Impact of contamination on training and test error rates in statistical clustering

C. Ruwet[1] and G. Haesbroeck

University of Liège

Abstract

The k-means algorithm is one of the most common nonhierarchical methods of clustering. It aims to construct clusters in order to minimize the within cluster sum of squared distances. However, as most estimators defined in terms of objective functions depending on global sums of squares, the k-means procedure is not robust with respect to atypical observations in the data. Alternative techniques have thus been introduced in the literature, e.g. the k-medoids method. The k-means and k-medoids methodologies are particular cases of the generalized k-means procedure. In this paper, focus is on the error rate these clustering procedures achieve when one expects the data to be distributed according to a mixture distribution. Two different definitions of the error rate are under consideration, depending on the data at hand. It is shown that contamination may make one of these two error rates decrease even under optimal models. The consequence of this will be emphasized with the comparison of the influence functions of these error rates and some simulations.

Key words and phrases: Clustering analysis, Error rate, Generalized k-means, Influence Function, Principal points, Robustness.

1 Introduction

Cluster analysis is useful when one wants to group together similar objects. Nonhierarchical procedures rely on distances between objects and on the fixed number of groups

[1]Department of Mathematics - Grande Traverse 12, B-4000 Liège, Belgium

cruwet@ulg.ac.be - tel : 04/366.94.06 - fax : 04/366.95.47

which must be chosen beforehand and which will be set to k throughout the text. Starting with an initial partition of the objects into k groups, nonhierarchical cluster algorithms proceed iteratively in order to assign each object to the closest cluster (the closeness being assessed by a given distance between the object and the center of the cluster). The usual output of these algorithms consists of k centers, each of which defines naturally a corresponding cluster as the region enveloping the points closest to that particular center. More about the basic ideas of nonhierarchical methods can be found in most textbooks on multivariate analysis (e.g. Johnson and Wichern, 2007).

As mentioned above, the result of a nonhierarchical clustering method can be given via a set of k points containing the k centers. The most common nonhierarchical clustering technique is probably the k-means algorithm which aims to find k centers in order to minimize the sum of the squared Euclidean distances between the observations assigned to a cluster and the mean of this cluster. However, in this paper, following García-Escudero and Gordaliza (1999), focus is on a generalization of this algorithm: the *generalized k-means* algorithm. The main idea is to replace the operator "mean" by another penalty function $\Omega : \mathbb{R}^+ \to \mathbb{R}^+$ which is assumed to be continuous, nondecreasing and such that $\Omega(0) = 0$ and $\Omega(x) < \lim_{y\to\infty} \Omega(y)$ for all $x \in \mathbb{R}^+$.

As this paper focuses on the study of influence functions, it is necessary to derive the functional forms of the estimators under consideration. As far as the k-means procedure is concerned, the functional form of the centers simply correspond to the principal points introduced by Flury in 1990. In the more general setting where the penalty function Ω is used, the statistical functionals characterizing the centers are given by $T_1(F), \ldots, T_k(F) \in \mathbb{R}^p$ which are solutions of the minimization problem

$$\{T_1(F), \ldots, T_k(F)\} = \underset{\{t_1,\ldots,t_k\}\subset\mathbb{R}^p}{\operatorname{argmin}} \int \Omega\left(\inf_{1\le j\le k} \|x - t_j\|\right) dF(x) \tag{1}$$

for those distributions F for which this integral exists. These statistical functionals $T_1(F)$, ..., $T_k(F)$ will be called *generalized principal points*. Taking $\Omega(x) = x^2$ yields the classical principal points of the distribution F while $\Omega(x) = x$ gives socalled k-medoids principal points of F. Existence, consistency and asymptotic normality of these generalized principal

points were treated by Cuesta-Albertos and Matrán (1988) and Pollard (1981 and 1982). García-Escudero and Gordaliza (1999) derived robustness properties of the generalized principal points in the particular case of univariate data to be clustered into two groups. For example, they showed that any Ω function with a bounded derivative yields a bounded influence function for the estimators $T_1(F)$ and $T_2(F)$.

Assuming that $T_1(F), \ldots, T_k(F)$ are the generalized principal points of the distribution F, clusters, denoted as $C_1(F), \ldots, C_k(F)$, can be constructed. The jth cluster consists of the region of points closer to $T_k(F)$ than to any other center, the closeness being assessed by the penalty function Ω. If $\omega(x)$ denotes the gradient of $\Omega(\|x\|)$ (when it exists), the first-order conditions corresponding to the minimization problem (1) are given by

$$\int_{C_i(F)} \omega(x - T_i(F)) dF(x) = 0 \qquad i = 1, \ldots, k \tag{2}$$

showing that the generalized principal points are the ω-means, in the sense of Brøns et al. (1969), of the corresponding clusters. This property will be referred to as property **(P)** in the sequel. For example, if $\Omega(x) = x^2$, $\omega(x) = 2x$ and the first order conditions simply imply that the principal points $T_i(F)$ are the means on the clusters $C_i(F)$ for $i = 1, \ldots, k$. When the gradient of $\Omega(\|x\|)$ does not exist for a finite number of points, the integral in (2) has to be split into a sum of integrals but property **(P)** still holds.

When clusters are constructed, one is often interested in describing the characteristics of the observations they contain. The *clustering rule* corresponding to the generalized k-means problem (1), denoted by R_F^{GkM}, associates any $x \in \mathbb{R}^p$ to a cluster $C_j(F)$ as follows:

$$R_F^{\mathrm{GkM}}(x) = j \Leftrightarrow j = \underset{1 \le i \le k}{\operatorname{argmin}} \, \Omega(\|x - T_i(F)\|). \tag{3}$$

Clustering is often confused with classification. However, there is some difference between the two. In classification, the objects to classify are assigned to pre-defined classes, whereas in clustering the classes are also to be defined. However, when a cluster analysis is performed on a mixture distribution, it is usually referred to as a *statistical cluster analysis* or *model-based clustering* (e.g. Fraley and Raftery, 2002, Gallegos and Ritter, 2005 or Qiu and Tamhane, 2007). This is the context considered in this paper. The underlying distribution F will therefore be assumed to be a mixture distribution of k distributions $F_1, \ldots, F_k$

with prior probabilities $\pi_1(F), \ldots, \pi_k(F)$, i.e. $F = \sum_{i=1}^{k} \pi_i(F) F_i$. In this setting, one hopes to end up with clusters representing the different components of the mixture. In this sense, an error rate might be defined to measure, as in classification, the performance of the clustering. Often, in practice, error rates are assessed on test data while the estimations are derived on training data, yielding the so-called *test set error rate*. In some applications however (e.g. Dučinskas, 1995 or Mansour and McAllester, 2002), *training set error rates* are computed, meaning that the same data are used for deriving the estimations and for measuring the classification performance. It is well known that in the latter case, the error rate is underestimated. The aim of the paper is to analyse the impact of contamination on this training set error rate by means of influence functions and compare the behaviour of the training set error rate with that of the test set error rate.

The paper is organized as follows. Section 2 derives the statistical functionals corresponding to the test and training error rates of the generalized k-means classification procedure and investigates their behaviour graphically. In Section 3, the first order influence function of the training error rate is derived and compared with the influence function of the test error rate computed in Ruwet and Haesbroeck (2010). Finally, Section 4 uses simulations to illustrate the comparison of the two performance measures on finite samples.

2 Training and test error rates

2.1 Statistical functionals

Any classification rule is bound to misclassify some objects. As mentioned in the Introduction, a measure of classification performance is often defined in terms of the error rate which corresponds to the probability of misclassifying observations distributed according to a given probability measure. When both the rule and the performance are derived on the same distribution, a training error rate is obtained while a test error rate is based on two distributions, a training distribution and a test (or model) distribution.

Let us now define the statistical functionals associated to the two types of error rates.

First, note that all the distributions considered in this paper are mixtures, with the implicit assumption that any observation x distributed according to the ith component of the mixture, F_i, belongs to an underlying group G_i.

The test error rate consists of setting up the clustering rule using a training distribution, $F = \sum_{i=1}^{k} \pi_i(F) F_i$, and testing the quality of this rule on a test (or model) distribution, $F_m = \sum_{i=1}^{k} \pi_i(F_m) F_{m,i}$. The corresponding statistical functional can therefore be written as follows:

$$\mathrm{ER}(F, F_m) = \sum_{j=1}^{k} \pi_j(F_m) \mathbb{P}_{F_{m,j}}[R_F^{\mathrm{GkM}}(X) \neq j]. \tag{4}$$

On the other hand, for the training error rate, the same distribution is used to compute and evaluate the rule leading to the following statistical functional:

$$\mathrm{ER}(F, F) = \sum_{j=1}^{k} \pi_j(F) \mathbb{P}_{F_j}[R_F^{\mathrm{GkM}}(X) \neq j]. \tag{5}$$

In the sequel, the test error rate will be denoted by TER while ER will refer to the training error rate. Moreover, as ER depends only on one distribution, it will be simply denoted as $\mathrm{ER}(F)$.

2.2 Optimality

A classification rule is said to be optimal if its error rate reaches the same error rate as the Bayes rule **(BR)** given by:

$$R_F^{\mathrm{BR}}(x) = j \Leftrightarrow \pi_j(F) f_j(x) > \pi_i(F) f_i(x) \, \forall i \neq j$$

where $f_1, \ldots, f_k$ are the densities (whose existence is assumed) corresponding to $F_1 \ldots, F_k$.

In the context of univariate data to be clustered into two groups, Qiu and Tamhane (2007) have proved that the generalized 2-means procedure with $\Omega(x) = x^2$ (classical 2-means method) is optimal under a mixture of two homoscedastic normal distributions with equivalent weight, i.e. under the model

$$\textbf{(N)}\ F_N \equiv 0.5\,\mathrm{N}(\mu_1, \sigma^2) + 0.5\,\mathrm{N}(\mu_2, \sigma^2) \text{ with } \mu_1 < \mu_2.$$

Model **(N)** will be refereed as the *optimal normal mixture*. This optimality result can be easily extended to the 2-medoids method. It is worth mentioning that adaptations of the k-means procedure have been suggested in the literature (see e.g. Symons, 1981) to reach optimality under unbalanced normal mixtures. As we do not wish to restrict the computation of the influence function to the normal setting, these adapted criteria are not used here.

2.3 Error rate under contamination

In ideal circumstances, the training and test distributions are identical and thus the two error rates (4) and (5) coincide. In practice however, data often contain outliers and these will affect the training distribution, while the test distribution may be assumed to remain unchanged. When contamination is present, the training distribution F would be better represented as a contaminated distribution $F_\varepsilon = (1-\varepsilon)F + \varepsilon H$, which corresponds to a proportion $1-\varepsilon$ of data distribution according to the model while the remaining fraction, ε, is contaminated (comes from another distribution). We consider that F_ε is a mixture distribution whose components will be explicitly given later. When working with the test error rate, even if the rule may be corrupted by being estimated on a contaminated training distribution, the error rate is evaluated on the test distribution assumed to be clean and with prior probabilities unaffected by the contamination (in practice, they are usually estimated assuming a prospective sampling scheme). Therefore, under contamination, the statistical functional TER, as defined in (4), becomes

$$\mathrm{TER}(F_\varepsilon, F_m) = \sum_{j=1}^{k} \pi_j(F_m) \mathbb{P}_{F_m,j}[R^{\mathrm{GkM}}_{F_\varepsilon} \neq j], \tag{6}$$

where only the rule shows a dependence on the contaminated distribution F_ε. When working with the training error rate, it is not only the rule which is corrupted but also the whole definition of the error rate, including the prior probabilities, as the following

expression, corresponding to equation (5), shows :

$$\mathrm{ER}(F_\varepsilon) = \sum_{j=1}^{k} \pi_j(F_\varepsilon) \mathbb{P}_{F_{\varepsilon,j}}[R_{F_\varepsilon}^{\mathrm{GkM}} \neq j]. \tag{7}$$

As in García-Escudero and Gordaliza (1999) and Qiu and Tahmane (2007), only univariate data naturally clustered into two groups ($k = 2$) will be considered in detail. In this simple setting, assuming that the penalty function Ω is strictly increasing, the clustering rule (3) leads to the following estimated clusters:

$$C_1(F) = \{x \in \mathbb{R} : |x - T_1(F)| < |x - T_2(F)|\} = \left]-\infty, \frac{T_1(F) + T_2(F)}{2}\right[\tag{8}$$

$$C_2(F) = \left]\frac{T_1(F) + T_2(F)}{2}, \infty\right[\tag{9}$$

assuming, w.l.o.g., that $T_1(F) < T_2(F)$. Classification based on the generalized principal points is therefore quite straightforward. One simply needs to check where an observation lies w.r.t. the threshold $C(F) = (T_1(F) + T_2(F))/2$, which under the optimal model **(N)** takes a close form as shown in Proposition 1 below (the proof is sketched in the Appendix).

Proposition 1 *The cut-off point of any generalized 2-means procedure which is optimal under model* ***(N)*** *is the midpoint of the true means, i.e.* $C(F_N) = \dfrac{\mu_1 + \mu_2}{2}$, *where* μ_1 *and* μ_2 *are the means of* F_1 *and* F_2 *respectively.*

Then, using (8) and (9), explicit expressions for (6) and (7) may be obtained:

$$\mathrm{TER}(F_\varepsilon, F_m) = \pi_1(F_m)\left\{1 - F_{m,1}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_m) F_{m,2}\left(C(F_\varepsilon)\right) \tag{10}$$

and

$$\mathrm{ER}(F_\varepsilon) = \pi_1(F_\varepsilon)\left\{1 - F_{\varepsilon,1}\left(C(F_\varepsilon)\right)\right\} + \pi_2(F_\varepsilon) F_{\varepsilon,2}\left(C(F_\varepsilon)\right). \tag{11}$$

In $\mathrm{ER}(F_\varepsilon)$, the contaminated conditional distributions $F_{\varepsilon,1}$ and $F_{\varepsilon,2}$, as well as the contaminated prior probabilities $\pi_1(F_\varepsilon)$ and $\pi_2(F_\varepsilon)$, are unknown.

A particular case of F_ε consists of taking as function H a Dirac distribution Δ_x having all its mass at the point x. From now on, focus will be on this kind of contamination. In this case, the contaminated prior probabilities are given by

$$\pi_i(F_\varepsilon) = (1 - \varepsilon)\pi_i(F) + \varepsilon\delta_i(x) \tag{12}$$

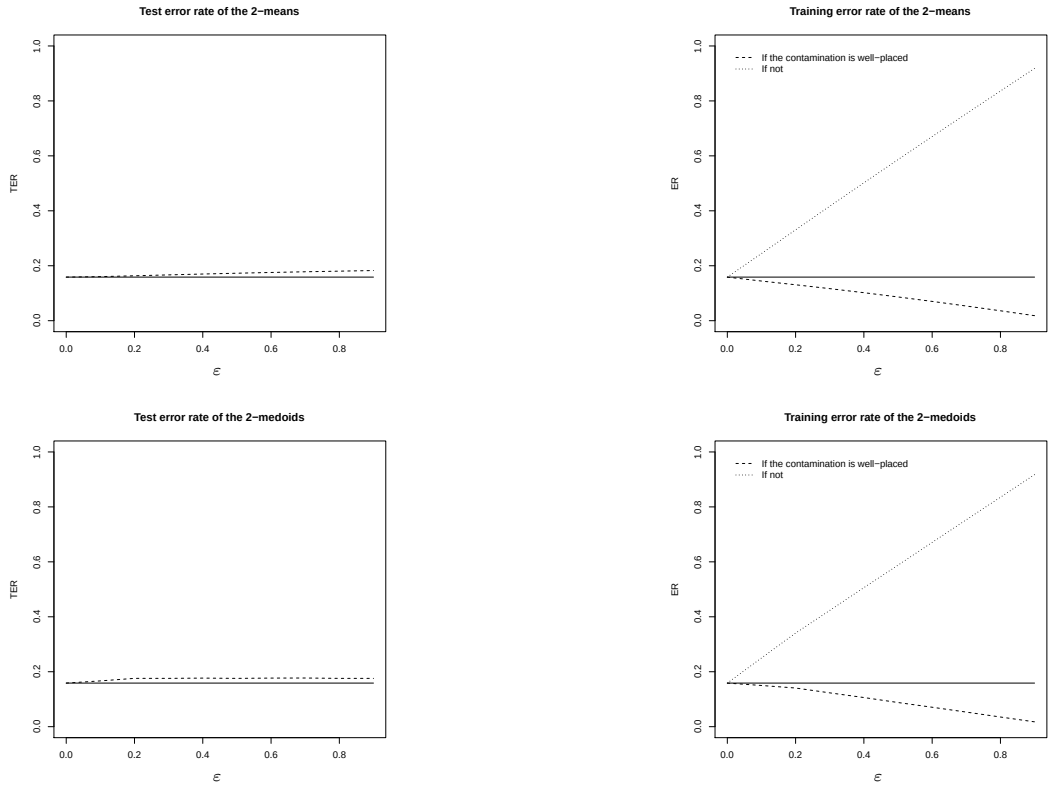

Figure 1: Optimal error rate (solid line) and contaminated error rate (dashed line) of the 2-means method (first row) and 2-medoids method (second row) when using TER (left panels) and ER (right panels) as a function of the mass of contamination ε under the optimal model **(N)** with $\mu_1 = -1$, $\mu_2 = 1$ and $\sigma = 1$. The contaminating observation is set at $x = -0.5$.

where $\delta_i(x) = \mathrm{I}\{x \in G_i\}$, which equals 1 if $x \in G_i$ and 0 otherwise. Indeed, since clean data arise with a probability $1 - \varepsilon$, the probability of getting clean data from the ith group is $(1 - \varepsilon)\pi_i(F)$. A mass ε is then added to the contaminated group from which the outlier comes. Also, following Croux, Filzmoser and Joossens (2008), the contaminated conditional distributions take the form

$$F_{\varepsilon,i} = (1 - \lambda_{i,x}(\varepsilon))F_i + \lambda_{i,x}(\varepsilon)\Delta_x \text{ with } \lambda_{i,x}(\varepsilon) = \frac{\varepsilon\delta_i(x)}{\pi_i(F_\varepsilon)} \tag{13}$$

yielding the following natural decomposition of F_ε: $F_\varepsilon = \pi_1(F_\varepsilon)F_{\varepsilon,1} + \pi_2(F_\varepsilon)F_{\varepsilon,2}$.

Plugging (12) and (13) in (11) yields explicit expressions of the contaminated training

error rate. Figures 1 and 2 illustrate the behavior of this error rate together with that of the test error rate (10), behavior w.r.t. the mass of contamination or w.r.t the position of the contamination under model **(N)** with $\mu_1 = -1$, $\mu_2 = 1$ and $\sigma = 1$. More precisely, Figure 1 represents (10) and (11) for x fixed to -0.5 while Figure 2 represents these quantities for ε fixed to 0.1. In both figures, the plots on the first row correspond to the 2-means clustering method while the second row is based on the 2-medoids procedure. Also, the first columns describe the behavior of the test error rate while the second columns concern the training one. In all plots, the solid line gives the optimal ER (equal to $\Phi(-\Delta/2)$ where Φ stands here for the standard normal cdf and Δ is the distance between the true means).

One clearly sees that contamination can only increase the error rate as soon as the test error rate is used. The behavior is quite different for the training error rate where contamination makes the error rate decrease as soon as it is well classified (this is clearly illustrated in Figure 2). Indeed, by Proposition 1, the cut-off between the two clusters is at 0. Thus, when $x < 0$, x is in $C_1(F_N)$ and is well classified. In this case, the training error rate is smaller than the optimal error rate. As soon as x becomes positive, i.e. turns out to be badly classified, the error rate increases and gets above the optimal one. When there is no contamination ($\varepsilon = 0$), the error rate of the k-medoids procedure attains the minimal error rate illustrating its optimality under F_N. It is also interesting to note from Figure 2 that the contaminated test error rate is the closest to the optimal error rate when x is closest to the generalized principal points, $T_1(F_N)$ and $T_2(F_N)$ and that the contaminated test error rate is symmetric with respect to the vertical axis $x = 0$. This means that contamination that is well or badly classified has the same impact of that error rate. For the training error rate, local minima are observed in the neighborhood of $T_1(F_N)$ and $T_2(F_N)$, but symmetry does not hold anymore. In all cases, the impact of contamination is more important on the training error rate than on test one.

As mentioned before, property **(P)** is dependent on the training distribution F. If $F = F_m$, the property holds on the test data. However, under contamination, $F = F_\varepsilon$, or for finite samples, $F = F_n$, and property **(P)** does not hold for F_m. Intuitively, in cluster analysis, one would use the training error rate since only in that case the generalized

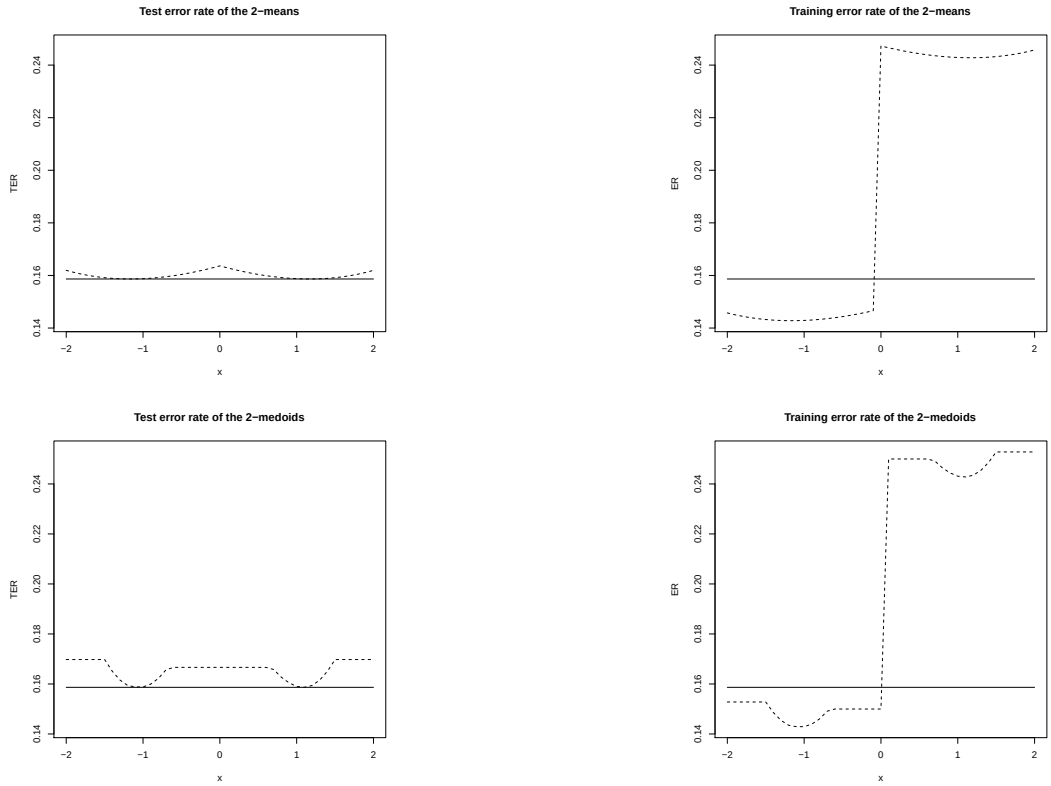

Figure 2: Optimal error rate (solid line) and contaminated error rate (dashed line) of the 2-means method (first row) and 2-medoids method (second row) when using TER (left panels) and ER (right panels) as a function of the position of the contaminating mass under model **(N)** with $\mu_1 = -1$, $\mu_2 = 1$ and $\sigma = 1$. The percentage of contamination is set at 10% and $x \in G_1$.

principal points fully characterize the observations in the clusters. However, it has the drawback that contamination may improve the classification performance w.r.t. the one achieved under optimal models.

3 Influence function of ER

Roughly speaking, influence functions (Hampel et al, 1986) measure the influence that an infinitesimal contamination placed on an arbitrary point has on an estimator of interest.

In Ruwet and Haesbroeck (2010), the influence function of TER is derived and is shown to vanish under the optimal model F_N. Under non optimal settings, it was observed that the influence function of TER is bounded as soon as the function ω is bounded. Moreover, as the cut-off point tends to move towards the center of the biggest group, outliers have a bigger influence when they are located in the smallest group. Also the closer the two groups are, the bigger the influence of contamination on TER is.

Focus here is on the derivation of the influence function of the training error rate which is more complicated because of the presence of ε everywhere in the expression (7). The influence function of the statistical functional ER at the model F is defined, for those distributions for which the derivative makes sense, by

$$\mathrm{IF}(x; \mathrm{ER}, F) = \lim_{\varepsilon \to 0} \frac{\mathrm{ER}(F_\varepsilon) - \mathrm{ER}(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} \mathrm{ER}(F_\varepsilon) \Big|_{\varepsilon=0}$$

where $F_\varepsilon = (1-\varepsilon)F + \varepsilon \Delta_x$ and Δ_x is the Dirac distribution having all its mass at the point x.

As the statistical functional ER is a smooth function of the functionals T_1 and T_2, its influence function exists (García-Escudero and Gordaliza, 1999) and can be derived in a straightforward way. The result is given in the following proposition (the proof is given in the Appendix).

Proposition 2 *For any mixture distribution F with prior probabilities π_1 and π_2 and conditional densities f_1 and f_2, the influence function of the training error rate of the generalized 2-means method with loss-function Ω is given by*

$$\mathrm{IF}(x; \mathrm{ER}, F) = \frac{1}{2}(\mathrm{IF}(x; T_1, F) + \mathrm{IF}(x; T_2, F))\{\pi_2(F) f_2(C(F)) - \pi_1(F) f_1(C(F))\} + \mathrm{I}\{x \leq C(F)\}(1 - 2\delta_1(x)) + \delta_1(x) - \mathrm{ER}(F) \quad (14)$$

for all $x \neq C(F)$ where $C(F) = (T_1(F) + T_2(F))/2$ and $\delta_1(x) = \mathrm{I}\{x \in G_1\}$.

The influence function (14) depends on the influence functions of $T_1(F)$ and $T_2(F)$. These were computed for strictly increasing Ω functions by García-Escudero and Gordaliza

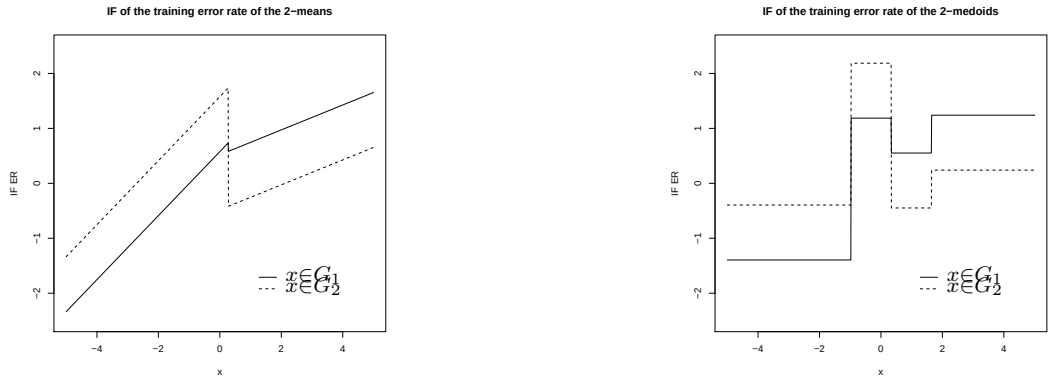

Figure 3: Influence functions of the training error rate of the 2-means (left panel) and 2-medoids (right panel) methods when the model distribution is the mixture $0.2\,\mathrm{N}(-1.5, 1) + (1 - 0.2)\,\mathrm{N}(1.5, 1)$, with a contaminating mass x assumed to come from group G_1 (solid line) or group G_2 (dashed line).

(1999). It also depends on the belonging group and on the relative position of x w.r.t. the cut-off point.

Figures 3 to 5 illustrate the behavior of the influence function (14) under some variations of the characteristics of the two sub-populations. A mixture of normal distributions (equivalent to model **(N)** but with the relaxation of the constraint $\pi_1 = \pi_2$) is selected for the representations. More specifically, $F = \pi_1\,\mathrm{N}(-\mu, \sigma^2) + \pi_2\,\mathrm{N}(\mu, \sigma^2)$, for given values of π_1, π_2 (with $\pi_1 + \pi_2 = 1$), μ and σ. Of interest is the impact on the influence function of the group of the mass x, of the prior probabilities, of the distance between the true centers $\Delta = 2\mu$ and of the (common) dispersion within the groups, σ.

As a first general comment, one can say that, as expected, ER based on the 2-means procedure has an unbounded influence function while that of the 2-medoids is bounded. The impact of infinitesimal contamination is thus less harmful on this last procedure. Moreover, the influence functions show some discontinuities: one at the cut-off $C(F)$ for both Ω functions and two additional ones at $T_1(F)$ and $T_2(F)$ for the 2-medoids procedure. These jumps come from the discontinuities already observed in the influence functions of the generalized principal points T_1 and T_2.

Now, more specifically, Figure 3 shows the effect of the group to which the contamination is assigned. One can see that a misclassified x, i.e. an x value on the right of $C(F)$ while belonging to G_1 or an x value on the left of the cut-off but belonging to G_2, results in a bigger influence than a correctly classified x. Indeed, the indicator functions in (14) add a value 1 to the influence function when the contamination is badly classified (the belonging group has no impact on $\mathrm{IF}(x; T_1, F)$ and $\mathrm{IF}(x; T_2, F)$). The sign of the influence is important to take into account here. Indeed, when the influence function gets negative for a given x, the Taylor expansion

$$\mathrm{ER}(F_\varepsilon) \approx \mathrm{ER}(F) + \varepsilon \mathrm{IF}(x; \mathrm{ER}, F), \text{ for all } \varepsilon \text{ small enough}$$

implies that ER is smaller when taking the contamination into account than under the model. This really illustrates the sensitivity of the procedure because badly classified observations with big negative values improve ER. The 2-medoids procedure shows a behavior which seems natural for the smallest group: well classified observations have a decreasing effect on the error rate while badly classified contamination increases the error rate. This behavior is also observable for the biggest group when the group masses are not too different ($\pi_1 \geq 0.3$). One distinguishes here two types of outliers: the well classified ones and the others. Well classified outliers have a positive effect on ER while badly classified ones have a negative effect on it. This is a similar phenomenon as in regression analysis where good leverage points are outliers that may improve some of the regression outputs. The positive impact of some outliers on the error rate has already been detected by Croux and Dehon (2001) in robust linear discriminant analysis.

Figure 4 illustrates how the influence function changes with respect to varying prior probabilities while assuming that x belongs to the first group. First, one can see that the position of the jump corresponding to the cut-off moves towards the center of the group with the highest prior probability, as expected. Then, as far as the 2-means is concerned, one can notice that the slope of the influence function is positive for small values of π_1 and negative for bigger values. Another comment concerns the magnitude of the slope which is bigger (in absolute value) in the smallest group. This implies that ER based on

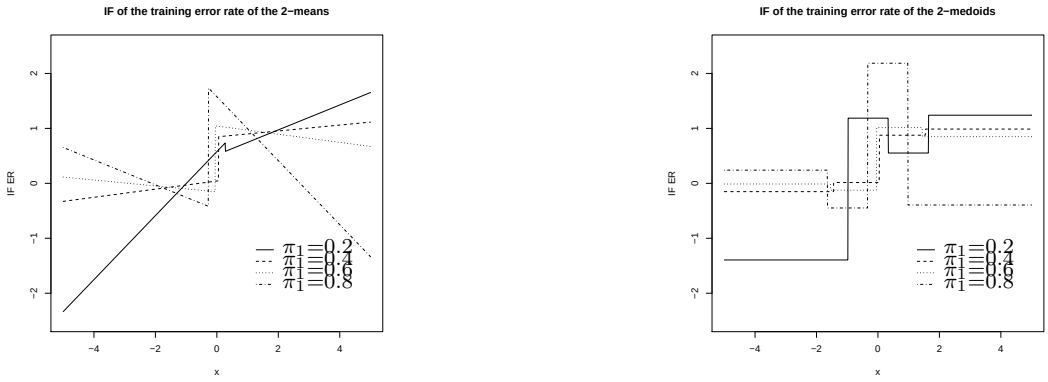

Figure 4: Influence functions of the training error rate of the 2-means (left panel) and 2-medoids (right panel) methods when the model distribution is the mixture $\pi_1\,\mathrm{N}(-1.5,1)+(1-\pi_1)\mathrm{N}(1.5,1)$ for different values of π_1 when x belongs to G_1.

the 2-means procedure is more sensitive to outliers in the smallest group. The 2-medoids method, on the other hand, keeps a similar behavior in all cases even if the IF gets bigger in the smallest group and smaller in the biggest group as π_1 increases.

It is easy to show that the generalized principal points are affine equivariant if and only if $\omega(ax) = \rho(a)\omega(x)$ for all $x \in \mathbb{R}$, with $\rho(a) \neq 0$ for all $a \neq 0$, and this holds for the 2-means and 2-medoids principal points. The error rate is then affine invariant and it does not change when one translates both centers of the two sub-populations. However, intuitively, the error rate should decrease when the distance between the two centers gets bigger and increase otherwise. It is then interesting to visualize the effect of the variation of the distance between the two centers, $\Delta = \mu_2 - \mu_1$, on the influence function. Assuming again that x belongs to the first group, Figure 5 considers the impact of this distance. It shows that any well classified x yields a negative influence (for both methods) while a badly classified x results in positive influence. Moreover, with the 2-means approach, the effect gets bigger in absolute value as the distance gets smaller while the 2-medoids procedure behaves differently when the observation is well or badly classified.

Similar comments hold for describing the effect of the within-group dispersion on the error rate. The corresponding figure is therefore omitted to save space.

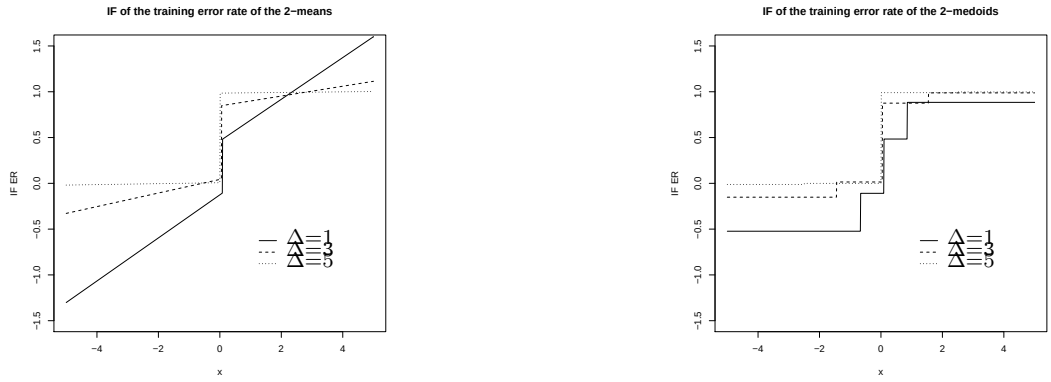

Figure 5: Influence functions of the training error rate of the 2-means (left panel) and 2-medoids (right panel) methods when the model distribution is $0.4\,\mathrm{N}(-\Delta/2, 1) + 0.6\,\mathrm{N}(\Delta/2, 1)$ for different values of Δ when x belongs to G_1.

To close the discussion on the influence function of the training error rate, it is interesting to note that the first term of (14) corresponds to the influence function of the test error rate which vanishes under optimality (Ruwet and Haesbroeck, 2010). As a consequence, the influence function of the training error rate simplifies under model **(N)** as Proposition 3 shows.

Proposition 3 *Under the optimal model **(N)**, the influence function of the training error rate of the generalized 2-means method reduces to*

$$\mathrm{IF}(x; \mathrm{ER}, F_N) = \begin{cases} \Phi\left(\frac{\mu_2-\mu_1}{2\sigma}\right) - \mathrm{I}\left\{x < \frac{\mu_1+\mu_2}{2}\right\} & \text{if } x \in G_1 \\ \mathrm{I}\left\{x < \frac{\mu_1+\mu_2}{2}\right\} - \Phi\left(\frac{\mu_1-\mu_2}{2\sigma}\right) & \text{if } x \in G_2 \end{cases}$$

where Φ denotes here again the standard normal cumulative distribution function.

Under optimality, the influence function of ER does not vanish. This is linked to the fact that contamination can improve ER. Furthermore, it does not depend on the Ω function under consideration anymore. As $C(F_N) = \frac{\mu_1+\mu_2}{2}$ (see Proposition 1), the influence function only depends on the distance between the two centers and on the fact that the point x is

well or badly classified. Indeed, one has

$$\mathrm{IF}(x;\mathrm{ER},F_N)=\begin{cases}-\Phi(\frac{-\Delta}{2\sigma}) & \text{if } x \text{ is well classified}\\ 1-\Phi(\frac{-\Delta}{2\sigma}) & \text{if } x \text{ is badly classified}\end{cases}$$

One can see, as before, that the influence is bigger (in absolute value) when the observation is badly classified (since $\Delta > 0$) while the influence is most extreme (0 and 1) when the distance between the groups is big. In this optimal case, the influence function of ER is bounded for all choices of Ω.

4 Comparison at finite samples

4.1 Training set versus test set error rates

Until now, the comparison of the two error rates was restricted to the population level. Following Qiu and Tamhane (2007), the behavior of these error rates on finite samples will now be investigated by means of simulations. The aim is to compare the so-called *test set* and *training set* error rates with and without contamination.

In this finite sample setting, the error rates are estimated by the proportion of misclassified observations (using the underlying data for the training set error rate or a test data of size 10^5 for the test set error rate). Both the 2-means and 2-medoids procedures are considered, both being optimal under a balanced mixture of normal distributions. The data are generated according to the contaminated mixture

$$F_\varepsilon=(1-\varepsilon)\big\{\pi_1 N(-1,1)+(1-\pi_1)N(1,1)\big\}+\varepsilon\Delta_x$$

with, w.l.o.g., $x \in G_1$. Samples of size 500 are considered with balanced ($\pi_1 = 0.5$) and unbalanced data ($\pi_1 = 0.3$ and $\pi_1 = 0.7$). The percentage of contamination is set at 0, 0.01 and 0.05 while the position of the contamination is ± 4 or ± 40. When $|x| = 4$ (resp. $|x| = 40$), one can say that there is moderate contamination (resp. extreme contamination). Moreover when x is negative, it is bound to be well classified by the classification rules

while x positive leads to a badly classified contamination. To simplify the report of the results, the different settings under study can be summarized as follows:

- Case 0: $\varepsilon = 0$ (no contamination);
- Case 1: $\varepsilon = 0.01$ and $|x| = 4$ (1% of moderate contamination);
- Case 2: $\varepsilon = 0.05$ and $|x| = 4$ (5% of moderate contamination);
- Case 3: $\varepsilon = 0.01$ and $|x| = 40$ (1% of extreme contamination);
- Case 4: $\varepsilon = 0.05$ and $|x| = 40$ (5% of extreme contamination).

Averages over 1000 simulations of the estimated training set and test set error rates are reported in Tables 1 and 2. For all error rates given in these tables, the maximal standard error is 0.045. In Table 1, the contamination is well classified while the reverse holds in Table 2. The optimal Bayes misclassification probability (denoted as *Bayes* in the tables) is also given for comparison.

Without contamination (Case 0), the 2-means procedure seems to be the best. On the other hand, the resistance to contamination of the 2-medoids method is slightly better. Indeed, this method resists to small amounts of extreme contamination (Case 3) which is not the case for the 2-means method. Unfortunately, the 2-medoids procedure also breaks down for bigger amounts of extreme contamination. This is not surprising since García-Escudero and Gordaliza (1999) have shown that this procedure can break down when a single outlier is located at a sufficiently extreme position.

As soon as any of the two clustering procedures has broken down, the cut-off coincides with x. If the contamination mass is well classified (Table 1), $x = -4$ or $x = -40$ and all observations from G_1 bigger than x are badly classified while observations from G_2 are well classified. If the error rate is computed on the training sample itself (ER), the contaminated points (x) are the only observations from G_1 which are well classified. Thus the error rate corresponds to $\pi_1 - \varepsilon$. If a non-contaminated test sample is used (TER), no observation to classify reaches x (with a probability close to 1) and all observations coming

Table 1: Simulated ER and TER of the 2-means and 2-medoids methods for balanced and unbalanced normal models with different kinds of badly classified contamination (case 0 to case 4 with negative values of x).

π_1	Bayes	Settings	ER		TER	
			2-means	2-medoids	2-means	2-medoids
0.3	0.1387	Case 0	0.1872	0.1983	0.1869	0.1985
		Case 1	0.1738	0.1938	0.1746	0.1948
		Case 2	0.1310	0.1718	0.1411	0.1780
		Case 3	0.2900	0.1946	0.3000	0.1958
		Case 4	0.2500	0.2500	0.3000	0.3000
0.5	0.1587	Case 0	0.1595	0.1600	0.1612	0.1616
		Case 1	0.1580	0.1583	0.1616	0.1617
		Case 2	0.1595	0.1522	0.1739	0.1634
		Case 3	0.4900	0.1583	0.5000	0.1617
		Case 4	0.4500	0.4500	0.5000	0.5000
0.7	0.1387	Case 0	0.1872	0.1983	0.1869	0.1985
		Case 1	0.1935	0.1988	0.1957	0.2017
		Case 2	0.2242	0.2023	0.2409	0.2164
		Case 3	0.6900	0.1992	0.7000	0.2022
		Case 4	0.6500	0.6500	0.7000	0.7000

from G_1 are badly classified, leading to an error rate of π_1. If the contamination mass is badly classified (Table 2), $x = 4$ or $x = 40$ and all observations from G_2 smaller than x are badly classified while observations from G_1 are well classified. As the contamination is also badly classified, ER equals $\pi_2 + \varepsilon$. This can lead to a worse classification than the classification obtain by chance (e.g. with a simple coin), illustrating the complete breakdown of the procedure.

Under optimal model ($\pi_1 = 0.5$), Figure 1 and Figure 2 showed that the impact of the contamination relies on the fact that it is well or badly classified. These simulations clearly show that this happens also at finite sample. When the procedure is not too affected by the contamination, one observes a smaller training set error rate under contamination than under the clean model. As expected, this is not the case for the test set error rate.

Under non-optimal models, the behavior is different if the contamination is allocated to the smallest or the biggest group. In the first case ($\pi_1 = 0.3$), the impact of contamination behaves as expected: well classified outliers make the error rate decrease while badly classified ones make it increase. When $\pi_1 = 0.7$, this is not always as such. This can be explained by the fact that the cut-off is always closer to the center of the biggest group (this was illustrated on Figure 4). For example, let us consider the case of well classified contamination (Table 1). Without contamination, the cut-off would be closer to the observations from G_1 than to those of G_2. Adding negative observations in G_1 draws even more that cut-off towards this group. Therefore, some well classified observations from G_1 located too close to the uncontaminated cut-off may now be above the new cut-off. Finally, the non-optimality of the generalized 2-means procedure is clearly illustrated by the difference of its error rates w.r.t. the error rate of the Bayes rule.

4.2 Error rate in clustering and in discrimination

Throughout this paper, it was assumed that the group membership was known. In practice though, other classification techniques would be more appropriate in that case, cluster analysis being mostly applied when there is no available information on the underlying

Table 2: Simulated ER and TER of the 2-means and 2-medoids methods for balanced and unbalanced normal models with different kinds of badly classified contamination (case 0 to case 4 with positive values of x).

π_1	Bayes	Settings	ER		TER	
			2-means	2-medoids	2-means	2-medoids
0.3	0.1387	Case 0	0.1872	0.1983	0.1869	0.1985
		Case 1	0.2073	0.2137	0.1981	0.2051
		Case 2	0.3052	0.2854	0.2596	0.2397
		Case 3	0.7100	0.2156	0.7000	0.2071
		Case 4	0.7500	0.7500	0.7000	0.7000
0.5	0.1587	Case 0	0.1595	0.1600	0.1612	0.1616
		Case 1	0.1686	0.1684	0.1617	0.1617
		Case 2	0.2201	0.2069	0.1757	0.1649
		Case 3	0.5100	0.1684	0.5000	0.1617
		Case 4	0.5500	0.5500	0.5000	0.5000
0.7	0.1387	Case 0	0.1872	0.1983	0.1869	0.1985
		Case 1	0.1814	0.1989	0.1729	0.1914
		Case 2	0.1873	0.2089	0.1407	0.1670
		Case 3	0.3100	0.1992	0.3000	0.1918
		Case 4	0.3500	0.3500	0.3000	0.3000

groups. Nevertheless, assuming that a classification is derived from a clustering technique, it would be of interest to compare its error rate with the one achieved by a classical discriminant analysis. In this section, simulations will be carried out to compare the classification performance of the generalized k-means procedure and that of Fisher discriminant analysis and a robust version of it (obtained by using the robust Minimum Covariance Determinant location and scatter estimators as advocated by Croux, Filzmozer and Joossens, 2008). The chosen models for the simulations are balanced mixtures of normal (denoted as N), Student (S) or lognormal (LN) distributions (translated in order to get a center of -1 for the first group and 1 for the second one). As before, contamination (both moderate and extreme and well and badly placed) is also considered. However, Table 3 only lists the results for extreme and well classified contamination.

Table 3: Simulated ER and TER of the 2-means and 2-medoids methods as well as the classical and robust Fisher discriminant rules for balanced models (normal, student and log-normal) with different kinds of well classified contamination (negative values of x).

Models	ER of	Settings	ER		TER		Fisher	
	Bayes		2-means	2-med	2-means	2-med	classical	robust
N	0.1587	Case 0	0.1595	0.1600	0.1612	0.1616	0.1607	0.1612
		Case 3	0.4900	0.1583	0.5000	0.1617	0.1781	0.1612
		Case 4	0.4500	0.4500	0.5000	0.5000	0.4149	0.1612
S	0.1955	Case 0	0.2082	0.1963	0.2095	0.1981	0.1977	0.1976
		Case 3	0.4900	0.1945	0.5000	0.1982	0.1977	0.1976
		Case 4	0.4500	0.4500	0.5000	0.5000	0.4140	0.1976
LN	0.0254	Case 0	0.0377	0.0453	0.0385	0.0459	0.0334	0.0526
		Case 3	0.4900	0.0450	0.5000	0.0464	0.0688	0.0540
		Case 4	0.4500	0.4500	0.5000	0.5000	0.5000	0.0570

Let us first consider the clean setting (case 0). Under normality, the clustering proce-

dures and Fisher analysis are equivalent. The symmetric S model leads to similar error rates of the 2-medoids procedure and the classical Fisher discriminant rule, while the 2-means method results in slightly bigger error rates due to the presence of some observations in the tails. The reverse holds under the lognormal model. Under contaminated models, one can see that the classical Fisher rule, as the 2-medoids one, resists to 1% of contamination (case 3) which is not the case of the 2-means procedure. However, only the robust version of the Fisher rule gets reasonable error rates under 5% of extreme contamination (case 4).

5 Conclusion

This paper studied the error rate as a measure of performance of a classification rule resulting from the generalized k-means algorithm, which is a generalization of the classical k-means procedure. Two definitions of the error rate were considered, depending on the knowledge or not of a reference distribution, the training and test error rates.

Their influence functions have been computed under general mixture distributions in order to measure their robustness w.r.t. small amounts of contamination. It was shown that contamination may make the training error rate decrease even under optimal models (balanced mixtures of normal distributions). This implies that the influence function of the training error rate does not vanish under optimality while that is the case for the test error rate (as expected under optimal models). The study of the influence function of the training error rate showed that observations that are badly classified have a bigger influence than well classified ones. Moreover, this influence is more important in the smallest group and/or when the groups are not too far away from each other. However, the main conclusion is that any penalty function with a bounded derivative leads to a bounded influence function of the training error rate.

Simulations were used to illustrate these comments on finite samples and to compare the generalized k-means clustering method with Fisher's linear discriminant analysis. The latter was found to yield error rates which are quite comparable to the ones attained

with the 2-medoids method. Only the robust version of the Fisher's linear discriminant rule resists to 5% percent of extreme oultiers. To improve the resistance of the generalized k-means procedure to contamination, Cuesta-Albertos, Gordaliza and Matrán (1997) introduced the trimmed k-means procedure.

Another use of the influence function of the training error rate would be to construct diagnostic measure and diagnostic plot allowing to detect outliers in a dataset. Some examples are given by Pison and Van Aelst (2004) for various statistical methods. However, these techniques are not very useful in our context of univariate analysis. They are used in Ruwet and Haesbroeck (2010) where the multivariate case is treated.

Appendix

Proof of Proposition 1

If a procedure leads to the minimal error rate under F_N, the influence function of the test error rate, whose expression is

$$\mathrm{IF}(x;\mathrm{TER},F_N)=\frac{1}{2}(\mathrm{IF}(x;T_1,F_N)+\mathrm{IF}(x;T_2,F_N))\{\pi_2(F_N)f_2(C(F_N))-\pi_1(F_N)f_1(C(F_N))\},$$

must be identically null (Ruwet and Haesbroeck, 2010). Since it is easy to verify that $\mathrm{IF}(x;T_1,F_N)+\mathrm{IF}(x;T_2,F_N)$ is not identically null, one must have $\pi_2(F_N)f_2(C(F_N))=\pi_1(F_N)f_1(C(F_N))$ where $\pi_1(F_N)=\pi_2(F_N)=\frac{1}{2}$. Using the traditional symmetry hypotheses of the model distribution, one gets

$$f_2(\mu_2+\{\mu_1-C(F_N)\})=f_2(\mu_2+\{C(F_N)-\mu_2\})=f_2(\mu_2+\{C(F_N)-\mu_1\})$$

Then, assuming first that the cut-off point is before μ_1, one has $C(F_N)-\mu_1$ and $C(F_N)-\mu_2$ which are negative and the first equality gives $C(F_N)-\mu_2=C(F_N)-\mu_1$ because f_2 is strictly increasing before μ_2. But this is impossible because $\mu_1<\mu_2$.

The case $C(F_N)>\mu_2$ is similar.

Let us now consider the case $\mu_1<C(F_N)<\mu_2$. Here, $\mu_1-C(F_N)$ and $C(F_N)-\mu_2$ are negative and the second equality gives $C(F_N)-\mu_2=\mu_1-C(F_N)$ because f_2 is strictly

increasing before μ_2. This leads to $C(F_N) = \frac{\mu_1+\mu_2}{2}$. □

Proof of Proposition 2

By definition, $\mathrm{IF}(x; \mathrm{ER}, F) = \dfrac{\partial}{\partial\varepsilon}\mathrm{ER}(F_\varepsilon)\Big|_{\varepsilon=0}$. The following derivatives are straightforward:

$$\frac{\partial}{\partial\varepsilon}\pi_i(F_\varepsilon)\Big|_{\varepsilon=0} = \delta_i(x) - \pi_i(F) \text{ and } \frac{\partial}{\partial\varepsilon}\lambda_{i,x}(\varepsilon)\Big|_{\varepsilon=0} = \frac{\delta_i(x)}{\pi_i(F)}.$$

This leads to the derivative of $F_{\varepsilon,i}(C(F_\varepsilon))$:

$$\frac{\delta_i(x)}{\pi_i(F)}\{\Delta_x(C(F)) - F_i(C(F))\} + f_i(C(F))\frac{1}{2}\{\mathrm{IF}(x; T_1, F) + \mathrm{IF}(x; T_2, F)\}.$$

As

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\mathrm{ER}(F_\varepsilon)\Big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}\pi_1(F_\varepsilon)\Big|_{\varepsilon=0}\{1 - F_1(C(F))\} - \pi_1(F)\,\frac{\partial}{\partial\varepsilon}F_{\varepsilon,1}(C(F_\varepsilon))\Big|_{\varepsilon=0} \\ + \frac{\partial}{\partial\varepsilon}\pi_2(F_\varepsilon)\Big|_{\varepsilon=0} F_2(C(F)) + \pi_2(F)\,\frac{\partial}{\partial\varepsilon}F_{\varepsilon,2}(C(F_\varepsilon))\Big|_{\varepsilon=0},\end{aligned}$$

one gets the influence function by plugging the derivatives of each component in this expression. □

References

Brøns HK, Brunk HD, Franck WE, Hanson DL (1969). Generalized means and associated families of distributions. The Annals of Mathematical Statistics 40:339-355.

Croux C, Dehon C (2001). Robust linear discriminant analysis using S-estimators. The Canadian Journal of Statistics 29:473-492.

Croux C, Filzmoser P, Joossens K (2008). Classification efficiencies for robust linear discriminant analysis. Statistica Sinica 18:581–599.

Cuesta-Albertos JA, Gordaliza A, Matrán C (1997). Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics 25:553-576.

Cuesta-Albertos JA, Matrán C (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. Probability theory and related fields 78:523-534.

Dučinskas K (1995). Optimal training sample allocation and asymptotic expansions for error rates in discriminant analysis. *Acta Applicandae Mathematicae* 38:3-11.

Flury BA (1990). Principal points. Biometrika 77:33-41.

Fraley C, Raftery AE (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97:611-631.

Gallegos MT, Ritter G (2005). A robust method for cluster analysis. The Annals of Statistics 33:347-380.

García-Escudero LA, Gordaliza A (1999). Robustness properties of k means and trimmed k means. Journal of the American Statistical Association 94:956-969.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986). Robust statistics: The Approach Based on Influence Functions. John Wiley and Sons, New-York

Johnson RA, Wichern DW (2007). *Applied multivariate statistical analysis,* 6^{th} *ed.* Pearson Prentice Hall, Upper Saddle River.

Mansour Y and McAllester D (2002). Boosting using branching programs. *Journal of computer and system sciences* 64:103-112.

Pison G, Van Aelst S. (2004). Diagnostic plots for robust multivariate methods. Journal of Computational and Graphical Statistics 13:310-329.

Pollard D (1981). Strong consistency of k-means clustering. The Annals of Statistics 9:135-140.

Pollard D (1982). A central limit theorem for k-means clustering. The Annals of Probability 10:919-926.

Qiu D, Tamhane AC (2007). A comparative study of the k-means algorithm and the normal mixture model for clustering: Univariate case. Journal of Statistical Planning and Inference 137:3722-3740.

Ruwet C, Haesbroeck G (2010). Optimality and classification performance of the k-means clustering. Preprint of the University of Liège.

Symons MJ (1981). Clustering criteria and multivariate normal mixtures. Biometrics 37:35-43