The Annals of Statistics
1999, Vol. 27, No. 3, 1061–1079

A CENTRAL LIMIT THEOREM FOR MULTIVARIATE GENERALIZED TRIMMED k-MEANS[1]

BY LUIS A. GARCÍA-ESCUDERO, ALFONSO GORDALIZA AND CARLOS MATRÁN

Universidad de Valladolid

A central limit theorem for generalized trimmed k-means is obtained in a very general framework that covers the multivariate setting, general penalty functions and general $k \geq 1$.

Several applications, including the location estimator case ($k = 1$) for elliptical distributions and the construction of multivariate (not necessarily connected) tolerance zones, are also given.

1. Introduction. Cuesta-Albertos, Gordaliza and Matrán [C-G-M, (1997)] introduced a robust clustering criteria, the trimmed k-means, consisting of the k-mean of the observations remaining after removing a fixed proportion of outlying observations: given a $\mathbb{R}^p$-valued random vector X, a suitable penalty function Φ and a trimming size α, we search for a Borel set B_0 in $\mathbb{R}^p$ and a k-set (a set with k points) $M_0 = \{m_1^0, m_2^0, \ldots, m_k^0\} \subset \mathbb{R}^p$ that are a solution to the constrained minimization problem,

$$\min_{B:\ P_X(B)\geq 1-\alpha} \ \min_{M\subset\mathbb{R}^p} \frac{1}{P_X(B)} \int_B \Phi\big(\inf_{i=1,\ldots,k} \|X - m_i\|\big)\, dP.$$

Trimmed k-means constitute the natural extension of the idea of the "impartial trimming," introduced in Gordaliza (1991a), to the clustering framework. Impartial trimming procedures include, as particular cases, Rousseeuw's (1983, 1984) least trimmed of squares estimator, LTS, and the least trimmed absolute deviations estimator, LTAD [Hössjer (1994), Tableman (1994)]. Other related estimators are D-estimators [Mili and Coakley (1996)] and the least trimmed log-likelihood estimators [Vandev and Neykov (1993)].

C-G-M (1997) established the existence and a characterization, without moment conditions, of trimmed k-means and proved their consistency for absolutely continuous multivariate distributions having a unique trimmed k - mean, but the important problem of determining their asymptotic distribution remains. Obtaining a central limit theorem (CLT) for trimmed k-means is the main objective of this paper.

To our knowledge, the only available results about this topic are for real valued random variables, $k = 1$, and mild penalty functions such as $\Phi(x) = x^2$ and $\Phi(x) = x$. Yohai and Maronna (1976) obtained a CLT for estimators such as the LTS estimator in the univariate case and for symmetric distributions,

Received November 1997; revised March 1999.

[1] Supported in part by DGICYT Grants PB95-0715-C02-00, 01 and by PAPIJCL VA08/97.

AMS 1991 *subject classifications*. Primary 60F05, 62H20; secondary 62G35, 62G15.

Key words and phrases. Central limit theorem, k-means, clustering, impartial trimming, robustness, tolerance zones.

by using ad hoc techniques based on linearity of rank statistics. Independently, Butler (1982) obtained a CLT, again for the LTS for univariate data, with empirical process techniques; these were utilizied for the LTAD estimator in Tableman (1994). In both cases, the existence of explicit expressions for the mean or the median is the key to the proofs, and this does not easily generalize to the multivariate setting nor to general penalty functions. García-Escudero, Gordaliza and Matrán (1997) obtained a generalization in the univariate setting for general k and $\Phi(x) = x^2$ or $\Phi(x) = x$. In the present paper we obtain a CLT under very general conditions: the multivariate framework, a general penalty function, Φ, and every $k \geq 1$. Interesting penalty functions are $\Phi(x) = x^r$, $1 \leq r < \infty$, which provide a robustified extension of classical r-means to the mixture setting. Moreover, when dealing with mixtures of spherical distributions, a natural choice of Φ would be based on the log-likelihood of the spherical distribution, analogously to the construction of location M-estimators.

In order to make a self-sufficient paper, Section 2 is devoted to background on the trimmed k-mean problem, stating the original definitions and preliminary results in C-G-M (1997).

The key step in the development of the paper is to transform the population trimmed k-means problem, which is a constrained double minimization problem, into a problem involving solution of implicit equations and then to apply Huber's (1967) classical result on multivariate M-estimators. The transformation of the problem is a consequence of the special probabilistic and geometrical structure of the restrictions. Also, geometrical properties are successfully used in the verification of Huber's conditions.

In Section 3.1, the original minimization problem is converted into a zero-property problem in the population setting. The finite sample counterpart is analyzed in Section 3.2. The CLT for trimmed k-means is given in Section 4, followed by a discussion of the imposed conditions and possible extensions.

The last section emphasizes the importance and applicability of the results. Firstly, we obtain an exact asymptotic limit law for multivariate elliptical unimodal distributions with $k = 1$. We also include a study on a mixture of bivariate normals. We apply the results to obtain multivariate tolerance zones (not necessarily connected) suitable when the parent distribution is a mixture. Although computational aspects of the procedure are not extensively considered, a brief sketch is given of a "simulated annealing" algorithm used to compute trimmed k-means.

C-G-M (1998) introduced trimmed k-nets as an extreme case of trimmed k-means arising from use of the L_∞-criterion, instead of the penalty functions, Φ. The obtained estimators are an extension of Rousseeuw's (1984) and Rousseeuw and Leroy's (1987) least median of squares estimator (LMS). However, their asymptotic behavior is different than that of trimmed k-means. For instance, consistency needs trimming sizes varying with the size of the sample [C-G-M (1998)] and the CLT fails because these estimators have a slower rate of convergence, $n^{1/3}$, to a nonnormal limit law. A detailed study of these estimators will be reported in a separate paper.

2. Notation and preliminaries.

Throughout this paper, (Ω, σ, P) denotes the probability space and X is an $\mathbb{R}^p$-valued random vector having probability law P_X in the σ-algebra $\mathfrak{B}^p$ of Borel sets in $\mathbb{R}^p$. A penalty function Φ: $\mathbb{R}^+ \to \mathbb{R}^+$ is considered and is assumed to be differentiable, nondecreasing and such that $\Phi(0) = 0$ and $\Phi(x) < \Phi(\infty)$ for all x. For obvious technical reasons, it will be assumed w.l.o.g. that $\Phi(\|\cdot\|) = \Phi_0(\|\cdot\|^2)$ and $\Phi_0'(\cdot) = \psi_0(\cdot)$.

For a set $B \subset \mathbb{R}^p$, $\overline{B}$ denotes its closure, B^c its complementary set and $Bd(B)$ its topological boundary. For $m \in \mathbb{R}^p$ and $r \geq 0$, $B(m, r)$ denotes the open ball with radius r centered at m. Given $M = \{m_1, \dots, m_k\} \subset \mathbb{R}^p$ and $r \geq 0$, denote the (generalized) ball centered at M and with radius r by

$$B(M, r) := \bigcup_{i=1}^{k} B(m_i, r).$$

Given a column vector $m_i \in \mathbb{R}^p$, its coordinates will be denoted as m_{ij}, $j = 1, \dots, p$.

Now we review the definitions, notations and needed results from C-G-M (1997). We refer to that paper for a more detailed description and proofs of the results.

For $\alpha \in (0, 1)$, τ_α denotes the nonempty set of trimming functions for X of level α, that is,

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

and $\tau_{\alpha-}$ denotes the set of trimming functions for levels less than or equal to α.

DEFINITION 1. A k-set $M_0 = \{m_1^0, m_2^0, \dots, m_k^0\}$ and a trimming function τ_0 are called a trimmed k-mean and an optimal trimming function if they satisfy

$$V_\Phi^{\tau_0}(M_0) = \inf_{\tau \in \tau_{\alpha-}} \inf_{M \subset \mathbb{R}^p,\, \#M = k} V_\Phi^{\tau}(M), \tag{1}$$

where $V_\Phi^\tau(M)$ is the variation about $M = \{m_1, m_2, \dots, m_k\}$ given τ,

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

Given any fixed k-set M and a $\beta \in (0, 1)$, define $r_\beta(M) = \inf\{r \geq 0\colon P_X(B(M, r)) \leq 1 - \beta \leq P_X(\overline{B}(M, r))\}$ and $\tau_{M, \beta}$ as the set of trimming functions

$$\tau_{M, \beta} = \big\{\tau \in \tau_\beta\colon I_{B(M, r_\beta(M))} \leq \tau \leq I_{\overline{B}(M, r_\beta(M))}, \text{ a.e. } P_X\big\}.$$

Following Remark 2.3 in C-G-M (1997), the double minimization problem in (1) can be restated as the selection of a k-set, M_0, satisfying

$$V_{\Phi, \alpha}(M_0) = \inf_{M \subset \mathbb{R}^p,\, \#M = k} V_{\Phi, \alpha}(M),$$

where

$$V_{\Phi,\alpha}(M) = \frac{1}{1-\alpha}\int \tau(X)\Phi\Big(\inf_{i=1,\dots,k} \|X - m_i\|\Big)\, dP,$$

with τ any function in $\tau_{M,\alpha}$.

If Φ is strictly increasing and τ_0 and M_0 are the solutions of (1), then there exits r_0 ($\equiv r_\alpha(M_0)$) such that

$$I_{B(M_0,r_0)} \le \tau_0 \le I_{\overline{B}(M_0,r_0)}, \quad P_X\text{-a.e.} \tag{2}$$

(r_0 will be called the optimal radius). For the remainder, we assume that Φ is strictly increasing.

$\overline{B}(M_0, r_0)$ is the optimal set except, possibly, for part of the boundary. Moreover, the optimal set can be partitioned into k clusters as follows: the cluster A_i consists of all the points in the optimal set which are closer to m_i^0 than the remaining $k-1$ points in the k-set M_0. Points at the same distance from two elements of M_0 are arbitrarily assigned.

The previous definition and remarks will also be used for the empirical probability measure $P_n^\omega(A) := n^{-1}\sum_{i=1,\dots,k} I_A[X_i(\omega)]$ (where $\{X_n\}_n$ is a sequence of independent, identically distributed random vectors with distribution P_X).

Definition 2. The trimmed k-mean of P_n^ω, $M_n := M_n^\omega = \{m_1^n, \dots, m_k^n\}$, will be called the sample trimmed k-mean, and the associated radius $r_n := r_n^\omega$ will be called the sample optimal radius.

Finally, we define the set-valued functions

$$\begin{aligned} A_1 &= A_1(m_1, m_2, \dots, m_k, r) \\ &:= \{x\colon \|x - m_1\| \le \|x - m_j\|,\ j = 1, \dots, k\} \cap \overline{B}(m_1, r) \end{aligned} \tag{3}$$

and

$$\begin{aligned} A_i &= A_i(m_1, m_2, \dots, m_k, r) \\ &:= \{x\colon \|x - m_i\| \le \|x - m_j\|,\ j = 1, \dots, k\} \\ &\qquad \cap \overline{B}(m_i, r) \cap A_1^c \cap \dots \cap A_{i-1}^c \end{aligned} \tag{4}$$

for $i = 2, \dots, k$. (Note that we needed a precise convention for allocating points common to the boundaries of two or more of the A_i's. We have supposed that the rule to break ties is to assign them to the set of lower index, but other rules are also possible).

The following proposition, obtained in C-G-M (1997), gives the main characterization of the trimmed k-means.

Proposition 2.1. *Let $M_0 = \{m_1^0, m_2^0, \dots, m_k^0\}$ be a trimmed k-mean of the distribution P_X, r_0 the optimal radius ($r_0 = r_\alpha(M_0)$) and $A_i^0 := A_i(m_1^0,$*

$m_2^0, \dots, m_k^0, r_0)$, $i = 1, \dots, k$, *be defined as in* (3) *and* (4). *Then* m_i^0 *must be the* Φ-*mean of* X *given the cluster* A_i^0 [*i.e.*, m_i^0 *minimizes* $\inf_{m \in \mathbb{R}^p} \int_{A_i^0} \Phi(\|X - m\|)\, dP$].

3. The trimmed k-means problem for absolutely continuous distributions and its sample version.

This section is devoted to demonstrating that the trimmed k-means problem for absolutely continuous distributions can be embedded into the general theory of estimators defined through a zero property (Z-estimators) and hence that Huber's result can be applied. As observed in the introduction, minimum (M-) problems can naturally be converted into zero (Z-) problems but it is rather surprising that a constrained minimization problem can be handled as a zero problem with even the Lagrange multipliers disappearing. This happens because of special restrictions that our trimming procedure imposes, as will be clarified in Remark 3.1.

We will assume that P_X is absolutely continuous with respect to Lebesgue measure in $(\mathbb{R}^p, \mathfrak{B}^p)$, with bounded density f. We will also suppose that f is not identically null in the boundary of the optimal zone.

3.1. *The population trimmed k-mean.*

From (2), it is not necessary to use trimming functions in order to get the trimmed k-mean and the optimal zone (i.e., $I_{\overline{B}(M_0, r_\alpha(M_0))} = \tau_0$, P_X-a.e.). Thus, we can consider the optimal zone as $\overline{B}(M_0, r_\alpha(M_0))$, and this zone must satisfy

$$\int I_{\overline{B}(M_0, r_\alpha(M_0))}\, dP_X(x) = 1 - \alpha. \tag{5}$$

Considering $M_0 = \{m_1^0, m_2^0, \dots, m_k^0\}$, the trimmed k-mean of the distribution P_X, and r_0, the optimal radius, we can split the optimal zone $\overline{B}(M_0, r_0)$ into k zones $A_i^0 := A_i(m_1^0, m_2^0, \dots, m_k^0, r_0)$, $i = 1, \dots, k$. By Proposition 2.1, we also have that m_i^0 must satisfy

$$\int I_{A_i^0}(x)(x_j - m_{ij}^0)\psi_0\big(\|x - m_i^0\|^2\big)\, dP_X(x) = 0, \tag{6}$$

$i = 1, \dots, k$; $j = 1, \dots, p$ and $x = (x_1, x_2, \dots, x_p)' \in \mathbb{R}^p$. Hence, combining the expressions (5) and (6) we have characterized the population trimmed k-mean by a zero property. Although we have transformed a constrained problem into an unconstrained problem, the following remark indicates why this is natural.

REMARK 3.1. For the sake of simplicity, assume first that $k = 1$. A standard Lagrange multiplier technique transforms the constrained minimization problem into the minimization of

$$H(m, r, \lambda) = \int_{\overline{B}(m,r)} \Phi_0\big(\|x - m\|^2\big) f(x)\, dx + \lambda\bigg(\int_{\overline{B}(m, r)} f(x)\, dx - (1 - \alpha)\bigg).$$

Denote, by Θ_1, Θ_2 and Θ_3, the partial derivatives of $H(m, r, \lambda)$ with respect to m, r and λ, respectively.

If $\sigma_{m,r}(x)$ is the surface measure on $Bd(\overline{B}(m,r))$ and $n(x,m,r)$ is the outward pointing unit vector normal to $B(m,r)$, then by applying tools of classical differential geometry [see, for instance, Baddeley (1977) and Section 5 in Kim and Pollard (1990)], we have that

$$\begin{aligned}\Theta_1 = &-\int_{\overline{B}(m,r)} 2(x-m)\psi_0(\|x-m\|^2)f(x)\,dx \\ &+\int_{Bd(\overline{B}(m,r))} n(x,m,r)'\Phi_0(\|x-m\|^2)f(x)\,d\sigma_{m,r}(x) \\ &+\lambda\int_{Bd(\overline{B}(m,r))} n(x,m,r)'f(x)\,d\sigma_{m,r}(x)\end{aligned}$$

and

$$\Theta_2 = \int_{Bd(\overline{B}(m,r))} \Phi_0(\|x-m\|^2)f(x)\,d\sigma_{m,r}(x) + \lambda\int_{Bd(\overline{B}(m,r))} f(x)\,d\sigma_{m,r}(x).$$

Taking into account that $\Phi_0(\|x-m\|^2) = \Phi_0(r^2)$ for all $x \in Bd(\overline{B}(m,r))$ and the fact that

$$\int_{Bd(\overline{B}(m_0,r_0))} f(x)\,d\sigma_{m_0,r_0}(x) > 0$$

(remember the regularity condition imposed on the density in the boundary of the optimal zone), we have that $\Theta_2 = 0$ implies that $\Phi_0(r_0^2)+\lambda_0 = 0$, where λ_0 is the optimal Lagrange multiplier. Placing this condition into the expression $\Theta_1 = 0$, we trivially see that m_0 and r_0 must satisfy

$$\int_{B(m_0,r_0)} (x-m_0)\psi_0(\|x-m_0\|^2)f(x)\,dx = 0,$$

which gives (6) for $j = 1,\dots,p$ with $k = 1$. Equation (5) follows automatically from $\Theta_3 = 0$.

In the case $k \geq 2$, we proceed analogously. The only apparent difference rests on the possible presence of the overlapping phenomena. If there exists a set $\Gamma_{ij}^0 := Bd(A_i^0) \cap Bd(A_j^0) \neq \varnothing$ for some $i \neq j$, then shifting m_i not only changes the integration region A_i^0 but also changes A_j^0. It can be easily proved that this fact is not a difficulty. Notice now the taking derivatives will yield two surface integrals in the zone Γ_{ij}^0, which will cancel because if $x \in \Gamma_{ij}^0$ then $\Phi_0(\|x-m_i\|^2) = \Phi_0(\|x-m_j\|^2)$ and the outward pointing unit normal vectors in Γ_{ij}^0, in each integral, have opposite directions. So we only have surface integrals in those regions included in the boundary of the whole set $\overline{B}(M_0,r_0)$ and the proof is similar to the $k = 1$ case.

So, in appealing to (5) and (6), if $\boldsymbol{\theta} = (m_1', m_2', \dots, m_k', r)' \in \mathbb{R}^{p\times k+1}$ with $m_i \in \mathbb{R}^p$ and $r \in \mathbb{R}^+$ and $A_i(\boldsymbol{\theta})$ is defined as in (3) and (4), we may consider the following function Ψ:

$$\Psi(x,\boldsymbol{\theta}) = \big(\Psi_{11}(x,\boldsymbol{\theta}),\dots,\Psi_{1p}(x,\boldsymbol{\theta}),\Psi_{21}(x,\boldsymbol{\theta}),\dots,\Psi_{kp}(x,\boldsymbol{\theta}),\Psi_R(x,\boldsymbol{\theta})\big)',$$

where

$$\begin{aligned}\Psi_{ij}(x,\boldsymbol{\theta}) &= I_{A_i(\boldsymbol{\theta})}(x)(x_j - m_{ij})\psi_0\bigl(\|x-m_i\|^2\bigr) \quad \text{and}\\ \Psi_R(x,\boldsymbol{\theta}) &= I_{\bigcup_{i=1}^k A_i(\boldsymbol{\theta})}(x) - (1-\alpha)\end{aligned} \tag{7}$$

to characterize the trimmed k-means as the solution of the zero-problem $E\Psi(X,\boldsymbol{\theta}) = 0$.

3.2. *The sample trimmed k-mean.* Given $X_1, X_2, \dots, X_n$, a sample of independent, identically distributed random vectors with distribution P_X, let

$$T_n = T_n(X_1, X_2, \dots, X_n) := \bigl((m_1^n)', \dots, (m_k^n)', r_n\bigr)'$$

be the sample estimators in Definition 2. In order to apply Huber's result, we need the consistency of the sequence of estimators $\{T_n\}_n$ and also to prove that $\sum_{l=1}^n \Psi(X_l, T_n) = o_P(n^{1/2})$.

The following lemma can be obtained through a straightforward modification of the proof of Theorem 3.6 in C-G-M (1997). Notice that the additional hypothesis of uniqueness of the optimal radius is included in order to obtain consistency of sample optimal radii.

LEMMA 3.1. *Let $\{M_n\}_n$, $M_n = \{m_1^n, \dots, m_k^n\}$, be the sequence of sample trimmed k-means and $\{r_n\}_n$ be the sequence of sample optimal radii. Assume that P_X is absolutely continuous and that there exists a unique* (*up to a relabeling*) *trimmed k-mean of the distribution P_X, $M_0 = \{m_1^0, \dots, m_k^0\}$, and a unique optimal radius r_0. Then*

$$m_i^n \to m_i^0, \qquad P\text{-a.e.}, \qquad i = 1, \dots, k$$

and

$$r_n \to r_0, \qquad P\text{-a.e.}$$

(*Properly, it should be stated as the existence of a relabeling of the set M_n such that the previous convergences are true. However, without loss of generality, in this paper we will assume that this relabeling is not necessary.*)

LEMMA 3.2. *Under the above conditions,*

$$\frac{1}{\sqrt{n}} \sum_{l=1}^n \Psi(X_l, T_n) = o_P(1).$$

PROOF. The result follows from the definition of T_n, because the sample optimal trimming function, based on a sample of size n, τ_n, satisfies

$$\int_{\overline{B}(M_n, r_n)} \tau_n(x)\, dP_n(x) = 1 - \alpha$$

and

$$\int_{A_i(T_n)} \tau_n(x)(x_j - m_{ij}^n)\psi_0(\|x - m_i^n\|^2)\, dP_n(x) = 0,$$

$$i = 1, \dots, k \text{ and } j = 1, \dots, p.$$

Then we have

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{l=1}^{n} \Psi_R(X_l, T_n) &= \frac{1}{\sqrt{n}} \left[n P_n\left(\bigcup_{i=1}^{k} A_i(T_n) \right) - n \int_{\overline{B}(M_n, r_n)} \tau_n(x)\, dP_n(x) \right] \\ &= \frac{1}{\sqrt{n}} \left[\sum_{\{l:\ X_l \in Bd(\overline{B}(M_n, r_n))\}} \bigl(1 - \tau_n(X_l)\bigr) \right] = o_P(1) \end{aligned} \tag{8}$$

because, for P_X absolutely continuous, there is a fixed number G which depends only on k and p such that the probability of the event that more than G points in the sequence $X_1, X_2, \dots, X_n$ lie on such a boundary is zero. To better understand this claim, think of the circumference $\mathscr{C}(\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3)$ determined in $\mathbb{R}^2$ by the random points X_1, X_2 and X_3. Since the Lebesgue measure in $\mathbb{R}^2$ of a circumference is zero, if f is the density of P_X (recall our assumption of absolute continuity) we have

$$\begin{aligned} &P\bigl(X_m \in \mathscr{C}(X_1, X_2, X_3)\bigr) \\ &\quad = \int_{\mathbb{R}^2} f(x_1) \int_{\mathbb{R}^2} f(x_2) \int_{\mathbb{R}^2} f(x_3) \int_{\mathscr{C}(x_1, x_2, x_3)} f(x_m)\, dx_m\, dx_3\, dx_2\, dx_1 = 0. \end{aligned}$$

Therefore the probability of the event "four random points of the sequence $X_1, X_2, \dots, X_n, \dots$ lie in the same circumference," that is,

$$P\left(\bigcup_{i,\, j,\, k,\, m} \bigl(X_m \in \mathscr{C}(X_i, X_j, X_k)\bigr) \right) \text{ for distinct } i, j, k \text{ and } m,$$

is zero.

Analogously for $i = 1, \dots, k$; $j = 1, \dots, p$ and $X_l = (X_{l1}, \dots, X_{lk})'$, we have

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{l=1}^{n} \Psi_{ij}(X_l, T_n) \\ &\quad = \frac{1}{\sqrt{n}} \Biggl[\sum_{l=1}^{n} I_{A_i(T_n)}(X_l)(X_{lj} - m_{ij}^n)\psi_0\bigl(\|X - m_i^n\|^2\bigr) \\ &\qquad\qquad - n \int_{A_i(T_n)} \tau_n(x)(x_j - m_{ij}^n)\psi_0\bigl(\|x - m_i^n\|^2\bigr)\, dP_n(x) \Biggr] \\ &\quad = \frac{1}{\sqrt{n}} \left[\sum_{\{l:\ X_l \in Bd(\overline{B}(M_n, r_n))\}} \bigl(1 - \tau_n(X_l)\bigr)(X_{lj} - m_{ij}^n)\psi_0\bigl(\|X_l - m_i^n\|^2\bigr) \right] \\ &\quad = o_P(1), \end{aligned}$$

by the previous comment about the number of points lying in the boundary of the optimal zone and the fact that $(X_{lj} - m^n_{ij})\psi_0(\|X_l - m^n_i\|^2) = O_P(1)$ (because $m^n_i \to m^0_i$, P-a.e.).

4. The main result. Before tackling the asymptotic normality proof, we begin with two lemmas corresponding to rather well-known results. The proof of the first lemma is a simple exercise in differential calculus while that of the second is not as straightforward, although the result is very intuitive [see Van der Vaart and Wellner (1996), page 163].

LEMMA 4.1. *Let A be an open subset in $\mathbb{R}^m$ and $f\colon A \to \mathbb{R}^m$ be differentiable at $z \in A$. If the differential at z is an automorphism of the vector space $\mathbb{R}^m$, then there exists two positive real numbers $a > 0$ and $d_0 > 0$ such that*

$$\|f(z+h) - f(z)\| \geq a\|h\| \quad \text{for all } h \in \mathbb{R}^m \text{ with } \|h\| \leq d_0.$$

Let $d(x, C) = \inf_{y \in C} d(x, y)$. For a given set C, let $C^\varepsilon = \{x\colon d(x, C) < \varepsilon\}$ and ${}_\varepsilon C = \{x\colon d(x, C^c) > \varepsilon\}$ denote the set of points within distance ε of C and the set of points that are at least a distance ε inside C, respectively.

LEMMA 4.2. *Let $\mathscr{C}$ be the class of all compact and convex subsets of a fixed bounded subset of $\mathbb{R}^m$. There exists a constant H depending only on $\mathscr{C}$, such that if λ is the Lebesgue measure in $\mathbb{R}^m$,*

$$\lambda(C^\varepsilon - {}_\varepsilon C) \leq H\varepsilon \quad \text{for every } \varepsilon > 0.$$

As we commented before, the proof of the asymptotic normality of trimmed k-means will be based on a result on M-estimators given in Section 4 of Huber (1967). This result will not be repeated here, but the notation in our normality proof will be chosen to match that of Huber's result. Given the probability space $(\mathbb{R}^p, \mathfrak{B}^p, P_X)$ and $\Theta = \mathbb{R}^{p\times k} \times \mathbb{R}^+ \subset \mathbb{R}^{p\times k+1}$, we consider the function $\Psi\colon \mathbb{R}^p \times \Theta \to \mathbb{R}^{p\times k+1}$ defined in (7). Let $\lambda(\boldsymbol{\theta}) = (\lambda_{11}(\boldsymbol{\theta}), \dots, \lambda_{1p}(\boldsymbol{\theta}), \lambda_{21}(\boldsymbol{\theta}), \dots, \lambda_{kp}(\boldsymbol{\theta}), \lambda_R(\boldsymbol{\theta}))'$ be the expectation of $\Psi(X, \boldsymbol{\theta})$ with respect to the true underlying distribution.

If M_0 and r_0 are the population trimmed k-mean and the optimal radius of the distribution P_X, then we have seen that

$$\boldsymbol{\theta}_0 = \big((m^0_1)', (m^0_2)', \dots, (m^0_k)', r_0\big)' \in \mathbb{R}^{p\times k+1} \tag{9}$$

satisfies $\lambda(\boldsymbol{\theta}_0) = 0$. With the notation introduced previously, the following theorem gives a CLT for the trimmed k-mean estimator:

THEOREM 4.1. *Let X be a random vector with bounded density, f, and such that $\lambda(\boldsymbol{\theta})$ admits a unique zero (up to a relabeling) at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ [$\boldsymbol{\theta}_0$ as in (9)]. Suppose that λ is differentiable at $\boldsymbol{\theta}_0$ with nonsingular derivative matrix Λ and that the density f is not identically null in the boundary of the optimal zone. Assume that $x\psi_0(\|x\|^2)$ is a Lipschitz function in $\overline{B}(0, R)$ with $R > r_0$.*

Under the above conditions, $\sqrt{n}(T_n - \boldsymbol{\theta}_0)$ is asymptotically normal with mean 0 and covariance matrix $\Lambda^{-1}C(\Lambda')^{-1}$, where C stands for the covariance matrix of the random vector $\Psi(X, \boldsymbol{\theta}_0)$.

PROOF. Conditions N-1 and N-2 from Huber's result are clearly satisfied. As a consequence of Lemma 4.1, there exist positive constants a and d_0 such that N-3(i) is also verified. Moreover, because of the trimming, assumption N-4 is always true. So, in order to apply Huber's result, we only need to prove conditions N-3(ii) and (iii).

First, we need $r_0 + 4d_0 < R$ and $d_0 < 1$. If this is not true, without loss of generality, we may consider a smaller d_0 satisfying these inequalities. Let us define

$$u(x, \boldsymbol{\theta}, d) = \sup_{\|\boldsymbol{\tau}-\boldsymbol{\theta}\| \le d} \big\|\Psi(x, \boldsymbol{\tau}) - \Psi(x, \boldsymbol{\theta})\big\|.$$

If $\boldsymbol{\theta} = (\theta_1', \theta_2', \dots, \theta_k', r_\theta)'$ and $\boldsymbol{\tau} = (\tau_1', \tau_2', \dots, \tau_k', r_\tau)'$ with $\theta_i,\ \tau_i \in \mathbb{R}^p$ and r_θ, $r_\tau \in \mathbb{R}^+$, then notice that $\|\boldsymbol{\theta} - \boldsymbol{\tau}\| \le d$ implies $\|\theta_i - \tau_i\| \le d$, $i = 1, \dots, k$ and $|r_\theta - r_\tau| \le d$ (note also that we have not made any notational distinction between norms in spaces with different dimensionality).

For fixed $\boldsymbol{\theta}$ and d satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + d \le d_0$ we will denote $A_i := A_i(\boldsymbol{\theta})$, as in (3) and (4), and we will split $\mathbb{R}^p$ into three zones $\mathscr{Z}_1 = \bigcup_{i=1}^k ({}_{2d}A_i)$, $\mathscr{Z}_2 = (\bigcup_{i=1}^k A_i)^{2d} - \bigcup_{i=1}^k ({}_{2d}A_i)$ and $\mathscr{Z}_3 = [(\bigcup_{i=1}^k A_i)^{2d}]^c$.

We will study the function $u(x, \boldsymbol{\theta}, d)$ in these three zones. So, suppose $\|\boldsymbol{\tau} - \boldsymbol{\theta}\| \le d$. Then we have:

(i) Let $x \in \mathscr{Z}_1$. Let us suppose that $x \in {}_{2d}A_i$, then we have

$$\|x - \tau_i\| \le \|x - \theta_i\| + \|\tau_i - \theta_i\| \le r_\theta - 2d + d \le r_\tau$$

and also

$$\|x - \tau_i\| = \inf_{j=1,\dots,k} \|x - \tau_j\|$$

(because $\|x - \tau_i\| \le \|x - \theta_i\| + d$ and $\|x - \tau_l\| > \|x - \theta_i\| + 2d - d$, when $l \neq i$). So, $x \in A_i(\boldsymbol{\tau})$ and this forces Ψ to satisfy

$$\Psi_{lj}(x, \boldsymbol{\theta}) - \Psi_{lj}(x, \boldsymbol{\tau}) = 0 \quad \text{for all } j = 1, \dots, p \text{ and } l \neq i$$

and

$$\Psi_{ij}(x, \boldsymbol{\theta}) - \Psi_{ij}(x, \boldsymbol{\tau}) = (x_j - \theta_{ij})\psi_0\big(\|x - \theta_i\|^2\big) - (x_j - \tau_{ij})\psi_0\big(\|x - \tau_i\|^2\big).$$

Therefore,

$$\begin{aligned}\big\|\Psi(x, \boldsymbol{\theta}) - \Psi(x, \boldsymbol{\tau})\big\| &= \big\|(x - \theta_i)\psi_0\big(\|x - \theta_i\|^2\big) - (x - \tau_i)\psi_0\big(\|x - \tau_i\|^2\big)\big\| \\ &\le L\|\theta_i - \tau_i\| \le L\,d \quad \text{for a constant } L,\end{aligned}$$

because of the Lipschitz character of the function $x\psi_0(\|x\|^2)$.

(ii) Now, let $x \in \mathscr{Z}_3$. We will suppose directly that x belongs to the interior of $\mathscr{Z}_3$ (this excludes also a Lebesgue measure zero set, which will be unimportant later). Then

$$\|x-\tau_i\| \geq \big|\|x-\theta_i\| - \|\theta_i-\tau_i\|\big| > r_\theta + 2d - d = r_\theta + d \geq r_\tau,$$

for every $i = 1, \dots, k$, and then $\Psi(x, \boldsymbol{\theta}) - \Psi(x, \boldsymbol{\tau}) = 0$ all over this zone.

(iii) Obviously, the most troublesome zone is $\mathscr{Z}_2$. But in this zone we will use Lemma 4.2. Trivially,

$$\left(\bigcup_{i=1}^{k} A_i\right)^{2d} - \bigcup_{i=1}^{k}({}_{2d}A_i) \subseteq \bigcup_{i=1}^{k}\big(\overline{A_i}^{2d} - {}_{2d}\overline{A_i}\big).$$

If $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + d \leq d_0$, then we have $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq d_0$ and it follows that

$$\overline{A_i} \subseteq \overline{B}(\theta_i, r_\theta) \subseteq \overline{B}(\theta_i, r_0 + d_0) \subseteq \overline{B}\big(m_i^0, r_0 + 2d_0\big) \subset \overline{B}(m_i^0, R) \subseteq \overline{B}(M_0, R).$$

Then we have $\overline{A_i}$, $i = 1, \dots, k$ a family of compact and convex sets (the Voronoi regions and the balls are trivially convex sets in $\mathbb{R}^p$) which are contained in the set $\overline{B}(M_0, R)$. Then for $d \leq d_0 \leq 1$, Lemma 4.2 holds for a constant H which does not depend on $\boldsymbol{\theta}$.

Also, if $x \in \overline{A_i}^{2d} - {}_{2d}\overline{A_i}$, we have that

$$\|x - \theta_i\| \leq r_\theta + 2d \leq r_0 + d_0 + 2d_0 \leq R$$

and

$$\|x - \tau_i\| \leq \|x - \theta_i\| + \|\tau_i - \theta_i\| \leq r_0 + 4d_0 \leq R.$$

Since $x\psi_0(\|x\|^2)$ is a Lipschitz function on the compact set $\overline{B}(0, R)$, it is bounded there; also it is continuous and hence $\|\Psi(x, \boldsymbol{\theta}) - \Psi(x, \boldsymbol{\tau})\|$ is also bounded by some constant D uniformly in $\bigcup_{i=1}^{k}(\overline{A_i}^{2d} - {}_{2d}\overline{A_i})$.

If M is a bound for the density f, joining all the previous bounds we will have that

$$Eu(X, \boldsymbol{\theta}, d) \leq LdP(X \in \mathscr{Z}_1) + DMkH2d \leq bd,$$

where $b = L + DMkH2$. Then, condition N-3(ii) is fulfilled.

To prove N-3(iii), we can proceed analogously. For instance, in the zone $\mathscr{Z}_1$ we can see that $u(x, \boldsymbol{\theta}, d)^2 \leq (Ld)^2 \leq (L^2 d_0)d$, and in the zone $\mathscr{Z}_2$ we can get the bound $u(x, \boldsymbol{\theta}, d) \leq D^2$. □

REMARK 4.1. The assumptions needed to apply the previous CLT are not very restrictive and are commonly assumed in the clustering setting. The nonsingularity of the matrix Λ is a common assumption for estimators defined by a minimum property (for instance, in the asymptotic theory for maximum likelihood estimators, the information matrix is usually just assumed nonsingular) and in k-means clustering. As far as we know, in the k-means setting, the unique paper where a possible singularity is dealt with is in Serinko and Babu (1992).

The requirement of uniqueness (up to a relabeling) of the solution of the minimum problem is usually hard to verify. In the asymptotics of (untrimmed)

k-means, this condition is usually assumed [see Pollard (1981, 1982), Hartigan (1978), Stute and Zhu (1995)] and only a few papers, such as Fleischer (1964) and Li and Flury (1995), consider this difficulty. We believe that, when dealing with "reasonable" mixtures for clustering, it is quite rare to find distributions where either condition fails and, even then, the lack of uniqueness could be only due to an improper choice of k or α. In the $k = 1$ case and for spherical distributions, nonsingularity of Λ and uniqueness in the solution of the minimization problem will be explicitly proved in Section 5. For $k > 1$, it is natural to hope that this property will be inherited by sufficiently distant mixtures of such distributions.

The boundedness of the density is really only needed in the boundary of the optimal sets. Of course, this is a logical regularity condition near the optimum.

REMARK 4.2. The condition about the Lipschitz property of $x\psi_0(\|x\|^2)$ can be notably weakened by an easy adaptation of the previous proof. For instance, in the univariate case only the Lipschitz property in each interval of the partition $[-R, a_1), [a_1, a_2), \dots, [a_s, R]$ of $\overline{B}(0, R)$ in $\mathbb{R}$ is needed (in $a_1, a_2, \dots, a_s$, the function $x\psi_0(|x|^2)$ may have discontinuities). This covers in $\mathbb{R}$ the case of the trimmed k-median (i.e., $\Phi(x) = x$ needs $\psi_0(x) = x^{-1/2}$, so $x\psi_0(|x|^2) = \operatorname{sign}(x)$, and this function is Lipschitz in $[-R, 0)$ and $[0, R]$ for every $R > 0$).

In $\mathbb{R}^p$ with $p \geq 2$, we need the Lipschitz condition of $x\psi_0(\|x\|^2)$ now over differences of balls $B(0, a_1), B(0, a_1)^c \cap B(0, a_2), \dots, B(0, a_s)^c \cap \overline{B}(0, R)$ with $0 < a_1 < \cdots < a_s \leq R$.

Note also that it is not difficult to give simple conditions based on the derivative of the function ψ_0 (if it exists up to a finite number of points) to guarantee these previous properties.

REMARK 4.3. The use of criteria based on minimization of the rth power deviation with $1 \leq r < 2$, appealing because of its robustness properties, has a long history. We can introduce a trimmed version of this procedure to this framework. Now, the function $x\psi_0(\|x\|^2)$ will be $x\|x\|^{r-2}$.

We can reproduce the proof of Theorem 4.1, but taking into account that there exist positive constants c_1 and c_2 such that $u(x, \theta, d) \leq c_1 d\|x - \theta\|^{r-2}$ and $u^2(x, \theta, d) \leq c_2 d\|x - \theta\|^{r-2}$ [these facts are mentioned in Example 1 in Huber (1967)]. We have to use that, for bounded densities in $\mathbb{R}^p$, if $1 \leq r < 2$ and $r + p > 2$, then there exists a constant F such that $E(\|x - \theta\|^{r-2}) \leq F < \infty$ for all $\theta \in \mathbb{R}^p$. The proof also needs a study of the zones $\mathscr{Z}_2$ and $\mathscr{Z}_3$, but this goes parallel to the proof of Theorem 4.1 (this result does not include the case $r = 1$ and $p = 1$, but this case is covered in Remark 4.2).

5. Examples and applications.

Trimmed Φ-means. The (impartial) trimmed Φ-mean constitutes the simplest setup for applicability of this work. These estimators, with general penalty functions in the multivariate setting, were introduced in Gordaliza

(1991a) and their high breakdown point property was studied in Gordaliza (1991b). We will study the applicability of Theorem 4.1 when $k = 1$ for elliptical distributions. Thus, suppose that the random vector X admits an elliptical unimodal density $f(x) = |\Sigma|^{-1/2} h((x-\theta_0)\Sigma^{-1}(x-\theta_0))$ where Σ is a p.d. matrix, $h(x) > 0$ and $h'(x) < 0$ for all $x > 0$. We will assume for simplicity that $\theta_0 = 0$.

Uniqueness. The proof of uniqueness will be based on a multivariate probability inequality in Davies (1987), Lemma 4.

LEMMA 5.1. *Let $\theta \in \mathbb{R}^p$ and Σ be a p.s.d. matrix and ξ and g: $\mathbb{R}^+ \to \mathbb{R}^+$ be nonincreasing functions such that $\int g(x'x)\,dx < \infty$. Then*

$$\int \xi\big((x-\theta)'\Sigma^{-1}(x-\theta)\big) g(x'x)\,dx \le \int \xi\big(x'\Sigma^{-1}x\big) g(x'x)\,dx.$$

Defining

$$V_{\Phi,\alpha}(m) = \frac{1}{1-\alpha}\int_{B(m,\,r_\alpha(m))} \Phi_0\big(\|x-m\|^2\big) f(x)\,dx,$$

we need to prove that $V_{\Phi,\alpha}(m) > V_{\Phi,\alpha}(0)$ for all $m \in \mathbb{R}^p$. If r_0 is the radius of the optimal ball [r_0 is equal to $r_\alpha(0)$], consider $\xi(\cdot) = h(\cdot)$ and $g(\cdot) = (\Phi_0(r_0^2) - \Phi_0(\cdot))I_{[0,\,r_0^2]}(\cdot)$ $(g(x'x) = (\Phi_0(r_0^2) - \Phi_0(\|x\|^2))I_{B(0,\,r_0)}(x))$. Applying Lemma 5.1 to the functions ξ and g previously defined, yields

$$\begin{aligned}
&\Phi_0(r_0^2)\int_{B(0,\,r_0)} f(x)\,dx - (1-\alpha)V_{\Phi,\alpha}(0)\\
&\qquad \ge \Phi_0(r_0^2)\int_{B(0,\,r_0)} f(x+m)\,dx - \int_{B(0,\,r_0)} \Phi_0\big(\|x\|^2\big) f(x+m)\,dx\\
&\qquad = \Phi_0(r_0^2)\int_{B(m,\,r_0)} f(x)\,dx - \int_{B(m,\,r_0)} \Phi_0\big(\|x-m\|^2\big) f(x)\,dx.
\end{aligned}$$

Now, add and subtract $(1-\alpha)V_{\Phi,\alpha}(m)$ in the second term of the previous inequality and rearrange terms to obtain

$$\begin{aligned}
(1-\alpha)\big[V_{\Phi,\alpha}(m) - V_{\Phi,\alpha}(0)\big] \ge{}& \Phi_0(r_0^2)\left[\int_{B(m,\,r_0)} f(x)\,dx - \int_{B(0,\,r_0)} f(x)\,dx\right]\\
&- \left[\int_{B(m,\,r_0)} \Phi_0\big(\|x-m\|^2\big) f(x)\,dx\right.\\
&\qquad \left. - \int_{B(m,\,r_\alpha(m))} \Phi_0\big(\|x-m\|^2\big) f(x)\,dx\right].
\end{aligned}$$

Taking into account that

$$\int_{B(0,\,r_0)} f(x)\,dx = \int_{B(m,\,r_\alpha(m))} f(x)\,dx,$$

we can see that $V_{\Phi,\alpha}(m) - V_{\Phi,\alpha}(0) > 0$ is true if $r_\alpha(m) > r_0$ and

$$\int_{B(m,r_0)^c \cap B(m,r_\alpha(m))} \Phi_0(\|x-m\|^2) f(x)\,dx \\ > \Phi_0(r_0^2) \int_{B(m,r_0)^c \cap B(m,r_\alpha(m))} f(x)\,dx,$$

but this is always true, because $\Phi_0(\|x-m\|^2) > \Phi_0(r_0^2)$, for all $x \in \overline{B(m,r_0)}^c \cap B(m, r_\alpha(m))$, $f(x) > 0$ and $r_\alpha(m) > r_0$ (this last fact is a simple exercise).

Nonsingularity of Λ. Let $\boldsymbol{\theta} = (m', r) = (m_1, \dots, m_p, r)' \in \mathbb{R}^p$, $\lambda(\cdot) = (\lambda_1, \dots, \lambda_p, \lambda_R)'(\cdot)$. Straightforward calculations give

$$\frac{\partial}{\partial m}\lambda_j(\boldsymbol{\theta})\bigg|_{m=0,\,r=r_0} = \frac{2}{\sqrt{|\Sigma|}} \int_{B(0,r)} x_j \psi_0(\|x\|^2) \\ \times h'\big((x+m)'\Sigma^{-1}(x+m)\big)\Sigma^{-1}(x+m)\,dx\bigg|_{m=0,\,r=r_0}$$

for $j = 1, \dots, p$, so we can write

$$\frac{\partial}{\partial m}\lambda_{(1,2,\dots,p)}(\boldsymbol{\theta})\bigg|_{m=0,\,r=r_0} = \frac{2}{\sqrt{|\Sigma|}}\Sigma^{-1} \int_{B(0,r)} \psi_0(\|x\|^2) h'(x'\Sigma^{-1}x) xx'\,dx.$$

Provided that $\psi_0(y) > 0$ and $h'(y) < 0$ for all $y > 0$, there exists a (negative) c_0 constant such that $f_0(y) = c_0\psi_0(\|y\|^2)h'(y'\Sigma^{-1}y)$ is a symmetric ($f_0(y) = f_0(-y)$) nondegenerate density in the ball $B(0, r_0) \subset \mathbb{R}^p$. Then $(\partial/\partial m)\lambda_{(1,2,\dots,p)}(\boldsymbol{\theta})|_{m=0,\,r=r_0}$ is equal to the matrix $|\Sigma|^{-1/2}2c_0^{-1}\Sigma^{-1}$ times the covariance matrix of a random vector with nondegenerate density f_0, so it is a nonsingular matrix. Due to symmetry considerations, it is trivial to see that $(\partial/\partial m)\lambda_R(\boldsymbol{\theta})|_{m=0,\,r=r_0} = 0'$ and $(\partial/\partial r)\lambda_{(1,2,\dots,p)}(\boldsymbol{\theta})|_{m=0,\,r=r_0} = 0$. The density f is strictly positive in the boundary of the optimal ball $B(0, r_0)$, so considering its surface integral we have that $(\partial/\partial r)\lambda_R(\boldsymbol{\theta})|_{m=0,\,r=r_0} \neq 0$. Thus, $\Lambda = (\partial/\partial m)\lambda_{(1,2,\dots,p,r)}(\boldsymbol{\theta})|_{m=0,\,r=r_0}$ is a nonsingular matrix.

Trimmed 2*-mean.* Consider the mixture of bivariate normals

$$1/2N_2(\mu_1, \mathrm{Id}) + 1/2N_2(\mu_2, \mathrm{Id}),$$

where $\mu_1 = (-1.5, 0)'$, $\mu_2 = (1.5, 0)'$ and Id is the identity matrix in $\mathbb{R}^2$. Given the trimming level $\alpha = 0.2$, we obtain the population trimmed 2-mean ($\Phi_0(x) = x$) equal to $M_0 = \{m_1^0, m_2^0\}$, where $m_1^0 = (m_{11}^0, m_{12}^0)' = (-1.42, 0)'$ and $m_2^0 = (m_{21}^0, m_{22}^0)' = (1.42, 0)'$. The optimal radius is equal to $r_0 = 1.69$, so the optimal zone is the union of two overlapped balls giving the clusters $A_1^0 = B(m_1^0, r_0) \cap \{(x, y)\colon x \le 0\}$ and $A_2^0 = B(m_2^0, r_0) \cap \{(x, y)\colon x > 0\}$.

Now, let $M_n = \{m_1^n, m_2^n\}$ with $m_1^n = (m_{11}^n, m_{12}^n)'$ and $m_2^n = (m_{21}^n, m_{22}^n)'$ be the empirical trimmed 2-mean and r_n the optimal empirical radius. We consider m_1^n and m_2^n as estimates of m_1^0 and m_2^0, respectively (i.e., $m_1^n \to m_1^0$ and $m_2^n \to m_2^0$, P-a.e. [C-G-M (1997)]). Then, Theorem 4.1 asserts that

$$\sqrt{n}\big((m_{11}^n, m_{12}^n, m_{21}^n m_{22}^n, r_n)' - (m_{11}^0, m_{12}^0, m_{21}^0, m_{22}^0, r_0)'\big)$$

is asymptotically normal with zero mean and an asymptotic covariance matrix which has been numerically approximated as

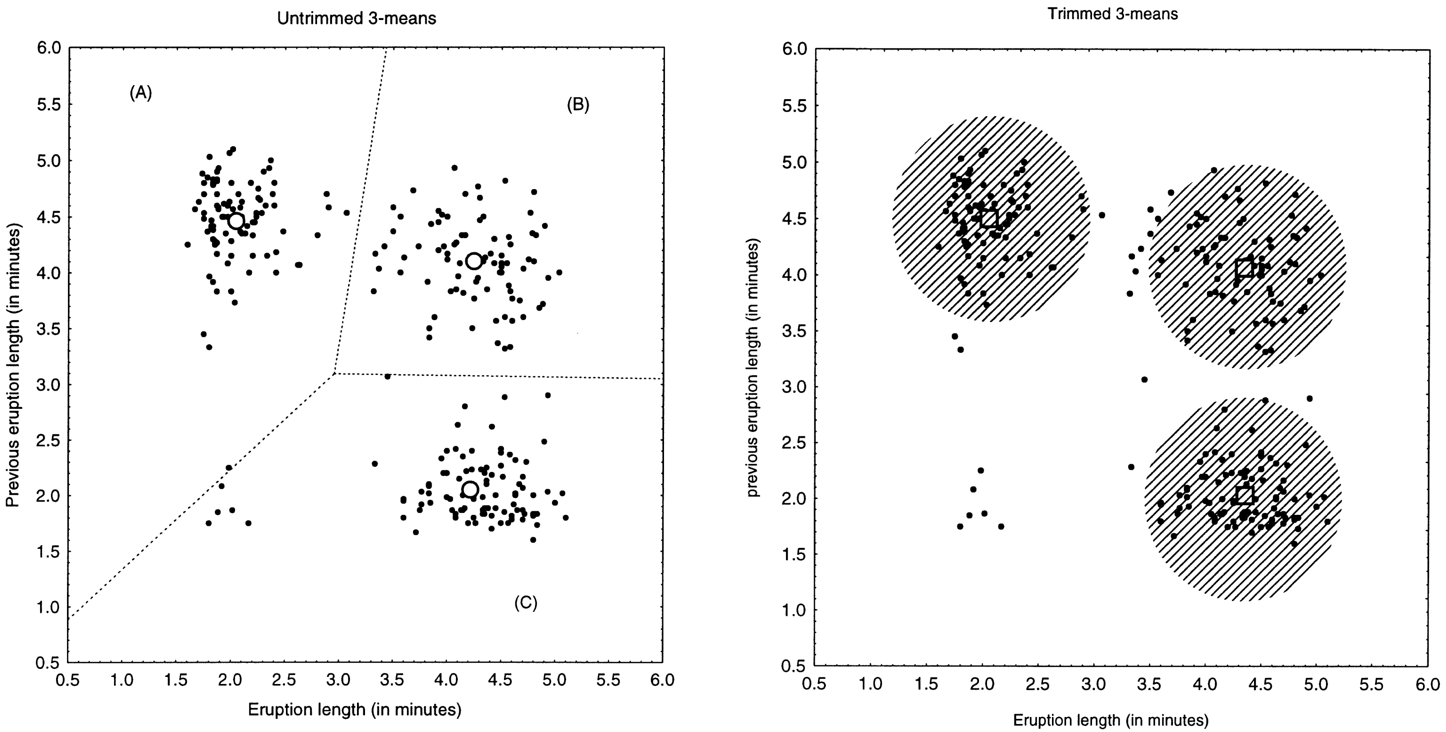
$$\begin{pmatrix} 6.95 & 0 & 3.83 & 0 & 0.05 \\ 0 & 4.85 & 0 & -0.32 & 0 \\ 3.83 & 0 & 6.95 & 0 & -0.05 \\ 0 & -0.32 & 0 & 4.85 & 0 \\ 0.05 & 0 & -0.05 & 0 & 1.10 \end{pmatrix}.$$

A study of the variance–covariance structure in the univariate case ($p = 1$), using the influence function for its computation, is given in García-Escudero and Gordaliza (1999). Also in that paper it is observed that, although the effect of the trimming is usually an increase in the asymptotic variances, for heavy-tailed distributions the situation may be reversed. Notice that heavy-tailed distributions are somehow associated with the presence of outliers, so this is another reason the method is especially suitable when outliers are suspected (remember also that no assumption concerning existence of population moments of the underlying distribution is needed in Theorem 4.1).

Tolerance regions. Many statistical procedures involve summarizing a probability distribution by a region of the sample space covering a previously specified probability. When we demand also a guarantee of attaining this covering probability, we may speak of tolerance regions. It seems plausible that a reasonable region should occupy the smallest possible volume and should consist of high density zones. The region satisfying both previous conditions obviously need not be a connected region (especially in the mixture setting).

Butler (1982) introduces, in the univariate case, a robust distribution-free tolerance interval based on minimization of the trimmed variance (i.e., based on the LTS estimator). Tableman (1994) considers a tolerance interval obtained from the LTAD estimator in the univariate case. Butler, Davies and Jhun (1993) consider an ellipsoidal tolerance zone based on the minimum covariance determinant estimator (MCD) apropriate for multivariate elliptical distributions. All the previous procedures give only connected tolerance regions.

The optimal zone provided by the trimmed k-mean method is not necessarily connected and can be used as a tolerance region. This possibility was noted in García-Escudero, Gordaliza and Matrán (1997) in the univariate setting. In



Untrimmed 3-mean (circles) and the corresponding data cluster allocations.

0.08-trimmed 3-mean (squares). The shaded region shows its associated tolerance region (95% guarantee of attaining a 90% coverage).

FIG. 1. *Tolerance region for the Old Faithful Geyser data.*

the multivariate setting and for general penalty functions, consider $\overline{B}(M_n, r_n)$, the sample optimal zone, and recall [from (8)] that $\sqrt{n}(P_n(\overline{B}(M_n, r_n)) - (1-\alpha)) = o_P(1)$. Arguing as in Butler, Davies and Jhun (1993), using standard empirical process theory, taking into account the fact that a finite union of closed balls in $\mathbb{R}^p$ is a Donsker class and using the convergence of $\overline{B}(M_n, r_n)$ to $\overline{B}(M_0, r_0)$, we obtain that

$$\sqrt{n}\big(P(\overline{B}(M_n, r_n)) - (1-\alpha)\big) \longrightarrow_D N\big(0, \alpha(1-\alpha)\big). \tag{10}$$

Figure 1 shows, for a well-known data set (eruptions and lagged eruptions from Old Faithful Geyser), that when we suspect that data arises from a mixture of distributions then the tolerance region based on a trimmed k-mean could provide a good performance. Making use of (10) with $n = 271$, we need a trimming size $\alpha = 0.08$ to reach a 0.95-guarantee of attaining a 0.9-coverage. The figure exhibits the tolerance zone associated with the 0.08-trimmed 3-mean ($\Phi(x) = x^2$).

Figure 1 also shows the classical (untrimmed) k-mean. We observe that the six anomalous data points in the lower left corner (short followed by short eruptions) are artificially assigned to clusters (C) and (A). This fact modifies slightly the centers of those clusters. If the anomalous data points were placed farther away, the untrimmed 3-mean could break down. For a detailed discussion about the robustness gain provided by the trimming procedure, see García-Escudero and Gordaliza (1999).

This methodology, when considering a suitable number of clusters, k, could be related to methods of support estimation [as in Cuevas and Fraiman (1997)] or with searching for highest density regions [Scott (1992) and Hyndman (1996)].

Sketch of the algorithm. We used a simulated annealing-based algorithm to obtain the trimmed k-mean and the corresponding optimal zone. A brief description of the algorithm is as follows: first, randomly partition the data into $k+1$ groups (k clusters and one additional group containing the trimmed observations). Then, in each step of the algorithm, one observation is allowed to be changed from one of the groups to another. Notice that, proceeding in this form, we are not guaranted to have the required trimming proportion α. To solve this drawback, a parameter γ which penalizes increasing the size of the trimmed observations group is also included in the "energy" function (i.e., bigger γ's will lead to lower trimmed solutions). The reason for introducing this parameter γ is to avoid local minimums which appear, even with the simulated annealing algorithm, if the trimming level is kept fixed. Finally, we must modify the parameter γ in an interactive way until the right trimming size is reached.

Acknowledgments. We thank an anonymous referee, an Associate Editor and the Editor for their valuable comments, which have improved the paper. We are also indebted to J. A. Cuesta-Albertos for his helpful suggestions during the preparation of the paper.

REFERENCES

BADDELEY, A. (1977). Integrals of a moving manifold and geometrical probability. *Adv. in Appl. Probab.* **9** 588–603.

BUTLER, R. W. (1982). Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Ann. Statist.* **10** 197–204.

BUTLER, R. W., DAVIES, P. L. and JHUN, M. (1993). Asymptotics for the minimum covarince determinant estimator. *Ann. Statist.* **21** 1385–1400.

CUESTA-ALBERTOS, J. A., GORDALIZA, A. and MATRÁN, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.* **25** 553–576.

CUESTA-ALBERTOS, J. A., GORDALIZA, A. and MATRÁN, C. (1998). Trimmed best k-nets: a robustified version of a L_∞-based clustering method. *Statist. Probab. Lett.* **36** 401–413.

CUEVAS, A. and FRAIMAN, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25** 2300–2312.

DAVIES, P.L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292.

FLEISCHER, P. (1964). Sufficient conditions for achieving minimum distorsion in a quantizer. *IEEE Int. Conv. Rec.* 104–111.

GARCÍA-ESCUDERO, L. A. and GORDALIZA, A. (1999). Robustness properties of k-means and trimmed k-means. *J. Amer. Statist. Assoc.* To appear.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A. and MATRÁN, C. (1997). Asymptotics for trimmed k-means and associated tolerance zones. *J. Statist. Plann. Inference* **77** 247–262.

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. A. (1978). Asymptotic distribution for clustering criteria. *Ann. Statist.* **6** 117–131.

HÖSSJER, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.* **89** 149–158.

HUBER, P. J. (1967). The behavior of maximum likelihood estimators under non-standard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 221–233. Univ. California press, Berkeley.

HYNDMAN, R. J. (1996). Computing and graphing highest density regions. *Amer. Statist.* **50** 120–126.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219.

LI, L. and FLURY, B. (1995). Uniqueness of principal points for univariate distributions. *Statist. Probab. Lett.* **25** 323–327.

MILI, M. and COAKLEY, C. (1996). Robust estimation in structured linear regression. *Ann. Statist.* **24** 2593–2607.

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

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

ROUSSEEUW, P. J. (1983). Multivariate estimation with high breakdown point. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics* (W. Grossman, G. Plufg, I. Vincze and W. Werttz, eds.) **B** 283–297. Reidel, Dordrecht.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880.

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

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

SERINKO, R. J. and BABU, G. J. (1992). Weak limit theorems for univariate k-means clustering inder nonregular conditions. *J. Multivariate Anal.* **49** 188–203.

STUTE, W. and ZHU, L. X. (1995). Asymptotics of k-means clustering based on projection pursuit. *Sankhyā* **57** 462–471.

TABLEMAN, M. (1994). The asymptotics of the least trimmed absolute deviation (LTAD) estimator. *Statist. Probab. Lett.* **19** 387–398.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Wiley, New York.

VANDEV, D. L. and NEYKOV, N. M. (1993). Robust maximum likelihood in the Gaussian case. In *New Directions in Statistical Data Analysis and Robustness* (S. Morgenthaler, E. Ronchetti and W. A. Stahel, eds.). Birkhäuser, Basel.

YOHAI, V. and MARONNA, R. (1976). Location estimators based on linear combinations of modified order statistics. *Comm. Statist. Theory Methods* **5** 481–486.

DEPARTAMENTO DE ESTADÍSTICA
E INVESTIGACIÓN OPERATIVA
UNIVERSIDAD DE VALLADOLID
C/ PRADO DE LA MAGDALENA S/N
VALLADOLID 47005
SPAIN
E-MAIL: langel@westad.eis.uva.es