phi(r_n + d(M_n, M_0)) - \Phi(r_n - d(M_n, M_0)) \to 0 \quad \text{as } n \to \infty. \qquad \square
\end{aligned}$$

PROPOSITION A.2. *Let* $M = \{m_1, \ldots, m_k\} \subset \Re^p$ *and* $\alpha \in (0, 1)$. *Then the following statements are equivalent*:

(a) $V_{\Phi, \alpha}(M) > 0$;
(b) *there exists* $m_0 \in \Re^p$ *such that* $V_{\Phi, \alpha}(M \cup \{m_0\}) < V_{\Phi, \alpha}(M)$.

PROOF. We only prove that (a) implies (b), because the other implication is obvious. To do this, suppose that $V_{\Phi, \alpha}(M) > 0$. Then we have that $r_\alpha(M) > 0$ and $P_X(M) < 1 - \alpha$. Moreover, for every $r < r_\alpha(M)$, we have that $P_X(\overline{B}(M, r)) < 1 - \alpha$ and therefore there exist $m_0 \in \Re^p$ and $r_0 > 0$ such that $B_0 = B(m_0, r_0)$ satisfies:

(i) $\int_{B_0} \tau_{\alpha, M}(X)\, dP > 0$;
(ii) $\min_{i=1,\ldots,k} \|m_i - m_0\| > \frac{2}{3} r_\alpha(M)$;
(iii) $r_0 < \frac{1}{3} r_\alpha(M)$.

Hence

$$
\begin{aligned}
V_{\Phi,\alpha}(M) &= \frac{1}{1-\alpha}\int \tau_{M,\alpha}(X)\Phi(d(X,M))\,dP \\
&= \frac{1}{1-\alpha}\int_{B_0} \tau_{M,\alpha}(X)\Phi(d(X,M))\,dP \\
&\quad + \frac{1}{1-\alpha}\int_{B_0^c} \tau_{M,\alpha}(X)\Phi(d(X,M))\,dP \\
&> \frac{1}{1-\alpha}\int_{B_0} \tau_{M,\alpha}(X)\Phi(d(X,m_0))\,dP \\
&\quad + \frac{1}{1-\alpha}\int_{B_0^c} \tau_{M,\alpha}(X)\Phi(d(X,M))\,dP \\
&\geq \frac{1}{1-\alpha}\int \tau_{M,\alpha}(X)\min\{\Phi(d(X,M)),\Phi(d(X,m_0))\}\,dP \\
&\geq \frac{1}{1-\alpha}\int \tau_{M\cup\{m_0\},\alpha}(X)\Phi(d(X,M\cup\{m_0\}))\,dP \\
&= V_{\Phi,\alpha}(M\cup\{m_0\}). \qquad \square
\end{aligned}
$$

Our next result is the existence of trimmed $k$-means. Note that, if $X$ is a random vector, by Proposition 2.3, there exists a sequence of $k$-sets $M_n = \{m_1^n,\dots,m_k^n\} \subset \Re^p$, $n = 0,1,2,\dots,$ such that

$$V_{\Phi,\alpha}(M_n) \downarrow V_{k,\Phi,\alpha}(X) \quad \text{as } n \to \infty. \tag{7}$$

The existence of trimmed $k$-means will be established in a two-step process: first, we prove the existence of convergent subsequences of $\{M_n\}_n$ and, second, we show that the limit sets are trimmed $k$-means of $X$. We begin with a lemma.

LEMMA A.3. *Let us denote $a_n = \min_{i=1,\dots,k} d(m_i^n, 0)$ and $r_n = r_\alpha(M_n)$. Then $\{a_n\}_n$ and $\{r_n\}_n$ are bounded sequences.*

PROOF. Let $\gamma < \infty$ such that $P_X(B(0,\gamma)) > 1-\alpha$. Then, for every $n = 1,2,\dots,$ we have

$$a_n - \gamma \leq r_n \leq a_n + \gamma.$$

Thus it suffices to prove that one of the mentioned sequences is bounded. First, note that from (2) and (7) we have

$$V_{\Phi,\alpha}(M_n) \downarrow V_{k,\Phi,\alpha}(X) \leq \Phi(\gamma) < \Phi(\infty). \tag{8}$$

Let $\{\varepsilon_n\}_n$ and $\{\gamma_n\}_n$ be two sequences of positive numbers such that $\varepsilon_n \downarrow 0$, $\gamma_n \uparrow \infty$ and $P[X \in B(0,\gamma_n)] \geq 1-\varepsilon_n$. If $\{a_n\}_n$ were not bounded, we could find a subsequence (which we denote as the initial one) such that $a_n > 2\gamma_n$ for every $n = 1, 2, \dots$ and then we would have

$$\begin{aligned} V_{\Phi,\alpha}(M_n) &\geq \frac{1}{1-\alpha}\int_{B_n} \tau_n(X)\Phi(d(X, M_n))\, dP \\ &\geq \frac{1}{1-\alpha}\int_{B_n} \tau_n(X)\Phi(\gamma_n)\, dP \\ &\geq \Phi(\gamma_n)\frac{1-\alpha-\varepsilon_n}{1-\alpha} \uparrow \Phi(\infty), \end{aligned}$$

which contradicts (8). □

PROOF OF THEOREM 3.1. After Lemma A.3 we have that there exists a nonempty set $I \subseteq \{1, \dots, k\}$ and a subsequence (which we denote as the initial one) such that:

$$(9) \qquad \begin{aligned} &\text{if } i \notin I, \text{ then } d(m_i^n, 0) \to \infty \quad \text{as } n \to \infty, \\ &\text{if } i \in I, \text{ there exists } m_i^0 \in \Re^p \text{ such that } m_i^n \to m_i^0 \quad \text{as } n \to \infty. \end{aligned}$$

We can assume, without loss of generality, that $I = \{1, \dots, h\}$ with $1 \leq h \leq k$. Let us use the notation $M_n^h = \{m_1^n, \dots, m_h^n\}$ and $r_n' = r_\alpha(M_n^h)$, $n = 1, 2, \dots,$ and note that $r_n' \geq r_n$, $n = 1, 2, \dots,$ and $\{r_n'\}_n$ is a bounded sequence. First, we will prove that

$$(10) \qquad V_{\Phi,\alpha}(M_n^h) \to V_{h,\Phi,\alpha} \quad \text{as } n \to \infty \quad \text{and} \quad V_{h,\Phi,\alpha} = v_{k,\Phi,\alpha}.$$

Let us take $\{\varepsilon_n\}_n$ and $\{\gamma_n\}_n$ as in Lemma A.3. After (9) we can assume, without loss of generality, that, for every $n \in \mathscr{N}$,

$$\begin{gathered} d(m_i^n, 0) > 2\gamma_n \quad \text{for } i = h+1, \dots, k, \\ \left(\bigcup_{i=1}^h \overline{B}(m_i^n, r_n)\right) \cap \left(\bigcup_{i=h+1}^k \overline{B}(m_i^n, r_n)\right) = \varnothing \end{gathered}$$

and

$$P_X\left(\bigcup_{i=h+1}^k \overline{B}(m_i^n, r_n)\right) \leq \varepsilon_n.$$

Hence we have

$$V_{\Phi,\alpha}(M_n^h) \leq \frac{1}{1-\alpha}\left[\int_{\overline{B}(M_n^h, r_n)} \tau_n(X)\Phi(d(X, M_n^h))\, dP + \Phi(r_n')\varepsilon_n\right]$$

and then

$$\begin{aligned} (1-\alpha)V_{\Phi,\alpha}(M_n) &\geq \int_{\overline{B}(M_n^h, r_n)} \tau_n(X)\Phi(d(X, M_n^h))\, dP \\ &\geq (1-\alpha)V_{\Phi,\alpha}(M_n^h) - \Phi(r_n')\varepsilon_n \\ &\geq (1-\alpha)V_{h,\Phi,\alpha}(X) - \Phi(r_n')\varepsilon_n. \end{aligned}$$

Now, $\lim_{n\to\infty} \Phi(r_n')\varepsilon_n = 0$ because $(r_n')_n$ is bounded, so that

$$\lim_{n\to\infty} V_{\Phi,\alpha}(M_n) \geq \lim_{n\to\infty} V_{\Phi,\alpha}\big(M_n^h\big) \geq V_{h,\Phi,\alpha}(X), \tag{11}$$

and from this and (7) we obtain

$$V_{k,\Phi,\alpha} = \lim_{n\to\infty} V_{\Phi,\alpha}(M_n) \geq V_{h,\Phi,\alpha}.$$

Then, necessarily, $V_{k,\Phi,\alpha} = V_{h,\Phi,\alpha}$ and (10) holds. Moreover, from Proposition A.1, we have

$$V_{\Phi,\alpha}\big(M_n^h\big) \to V_{\Phi,\alpha}\big(M_0^h\big) \quad \text{as } n\to\infty \tag{12}$$

and from (11) and (12) it follows that

$$V_{\Phi,\alpha}\big(M_0^h\big) = V_{h,\Phi,\alpha}(X),$$

and then $M_0^h = \{m_1^0,\dots,m_h^0\}$ is a trimmed $h$-mean of X.

Now, if $h = k$, the proof is complete. If $h < k$, Proposition A.2 and (10) imply that $V_{\Phi,\alpha}(M_0^h) = 0$ and then the existence is obviously guaranteed for every $k \geq h$. $\square$

Finally, we are going to prove Theorem 3.4. We employ the same notation as in Section 3. That is, $\{x_n\}_n$ is a sequence of $\Re^p$-valued random vectors defined on $(\Omega,\sigma,P)$ and $M_n = \{m_1^n,\dots,m_k^n\}$, $n = 0,1,2,\dots,$ is a trimmed $k$-mean of $X_n$ with associated optimal trimming function $\tau_n$ and optimal radius $r_n$. Moreover, $V_n(= V_{k,\Phi,\alpha}(X_n))$, $n = 0,1,2,\dots,$ denotes the trimmed $k$-variation of $X_n$.

We begin with the following lemma. Its proof is somewhat related to that of Lemma A.3.

Lemma A.4. *If $X_n \to X_0$, $P$-a.e., and we denote $a_n = \min_{i=1,\dots,k} d(m_i^n, 0)$ for $n = 1,2,\dots,$ then $\{a_n\}_n$ and $\{r_n\}_n$ are bounded sequences.*

Proof. The sequence $\{X_n\}_n$ is tight. Thus there exists a ball $B(0,\gamma)$, $\gamma < \infty$, such that $P_{X_n}[B(0,\gamma)] > 1-\alpha$ for every $n = 1,2,\dots$. Then, for every $n = 1,2,\dots,$ we have

$$a_n - \gamma \leq r_n \leq a_n + \gamma,$$

so that is suffices to show that one of the sequences is bounded. First, note that

$$\begin{aligned} V_n &\leq \frac{1}{P_{X_n}(B(0,\gamma))} \int I_{B(0,\gamma)}(X_n)\Phi\big(d(X_n,0)\big)\,dP \\ &\leq \Phi(\gamma) < \Phi(\infty). \end{aligned} \tag{13}$$

Now, let $\{\varepsilon_n\}_n$ and $\{\gamma_n\}_n$ be sequences such that $\varepsilon_n \downarrow 0$, $\gamma_n \uparrow \infty$ and $P[X_n \in B(0,\gamma_n)] \geq 1-\varepsilon_n$.

If $\{a_n\}_n$ were not bounded, we could obtain a subsequence (which we denote as the initial one) such that $a_n > 2\gamma_n$ for every $n = 1,2,\dots$ and then we would have

$$\begin{aligned} V_n &\geq \frac{1}{1-\alpha} \int_{(X_n \in B_n)} \tau_n(X_n) \Phi(d(X_n, M_n))\, dP \\ &> \frac{1}{1-\alpha} \int_{(X_n \in B_n)} \tau_n(X_n) \Phi(\gamma_n)\, dP \\ &\geq \Phi(\gamma_n) \frac{1-\alpha-\varepsilon_n}{1-\alpha} \uparrow \Phi(\infty), \end{aligned}$$

which contradicts (13). □

PROOF OF THEOREM 3.4. It suffices to prove that every subsequence of $\{M_n\}_n$ (resp. $\{V_n\}_n$) admits a new subsequence which converges to $M_0$ (resp. to $V_0$).

For every $n = 1, 2, \ldots,$ let us denote by $\tau'_n$ any trimming function in $\tau_\alpha(X_n)$ based on the ball centered at $M_0$. Moreover, let us denote by $r'_n$, $n = 1, 2, \ldots,$ the radius associated with $\tau'_n$, that is,

$$I_{B(M_0, r'_n)} \leq \tau'_n \leq I_{\overline{B}(M_0, r'_n)}.$$

Obviously, $\{r'_n\}_n$ is a bounded sequence, and we can assume, without loss of generality, that $\lim_{n\to\infty} r'_n = r'_0$ for some $r'_o \in \Re$. Then, because of the continuity of $P_{X_0}$, we have

$$\tau'_n(X_n) \to I_{B(M_0, r'_0)}(X_0), \quad P\text{-a.e.},$$

and then, taking into account that $|\tau'_n| \leq 1$ for every $n \in N$, we may write

$$1 - \alpha = \int \tau'_n(X_n)\, dP \to \int I_{B(M_0, r'_0)}(X_0)\, dP \quad \text{as } n \to \infty.$$

Hence

$$\int I_{B(M_0, r'_0)}(X_0)\, dP = 1 - \alpha$$

and

$$I_{B(M_0, r'_0)} = \tau_0, \quad P_{X_0}\text{-a.e.}$$

Moreover, we have

$$\tau'_n(X_n)\Phi(d(X_n, M_0)) \to \tau_0(X_0)\Phi(d(X_0, M_0)), \quad P\text{-a.e.},$$

and $\{\tau'_n(X_n)\Phi(d(X_n, M_0))\}_n$ is uniformly bounded. Thus

$$\begin{aligned} V_n &\leq \frac{1}{1-\alpha} \int \tau'_n(X_n)\Phi(d(X_n, M_0)\, dP \\ &\to \frac{1}{1-\alpha} \int \tau_0(X_0)\Phi(d(X_0, M_0))\, dP \quad \text{as } n \to \infty \end{aligned}$$

and, consequently,

$$\limsup_n V_n \leq V_0. \tag{14}$$

By Lemma A.4 there exist a nonempty set $I \subseteq \{1, \dots, k\}$ and a subsequence of $\{M_n\}_n$ (which we denote as the initial one) such that:

$$\text{(15)} \qquad \begin{aligned} &\text{if } i \notin I, \text{ then } d(m_i^n, 0) \to \infty \quad \text{as } n \to \infty, \\ &\text{if } i \in I, \text{ there exists } m_i \in \Re^P \text{ such that } m_i^n \to m_i \quad \text{as } n \to \infty. \end{aligned}$$

We can assume, without loss of generality, that $I = \{1, \dots, h\}$ with $1 \le h \le k$. Let us use the notation $M^{(h)} = \{m_1, \dots, m_h\}$ and $M_n^{(h)} = \{m_1^n, \dots, m_h^n\}$, $n = 1, 2, \dots$. We can also assume that $\{r_n\}_n$ is a convergent sequence with limit, say, $r$. Then, for $n$ large enough,

$$\text{(16)} \qquad \begin{aligned} &I_{B(M_n^{(h)}, r_n)}(X_n) + I_{B(M_n - M_n^{(h)}, r_n)}(X_n) \\ &\qquad \le \tau_n(X_n) \le I_{\overline{B}(M_n^{(h)}, r_n)}(X_n) + I_{\overline{B}(M_n - M_n^{(h)}, r_n)}(X_n). \end{aligned}$$

Moreover,

$$I_{\overline{B}(M_n - M_n^{(h)}, r_n)}(X_n) \to 0, \quad P\text{-a.e.}$$

Thus, we obtain from (16) that

$$\lim_n \tau_n(X_n) = I_{B(M^{(h)}, r)}(X_0), \quad P\text{-a.e.}$$

Then, by taking into account that $|\tau_n| \le 1$ for every $n \in N$, we have

$$1 - \alpha = \int \tau_n(X_n)\, dP \to \int I_{B(M^{(h)}, r)}(X_0)\, dP \quad \text{as } n \to \infty,$$

so that $I_{B(M^{(h)}, \tau)}$ is a trimming function of level $\alpha$ for $X_0$. Furthermore,

$$\begin{aligned} \liminf_n V_n &= \frac{1}{1-\alpha} \liminf_n \int \tau_n(X_n) \Phi(d(X_n, M_n)\, dP \\ &\ge \frac{1}{1-\alpha} \int \liminf_n \big(\tau_n(X_n) I_{\overline{B}(M_n^{(h)}, r_n)}(X_n) \Phi(d(X_n, M_n)\big)\, dP \\ &\quad + \frac{1}{1-\alpha} \int \liminf_n \big(\tau_n(X_n) I_{\overline{B}(M_n - M_n^{(h)}, r_n)}(X_n) \Phi(d(X_n, M_n)\big)\, dP. \end{aligned}$$

Therefore, in view of

$$\frac{1}{1-\alpha} \int \liminf_n \big(\tau_n(X_n) I_{\overline{B}(M_n - M_n^{(h)}, r_n)}(X_n) \Phi(d(X_n, M_n)\big)\, dP = 0,$$

we obtain

$$\begin{aligned} \liminf_n V_n &\ge \frac{1}{1-\alpha} \int I_{B(M^{(h)}, r)}(X_0) \Phi\big(d\big(X_0, M^{(h)}\big)\big)\, dP \\ &\ge V_{h, \Phi, \alpha}(X_0). \end{aligned}$$

This and (14) imply

$$V_{h, \Phi, \alpha}(X_0) = V_{k, \Phi, \alpha}(X_0) = \lim_n V_n,$$

and the continuity of $P_{X_0}$ together with the uniqueness of $M_0$ shows that $I = \{1, \dots, k\}$ and $\{m_1, \dots, m_h\} = M_0$. □

**Acknowledgments.** The suggestion of using simulated annealing as well as the design of the algorithm and the software employed in the simulations are due to Professor Marc Lavielle (Univ. Paris V). We want to thank an anonymous referee and an Associate Editor for their valuable comments which have considerably improved the readability of the paper.

## REFERENCES

ARCONES, M. A. and GINÉ, E. (1992). On the bootstrap of $M$-estimators and other statistical functions. In *Exploring the Limits of Bootstrap* (R. Lepage and L. Billard, eds.) 13–47. Wiley, New York.

CAMBANIS, S. and GERR, N. L. (1983). A simple class of asymptotically optimal quantiziers. *IEEE Trans. Inform. Theory* **IT-29** 664–676.

CUESTA-ALBERTOS, J. A. and MATRÁN, C. (1988). The strong law of large numbers for $k$-means and best possible nets of Banach valued random variables. *Probab. Theory Related Fields* **78** 523–534.

CUESTA-ALBERTOS, J. A., GORDALIZA, A. and MATRÁN, C. (1995). On the Cauchy mean value property for $\Phi$-means. *Multivariate Statistics. Proceedings of the Fifth Tartu Conference on Multivariate Statistics* 247–265. VSP/TEV, Vilnius, Lithuania.

CUESTA-ALBERTOS, J. A., GORDALIZA, A. and MATRÁN, C. (1996). Trimmed $k$-nets. Preprint.

GORDALIZA, A. (1991a). Best approximations to random variables based on trimming procedures. *J. Approx. Theory* **64** 162–180.

GORDALIZA, A. (1991b). On the breakdown point of multivariate location estimators based on trimming procedures. *Statist. Probab. Lett.* **11** 387–394.

HARTIGAN, J. (1975). *Clustering Algorithms*. Wiley, New York.

HARTIGAN, J. (1978). Asymptotic distributions for clustering criteria. *Ann. Statist.* **6** 117–131.

IEEE (1982). *IEEE Trans. Inform. Theory* **IT-28**.

KAUFMAN, L. and ROUSSEEUW, P. J. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley, New York.

POLLARD, D. (1981). Strong consistency of $k$-means clustering. *Ann. Statist.* **9** 135–140.

POLLARD, D. (1982). A central limit theorem for $k$-mean clustering. *Ann. Probab.* **10** 919–926.

ROUSSEEUW, P. J. and LEROY, A. (1987). *Robust Regression and Outliers Detection*. Wiley, New York.

SERINKO, R. J. and BABU, G. J. (1992). Weak limit theorems for univariate $k$-mean clustering under a nonregular condition. *J. Multivariate Anal.* **41** 273–296.

SVERDRUP-THYGESON, H. (1981). Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metrics spaces. *Ann. Statist.* **9** 141–145.

TARPEY, T., LI, L. and FLURY, B. (1995). Principal points and self-consistent points of elliptical distributions. *Ann. Statist.* **23** 103–112.

J. A. CUESTA-ALBERTOS
DEPARTAMENTO DE MATEMÁTICAS
ESTADÍSTICA Y COMPUTACIÓN
UNIVERSIDAD DE CANTABRIA
SAN TANDER
SPAIN

A. GORDALIZA
C. MATRÁN
DEPARTAMENTO DE ESTADÍSTICA
E INVESTIGACIÓN OPERATIVA
UNIVERSIDAD DE VALLADOLID
47005 VALLADOLID
SPAIN