Classification performance resulting from a 2-means

C. Ruwet[a,*], G. Haesbroeck[a]

[a] *University of Liège, Department of Mathematics, Liège, Belgium*

Abstract

The k-means procedure is probably one of the most common nonhierachical clustering techniques. From a theoretical point of view, it is related to the search for the k principal points of the underlying distribution. In this paper, the classification resulting from that procedure for $k = 2$ is shown to be optimal under a balanced mixture of two spherically symmetric and homoscedastic distributions. Then, the classification efficiency of the 2-means rule is assessed using the second order influence function and compared to the classification efficiencies of the Fisher and logistic discriminations. Influence functions are also considered here to compare the robustness to infinitesimal contamination of the 2-means method w.r.t. the generalized 2-means technique.

Keywords: Asymptotic loss, Cluster analysis, Error rate, k-means, Influence Function, Principal points, Robustness.

1. Introduction

Generally, clustering methods aim to classify observations into several groups on the basis of some distances. Cluster analysis differs from discrim-

*Corresponding author
Email address: cruwet@ulg.ac.be (C. Ruwet)

inant analysis because in clustering, there is no training sample (for which the source population is known for each observation) to set up the rule which will be used afterwards to classify the observations. This implies that the prior probabilities to belong to each group cannot be estimated by proportions observed on the training sample. Statistical clustering (e.g. Vermunt and Magidson, 2002; Fraley and Raftery, 2002; Gallegos and Ritter, 2005; Qiu and Tamhane, 2007; García-Escudero et al., 2008) is somewhat between these two kinds of analysis. The underlying distribution F is 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. $F = \sum_{i=1}^{k} \pi_i(F) F_i$. Each of the mixture components represents a sub-population which is denoted by $G_i, i = 1, \ldots, k$. These sets G_i are easier to understand when assuming the presence of a latent variable, Y, which gives the membership. Then, G_i is the set $\{x : Y(x) = i\}$. In this setting, one hopes to end up with clusters representing the different sub-groups. In this sense, an error rate might be defined to measure, as in classification, the performance of the clustering.

A well-known clustering technique is the k-means procedure which consists of looking for 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. At the population level, the name "k principal points" instead of "k-means" has been introduced by Flury (1990). Principal points have already been extensively studied in the literature, even in recent years. For example, the uniqueness of principal points has been shown for univariate models (Li and Flury, 1995), for univariate location mixtures (Yamamoto and Shinozaki, 2000a) and for multivariate location mixtures of spherically

symmetric distributions (Yamamoto and Shinozaki, 2000b). Also, the position of the principal points of elliptical distributions (Tarpey et al., 1995), of mixtures of spherically symmetric distributions (Yamamoto and Shinozaki, 2000b; Kurata, 2008; Kurata and Qiu, 2011) or of general mixtures (Matsuura and Kurata, 2011) have been derived. In this paper, when no confusion is possible, the terminology "k-means" refers to the empirical problem as well as to its population version.

The aim of this paper is to study the classification performance resulting from a k-means procedure when $k = 2$, refereed to as "2-means" procedure. First, as Qiu and Tamhane (2007) and Qiu (2010) did in the univariate and bivariate normal cases, the 2-means procedure is shown here to be optimal (in the sense of achieving the smallest error rate) under multivariate mixtures of spherically symmetric distributions. This part of the work can be viewed as an extension of their previous works. Then, as Croux et al. (2008a) and Croux et al. (2008b) did in the context of discriminant analysis, influence functions (e.g. Hampel et al., 1986) are used to compute the asymptotic loss (Efron, 1975) of 2-means classification. This asymptotic loss is then used to compare the 2-means method to the Fisher and logistic discriminant analyses. This second part of the work is based on the paper of García-Escudero and Gordaliza (1999) who derived the influence functions of the k-means centers in the particular case of univariate data to be clustered into two groups ($k = 2$).

Here are some notations used throughout the paper. The set of all real vectors of dimension p is denoted by $\mathbb{R}^p$ while $\mathbb{R}^p_n$ denotes the set of all real matrices of dimensions $p \times n$. The vector e_1 represents the unity vector

$(1, 0, \ldots, 0)^t$, whatever the dimension.

This paper is organized as follows: Section 2 presents the 2-means clustering methodology and the setting in which it will be used. Section 3 defines the error rate as a measure of performance. In Section 4, the expressions of the first and second order influence functions are derived under a general mixture model. Some particular cases are also emphasized and some influence functions are represented. Section 5 introduces the asymptotic classification efficiency of the 2-means procedure w.r.t. Fisher linear discrimination and logistic discriminant analysis. In Section 6, some simulations illustrate the finite sample behaviors of all these procedures while Section 7 outlines some conclusions.

2. The 2-means procedure

The result of a clustering method can be provided via a set of two points (simply called a 2-set from now on) containing the two centers. Letting F denote the distribution of interest, the population version of the 2-means procedure is the 2-set $\{T_1(F), T_2(F)\} \subset \mathbb{R}^p$ which is solution of the following minimization problem

$$\{T_1(F), T_2(F)\} = \underset{\{t_1,t_2\}\subset\mathbb{R}^p}{\operatorname{argmin}} \int_{\mathbb{R}^p} \left(\inf_{1\leq j\leq 2} \|x - t_j\| \right)^2 dF(x)$$

for those distributions for which this integral exists.

A generalization of this method is the generalized 2-means procedure: the main idea is to replace the quadratic penalty function by another penalty function, denoted by Ω, which is assumed to be non-decreasing. Its population version is defined as the 2-set $\{T_1(F), T_2(F)\}$ in $\mathbb{R}^p$ which is solution of

the following minimization problem

$$\{T_1(F), T_2(F)\} = \underset{\{t_1,t_2\}\subset\mathbb{R}^p}{\operatorname{argmin}} \int_{\mathbb{R}^p} \Omega\left(\inf_{1\leq j\leq 2} \|x - t_j\|\right) dF(x) \tag{1}$$

for those distributions for which this integral exists. Taking $\Omega(x) = x^2$ leads to the classical 2-means estimator while $\Omega(x) = x$ gives the 2-medoids estimator. García-Escudero and Gordaliza (1999) derived robustness properties of the generalized 2-means procedure in the univariate case. 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 $T_1(F)$ and $T_2(F)$ are the outputs of the generalized 2-means analysis, corresponding clusters, denoted as $C_1(F)$ and $C_2(F)$ can be constructed. The jth cluster consists of the region of points closer to $T_j(F)$ than to the other center, the closeness being assessed by the penalty function. A *clustering rule* can then be defined as

$$R_F(x) = C_j(F) \Leftrightarrow j = \underset{1\leq i\leq 2}{\operatorname{argmin}}\, \Omega(\|x - T_i(F)\|),$$

for any $x \in \mathbb{R}^p$. For a strictly increasing penalty function, the allocation of an observation x to a cluster depends on its position in $\mathbb{R}^p$ with respect to a hyperplane and the previous rule can be written as:

$$\begin{cases} R_F(x) = C_1(F) & \text{if } A(F)^t x + b(F) > 0 \\ R_F(x) = C_2(F) & \text{otherwise} \end{cases} \tag{2}$$

where

$$A(F) = T_1(F) - T_2(F) \quad \text{and} \quad b(F) = -\frac{1}{2}\left(\|T_1(F)\|^2 - \|T_2(F)\|^2\right).$$

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

$$\int_{C_i(F)} \omega(x - T_i(F))dF(x) = 0 \qquad i = 1, 2 \tag{3}$$

showing that the generalized principal points are the ω-means, in the sense of Brøns et al. (1969), of the corresponding clusters. 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 $i = 1, 2$. When the gradient of $\Omega(\|x\|)$ does not exist for a finite number of points, the integral in (3) has to be split into a sum of integrals but the property still holds.

The set of centers resulting from a 2-means procedure is a maximum likelihood estimate obtained under a model which assumes that the two populations are normally distributed with the same spherical covariance matrix (Scott and Symons, 1971). Then, only mixture distributions with spherically and equally scattered components will be considered in the sequel.

Let F_{μ,σ^2} denote a spherically symmetric distribution with center $\mu \in \mathbb{R}^p$ and scatter $\sigma^2 I_p \in \mathbb{R}^p_p$. Its density function can be written as

$$f_{\mu,\sigma^2}(x) = \frac{K}{\sigma^p}\, g\left(\frac{(x-\mu)^t(x-\mu)}{\sigma^2}\right)$$

with K a constant such that the honesty condition holds and where g is a non-increasing generator function. For example, the multinormal distribution with spherical covariance is defined by the function $g(r) = \exp(-\frac{r}{2})$ while the function $g(r) = \left(1 + \frac{r}{\nu}\right)^{-\frac{\nu+p}{2}}$ defines the multivariate Student distribution with ν degrees of freedom. See e.g. Serfling (2006) for more information about spherically symmetric distributions.

With that notation, the mixture distribution under consideration here is given by $F = \pi_1 F_{\mu_1,\sigma^2} + \pi_2 F_{\mu_2,\sigma^2}$. W.l.o.g., one can assume that the means of the distributions are located on the first axis, symmetrically w.r.t. the origin, yielding the following model

(M) $F_M \equiv \pi_1 F_{-\mu,\sigma^2} + \pi_2 F_{\mu,\sigma^2}$ where w.l.o.g. $\mu = \mu_1 e_1$ and $\mu_1 > 0$.

Under this particular setting, Yamamoto and Shinozaki (2000b) showed that the 2-means centers are on the first axis. Although a formal proof could not be worked out, the symmetry of the problem makes us believe that the same property holds for any generalized 2-means procedure, as the following conjecture states:

Conjecture 1. *Under model **(M)**, the generalized 2-means centers $T_i(F_M)$, $i = 1, 2$, are given by $t_i\, e_1$ for some t_1 and t_2 in $\mathbb{R}$.*

Simulations computing the distance between the centers and the first axis have been conducted and support this conjecture. Under model **(M)**, it is easy to check that the multivariate 2-means analysis reduces to a univariate one since the first coordinates t_i, $i = 1, 2$, are simply the centers of the one dimensional 2-means problem based on the univariate mixture distribution $F_{M,1D} = \pi_1 F_{-\mu_1,\sigma^2} + \pi_2 F_{\mu_1,\sigma^2}$.

3. Error rate

Any classification rule is bound to misclassify some objects. A measure of classification performance may be defined in terms of the error rate which corresponds to the probability of misclassifying observations distributed according to a given model. Assuming that the model distribution, F_m say, is a

mixture of two distributions, $F_{m,1}$ and $F_{m,2}$, with respective proportions π_1, and π_2, i.e. $F_m = \pi_1 F_{m,1} + \pi_2 F_{m,2}$, while F still represents the distribution under which the 2-means centers are derived, the error rate takes the form

$$\mathrm{ER}(F, F_m) = \sum_{j=1}^{2} \pi_j \mathbb{P}_{F_{m,j}} \Big[R_F(X) \neq C_j(F) \Big]. \tag{4}$$

Using the clustering rule (2), the error rate can be written as

$$\mathrm{ER}(F, F_m) = \pi_1 \mathbb{P}_{F_{m,1}} \Big[A(F)^t X + b(F) < 0 \Big] + \pi_2 \mathbb{P}_{F_{m,2}} \Big[A(F)^t X + b(F) > 0 \Big]. \tag{5}$$

As the error rate is only based on the clustering rule, it remains the same for any generalized 2-means procedure based on a strictly increasing penalty function.

In ideal circumstances, the distribution used to compute the clustering rule is the same as the one on which the quality of the rule is assessed. However, as will be further discussed in Section 4.1, this is not always the case.

In classification, the Bayes rule gives the smallest error rate; this is the gold-standard. This rule is defined by $C_1(F) = \{x \in \mathbb{R}^p : \pi_1 f_1(x) > \pi_2 f_2(x)\}$ and $C_2(F) = \mathbb{R}^p \backslash C_1(F)$. The error rate of the 2-means procedure can reach the minimal error rate of the Bayes rule for some particular models. For example, this holds under a balanced mixture of spherically symmetric and homoscedastic distributions, i.e.

(O) $F_O \equiv 0.5 F_{-\mu,\sigma^2} + 0.5 F_{\mu,\sigma^2}$ where w.l.o.g. $\mu = \mu_1 \, e_1$ with $\mu_1 > 0$

as the following proposition, proved in the Appendix, shows.

Proposition 1. *Under model* ***(O)****, the error rate of the 2-means clustering procedure is equal to the one of the Bayes rule,* ER^{BR}*. This implies the optimality of the 2-means procedure under this model.*

Moreover, under Conjecture 1, the error rate of any generalized 2-means procedure based on a strictly increasing penalty function also reaches the error rate of the Bayes rule, leading to their optimality under model ***(O)****.*

Proposition 1 provides an extension to any multivariate and spherically symmetric distribution of the univariate and bivariate cases proved under normality by Qiu and Tamhane (2007) and Qiu (2010).

4. Influence function of the error rate

4.1. Contamination model

In practice, data often contain outliers, in which case the distribution yielding the clustering rule would be better represented as a distribution F_ε defined as $F_\varepsilon = (1-\varepsilon)F_m + \varepsilon G$, i.e. a proportion $1-\varepsilon$ comes from the true model while the remaining fraction, ε, comes from another distribution G. It is usually assumed that contamination cannot affect the test dataset, nor the prior probabilities (which are estimated assuming a prospective sampling scheme). Nevertheless, contamination will have an impact on the error rate through the corruption of the classification rule as definition (4) adapted to F_ε clearly shows

$$\mathrm{ER}(F_\varepsilon, F_m) = \sum_{j=1}^{2} \pi_j \mathbb{P}_{F_{m,j}}\Big[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon)\Big].$$

4.2. First and second order influence functions

Let us now turn to the derivation of the influence function of the error rate. Roughly speaking, influence functions (Hampel et al., 1986) measure the influence that an infinitesimal contamination placed on an arbitrary point has on the estimator of interest. More formally, when existing, the influence function of the statistical functional ER at the model F_m is defined by

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

where $F_\varepsilon=(1-\varepsilon)F_m+\varepsilon\Delta_x$ and Δ_x is the Dirac distribution having all its mass at the point $x\in\mathbb{R}^p$. In this classification setting, the contaminated distribution F_ε can be written as the natural mixture $F_\varepsilon=\pi_1F_{1,\varepsilon}+\pi_2F_{2,\varepsilon}$ where $F_{j,\varepsilon}=(1-\varepsilon)F_{m,j}+\varepsilon\delta_j(x)\Delta_x$, $\delta_j(x)$ being equal to 1 if x comes from the jth population and 0 otherwise.

Before considering Proposition 2 which gives the influence function of the error rate, let us introduce some additional notations. Under the model distribution F_m, the 2-means centers $T_j(F_m)$ will be denoted by τ_j for $j=1,2$ and the functionals A and b evaluated at F_m by α and β. Thus, one has $\alpha=\tau_1-\tau_2$ and $\beta=-\frac{1}{2}\left(\|\tau_1\|^2-\|\tau_2\|^2\right)$. Moreover, the conditional distributions $F_{m,1}$ and $F_{m,2}$ correspond to some densities $f_{m,1}$ and $f_{m,2}$. For any p-dimensional vector y, the notation $y=(y_1,y_2^t)^t$ distinguishes its first component, $y_1\in\mathbb{R}$, and the vector of its last components, $y_2\in\mathbb{R}^{p-1}$. This decomposition will also be used for the 2-means centers T_1 and T_2 as well as for A, leading to T_{11}, T_{21} and A_1 in $\mathbb{R}$ and to T_{12}, T_{22} and A_2 in $\mathbb{R}^{p-1}$.

Proposition 2. *With the previous notations and the hypothesis that the*

centers τ_1 *and* τ_2 *satisfy*

$$-\tau_1 = \tau_2 = \tau\, e_1, \;\; \textit{with } \tau > 0, \tag{6}$$

the influence function of the error rate of any generalized 2*-means procedure based on a strictly increasing penalty function, under model* F_m, $\mathrm{IF}(x; \mathrm{ER}, F_m)$, *is given by*

$$\int_{\mathbb{R}^{p-1}} \left(\frac{\mathrm{IF}(x; b, F_m) + y_2^t \mathrm{IF}(x; A_2, F_m)}{\alpha_1} \right) \Big(\pi_1 f_{m,1}\,(0, y_2) - \pi_2 f_{m,2}\,(0, y_2) \Big) dy_2 \tag{7}$$

where

$$\mathrm{IF}(x; b, F_m) = \tau \Big(\mathrm{IF}(x; T_{21}, F_m) + \mathrm{IF}(x; T_{11}, F_m) \Big)$$
$$\mathrm{IF}(x; A_2, F_m) = \mathrm{IF}(x; T_{12}, F_m) - \mathrm{IF}(x; T_{22}, F_m)$$

with $\mathrm{IF}(x; T_1, F_m)$ *and* $\mathrm{IF}(x; T_2, F_m)$ *the influence functions of the two generalized* 2*-means centers.*

The proof is in the Appendix. The result of Yamamoto and Shinozaki (2000b) and Conjecture 1 show that the assumption (6) is fulfilled under model **(M)** up to a translation. Under another model for which the true centers are not located on the first axis, the orthogonal equivariance of the generalized k-means procedure allows to modify the distribution in order to satisfy (6). Indeed, one can always construct an orthogonal matrix Γ such that $\Gamma\tau_i = \tau_{i1}\, e_1$ for $i = 1, 2$, and translate the data such that (6) holds. Let F'_m be the distribution of $\Gamma X + \gamma$ so that F'_m satisfies (6). Following Hampel et al. (1986, p. 259), one gets $\mathrm{IF}(x; \mathrm{ER}, F_m) = \mathrm{IF}(\Gamma x + \gamma; \mathrm{ER}, F'_m)$, where the IF on the right hand-side is given by (7).

The influence function of the error rate relies on the influence functions of the 2-means centers T_1 and T_2 which were computed by García-Escudero and Gordaliza (1999) for any generalized 2-means procedure based on a strictly increasing penalty function. These influence functions can be written in the form

$$\begin{pmatrix} \mathrm{IF}(x; T_1, F_m) \\ \mathrm{IF}(x; T_2, F_m) \end{pmatrix} = M^{-1} \begin{pmatrix} \omega_1(x) \\ \omega_2(x) \end{pmatrix}$$

where $\omega_i(x) = -\operatorname{grad}_y \Omega(\|y\|)\big|_{y=x-T_i(F_m)} \mathrm{I}(x \in C_i(F_m))$ and where the matrix M depends only on the distribution F_m. This implies that these influence functions are bounded as soon as the inverse of the matrix M exists and the gradient of the penalty function is bounded. On the other hand, it is clear that the influence function of the error rate (7) is bounded as soon as the influence functions of the functionals T_1 and T_2 are bounded and the first moment of the model distribution exits.

Two special cases where expression (7) simplifies further are worth considering:

The spherical mixture model ***(M)***: As explained above, condition (6) holds under model **(M)**. Moreover, it is easy to check that

$$f_{M,1}(0, y_2) = f_{M,2}(0, y_2) \tag{8}$$

for all $y_2 \in \mathbb{R}^{p-1}$. Then, the influence function of the error rate reduces to

$$\mathrm{IF}(x; \mathrm{ER}, F_M) = \frac{\pi_1 - \pi_2}{\alpha_1} f_{\mu,\sigma^2}(0)\, \mathrm{IF}(x; b, F_M).$$

When the mixture probabilities are equal ($\pi_1 = \pi_2 = 0.5$), i.e. under model **(O)**, it is clear that the first order influence function vanishes. This

is a consequence of the optimality of the generalized 2-means procedure. Indeed, $\mathrm{ER}(F_\varepsilon, F_O) > \mathrm{ER}(F_O, F_O) = \mathrm{ER}^{\mathrm{BR}}(F_O)$ by optimality. Therefore, the IF of ER has to be always positive or null for all x since, for any ε small enough, the first order Taylor's expansion of ER states that

$$\mathrm{ER}(F_\varepsilon, F_O) \approx \mathrm{ER}(F_O, F_O) + \varepsilon \mathrm{IF}(x; \mathrm{ER}, F_O). \tag{9}$$

As the expected value of the influence function must equal zero (Hampel et al., 1986), the influence function of the error rate vanishes under model **(O)**. A second order influence function needs then to be computed. It is defined here as

$$\mathrm{IF2}(x; \mathrm{ER}, F_m) = \frac{\partial^2}{\partial \varepsilon^2} \mathrm{ER}(F_\varepsilon, F_m)\Big|_{\varepsilon=0}$$

(when this derivative exists) and is derived in Proposition 3.

Proposition 3. *Under model **(O)** and with the same notations as in Proposition 2, the second order influence function of the error rate of any generalized 2-means procedure based on a strictly increasing penalty function,* $\mathrm{IF2}(x; \mathrm{ER}, F_O)$*, is given by*

$$-2\mu_1 \frac{K}{\sigma^{p+2}} \int_{\mathbb{R}^{p-1}} \left(\frac{\mathrm{IF}(x; b, F_O)}{\alpha_1} + y_2^t \frac{\mathrm{IF}(x; A_2, F_O)}{\alpha_1} \right)^2 g' \left(\frac{\mu_1^2 + y_2^t y_2}{\sigma^2} \right) dy_2 \tag{10}$$

where the function g' is the derivative of the generator function of the spherically symmetric distribution under consideration and

$$\mathrm{IF}(x; b, F_O) = \tau \Big(\mathrm{IF}(x; T_{21}, F_O) + \mathrm{IF}(x; T_{11}, F_O) \Big)$$
$$\mathrm{IF}(x; A_2, F_O) = \mathrm{IF}(x; T_{12}, F_O) - \mathrm{IF}(x; T_{22}, F_O)$$

with $\mathrm{IF}(x; T_1, F_O)$ *and* $\mathrm{IF}(x; T_2, F_O)$ *the influence functions of the two generalized 2-means centers.*

The proof is in the Appendix. Since the first order influence function vanishes under model **(O)**, the Taylor's expansion (9) becomes

$$\mathrm{ER}(F_\varepsilon, F_O) \approx \mathrm{ER}(F_O, F_O) + \frac{\varepsilon^2}{2}\mathrm{IF2}(x; \mathrm{ER}, F_O).$$

By optimality under model **(O)**, $\mathrm{ER}(F_\varepsilon, F_O) > \mathrm{ER}(F_O, F_O)$ and this implies that the second order influence function must be positive (which is indeed the case since g is non-increasing).

Under normality (model F_N say), $g'(r) = -e^{-\frac{r}{2}}/2$ and expression (10) can be written in a more explicit way:

$$\mathrm{IF2}(\mathrm{ER}; x, F_N) = \frac{\mu_1}{\sigma^3\alpha_1^2}\varphi\left(\frac{\mu_1}{\sigma}\right)\Big(\mathrm{IF}(x; b, F_N)^2 + \sigma^2 \mathrm{IF}(x; A_2, F_N)^t \mathrm{IF}(x; A_2, F_N)\Big)$$

where φ is the pdf of the standard normal distribution.

The univariate case (p=1): In this case, (7) cannot be used but it is straightforward to compute the influence function of the error rate:

$$\mathrm{IF}(x; \mathrm{ER}, F_m) = \frac{1}{2}\Big(\mathrm{IF}(x; T_1, F_m) + \mathrm{IF}(x; T_2, F_m)\Big) \Big(\pi_2 f_{\mu_2,\sigma^2}(C(F_m)) - \pi_1 f_{\mu_1,\sigma^2}(C(F_m))\Big)$$

where $C(F_m) = \Big(T_1(F_m) + T_2(F_m)\Big)/2$ is the cut-off point between the two clusters. Under an optimal model (balanced mixture of spherically symmetric distributions), one gets

$$\mathrm{IF2}(x; \mathrm{ER}, F_O) = -\frac{1}{4} f'_{-\mu,\sigma^2}(0)\Big(\mathrm{IF}(x; T_1, F_O) + \mathrm{IF}(x; T_2, F_O)\Big)^2$$

The univariate generalized 2-means and the influence function of its error rate are studied in details in Ruwet and Haesbroeck (2011).

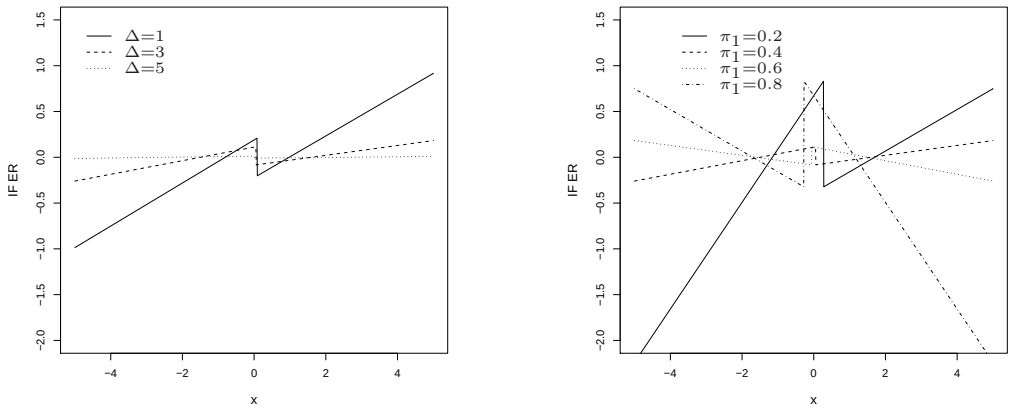

Figure 1: First order influence function of the error rate based on the 2-means procedure under the univariate mixture model $F = \pi_1 N(-\Delta/2, 1) + \pi_2 N(\Delta/2, 1)$, with varying values of Δ and $\pi_1 = 0.4$ (left panel) or varying values of π_1 and $\Delta = 3$ (right panel).

4.3. Graphical representations

In this Section, the spherically symmetric distribution under consideration is the normal distribution.

First, let us consider the univariate case ($p = 1$) where visual analysis of the influence function is much easier. Figure 1 gives the influence functions of the error rate derived from the 2-means methodology under the following mixture of normal distributions: $F = \pi_1 N(-\Delta/2, 1) + \pi_2 N(\Delta/2, 1)$. The left panel shows the impact of varying values of the distance between the means of the two components of the mixture, denoted as Δ, on the first order influence function of the error rate when $\pi_1 = 0.4$. The right panel represents the changes in the first order influence function of the error rate when the weight of the components of the mixture varies while $\Delta = 3$. A first general comment concerns the unbounded characteristic of the influence

function. This is a result of the computation of the influence functions of the clusters centers $T_1(F)$ and $T_2(F)$ which are unbounded too (García-Escudero and Gordaliza, 1999). Secondly, the discontinuity in the function comes from a discontinuity in $\mathrm{IF}(x, T_1, F)$ and $\mathrm{IF}(x, T_2, F)$ at the cut-off point $C(F)$. As far as the impact of the distance Δ between the two groups is concerned, the impact of contamination is bigger when the groups overlap more (small value of Δ), as expected. When x lies before the cut-off point, i.e. in the cluster corresponding to the smallest group, the influence function is mainly negative, yielding a decrease of the error rate (see Taylor's expansion (9)). The impact of the standard deviation on the influence function is similar to the one of the distance between the means and the corresponding plots are omitted to save space. For varying values of the prior probabilities, one observes first that the position of the jump corresponding to the cut-off moves towards the center of the group with the highest prior probability. This illustrates the fact that the k-means procedure tries to get groups of similar weights. Furthermore, 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 the error rate based on the 2-means procedure is more sensitive to outliers in the smallest group.

Under the optimal setting, the first order influence function vanishes and one needs to look at the second order one to measure the impact of contamination. Figure 2 shows this second order influence function under the model $F = 0.5\,N(-\Delta/2, 1) + 0.5\,N(\Delta/2, 1)$ for different values of Δ. Since the 2-means method is optimal under the given model, this second order influence

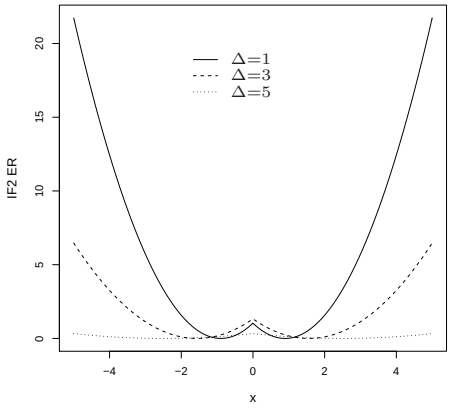

Figure 2: Second order influence function of the error rate based on the 2-means procedure under the optimal model $F = 0.5\,N(-\Delta/2, 1) + 0.5\,N(\Delta/2, 1)$ for varying values of the distance between the means of the two components.

function is always positive. Of course, it is still unbounded, leading to a possible harmful impact of infinitesimal contamination on the error rate of the 2-means clustering method. As under non-optimal models, the influence of outliers becomes smaller when the distance between the groups becomes bigger.

Let us now consider the bivariate case ($p = 2$) with the distribution $F = \pi_1\, N_2(-2\, e_1, I_2) + \pi_2\, N_2(2\, e_1, I_2)$. The left panel of Figure 3 shows the behavior of the first order influence function ($\pi_1 = 0.4$) which is quite similar to the one observed in the one dimensional case: there is a discontinuity in the function corresponding this time to the plane separating the two clusters and the influence function of the error rate based on the 2-means procedure is unbounded. The right panel pictures the second order influence function ($\pi_1 = 0.5$) and the similarity with the one-dimensional case still holds.

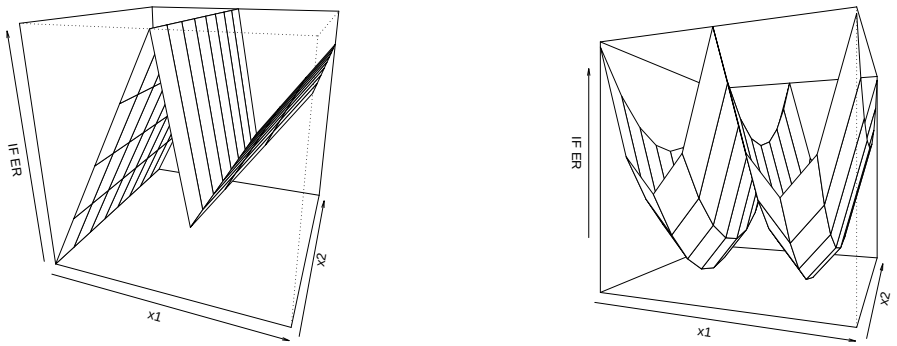

Figure 3: First order influence function of the error rate based on the 2-means procedure under the multivariate mixture $F = 0.4\,N_2(-2\,e_1, I_2) + 0.6\,N_2(2\,e_1, I_2)$ (left panel) and second order influence function of the error rate based on the 2-means procedure under the optimal multivariate model $F = 0.5\,N_2(-2\,e_1, I_2) + 0.5\,N_2(2\,e_1, I_2)$ (right panel).

Proposition 2 being valid for other penalty function than $\Omega(x) = x^2$, one can also look at the influence function of the error rate of the 2-medoids procedure ($\Omega(x) = x$). The left panel of Figure 4 represents the one-dimensional influence function of the error rate based on the 2-medoids method under the model $F = 0.4\,N(-\Delta/2, 1) + 0.6\,N(\Delta/2, 1)$ with varying values of Δ while the right part illustrates the multivariate model $F = 0.4\,N_2(-2\,e_1, I_2) + 0.6\,N_2(2\,e_1, I_2)$. The most important feature of these two graphs is the bounded behavior of the influence functions of the error rate related to the 2-medoids procedure. The impact of infinitesimal contamination is thus less harmful on the 2-medoids method than on the 2-means one. However, as García-Escudero and Gordaliza (1999) show, the 2-medoids procedure can still break down when faced with a single outlier, leading to

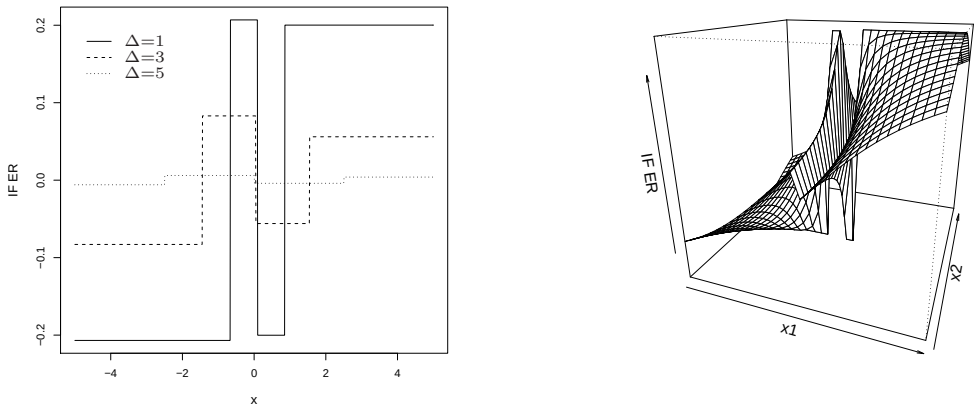

Figure 4: First order influence function of the error rate based on the 2-medoids procedure under the model $F = 0.4\,N(-\Delta/2, 1) + 0.6\,N(\Delta/2, 1)$ (left panel) and under the multivariate model $F = 0.4\,N_2(-2\,e_1, I_2) + 0.6\,N_2(2\,e_1, I_2)$ (right panel).

a breakdown point asymptotically equal to zero. Besides the discontinuity corresponding to the cut-off point (in one dimension) or to the hyperplane between the clusters (in higher dimension), there are two other discontinuities coming from the influence functions of the clusters centers (García-Escudero and Gordaliza, 1999) and corresponding to the 2-medoids centers τ_1 and τ_2.

5. Asymptotic classification efficiencies

As already mentioned in the Introduction, cluster analysis is not the most natural tool to use in order to classify mixture data. Other more appropriate methods would be preferred, as Fisher discrimination or logistic discrimination, both being optimal under a different set of models.

Nevertheless, these three classification procedures are optimal under a balanced mixture ($\pi_1 = \pi_2 = 0.5$) of spherically and equally scattered nor-

mal distributions (denoted F_N as before) and it is of interest to determine whether the 2-means clustering compares favorably to the other methodologies under this particular setting. From now on, the focus is on the 2-means procedure instead of the generalized 2-means one since the following results are really dependent on the optimality result which relies on conjecture 1 in the generalized case.

To characterize the classification performance of the different procedures, the same approach as the one advocated by Croux et al. (2008a) and Croux et al. (2008b) will be considered. They suggest computing Asymptotic Relative Classification Efficiencies (ARCE) of a discrimination method (Method 1, say) with respect to another one (Method 2, say) by means of the ratio

$$\text{ARCE}(\text{Method 1}, \text{Method 2}) = \frac{\text{A-Loss}(\text{Method 2})}{\text{A-Loss}(\text{Method 1})},$$

where A-Loss stands for the asymptotic loss which is defined as

$$\text{A-Loss} = \lim_{n\to\infty} nE_{F_m}[\text{ER}_n - \text{ER}_{\text{opt}}].$$

There, ER_n stands for the error rate of an optimal discriminant rule based on a training sample of size n drawn from the model. Proposition 3 in Croux et al. (2008a) shows that this finite-sample error rate converges at the n^{-1} rate to the optimal error rate. Hence, the asymptotic loss measures how much increase in error rate is to be expected by estimating the optimal discriminant rule from a finite training sample. Under consistency and asymptotic normality of the estimators appearing in the definition of the error rate (here T_1 and T_2), Proposition 3 in Croux et al. (2008a) also shows that the A-Loss of an optimal rule is related to the second order influence function in this

way:

$$\text{A-Loss} = \frac{1}{2} E_{F_m}\Big[\text{IF2}(X; \text{ER}, F_m)\Big].$$

Conditions stated in Pollard (1981, 1982) to ensure consistency and asymptotic normality of T_1 and T_2 are satisfied in the present setting.

See Efron (1975) and Croux et al. (2008a) for more details on classification efficiencies.

The A-Loss corresponding to the 2-means procedure under a balanced mixture of spherically symmetric and homoscedastic normal distributions is given Proposition 4.

Proposition 4. *Under the optimal mixture of normal distributions F_N with $\mu = \Delta/2\, e_1$ and using the notations of Proposition 2, the asymptotic loss of the 2-means procedure is given by*

$$\text{A-Loss} = \frac{\Delta}{4\sigma^3\alpha_1^2}\varphi\left(\frac{\Delta}{2\sigma}\right)\Big(\tau^2[\text{ASV}(T_{21}) + \text{ASV}(T_{11}) + 2\text{ASC}(T_{11}, T_{21})] + \sigma^2[\text{ASV}(T_{12}) + \text{ASV}(T_{22}) - 2\text{ASC}(T_{12}, T_{22})]\Big)$$

where ASV *and* ASC *stand for the asymptotic variance and covariance of their component (at the model distribution).*

Figure 5 shows how the asymptotic loss and the asymptotic relative classification efficiency vary with the distance between the means of the two components, Δ. The left panel of Figure 5 represents the A-Loss of the three classification procedures while the right part yields the ARCE of the 2-means clustering method w.r.t. the Fisher and logistic discriminations. As expected, the loss in classification performance decreases as the distance between the mixture components increases. For the 2-means clustering and

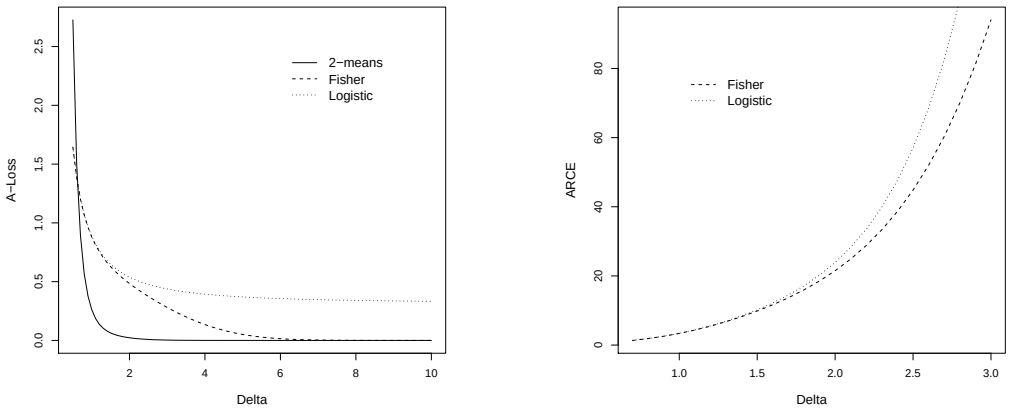

Figure 5: A-Loss of the 2-means clustering procedure, the Fisher discriminant analysis and the logistic discrimination (left panel) and ARCE of the 2-means clustering method w.r.t. the Fisher and logistic discriminations under the multivariate mixture model $F = 0.5\, N_2(-\Delta/2\, e_1, I_2) + 0.5\, N_2(\Delta/2\, e_1, I_2)$.

Fisher discrimination, the A-loss tends towards 0 when Δ increases (while this is not the case for logistic discrimination). Moreover, the A-loss of the 2-means goes to zero much faster than the other one. This corresponds to a better efficiency to classify observations coming from F_N in favor of the 2-means. Therefore, even if the 2-means method is optimal under fewer models than the other two methods, when the model distribution is a balanced mixture of spherically symmetric and homoscedastic normal distributions, the 2-means method performs better as far as this ARCE measure is concerned.

6. Empirical results

Section 5 compared the performance of the 2-means clustering method w.r.t. the Fisher and logistic discriminant analyses under optimality. In

this Section, simulations are conducted to illustrate their behavior under other settings. The idea is to measure the impact of some deviation from the optimal model **(O)**. First, the tails of the distributions are modified by using the multivariate Student distribution. Then, correlation between the covariates will be considered. Finally, skewness is introduced in the distributions following the idea of Azzalini (2005): if $X \in \mathbb{R}$ and $Y \in \mathbb{R}^p$ are independent normal variables, then

$$W = \begin{cases} Y & \text{if } X < \alpha^t Y \\ -Y & \text{otherwise} \end{cases}$$

follows a skew-normal distribution with skewness $\alpha \in \mathbb{R}^p$. More formally, the chosen models for the simulations are balanced mixtures ($\pi_1 = \pi_2 = 0.5$) of

(IN) normal distributions with means $\pm\Delta/2\, e_1$ and an identity covariance matrix,

(T) translated Student distributions centered at $\pm\Delta/2\, e_1$ with $\nu = 4$ degrees of freedom and an identity covariance matrix,

(DN) normal distributions with means $\pm\Delta/2\, e_1$, standard deviations 1 and correlations $\rho = 0.3$,

(SN) skew-normal distributions with skewness parameter e_1, an identity covariance matrix and location parameters $(\pm\Delta/2 - \sqrt{1/\pi})\, e_1$ (in order to get means of $\pm\Delta/2\, e_1$).

It is important to note that the tails of the multivariate Student distribution used here are thinner than the tails of the normal distribution with the same covariance structure (Kotz and Nadarajah (2004)).

Several sampling schemes are considered with $p = 5$ dimensions and $N = 1000$ training samples of size $n = 100$. Other values of these inputs have been considered as well and the results were similar. The classification rule derived from each training sample is assessed on a test sample of size 10^5 (which can be assumed to be a good representation of the population) by computing the error rate of the classification rule. Average error rates over the N simulations ($\pm$ standard deviations) are reported in Table 1. The gold standard given by the error rate of the Bayes rule is also reported in Table 1.

In the case of a mixture of independent normal distributions (IN), all three methods are optimal. One can observe that the finite-sample performances are comparable even if the 2-means procedure tends to achieve the smallest error rate as the distance between the groups increases.

Under the multivariate Student distribution (T), the Fisher discriminant rule seems to be the best choice while the logistic discrimination and 2-means method are either at the second or the third position depending on the distance between the groups. The nice behavior of the Fisher analysis is due to the fact that there are less observations in the tails than under normality.

When the covariates are dependent (DN), Fisher classification still gives the best overall results but the 2-means method performs better than the logistic discrimination when the distance between the means is big enough.

Finally, when there is skewness in the data (SN), the 2-means procedure is able to do as well as Fisher analysis when the components of the mixture are sufficiently separated.

Table 1: Simulated error rates of the 2-means, Fisher and logistic methods for balanced models (independent normal, student, dependent normal and skew-normal) for different values of Δ.

Models	Δ	Bayes	2-means	Fisher	Logistic
(IN)	1	.3085	.3837 ($\pm$ .0528)	.3261 ($\pm$ .0127)	.4256 ($\pm$ .0129)
	2	.1587	.1779 ($\pm$ .0299)	.1710 ($\pm$ .0077)	.1726 ($\pm$ .0090)
	3	.0668	.0720 ($\pm$ .0037)	.0744 ($\pm$ .0042)	.0812 ($\pm$ .0094)
	4	.0228	.0246 ($\pm$ .0013)	.0266 ($\pm$ .0024)	.0404 ($\pm$ .0121)
	5	.0062	.0068 ($\pm$ .0004)	.0078 ($\pm$ .0009)	.0167 ($\pm$ .0087)
(T)	1	.2593	.3689 ($\pm$ .0777)	.2815 ($\pm$ .0162)	.2828 ($\pm$ .0168)
	2	.1151	.1457 ($\pm$ .0832)	.1280 ($\pm$ .0101)	.1329 ($\pm$ .0134)
	3	.0506	.0610 ($\pm$ .0615)	.0578 ($\pm$ .0062)	.0680 ($\pm$ .0136)
	4	.0237	.0305 ($\pm$ .0517)	.0280 ($\pm$ .0041)	.0406 ($\pm$ .0129)
	5	.0121	.0142 ($\pm$ .0341)	.0139 ($\pm$ .0022)	.0234 ($\pm$ .0001)
(DN)	1	.2893	.4314 ($\pm$ .0156)	.3054 ($\pm$ .0118)	.3056 ($\pm$ .0117)
	2	.1333	.2947 ($\pm$ .0473)	.1449 ($\pm$ .0068)	.1473 ($\pm$ .0085)
	3	.0478	.1008 ($\pm$ .0291)	.0542 ($\pm$ .0039)	.0637 ($\pm$ .0110)
	4	.0132	.0273 ($\pm$ .0065)	.0161 ($\pm$ .0017)	.0287 ($\pm$ .0111)
	5	.0027	.0070 ($\pm$ .0017)	.0035 ($\pm$ .0005)	.0092 ($\pm$ .0065)
(SN)	1	.2710	.4350 ($\pm$ .0475)	.2883 ($\pm$ .0115)	.2884 ($\pm$ .0116)
	2	.1120	.1513 ($\pm$ .0591)	.1227 ($\pm$ .0060)	.1257 ($\pm$ .0077)
	3	.0342	.0391 ($\pm$ .0135)	.0396 ($\pm$ .0032)	.0515 ($\pm$ .0117)
	4	.0079	.0090 ($\pm$ .0007)	.0097 ($\pm$ .0010)	.0197 ($\pm$ .0095)
	5	.0013	.0016 ($\pm$ .0002)	.0019 ($\pm$ .0003)	.0051 ($\pm$ .0043)

7. Conclusion

This paper has shown the optimality (in the sense of reaching the smallest possible error rate) of the 2-means clustering procedure when the model distribution is a balanced mixture of spherically symmetric distributions with the same covariance matrix. This result is an extension of the univariate and bivariate cases proved by Qiu and Tamhane (2007) and Qiu (2010) under normality. Unfortunately, the same result concerning the generalized 2-means is still based on the conjecture that the generalized centers are on the axis of symmetry of the distribution. Due to the symmetry, this hypothesis seems natural. Furthermore, it is supported by simulation results. However, a formal proof is still lacking.

The computation of the first and second order influence functions of the error rate of the 2-means method has extended to the multivariate setting the work done in Ruwet and Haesbroeck (2011). Influence functions have been derived for any generalized 2-means procedure defined with a strictly increasing penalty function and were shown to be bounded as soon as the corresponding penalty function has a bounded derivative.

Under balanced mixtures of spherically and equally scattered normal distributions, the classification performance of the 2-means method has been compared with that of the Fisher and logistic discriminant analyses, all of these methods being optimal under this model. The tool used for this comparison is the asymptotic loss which is based on the second order influence function of the error rate, as in Croux et al. (2008a). The loss in classification efficiency resulting from the use of an empirical rule instead of the optimal one is smaller with the 2-means procedure than with the two others, yielding

a better efficiency to classify data under this model distribution.

Finally, a simulation study has compared the finite sample error rates of these three classification procedures. Under the setting for which all three procedures are optimal, it is the 2-means procedure which achieves the smallest error rate as soon as the distance between the means of the two components is big enough. Under other model distributions, the 2-means procedure is a good alternative to the logistic discrimination. It is even able to compete with the Fisher discriminant analysis in presence of skewness.

Although presented here with only 2 clusters, the definition of the clustering rule (2) can be adapted to the general case of k clusters by considering $k(k-1)/2$ hyperplanes and clusters defined by intersections of half-spaces. In this case, the error rate (4) becomes

$$\sum_{j=1}^{k} \pi_j \sum_{l=1}^{k-1} (-1)^{l+1} \sum_{\mathcal{I}_j^l} \mathbb{P}_{F_{m,j}} \left[\bigcap_{i=i_1}^{i_l} \left(A_{ji}(F_\varepsilon)^t X + b_{ji}(F_\varepsilon) < 0 \right) \right]$$

with $\mathcal{I}_j^l = \left\{ (i_1, \ldots, i_l) \in \{1, \ldots, k\} \backslash \{j\} : i_1 < \ldots < i_l \right\}$, $A_{ji}(F) = T_j(F) - T_i(F)$ and $b_{ji}(F) = -\frac{1}{2} \left(\|T_j(F)\|^2 - \|T_i(F)\|^2 \right)$. In order to get the influence function, one has to derive this expression. Under a more general form, the derivation of the IF implies the computation of the following derivative:

$$\frac{\partial}{\partial \varepsilon} \int_{\{x : \mathcal{A}(\varepsilon)^t x + \mathcal{B}(\varepsilon) > 0_l\}} f(x)\, dx$$

with $\mathcal{A} \in \mathbb{R}_l^p$, $\mathcal{B} \in \mathbb{R}^p$, 0_l the null vector in l dimensions ($l = 1, \ldots, k-1$) and f a density function. This is not a trivial problem and, up to our knowledge, no general solution has been proposed yet in the literature. Schechter (1998) developed a method to replace the integral in $\mathbb{R}^p$ by p integrals in $\mathbb{R}$. Then,

the derivative can be passed through the different integrals. However, this technique is not applicable to general values of k and p.

Appendix

Proof of Proposition 1 Model **(O)** satisfies the condition of Yamamoto and Shinozaki (2000b) or Conjecture 1 leading to $T_i(F_O) = t_i\, e_1$, $t_i \in \mathbb{R}$ for $i = 1, 2$. Then the first order conditions (3) become

$$\begin{cases} \displaystyle\int_{\{x\in\mathbb{R}^p : x_1 < (t_1+t_2)/2\}} \omega(x - T_1(F_O)) f(x)\, dx = 0 & (.1) \\ \displaystyle\int_{\{x\in\mathbb{R}^p : x_1 > (t_1+t_2)/2\}} \omega(x - T_2(F_O)) f(x)\, dx = 0 & (.2) \end{cases}$$

where the function ω is odd and f is symmetric w.r.t. the origin. Changing x into $-x + T_1(F_O) + T_2(F_O)$ in equation (.1) leads to

$$\int_{\{x\in\mathbb{R}^p : x_1 > (t_1+t_2)/2\}} \omega(x + T_1(F_O)) dF(x) = 0.$$

The resulting system implies that the generalized 2-means must be such that $T_2(F_O) = -T_1(F_O) = t\, e_1$ with $t > 0$. This leads to $b(F_O) = 0$ and $A(F_O) = -2\, T_2(F_O) = -2t\, e_1$. Thus, the error rate of the generalized 2-means procedure is given by

$$\begin{aligned} \mathrm{ER}(F_O, F_O) &= \frac{1}{2}\Big(\mathbb{P}_{F_{O,1}}[-2tX_1 < 0] + \mathbb{P}_{F_{O,2}}[-2tX_1 > 0] \Big) \\ &= \frac{1}{2}\Big(\mathbb{P}_{F_{O,1}}[X_1 > 0] + \mathbb{P}_{F_{O,2}}[X_1 < 0] \Big). \end{aligned}$$

On the other hand, the Bayes rule is based on the clusters

$$C_1(F) = \{x \in \mathbb{R}^p : f_1(x) > f_2(x)\} \text{ and } C_2(F) = \mathbb{R}^p \backslash C_1(F).$$

Since the generator function g of the spherically symmetric distribution is non-increasing,

$$\begin{aligned} f_1(x) > f_2(x) &\Leftrightarrow f_{-\mu,\sigma^2}(x) > f_{\mu,\sigma^2}(x) \\ &\Leftrightarrow g\left(\frac{(x+\mu)^t(x+\mu)}{\sigma^2}\right) > g\left(\frac{(x-\mu)^t(x-\mu)}{\sigma^2}\right) \\ &\Leftrightarrow (x+\mu)^t(x+\mu) < (x-\mu)^t(x-\mu) \\ &\Leftrightarrow 4\mu^t x < 0 \Leftrightarrow \mu_1 x_1 < 0 \end{aligned}$$

where $\mu_1 > 0$. The error rate of this classification rule is then given by

$$\begin{aligned} \mathrm{ER}^{\mathrm{BR}}(F_O, F_O) &= \frac{1}{2}\Big(\mathbb{P}_{F_{O,1}}[X \in C_2(F)] + \mathbb{P}_{F_{O,2}}[X \in C_1(F)]\Big) \\ &= \frac{1}{2}\Big(\mathbb{P}_{F_{O,1}}[X_1 > 0] + \mathbb{P}_{F_{O,2}}[X_1 < 0]\Big) \\ &= \mathrm{ER}(F_O, F_O). \end{aligned}$$

This prove that the error rate of the generalized 2-means procedure under model **(O)** (and under Conjecture 1 when $\Omega(x) \neq x^2$) reaches the smallest value and is thus optimal.

Proof of Proposition 2 Let us consider the contaminated model $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$ and the shorthand notations $\alpha_\varepsilon = A(F_\varepsilon)$ and $\beta_\varepsilon = b(F_\varepsilon)$. From (5), one has

$$\mathrm{ER}(F_\varepsilon, F_m) = \pi_1 \int_{\beta_\varepsilon + \alpha_\varepsilon^t y < 0} f_{m,1}(y)dy + \pi_2 \int_{\beta_\varepsilon + \alpha_\varepsilon^t y > 0} f_{m,2}(y)dy. \qquad (.3)$$

As one decomposes every vector of $y \in \mathbb{R}^p$ such as $y = (y_1, y_2^t)^t$, one does the same with α, β, α_ε and β_ε. Under the hypothesis $-\tau_1 = \tau_2 = \tau\, e_1$ with $\tau > 0$, one has $\alpha = \alpha_1\, e_1$ with $\alpha_1 < 0$ and $\beta = 0$. Since it implies that, for ε small enough, $\alpha_{\varepsilon,1} < 0$, one introduces the notation

$$k(y_2, \varepsilon) = \frac{-\beta_\varepsilon - y_2^t \alpha_{\varepsilon,2}}{\alpha_{\varepsilon,1}}.$$

With this notation, the integral (.3) can be written as the double integral:

$$\pi_1 \int_{\mathbb{R}^{p-1}} \int_{k(y_2,\varepsilon)}^{+\infty} f_{m,1}(y_1, y_2) dy_1 dy_2 + \pi_2 \int_{\mathbb{R}^{p-1}} \int_{-\infty}^{k(y_2,\varepsilon)} f_{m,2}(y_1, y_2) dy_1 dy_2.$$

Derivation w.r.t. ε results in

$$\frac{\partial}{\partial \varepsilon} \mathrm{ER}(F_\varepsilon, F_m) = \int_{\mathbb{R}^{p-1}} -\Big(\pi_1 f_{m,1}(k(y_2,\varepsilon), y_2) - \pi_2 f_{m,2}(k(y_2,\varepsilon), y_2)\Big) \frac{\partial k(y_2,\varepsilon)}{\partial \varepsilon} dy_2. \tag{.4}$$

Taking this last expression for $\varepsilon = 0$, one has

$$\mathrm{IF}(x; ER, F_m) = \int_{\mathbb{R}^{p-1}} -\Big(\pi_1 f_{m,1}(0, y_2) - \pi_2 f_{m,2}(0, y_2)\Big) \left.\frac{\partial k(y_2,\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0} dy_2$$

since $k(y_2, 0) = -\beta/\alpha_1 = 0$ in this setting. The derivation of $k(y_2, \varepsilon)$ is straightforward and gives

$$\begin{aligned} \left.\frac{\partial k(y_2,\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0} &= \left. \frac{-\left(\frac{\partial \beta_\varepsilon}{\partial \varepsilon} + y_2^t \frac{\partial \alpha_{\varepsilon,2}}{\partial \varepsilon}\right) \alpha_{\varepsilon,1} + \left(\beta_\varepsilon + y_2^t \alpha_{\varepsilon,2}\right) \frac{\partial \alpha_{\varepsilon,1}}{\partial \varepsilon}}{\alpha_{\varepsilon,1}^2} \right|_{\varepsilon=0} \\ &= -\frac{\mathrm{IF}(x; b, F_m)}{\alpha_1} - y_2^t \frac{\mathrm{IF}(x; A_2, F_m)}{\alpha_1}. \end{aligned} \tag{.5}$$

The computation of the influence functions of A_1 and A_2 are immediate and, for b, one has

$$\begin{aligned} \mathrm{IF}(x; b, F_m) &= -\frac{1}{2} \left(\left.\frac{\partial \|T_1(F_\varepsilon)\|^2}{\partial \varepsilon}\right|_{\varepsilon=0} - \left.\frac{\partial \|T_2(F_\varepsilon)\|^2}{\partial \varepsilon}\right|_{\varepsilon=0} \right) \\ &= -\frac{1}{2} \left(\left.\frac{\partial T_1(F_\varepsilon)^t T_1(F_\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0} - \left.\frac{\partial T_2(F_\varepsilon)^t T_2(F_\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0} \right) \\ &= -\frac{1}{2} \left. \left(2T_1(F_\varepsilon)^t \frac{\partial T_1(F_\varepsilon)}{\partial \varepsilon} - 2T_2(F_\varepsilon)^t \frac{\partial T_2(F_\varepsilon)}{\partial \varepsilon} \right) \right|_{\varepsilon=0} \\ &= T_2(F_m)^t \mathrm{IF}(x; T_2, F_m) - T_1(F_m)^t \mathrm{IF}(x; T_1, F_m) \\ &= \tau \mathrm{IF}(x; T_{21}, F_m) + \tau \mathrm{IF}(x; T_{11}, F_m). \end{aligned}$$

Proof of Proposition 3 One uses the notation f^i to denote the derivative of the function $f : \mathbb{R}^p \to \mathbb{R}$ w.r.t. the ith component.

Let us start with the expression (.4) of the previous proof which is now derived once more to give $\dfrac{\partial^2}{\partial\varepsilon^2}\mathrm{ER}(F_\varepsilon, F_m) =$

$$\int_{\mathbb{R}^{p-1}} \left(-\left[\pi_1 f^1_{m,1}(k(y_2,\varepsilon),y_2)\frac{\partial k(y_2,\varepsilon)}{\partial\varepsilon} - \pi_2 f^1_{m,2}(k(y_2,\varepsilon),y_2)\frac{\partial k(y_2,\varepsilon)}{\partial\varepsilon}\right]\frac{\partial k(y_2,\varepsilon)}{\partial\varepsilon} \right.$$
$$\left. -\Big[\pi_1 f_{m,1}(k(y_2,\varepsilon),y_2) - \pi_2 f_{m,2}(k(y_2,\varepsilon),y_2)\Big]\frac{\partial^2 k(y_2,\varepsilon)}{\partial\varepsilon^2}\right) dy_2.$$

Taking $\varepsilon = 0$, $\pi_1 = \pi_2 = 0.5$ and $F_m = F_O$ one has

$$-0.5\int_{\mathbb{R}^{p-1}} \Big[f^1_{O,1}(k(y_2,0),y_2) - f^1_{O,2}(k(y_2,0),y_2)\Big]\left(\frac{\partial k(y_2,\varepsilon)}{\partial\varepsilon}\bigg|_{\varepsilon=0}\right)^2 dy_2$$
$$-0.5\int_{\mathbb{R}^{p-1}} \Big[f_{O,1}(k(y_2,0),y_2) - f_{O,2}(k(y_2,0),y_2)\Big]\frac{\partial^2 k(y_2,\varepsilon)}{\partial\varepsilon^2}\bigg|_{\varepsilon=0} dy_2$$

where one has already seen before that $k(y_2, 0) = 0$ and that the second term vanishes under F_O (equation (8)). Using (.5) leads to

$$-0.5\int_{\mathbb{R}^{p-1}} \Big(f^1_{O,1}(0,y_2) - f^1_{O,2}(0,y_2)\Big)\left(-\frac{\mathrm{IF}(x;b,F_O)}{\alpha_1} - y_2^t\frac{\mathrm{IF}(x;A_2,F_O)}{\alpha_1}\right)^2 dy_2. \tag{.6}$$

Using the definition of spherically symmetric densities, it follows that

$$f^1_{O,1}(0,y_2) = D_{y_1}\, f_{-\mu,\sigma^2}(y_1,y_2)|_{y_1=0} = D_{y_1}\,\frac{K}{\sigma^p}\, g\left(\frac{(y_1+\mu_1)^2 + y_2^t y_2}{\sigma^2}\right)\bigg|_{y_1=0}$$

since $\mu = \mu_1\, e_1$. Then

$$f^1_{O,1}(0,y_2) = \frac{K}{\sigma^{p+2}}\, g'\left(\frac{\mu_1^2 + y_2^t y_2}{\sigma^2}\right) 2\mu_1$$

and

$$f^1_{O,2}(0,y_2) = D_{y_1}\ f_{\mu,\sigma^2}(y_1,y_2)\big|_{y_1=0} = D_{y_1}\ \frac{K}{\sigma^p}\ g\left(\frac{(y_1-\mu_1)^2+y_2^t y_2}{\sigma^2}\right)\Bigg|_{y_1=0}$$
$$= -\frac{K}{\sigma^{p+2}}\ g'\left(\frac{\mu_1^2+y_2^t y_2}{\sigma^2}\right)2\mu_1.$$

Introducing this in (.6), one finally gets IF2(ER; x, F_O)=

$$-2\mu_1\frac{K}{\sigma^{p+2}}\int_{\mathbb{R}^{p-1}}\left(\frac{\mathrm{IF}(x;b,F_O)}{\alpha_1}+y_2^t\frac{\mathrm{IF}(x;A_2,F_O)}{\alpha_1}\right)^2 g'\left(\frac{\mu_1^2+y_2^t y_2}{\sigma^2}\right)dy_2.$$

References

Azzalini, A., 2005. The skew-normal distribution and related multivariate families. Scand. J. Statist. 32, 159–200. With discussion by Marc G. Genton and a rejoinder by the author.

Brøns, H.K., Brunk, H.D., Franck, W.E., Hanson, D.L., 1969. Generalized means and associated families of distributions. Ann. Math. Statist. 40, 339–355.

Croux, C., Filzmoser, P., Joossens, K., 2008a. Classification efficiencies for robust linear discriminant analysis. Statist. Sinica 18, 581–599.

Croux, C., Haesbroeck, G., Joossens, K., 2008b. Logistic discrimination using robust estimators: an influence function approach. Canad. J. Statist. 36, 157–174.

Efron, B., 1975. The efficiency of logistic regression compared to normal discriminant analysis. J. Amer. Statist. Assoc. 70, 892–898.

Flury, B.A., 1990. Principal points. Biometrika 77, 33–41.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Amer. Statist. Assoc. 97, 611–631.

Gallegos, M.T., Ritter, G., 2005. A robust method for cluster analysis. Ann. Statist. 33, 347–380.

García-Escudero, L.Á., Gordaliza, A., 1999. Robustness properties of k means and trimmed k means. J. Amer. Statist. Assoc. 94, 956–969.

García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., 2008. A general trimming approach to robust cluster analysis. Ann. Statist. 36, 1324–1345.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust statistics. The approach based on influence functions. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York.

Kotz, S., Nadarajah, S., 2004. Multivariate t distributions and their applications. Cambridge University Press, Cambridge.

Kurata, H., 2008. On principal points for location mixtures of spherically symmetric distributions. J. Statist. Plann. Inference 138, 3405–3418.

Kurata, H., Qiu, D., 2011. Linear subspace spanned by principal points of a mixture of spherically symmetric distributions. Communications in Statistics - Theory and Methods forthcoming.

Li, L., Flury, B., 1995. Uniqueness of principal points for univariate distributions. Statist. Probab. Lett. 25, 323–327.

Matsuura, S., Kurata, H., 2011. Principal points of a multivariate mixture distribution. Journal of Multivariate Analysis 102, 213–224.

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.

Qiu, D., 2010. A comparative study of the K-means algorithm and the normal mixture model for clustering: bivariate homoscedastic case. J. Statist. Plann. Inference 140, 1701–1711.

Qiu, D., Tamhane, A.C., 2007. A comparative study of the K-means algorithm and the normal mixture model for clustering: univariate case. J. Statist. Plann. Inference 137, 3722–3740.

Ruwet, C., Haesbroeck, G., 2011. Impact of contamination on training and test error rates in statistical clustering analysis. Communications in Statistics - Simulation and computation 40, 394–411.

Schechter, M., 1998. Integration over a polyhedron: an application of the Fourier-Motzkin elimination method. Amer. Math. Monthly 105, 246–251.

Scott, A.J., Symons, M.J., 1971. Clustering methods based on likelihood ratio criteria. Biometrics 27, pp. 387–397.

Serfling, R., 2006. Multivariate symmetry and asymmetry, in: Kotz, S., Balakrishnan, C., Read, B., Vidakovic, B. (Eds.), Encyclopedia of Statistical Sciences. Wiley. volume 8, pp. 5338–5345.

Tarpey, T., Li, L., Flury, B.D., 1995. Principal points and self-consistent points of elliptical distributions. Ann. Statist. 23, 103–112.

Vermunt, J.K., Magidson, J., 2002. Latent class cluster analysis, in: Applied latent class analysis. Cambridge Univ. Press, Cambridge, pp. 89–106.

Yamamoto, W., Shinozaki, N., 2000a. On uniqueness of two principal points for univariate location mixtures. Statist. Probab. Lett. 46, 33–42.

Yamamoto, W., Shinozaki, N., 2000b. Two principal points for multivariate location mixtures of spherically symmetric distributions. J. Japan Statist. Soc. 30, 53–63.