DEPARTAMENTO DE MATEMÁTICA
DOCUMENTO DE TRABAJO

"Impartial Trimmed k-means for Functional Data"

Juan Antonio Cuesta-Albertos, Ricardo Fraiman

D.T.: N° 32 Marzo 2005

Vito Dumas 284, (B1644BID) Victoria, Pcia. de Buenos Aires, Argentina
Tel: 4725-7053 Fax: 4725-7010
Email: matematica@udesa.edu.ar

Impartial Trimmed k-means for Functional Data

Juan Antonio Cuesta-Albertos [a 1], Ricardo Fraiman [b,*]

[a] *Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Spain.*

[b] *Departamento de Matemáticas, Universidad de San Andrés, Argentina and Centro de Matemática, Universidad de la República, Uruguay*

Abstract

A robust cluster procedure for functional data is introduced. It is based on the notion of impartial trimming. Existence and consistency results are obtained. Furthermore, a feasible algorithm is proposed and implemented in a real data example, where patterns of electrical power consumers are observed.

Key words: Cluster Analysis, Functional Data, Impartial Trimming.

1 Introduction

Resistant methods for cluster analysis, based on the concept of "impartial trimming" are proposed. The framework under consideration will include high dimensional problems and the case of functional data (also called longitudinal data in health and social sciences) which arise nowadays in a number of increasing scientific fields, associated with continuous-time monitoring processes that provides samples of functions.

The impartial trimming (where those data to be trimmed are self–determined by the whole data set) was introduced in [7] as a robust alternative tech-

[1] This author has been partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant BFM2002-04430-C02-02

* Corresponding author. Postal address: Departamento de Matemática y Ciencias, Universidad de San Andrés, Vito Dumas 284, 1644, Victoria, Argentina. *Tel.*:5411-47257062 . *Fax*: 5411-47257010. Email address: rfraiman@udesa.edu.ar

nique for the multi-dimensional location problem. Notice that the usual one-dimensional trimming methods fail in the multi–dimensional case because there is no more a natural zone (a neighborhood of ∞) where the observations to be trimmed should be chosen. The idea of impartial trimming have been particularly useful for cluster analysis in finite dimensional spaces.

For cluster analysis, impartial trimming techniques offer a resistant alternative to k-means, one of the most widely used cluster methods. A quick description of k-means is the following:

- For a fixed value of k the method search for the centers of the k groups into which the data will be classified.
- The criteria for this search is to minimize the dispersion within groups.
- Once the centers are founded, each data point is assigned to the group of its nearest center.

However, if $k > 1$, k-means is very sensitive to the effect of a small group of atypical observations (outliers). Effectively, as pointed in [3], a single data point far away from the data cloud, will produce an artificial group center and the estimate will breakdown in this sense. In order to avoid this problem, the notion of impartial trimmed k–means (ITkM in what follows) was introduced in [3]. Roughly speaking, the method drops out a small proportion of the data before starting the search of the centers of the groups (see a more precise description of the procedure in Section 2). The deletion step is impartial in the sense that the deleted points are self-determined by the sample.

In order to apply the ITkM method, it is necessary to provide an effective algorithm to find the ITkM centers for a given data set, a problem which has shown to be very involved. This is particularly important for high dimensional problems, which will be the case where will be focused this work. In [2] an approximate algorithm was proposed to handle this problem in the case $k = 1$. We extend the same idea to approximate the ITkM centers, which, broadly speaking, restrict the search for the ITkM centers to the points in the sample. An important computational advantage of this algorithm is that it only requires to calculate the matrix of distances between the points in the sample. Another property of this estimate is that the selected centers are already data points in the sample and not averages of data points. This allows to measure additional variables which were missed in a first analysis for the centers which can be thought as typical representatives of their cluster group.

The paper is organized as follows. In Section 2 we state the ITkM problem in a general setting; we introduce the estimates and extend the algorithm proposed in [2] in order to calculate them. In Section 3 we show the existence of ITkM and present some asymptotic results. The proofs are rather technical and combine techniques developed in [2] and [3]. A sketch of the proofs is given

in the Appendix. In Section 4 patterns of electricity consumers are analyzed by the ITkM technique. This is a real data set with a large amount of outliers (atypical data), where the method seems to do well its job.

In this paper we will consider the data in a quite general setting, the feature space (see, for instance, [10]) in order to include the case of functional data, or more general data structures. This objective will be tackled by considering the case where we have a sample of random elements taken from a random process taking values in an abstract Banach space E. We chose this framework because sometimes a Hilbert space setting could be very restrictive in the context of Functional Data. For instance, even if our data are real functions in L^2, quite often we are interested in considering the L^1 distance, or the L^∞ distance. Other times we are interested in the curve and its derivative. In this case, we consider a norm that takes both aspects into consideration falling in a Sobolev space context. Other simple examples arise when considering functional linear models.

Unless otherwise is stated $(E, \|-\|)$ will stand for a uniformly convex Banach space, β will denote the associated Borel σ-algebra and we will assume that all the random elements are E-valued and defined on the same rich enough probability space (Ω, σ, μ).

Given $h \in E$ and $r > 0$, $B(h, r)$ (respectively $\overline{B}(h, r)$) will denote the open (resp. closed) ball with center at h and radius r.

We will need to handle convergence of sets which is defined as follows. Given $k \geq 1$ and a sequence of subsets of E with cardinal k, $H_n = \{h_1^n, ..., h_k^n\}$ we will say that the sequence $\{H_n\}_n$ converges (weakly) to the set $H = \{h_1, ..., h_m\}$ with $m \leq k$ if there exists a labeling such that if we denote $H_n = \{h_{i_1}^n, ..., h_{i_k}^n\}$, then, for every $j = 1, ..., m$, we have that $\lim_n h_{i_j}^n = h_j$ (weakly) and, if $j > m$ then $\lim_n \|h_{i_j}^n\| = \infty$.

2 Trimmed k-M-parameters and estimates.

The underlying model assumes that we are handling a population which is split into k clusters although this is not explicitly stated. The goal is to estimate the centers of those clusters. We also allow to the sample to be contaminated with a proportion less or equal than α of points which belong to none of the clusters.

2.1 k-M-parameters and estimates.

We start defining the population *k-M-mean* parameter (or *k-M-cluster centers*) whose trimmed version we want to estimate. Here, letter M refers to the fact that 1-M-estimates (see Subsection 2.2) are, in fact, M-estimates which are based on the same methodology. Given a continuous and strictly increasing score function $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$, an E-valued random element X with distribution P and $k \in \mathbb{N}$, let us denote

$$I_\Phi^k(P) := I_\Phi^k(X) := \inf_{h_1,...,h_k \in E} \int \Phi \left[\inf_{i=1,...,k} \|x - h_i\| \right] P(dx). \tag{1}$$

$H_{P,\Phi}^k$ will stand for any set with k elements $\{h_1^P, ..., h_k^P\} \subset E$ for which equality in (1) is reached.

When $\Phi(t) = t^2$ and E is the p-dimensional euclidean space, $H_{P,\Phi}^k$ coincides with the usual k-means parameter of P. If we take Φ as the identity function and $k = 1$ we have the median for $E = \mathbb{R}$ and the spatial median in general. Moreover, if P is symmetrical and unimodal, then $H_{P,\Phi}^1$ coincides with the symmetry center of P.

If we allow Φ to denote a general penalty function, then the set $H_{P,\Phi}^k$ is a so-called *k-Φ-mean of the distribution P* or, simply, a k-mean if no confusion is possible. The k-Φ-means were introduced in [11] and [1] for euclidean spaces and in [12] and [4] in more general spaces and, obviously, can be considered as a generalization of usual k-means.

In order to estimate the parameters, let us assume that we have a sample $\{X_n\}$ of random elements with the same distribution as X. For $n \in \mathbb{N}$, let P_n be the empirical probability distribution, i.e., given $A \in \beta$

$$P_n[A] = \frac{1}{n} \sum_{i=1}^{n} I_A[X_i],$$

where I_A denotes the indicator of the set A.

Natural plug–in estimates of the set $H_{P,\Phi}^k$ and the value $I_\Phi^k(P)$ are $H_{P_n,\Phi}^k$ (which will be called a k-M-estimate) and $I_\Phi^k(P_n)$.

2.2 Impartial trimmed k-M-estimates.

As stated in the introduction, it is well known that, for any score function Φ, if we replace a single point in the sample by another point X_0 which is

located far away from the data cloud, and $k > 1$, then $X_0 \in H^k_{P_n,\Phi}$. Taking into account that this point may be located as far as desired, we have that, if the sample size is n, the replacement of a proportion of n^{-1} points in the sample could lead to that at least one of the elements in $H^k_{P_n,\Phi}$ goes to infinity, and we can say that the breakdown point of the k-M-estimate is n^{-1}.

To get more robust estimates, if $E = I\!R^p$, $p \geq 1$, in [3] (see also [7] for the case $k = 1$ and [2] for $k = 1$ and $E = L^2[0,1]$) the *impartial trimmed k-means* were introduced. The idea of ITkM is to choose an (usually small) number $\alpha \in (0,1)$, and then change a little the function to be minimized, in order to allow a proportion α of the data points in the sample to be dropped away.

To be precise, let $\mathcal{P}_{\alpha,n}$ be the family of all measurable functions $\tau : E \to [0,1]$ such that $\int \tau(x) P_n(dx) \geq 1 - \alpha$. Thus, if $\tau \in \mathcal{P}_{\alpha,n}$, then τ trims at most a proportion α of points in the sample. However, τ does not necessarily trim complete points (giving weight 0 or 1 to all data points), but may give them a weight in $(0,1)$. This is required to obtain an exact α trimming level and, then, to show the existence of optimal trimmings functions exactly in the same way as randomized tests are required to obtain uniformly more powerful tests in the Neyman-Pearson theory.

Obviously, trimming functions are not necessarily limited to sample distributions. In general, given a probability distribution P, $\mathcal{P}_\alpha$ will denote the family of all measurable functions $\tau : E \to [0,1]$ such that $\int \tau(x) P(dx) \geq 1 - \alpha$.

Now, given $k \geq 1$, a distribution P on β, $\tau \in \mathcal{P}_\alpha$, and $h_1, ..., h_k \in E$, we define

$$g^k(\tau, h_1, ..., h_k, P) := \int \Phi \left[\inf_{i=1,...,k} \|x - h_i\| \right] \tau(x) P(dx),$$

and

$$I^k_{\alpha,\Phi}(P) := \inf_{\tau \in \mathcal{P}_\alpha} \inf_{h_1,...,h_k \in E} g^k(\tau, h_1, ..., h_k, P), \tag{2}$$

and let $H^k_{P,\Phi,\alpha}$ be any set in which this infimum is reached.

The empirical version is defined in the natural way as follows. Define

$$I^k_{\alpha,\Phi}(P_n) := \inf_{\tau \in \mathcal{P}_{\alpha,n}} \inf_{h_1,...,h_k \in E} g^k(\tau, h_1, ..., h_k, P_n), \tag{3}$$

and let now $H^k_{P_n,\Phi,\alpha}$ be any set in which the infimum in (3) is reached.

Functions in $\mathcal{P}_\alpha$ are often called *α-trimming functions*. We will denote $\tau_{\alpha,\Phi}$ to any empirical trimming function in which the infimum in (2) is reached and

we will refer to it as an *optimal α-trimming function*.

Functions in $\mathcal{P}_{\alpha,n}$ are called *empirical α-trimming functions*. We will denote $\tau_{n,\alpha,\Phi}$ to any trimming function in which the infimum in (3) is reached and we will refer to it as an *optimal empirical α-trimming function*.

Broadly speaking, we can say that the difference between $H^k_{P,\Phi}$ and $H^k_{P,\Phi,\alpha}$ is that in the computation of the last quantity we are allowed to get rid of a proportion α of the points in E. The points to be trimmed are those for which the value of $g^k(\tau, h_1, ..., h_k, P)$ is as low as possible. But this difference is more deep than it seems. For instance, even in the case $E = I\!R^2$, $k = 1$ and $\Phi(t) = t^2$, in [2] appears an example of a symmetrical and unimodal distribution P such that $H^1_{P,\Phi}$ does not contain the symmetry center of P. However, the same paper also includes a result with an additional sufficient condition to avoid this striking behavior.

Several properties of the ITkM-estimates are known in the euclidean case, being, as far as we know, [2] the only paper containing results on the application of those ideas to the functional framework and even this paper is restricted to the case $E = L^2[0, 1]$ and $k = 1$. In particular, in the situations above described it is known that the sets $\{H^k_{P_n,\Phi,\alpha}\}_n$ converge a.s. to the set $H^k_{P,\Phi,\alpha}$.

A well known problem is that, even in the one-dimensional, non-trimmed, k-means case, $k \geq 2$, the only effective algorithm to compute the set $H^k_{P_n,\Phi}$ needs to check all possible partitions of the sample obtained with $(k-1)$ hyperplanes; the situation being even worst if you try to apply trimming because, in this case, it is also required to check all possible trimmings. To circumvent this, several procedures have been proposed (see, for instance, [6]) which, in fact, do not provide the absolute minimum in (3) but a stationary point. Thus, the consistency of those algorithms is not guaranteed unless the objective function contains only a stationary point.

Following the idea proposed in [2] for the case $k = 1$, here we propose to restrict the search of the centers set $H^k_{P_n\Phi,\alpha}$ to the family of all possible subsets of the sample with cardinal k. Thus, it is only required to make the search in a family of $\binom{n}{k}$ possible candidates independently of the dimension of E. This idea, in fact, reduces to replace in equation (3) "E" in the second infimum by $\{X_1, ..., X_n\}$, i.e. to minimize

$$\widehat{I^k_{\alpha,\Phi}}(P_n) := \inf_{\tau\in\mathcal{P}_{\alpha,n}} \inf_{h_1,...,h_k\in\{X_1,...,X_n\}} g^k(\tau, h_1, ..., h_k, P_n). \tag{4}$$

As it will be stated in the next section (Theorem 3.2), there always exists a set $\widehat{H}^k_{P_n,\Phi,\alpha}$ in which the restricted optimum is reached since essentially, we are restricting our search to a finite set of candidates (see Proposition 3.1).

On the other hand, in Theorem 3.4, we show that, if H_P^k is unique, under mild conditions on the support of P, then $\widehat{H}^k_{P_n,\Phi,\alpha}$ is a strongly consistent estimate of H_P^k. Although both estimates $\widehat{H}^k_{P_n,\Phi,\alpha}$ and $H^k_{P_n,\Phi,\alpha}$ are consistent, we may have a loss of efficiency when using $\widehat{H}^k_{P_n,\Phi,\alpha}$ instead of $H^k_{P_n,\Phi,\alpha}$. At this point we have not been able to find the asymptotic distribution of both estimates.

An important advantage of the restricted minimum is its low computational time. In Section 4, once the distances matrix is computed, the required time to obtain the set $\widehat{H}^k_{P_n,\Phi,\alpha}$ is around .4 seconds if $k = 2$ and around 15 seconds if $k = 3$ (running times obtained with MatLab on a PowerPC G5 at 1.8GHz). This speed suggest the possibility to use $\widehat{H}^k_{P_n,\Phi,\alpha}$ as a starting point in the search of the infimum $H^k_{P_n,\Phi,\alpha}$ in (3).

In the rest of the paper the function Φ and the value of $\alpha \in [0, 1)$ will remain fixed and, often, will be omitted in the notation. In particular, H_P^k and $H_{P_n}^k$ will stand for $H^k_{P,\Phi,\alpha}$ and $H^k_{P_n,\Phi,\alpha}$ respectively.

3 Existence and Asymptotic Results.

The existence and consistency proofs are given in the Appendix. They follow the corresponding ones for the case $k = 1$ and E a Hilbert space which were given in [2].

An important property of the optimal trimming functions is that they are, essentially, a union of k balls with the same radius.

Proposition 3.1 *Let P be a Borel probability on E and let $G = \{g_1, ..., g_m\} \subset E$. Let us denote*

$$r_P(G) := \inf\{r > 0 : P[\cup_{i\leq m} B(g_i, r)] \geq 1 - \alpha\}.$$

Let $\tau_G \in \mathcal{P}_\alpha$ such that

$$\int \tau_G(y)P(dy) = 1 - \alpha \text{ and } I_{\cup_{i\leq m} B(g_i, r_P(G))} \leq \tau_G \leq I_{\cup_{i\leq m} \overline{B}(g_i, r_P(G))}. \qquad (5)$$

Then, we have that for every $\tau \in \mathcal{P}_\alpha$,

$$\int \inf_{i=1,...,m} \Phi[\|x - g_i\|]\tau_G(x)dP(x) \leq \int \inf_{i=1,...,m} \Phi[\|x - g_i\|]\tau(x)dP(x).$$

Given $G = \{g_1, ..., g_m\}$ we will denote by $\tau_{G,P}$ to any function in $\mathcal{P}_\alpha$ which

satisfies (5). Obviously, $\tau_{G,P}$ always exists. We also denote

$$D(G,P) := \int \inf_{i=1,\ldots,m} \Phi[\|x - g_i\|]\tau_{G,P}(x)dP(x).$$

Therefore, according to the definition of the set $\widehat{H}^k_{P_n}$ given in the paragraph after (4), we have that $\widehat{H}^k_{P_n} \subset \{X_1, ..., X_n\}$ and

$$D(\widehat{H}^k_{P_n}, P_n) = \inf_{\{h_1,\ldots,h_k\}\subset\{X_1,\ldots,X_n\}} D(\{h_1, ..., h_k\}, P_n). \tag{6}$$

Theorem 3.2 (Existence of optimal trimming functions and sets) *Let P be a Borel probability on E. Given $\alpha \in [0,1)$ and $k \geq 1$, there exists $H_P = \{h^P_1, ..., h^P_k\} \subset E$ such that*

$$I^k(P) = \int \Phi\left[\inf_{i=1,\ldots,k} \left\|x - h^P_i\right\|\right] \tau_{H_P,P}(x)P(dx). \tag{7}$$

Consistency results for $\{H^k_{P_n}\}$ and $\left\{\widehat{H}^k_{P_n}\right\}_n$ are given in the following two Theorems.

Theorem 3.3 (Consistency of ITkM-estimates) *Let $\alpha \in (0,1)$, $k \geq 1$ and let P be a Borel probability measure on E such that the set H^k_P is unique. Then the sequence $\{H^k_{P_n}\}$ converges (in norm) to H^k_P μ-a.s.*

Theorem 3.4 (Consistency of approximate ITkM-estimates) *Let $\alpha \in (0,1)$, $k \geq 1$ and let P be a probability on E such that the set H^k_P is unique and is contained in the support of P.*

Then, the sequence $\left\{\widehat{H}^k_{P_n}\right\}_n$ converges almost surely to H^k_P.

Remark 3.5 It is easy to find counterexamples to Theorem 3.4 if the hypothesis that H^k_P is contained in the support of P is removed.

Remark 3.6 The uniqueness assumption we handle in this paper is not always guaranteed even if $\alpha = 0$. This assumption has been discussed very often in the k-means literature (see, for instance, [2]). Some results are known in the case that $E = I\!R$ and $\alpha = 0$ (see, for instance, [13], [8], [5], [14], and [9]), but, at our best acknowledge, yet there is no satisfactory result even in the case $E = I\!R^2$, $k = 2$ and $\alpha = 0$.

However, if this hypothesis is suppressed, the proofs of all consistency results, with obvious modifications, work to show that with μ-probability one, the sequence $\left\{\widehat{H}^k_{P_n}\right\}_n$ is sequentially compact and its adherence values are trimmed k-means of P.

Finally, the following result shows that our method also can be used to estimate $I^k(P)$. Observe that the uniqueness assumption is not required on it.

Corollary 3.7 *Let $\alpha \in (0,1)$, $k \geq 1$ and let P be a Borel probability on E such that the set H_P^k is contained in the support of P. Then we have that,*

$$\lim_n D\left[\widehat{H}_{P_n}^k, P_n\right] = I^k(P),\ \mu - a.s.$$

4 A real data example: looking for patterns of electric power consumers

The study was oriented to find patterns in the behavior of the electric power home-consumers at Buenos Aires, Argentina in 2001. For each individual home in the sample, measurements were taken at each of the 96 sub–intervals of 15 minutes in every week day, Monday to Friday, during January 2001. The analyzed data were monthly averages over week days for each individual home. Thus, every data is a vector of dimension 96.

We have taken a sample of 111 individuals. Since we were only interested in the shape of the curves, the data were normalized in such a way that the maximum of each curve was equal to one and the minimum equal to zero.

A first visual inspection of the data shows that there is an important group of outliers that have no typical pattern behavior (see, as an illustration, the two lower rows in Figure 1). We have performed the method described in the previous section in the following way.

Concerning the mathematical framework, we have considered that our sample is composed by square integrable real functions defined on the interval $[0, 24]$ with the L^2-distance. Thus, given the function x in this space,

$$\|x\| = \left[\frac{1}{24}\int_0^{24} x^2(t)dt\right]^{1/2}.$$

Moreover, we take advantage of the properties of the least squares method by taking $\Phi(t) = t^2$.

Let us begin by computing the trimmed 2-means with a trimming level $\alpha = 13/111$ which allows to trim exactly 13 functions. Several computations (to be described below) suggested us that the anomalous observations (with very atypical pattern behaviour) in the sample are around 9. For this reason we

have chosen the value of $\alpha = 13/111$. The resulting trimmed 2-mean functions as well as four trimmed functions are shown in Figure 1.

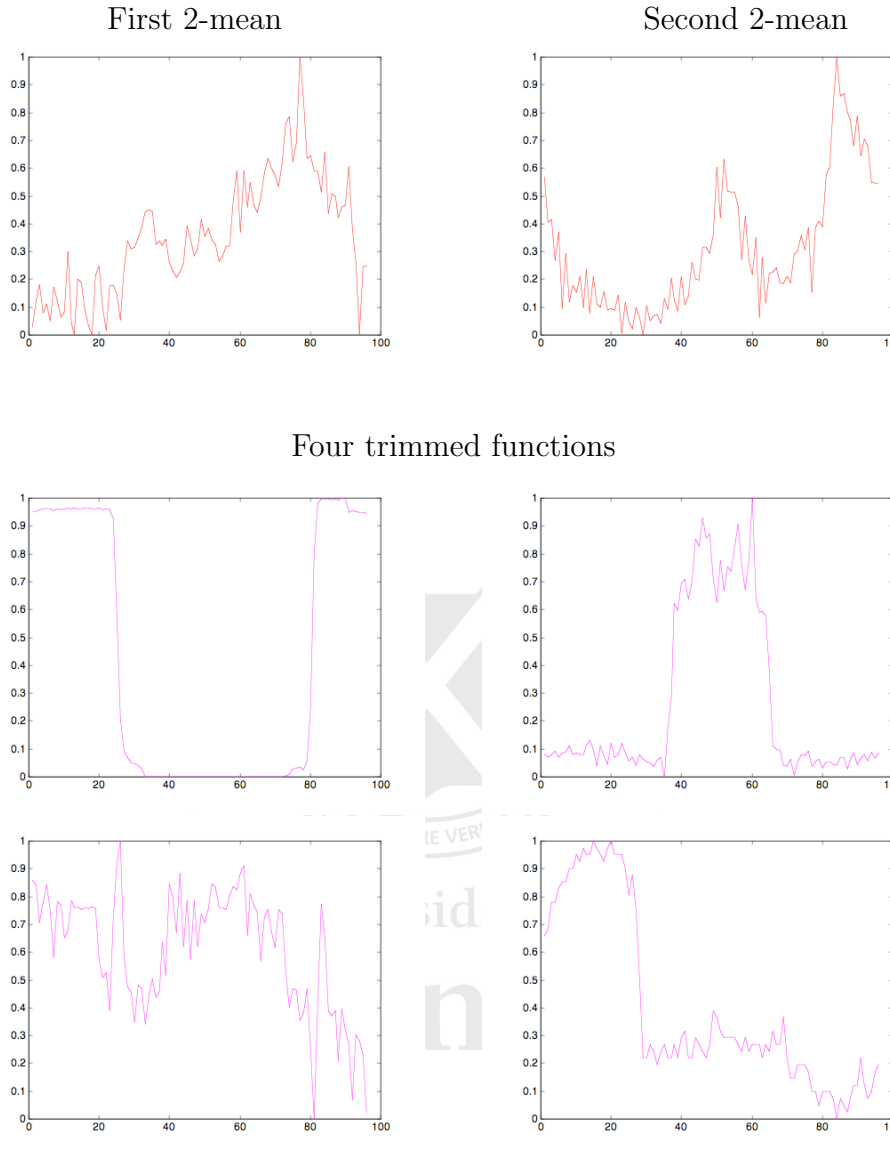

Fig. 1. Result to apply the 2-means procedure with $\alpha = 13/111$ and the L_2-distance to the electrical consumption data set. In the figure are shown the α-2-means of the data (first row) and a sample of 4 from the 13 trimmed data in the parent sample.

The function in the graph labeled "First 2-mean" has a pike around 8pm while the "Second 2-mean" function has two pikes: one around noon and second one around 10pm. Thus, we can consider that the first function is a representative of the families which are mainly outside during day time, while the second function represents the group of families that have some activity at home during day time (probably having lunch at home). It is quite reasonable to assume that, from the point of view of the pattern of the electrical power consumption, this is a key division.

Let us have a look to the four trimmed functions represented in Figure 1. They were chosen at random from the set of 13 trimmed functions, except for the fact that in the initial sample there was another function quite similar to the one in the right in the first row, and then, it was replaced by the one on the lower right. All of them exhibit patterns quite difficult to be explained from a home-consumer point of view. In particular, the one in the upper left is more appropriate for a night business, the one in the upper right could correspond to a day-business. The ones in the lower row may correspond to closed apartments but with some electrical equipment running (like for instance a refrigerator). Then, after normalization, small (in absolute value) differences are increased showing a kind of crazy pattern.

Once the centers of the groups have been chosen, we have to assign each non-trimmed function to the closest center. With this criteria we obtain two clusters, first one composed of 37 elements and second one of 61.

In spite of the fact that previous explanation looks quite plausible, we need to get some confidence on that this is the right explanation. In particular, we need to have some hint to be sure that there are just two groups in the data and to know (approximately) how many anomalous observations contains our data set.

In order to try to solve these questions, first we have selected at random some members of each cluster and, in Figure 2, we have represented them jointly with the center of the group in which they have been included.

All the chosen curves have one or two picks, according to the group in which they have been included except for the curve represented in the upper left graph. However, as shown in Figure 3, this curve is much more similar to the center in the first group than to the center of the second group. The problem with this home, is possibly, that the time-schedule does not match with those of the two-picks group (having each of both picks earlier).

However, the most sound reason to accept that this data set contains two and only two groups appeared when we computed the 3-means of this data set with trimming sizes $\alpha = 3/111, 4/111, ..., 15/111$. This set offers 13 different possibilities to trim, and in consequence, 13 different possibilities for the 3-means. But, we obtained just two different 3-means sets (see Figure 4). The first 3-means set appeared for trimming sizes $\alpha = 3/111, ..., 9/111$. The second one appeared for trimming sizes $\alpha = 10/111, ..., 15/111$.

Functions in the left hand side in Figure 4, corresponds to the first 3-means set. Again, the first curve seems to represent the group of consumers with just a pick (upper graph), the second one to the group with two picks (middle) while the third one to a group of anomalous curves (because this function has just a pick around noon). Observe that the first two curves (centers) are

Fig. 2. Three randomly chosen elements in each cluster (green curves) represented jointly with the curve which the trimmed 2-means procedure chosen as center of the cluster (magenta curves). All curves share the main characteristic of the corresponding center (one or two modes) except for the one represented in the upper left graph.

exactly the same curves that we got when we search for the 2-means centers. On the other hand, in spite of the fact the functions in the right hand side represent not-so-strange patterns, it is not evident for us whether they have one or two picks.

Our explanation for this fact is that, while it is allowed to trim up to 9 functions, and we try to find three groups in the data set, we find the same two groups as in the 2-means case plus a group of anomalous observations. However, if we trim 10 or more functions, no more groups of anomalous functions are left. However, since we look for 3-means, it is compulsory to construct exactly three groups and now the non-anomalous families are split in three

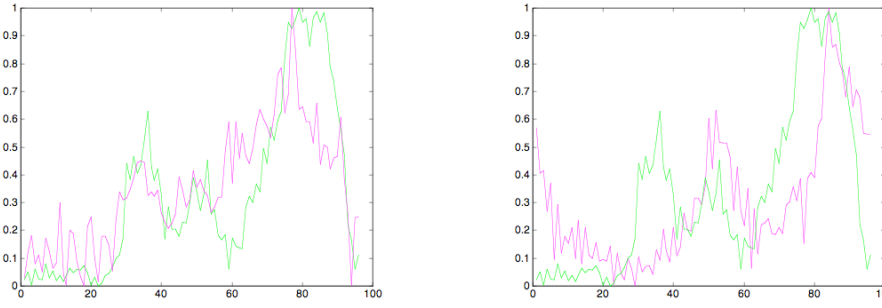

Fig. 3. The element represented in the upper left graph in Figure 2, appears here (green curve in both graphics) jointly with the centers of both two clusters (magenta curves).

Fig. 4. 3-means of the electric consumption data when it is allowed to trim 3 to 15 functions takes only two values. First 3-means are shown in the graphs in the left hand side and corresponds to trim from 3 to 9 functions. The 3-means, when the number of trimmed functions lies between 10 and 15, are drawn in the graphs in the right hand side.

(more or less arbitrary) groups. This fact suggests that the number of the anomalous observations in the sample is 9.

An additional reason to argue for this interpretation of the data is that the 2-means curves that we obtain if we compute them with trimming sizes varying in the set $\{3/111, 4/111, ..., 15/111\}$, are always the same except for the case $\alpha = 6/111$ in which the one pick function is replaced by another (more noisy) function which mainly exhibits just a pick at approximately the same hour (see Figure 5).

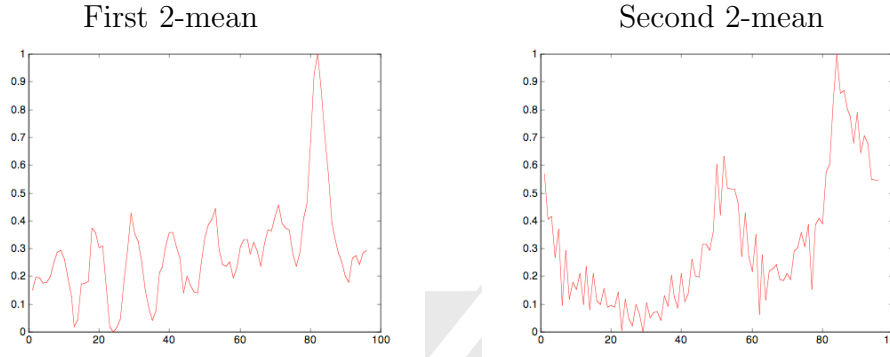

Fig. 5. 2-means of the electric consumption data when it is allowed to trim exactly 6 functions.

5 Appendix. Sketch of the Proofs

PROOF OF PROPOSITION 3.1.- It is similar to that of Proposition 3.1 in [2], and we omit it here. •

PROOF OF THEOREM 3.2.- Let us first show that if $H_n = \{h_1^n, ..., h_k^n\} \subset E$, $n \in I\!N$ satisfy that

$$I^k(P) = \lim_n D(H_n, P), \tag{8}$$

then, the set

$$I := \{i : \liminf_n \|h_i^n\| < \infty\}$$

is not empty.

Let us assume that, on the contrary $I = \emptyset$. Obviously there exists $R > 0$ such

that $P[B(0,R)] \geq 1-\alpha$ and, in consequence

$$I^k(P) \leq \Phi(R)(1-\alpha). \tag{9}$$

On the other hand, we have that

$$\limsup P\left[\cup_{i=1}^k B(h_i^n, 2R)\right] = 0$$

and, in consequence, from an index onward $\cup_{i=1}^k B(h_i^n, 2R) \subset \cup_{i=1}^k B(h_i^n, r_P(H_n))$. From here

$$\begin{aligned}\lim_n D(H_n, P) &\geq \lim_n \Phi(2R)\left(1-\alpha-P\left[\cup_{i=1}^k B(h_i^n, 2R)\right]\right)\\ &= \Phi(2R)(1-\alpha) > \Phi(R)(1-\alpha),\end{aligned}$$

since Φ is strictly increasing. But this, contradicts (8) and (9). In consequence, we have that $I \neq \emptyset$.

Therefore, there exist $I^* \supset I$, $H = \{h_1, ..., h_{k^*}\} \subset E$ with the same cardinal as I^* and a subsequence $\{H_{n_j}\}$ which converges weakly to H. Now, the reasoning in the first part of the proof of Theorem 3.6 in [2] allows to conclude that we can take the sequence in such a way that there exists $r = \lim_j r_P(H_{n_j})$.

Given $n \in I\!N$, let us consider the disjoint sets

$$\begin{aligned}A_i^n := \Big\{ x \in E :\ &\|x - h_i^n\| < \inf_{j=1,...,i-1} \|x - h_j^n\| \\ &\text{and } \|x - h_i^n\| \leq \inf_{j=i+1,...,k^*} \|x - h_j^n\| \Big\},\ \text{for } i = 1, ..., k.\end{aligned}$$

Let us denote $\overline{C}_{i,j} := \overline{B}(h_i^{n_j}, r_P(H_{n_j}))$. Taking into account that

$$\lim_j \int_{\cup_{i\in I^*} \overline{C}_{i,j}} \tau_{H_{n_j}}(x) P(dx) = \alpha,$$

there exists a non-empty set $J \subset I^*$ and a new subsequence such that

$$\begin{aligned}&\text{if } i \in J \text{ then } \lim_m \int_{A_i^{n_m}} \tau_{H_{n_m}}(x) P(dx) = p_j > 0 \qquad (10)\\ &\text{if } j \notin J \text{ then } \lim_m \int_{A_i^{n_m}} \tau_{H_{n_m}}(x) P(dx) = 0.\end{aligned}$$

Without loss of generality, we can assume that the initial sequence coincides with the last subsequence we have obtained.

From the boundness of the sequence $\{r_P(H_n)\}$, it follows that

$$I^k(P) = \lim_n D(H_n, P) = \lim_n D\left(\{h_j^n : j \in J\}, P\right).$$

If for every $j \in J$, $\lim_n \|h_j^n - h_j\| = 0$, then we can repeat the reasoning given for the corresponding case in Theorem 3.5 in [2] to obtain that

$$I^k(P) = D(H, P)$$

(notice that, if $m := \#H < k$ we only need to add $k - m$ arbitrary points to H to get the set we are looking for).

Thus, if we show that for every $j \in J$, $\lim_n \|h_j^n - h_j\| = 0$, the result will be proved. Let us assume that, on the contrary, there exists $j \in J$ such that $\limsup_n \|h_j^n - h_j\| > 0$.

Since (10) holds, we can assume that there exists $p > 0$ such that

$$P\left[B\left(h_j^n, r_P(H_n)\right)\right] > p, \text{ for every } n \in I\!N,\ j \in J.$$

Now, let $\delta > 0$, $n \in I\!N$ and $i \in \{1, ..., k^*\}$, and define the sets

$$B_\delta^i := \{y : \liminf \Phi\left[||y - h_i^n||\right] > \Phi\left[||y - h_i||\right] + \delta\}$$
$$B_\delta^{n,i} := \{y : \Phi\left[||y - h_i^k||\right] > \Phi\left[||y - h_i||\right] + \delta, \text{ for every } k \geq n\}.$$

We can repeat the last part of the argument in the proof of Theorem 3.6 in [2] just taking the value δ_0 as a positive value such that $P[B_{\delta_0}^i] > 1 - p + \eta_0$, for $i = 1, ..., k^*$ where $\eta_0 > 0$, $\delta < \delta_0$ such that $P[B_\delta^i] > 1 - \epsilon$ for $i = 1, ..., k^*$ and

$$R = \sup_{i=1,...,k^*} \Phi\left[2 \sup_n ||h_i^n|| + \sup r_\alpha(H_n)\right],$$

to get that

$$\lim \int \left(\inf_{i=1,...,k^*} \Phi\left[||y - h_i^n||\right] - \inf_{i=1,...,k^*} \Phi\left[||y - h_i||\right]\right) \tau_n(y) dP(y)$$

$$
\begin{aligned}
&\geq \sum_{i=1,\ldots,k^*} \lim \int_{A_i^n} \left(\Phi\left[||y-h_i^n||\right] - \Phi\left[||y-h_i||\right]\right) \tau_n(y) dP(y) \\
&= \sum_{i=1,\ldots,k^*} \lim \left[\int_{A_i^n \cap B_{\delta_0}^{n,i}} \left(||y-x_n||^2 - ||y-x_0||^2\right) \tau_n(y) dP(y) \right. \\
&\qquad + \int_{A_i^n \cap \left(B_{\delta_0} n,i\right)^c \bigcap B_\delta n,i} \left(||y-x_n||^2 - ||y-x_0||^2\right) \tau_n(y) dP(y) \\
&\qquad \left. + \int_{A_i^n \cap \left(B_{\delta_0} n,i\right)^c \bigcap (B_\delta n,i)^c} \left(||y-x_n||^2 - ||y-x_0||^2\right) \tau_n(y) dP(y) \right] \\
&\geq \lim_n \left[\delta_0 \eta_0 - R\epsilon\right],
\end{aligned}
$$

and taking limits on ϵ, we finally have that

$$
\begin{aligned}
I^k(P) &= \lim \int \inf_{i=1,\ldots,k^*} \Phi\left[||y-h_i^n||\right] \tau_n(y) dP(y) \\
&\geq \delta_0 \eta_0 + \lim \int \inf_{i=1,\ldots,k^*} \Phi\left[||y-h_i||\right] \tau_n(y) dP(y) \\
&\geq \delta_0 \eta_0 + \int \inf_{i=1,\ldots,k^*} \Phi\left[||y-h_i||\right] \tau_{x_0}(y) dP(y),
\end{aligned}
$$

what contradicts the definition of $I^k(P)$. •

The consistency result (given in Theorem 3.4) is based on the following result.

Proposition 5.1 *It happens that* $\limsup_n I^k(P_n) \leq I^k(P)$, $\mu - a.s.$

PROOF.- It is similar to that of Proposition 7.4 in [2], and we omit it here.•

PROOF OF THEOREM 3.3.- Let us denote $H_{P_n}^k = \{h_1^n, ..., h_k^n\}$. Following the argument in the first part of Theorem 3.6 in [2] it is possible to prove that μ-a.s. the set $\cup_n H_{P_n}^k$ and the sequence $\{r_{P_n}(H_n)\}$ are bounded. Let us fix a point ω_0 in the set in which those facts plus the Glivenko-Cantelli Theorem are satisfied.

If we prove that for this ω_0, every subsequence $\left\{H_{P_{i_n}}^k\right\}$ of $\left\{H_{P_n}^k\right\}$ admits a new subsequence $\left\{H_{P_{i_n'}}^k\right\}$ which converges in norm to $\left\{H_P^k\right\}$ the result will be proved.

From a similar argument to that used in the proof of Theorem 3.2 we obtain that every subsequence $\left\{H^k_{P_{i_n}}\right\}$ contains a further subsequence $\left\{H^k_{P'_{i_n}}\right\}$ which satisfies that there exist $J \subset \{1,...,k\}$, $H = \{h_j, j \in J\} \subset E$ and $p > 0$ such that if we define the sets A^n_i similarly as in the proof of Theorem 3.2 then

(1) If $j \in J$, then the sequence $\left\{h^{i'_n}_j\right\}$ converges weakly to h_j and

$$\inf_n P_n\left[A^n_j\right] > p.$$

(2) If $j \notin J$, then $\lim_n \left\|h^{i'_n}_j\right\| = \infty$ or, else, $\lim_n P_n\left[A^n_i\right] = 0$.

From here it is possible to repeat the reasoning in the proof of Theorem 3.7 in [2] to get the desired result. •

The consistency of the approximate ITkM $\widehat{H}^k_{P_n}$ follows from the following Proposition.

Proposition 5.2 *Let us assume that the hypotheses in Theorem 3.3 hold. Let $G_n = \{g^n_1, ..., g^n_k\}$, $n \in I\!N$, be a sequence of sets with cardinal k which converges (in norm) to the set H^k_P. Then, we have that*

$$\limsup D(G_n, P_n) \leq I^k(P),\ \mu - a.s.$$

PROOF.- Let us denote the trimmed k-M estimates by $H_n = \{h^n_1, ..., h^n_k\}$ $n \in I\!N$. Notice that the set of real numbers

$$I\!H := \left\{\|y - g^n_i\|, \|y - h^n_i\| : y \in \cup^k_{i=1}\overline{B}\left[h^n_i, r_{P_n}(H_n)\right],\ n \in I\!N\right\}$$

is bounded. Therefore, the map Φ is uniformly continuous on $I\!H$. By Theorem 3.3, we have that (possibly after a re–labeling)

$$\lim \sup_{i=1,...,k} \|g^n_i - h^n_i\| = 0,\ \mu - \text{a.s.}$$

Let us define the family of sets $\{A^n_i,\ i = 1, ..., k, n \in I\!N\}$ like in Theorem 3.2. We obtain that

$$\begin{aligned} D(G_n, P_n) &\leq \int \Phi[\inf_{i=1,...,k} \|y - g^n_i\|]\tau_{H_n,P_n}(y)P_n(dy) \\ &\leq I^k(P_n) + \sum_{i=1}^k \int_{A^n_i} \left(\Phi[\|y - g^n_i\|] - \Phi[\|y - h^n_i\|]\right)\tau_{H_n,P_n}(y)P_n(dy) \end{aligned}$$

and the result is proved taking into account Proposition 5.1 and that the second term in the right hand side converges to zero because of the uniform continuity of Φ. •

PROOF OF THEOREM 3.4. It is similar to that of Theorem 5.1 in [2], and we omit it here •

Acknowledgements

We are very grateful to Daniel Fraiman (Universidad de San Andrés), for providing us with the electric power consumer's data. We would also like to thank three referee's for their constructive comments on a first version of this paper.

References

[1] Cuesta-Albertos, J.A. (1984) Medidas de Centralización multidimensionales (ley fuerte de los grandes números), *Trabajos Estadíst. Investigación Oper.* **35**, 1, 1–16.

[2] Cuesta-Albertos, J.A and Fraiman, R. (2005). Impartial trimmed means for functional data. To appear in *Data Depth: Robust Multivariate Statistical Analysis, Computational Geometry and Applications.* Eds. R. Liu, R. Serfling and D. Souvaine. American Mathematical Society in DIMACS Series.

[3] Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.* **25**, 2, 553–576.

[4] 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** , 4 523–534.

[5] Eubank, R.L. (1988) Optimal grouping, spacings, stratification, and piecewise constant approximation. *SIAM Review* **30**, 3 404–420.

[6] García-Escudero, L.A. and Gordaliza, A. and Matrán, C. (2003) Trimming tools in exploratory data analysis. *J. Computat. Graph. Statist.* **12**, 2 434–459.

[7] Gordaliza, A. (1991) Best approximations to random variables based on trimming procedures. *J. Approx. Theory* **64**, 2, 162–180

[8] Kiefer, J.C. (1983) Uniqueness of locally optimal quantizer for long-concave density and convex error weighting function. *IEEE Trans. Inform. Theory* **29**, 1, 42–47.

[9] Li, L. and Flury, B. (1995) Uniqueness of principal points for univariate distributions. *Statist. Probab. Letters* **25**, 323–327.

[10] Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T. and Cohen, K.L. (1999) Robust principal components for functional data. *Test.* **8**, 1–73.

[11] Pollard, D. (1981) Strong consistency of k-means clustering. *Ann. Statist.* **9**, 1, 135–140.

[12] Sverdrup-Thygeson, H. (1981) Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces. *Ann. Statist.* **9**, 1, 141–145.

[13] Trushkin, A. V. (1982) Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Trans. Inform. Theory* **28**, 2, 187–198.

[14] Zoppé, A. (1994) On uniqueness and symmetry of self-consistent point of univariate continuous distributions. *J. Classification* **14**, 147–158.