Cluster Identification Using Projections

Daniel PEÑA and Francisco J. PRIETO

This article describes a procedure to identify clusters in multivariate data using information obtained from the univariate projections of the sample data onto certain directions. The directions are chosen as those that minimize and maximize the kurtosis coefficient of the projected data. It is shown that, under certain conditions, these directions provide the largest separation for the different clusters. The projected univariate data are used to group the observations according to the values of the gaps or spacings between consecutive-ordered observations. These groupings are then combined over all projection directions. The behavior of the method is tested on several examples, and compared to k-means, MCLUST, and the procedure proposed by Jones and Sibson in 1987. The proposed algorithm is iterative, affine equivariant, flexible, robust to outliers, fast to implement, and seems to work well in practice.

KEY WORDS: Classification; Kurtosis; Multivariate analysis; Robustness; Spacings.

1. INTRODUCTION

Let us suppose we have a sample of multivariate observations generated from several different populations. One of the most important problems of cluster analysis is the partitioning of the points of this sample into nonoverlapping clusters. The most commonly used algorithms assume that the number of clusters, G, is known and the partition of the data is carried out by maximizing some optimality criterion. These algorithms start with an initial classification of the points into clusters and then reassign each point in turn to increase the criterion. The process is repeated until a local optimum of the criterion is reached. The most often used criteria can be derived from the application of likelihood ratio tests to mixtures of multivariate normal populations with different means. It is well known that (i) when all the covariance matrices are assumed to be equal to the identity matrix, the criterion obtained corresponds to minimizing $\mathrm{tr}(\mathbf{W})$, where $\mathbf{W}$ is the within-groups covariance matrix, this is the criterion used in the standard k-means procedure; (ii) when the covariance matrices are assumed to be equal, without other restrictions, the criterion obtained is minimizing $|\mathbf{W}|$ (Friedman and Rubin 1967); (iii) when the covariance matrices are allowed to be different, the criterion obtained is minimizing $\sum_{j=1}^{G} n_j \log\left|\mathbf{W}_j/n_j\right|$, where $\mathbf{W}_j$ is the sample cross-product matrix for the jth cluster (see Seber 1984, and Gordon 1994, for other criteria). These algorithms may present two main limitations: (i) we have to choose the criterion *a priori*, without knowing the covariance structure of the data and different criteria can lead to very different answers; and (ii) they usually require large amounts of computer time, which makes them difficult to apply to large data sets.

Banfield and Raftery (1993), and Dasgupta and Raftery (1998) have proposed a model-based approach to clustering that has several advantages over previous procedures. They assume a mixture model and use the EM algorithm to estimate the parameters. The initial estimation is made by hierarchical agglomeration. They make use of the spectral decomposition of the covariance matrices of the G populations to allow some groups to share characteristics in their covariance matrices (orientation, size, and shape). The number of groups is chosen by the BIC criterion. However, the procedure has several limitations. First, the initial values have all the limitations of agglomerative hierarchical clustering methods (see Bensmail and Celeux 1997). Second, the shape matrix has to be specified by the user. Third, the method for choosing the number of groups relies on regularity conditions that do not hold for finite mixture models.

More flexibility is possible by approaching the problem from the Bayesian point of view using normal mixtures (Binder 1978) and estimating the parameters by Markov Chain Monte Carlo methods (see Lavine and West 1992). These procedures are very promising, but they are subject to the label switching problem (see Stephens 2000 and Celeux, Hurn, and Robert 2000 for recent analysis of this problem) and more research is needed to avoid the convergence problems owing to masking (see Justel and Peña 1996) and to develop better algorithms to reduce the computational time. The normality assumption can be avoided by using nonparametric methods to estimate the joint density of the observations and identifying the high density regions to split this joint distribution. Although this idea is natural and attractive, nonparametric density estimation suffers from the curse of dimensionality and the available procedures depend on a number of parameters that have to be chosen *a priori* without clear guidance. Other authors (see Hardy 1996) have proposed a hypervolume criterion obtained by assuming that the points are a realization of a homogeneous Poisson process in a set that is the union of G disjoint and convex sets. The procedure is implemented in a dynamic programming setting and is again computationally very demanding.

An alternative approach to cluster analysis is projection pursuit (Friedman and Tukey 1974). In this approach, low-dimensional projections of the multivariate data are used to provide the most interesting views of the full-dimensional data. Huber (1985) emphasized that interesting projections are those that produce nonnormal distributions (or minimum entropy) and, therefore, any test statistic for testing nonnormality could be used as a projection index. In particular, he suggested that the standardized absolute cumulants can be useful for cluster detection. This approach was followed by Jones and Sibson (1987) who proposed to search for clusters by

Daniel Peña (E-mail: *dpena@est-econ.uc3m.es*) is Professor and Francisco J. Prieto (E-mail: *fjp@est-econ.uc3m.es*) is Associate Professor in Dept. Estadística y Econometrica, Univ. Carlos III de Madrid, Spain. We thank the referees and the Associate Editor for their excellent comments and suggestions, that have improved the contents of this article. This research was supported by Spanish grant BEC2000-0167.

maximizing the projection index

$$I(d) = \kappa_3^2(\mathbf{d}) + \kappa_4^2(\mathbf{d})/4,$$

where $\kappa_j(\mathbf{d})$ is the jth cumulant of the projected data in the direction $\mathbf{d}$. These authors assumed that the data had first been centered, scaled, and sphered so that $\kappa_1(\mathbf{d}) = 0$ and $\kappa_2(\mathbf{d}) = 1$. Friedman (1987) indicated that the use of standardized cumulants is not useful for finding clusters because they heavily emphasize departure from normality in the tails of the distribution. As the use of univariate projections based on this projection index has not been completely successful, Jones and Sibson (1987) proposed two-dimensional projections, see also Posse (1995). Nason (1995) has investigated three-dimensional projections, see also Cook, Buja, Cabrera, and Hurley (1995).

In this article, we propose a one-dimensional projection pursuit algorithm based on directions obtained by both maximizing and minimizing the kurtosis coefficient of the projected data. We show that minimizing the kurtosis coefficient implies maximizing the bimodality of the projections, whereas maximizing the kurtosis coefficient implies detecting groups of outliers in the projections. Searching for bimodality will lead to breaking the sample into two large clusters that will be further analyzed. Searching for groups of outliers with respect to a central distribution will lead to the identification of clusters that are clearly separated from the rest along some specific projections. In this article it is shown that through this way we obtain a clustering algorithm that avoids the curse of dimensionality, is iterative, affine equivariant, flexible, fast to implement, and seems to work well in practice.

The rest of this article is organized as follows. In Section 2, we present the theoretical foundations of the method, discuss criteria to find clusters by looking at projections, and prove that if we have a mixture of elliptical distributions the extremes of the kurtosis coefficient provide directions that belong to the set of admissible linear rules. In the particular case of a mixture of two multivariate normal distributions, the direction obtained include the Fisher linear discriminant function. In Section 3, a cluster algorithm based on these ideas is presented. Section 4 presents some examples and computational results, and a Monte Carlo experiment to compare the proposed algorithm with k-means, the Mclust algorithm of Fraley and Raftery (1999) and the procedure proposed by Jones and Sibson (1987).

2. CRITERIA FOR PROJECTIONS

We are interested in finding a cluster procedure that can be applied for exploratory analysis in large data sets. This implies that the criteria must be easy to compute, even if the dimension of the multivariate data, p, and the sample size, n, are large. Suppose that we initially have a set of data $S = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. We want to apply an iterative procedure where the data are projected onto some directions and a unidimensional search for clusters is carried out along these directions. That is, we first choose a direction, project the sample onto this direction, and we analyze if the projected points can be split into clusters along this first direction. Assuming that the set S is split into k nonoverlapping sets $S = S_1 \cup S_2 \cup \cdots \cup S_k$, where $S_i \cap S_j = \varnothing \, \forall \, i, j$, the sample data is projected over a second direction and we check if each cluster $S_i, i = 1, \ldots, k$, can be further split. The procedure is repeated until the data is finally split into m sets. Formal testing procedures can then be used to check if two groups can be combined into one. For instance, in the normal case we check if the two groups have the same mean and covariance matrices. In this article, we are mainly interested in finding interesting directions useful to identify clusters.

An interesting direction is one where the projected points cluster around different means and these means are well separated with respect to the mean variability of the distribution of the points around their means. In this case we have a bimodal distribution, and therefore a useful criterion is to search for directions which maximize the bimodality property of the projections. This point was suggested by Switzer (1985). For instance, a univariate sample of zero-mean variables $(x_1, \ldots, x_n)$ will have maximum bimodality if it is composed of $n/2$ points equal to $-a$ and $n/2$ points equal to a, for any value a. It is straightforward to show that this is the condition required to minimize the kurtosis coefficient, as in this case it will take a value of one. Now assume that the sample of size n is concentrated around two values but with different probabilities, for instance, n_1 observations take the value $-a$ and n_2 take the value a, with $n = n_1 + n_2$. Let $r = n_1/n_2$, the kurtosis coefficient will be $(1 + r^3)/r(1 + r)$. This function has its minimum value at $r = 1$ and grows without limit either when $r \to 0$ or when $r \to \infty$. This result suggests that searching for directions where the kurtosis coefficient is minimized will tend to produce projections in which the sample is split into two bimodal distributions of about the same size. Note that the kurtosis coefficient is affine invariant and verifies the condition set by Huber (1985) for a good projection index for finding clusters. On the other hand, maximizing the kurtosis coefficient will produce projections in which the data is split among groups of very different size: we have a central distribution with heavy tails owing to the small clusters of outliers. For instance, Peña and Prieto (2001) have shown that maximizing the kurtosis coefficient of the projections is a powerful method for searching for outliers and building robust estimators for covariance matrices. This intuitive explanation is in agreement with the dual properties of the kurtosis coefficient for measuring bimodality and concentration around the mean, see Balanda and MacGillivray (1988).

To formalize this intuition, we need to introduce some definitions. We say that two random variables on $\mathbb{R}^p$, $(\mathbf{X}_1, \mathbf{X}_2)$, with distribution functions F_1 and F_2, can be linearly separated with power $1 - \varepsilon$ if we can find a partition of the space into two convex regions, A_1 and A_2, such that $P(\mathbf{X}_1 \in A_1) \geq 1 - \varepsilon$, and $P(\mathbf{X}_2 \in A_2) \geq 1 - \varepsilon$. This is equivalent to saying that we can find a unit vector $\mathbf{d} \in \mathbb{R}^p$, $\mathbf{d}'\mathbf{d} = 1$, and a scalar $c = c(F_1, F_2)$ such that $P(\mathbf{X}_1'\mathbf{d} \leq c) \geq 1 - \varepsilon$ and $P(\mathbf{X}_2'\mathbf{d} \geq c) \geq 1 - \varepsilon$. For example, given a hyperplane separating A_1 and A_2, one such vector $\mathbf{d}$ would be the unit vector orthogonal to this separating hyperplane. From the preceding definition it is clear that (trivially) any two distributions can be linearly separated with power 0.

Now assume that the observed multivariate data, $S = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ where $\mathbf{X} \in \mathbb{R}^p$, have been generated from a mixture defined by a set of distribution functions $F = (F_1, \ldots, F_k)$

with finite means, $\boldsymbol{\mu}_i = E(\mathbf{X}|\mathbf{X} \sim F_i)$ and covariance matrices $\mathbf{V}_i = \mathrm{Var}(\mathbf{X}|\mathbf{X} \sim F_i)$, and mixture probabilities $\alpha = (\alpha_1, \ldots, \alpha_k)$, where $\alpha_i \geq 0$ and $\sum_{i=1}^k \alpha_i = 1$. Generalizing the previous definition, we say that a distribution function F_i can be linearly separated with power $1-\varepsilon_i$ from the other components of a mixture (F, α) if given $\varepsilon_i > 0$ we can find a unit vector $\mathbf{d}_i \in \mathbb{R}^p$, $\mathbf{d}_i'\mathbf{d}_i = 1$, and a scalar $c_i = g_i(F, \alpha, \varepsilon_i)$ such that

$$P(\mathbf{X}'\mathbf{d}_i \leq c_i | X \sim F_i) \geq 1 - \varepsilon_i$$

and

$$P(\mathbf{X}'\mathbf{d}_i \geq c_i | X \sim F_{(i)}) \geq 1 - \varepsilon_i,$$

where $F_{(i)} = \sum_{j \neq i} \alpha_j F_j / \alpha_i$. Defining $\varepsilon = \max_i \varepsilon_i$, we say that the set is linearly separable with power $1-\varepsilon$.

For instance, suppose that F_i is $N_p(\boldsymbol{\mu}_i, \mathbf{V}_i)$, $i = 1, \ldots, k$. Then, if $\boldsymbol{\Phi}$ denotes the distribution function of the standard normal, the distributions can be linearly separated at level .05 if for $i = 1, \ldots, k$, we can find c_i such that $1 - \boldsymbol{\Phi}((c_i - m_i)\sigma_i^{-1}) \leq .05$ and $\sum_{j \neq i}^k \boldsymbol{\Phi}((c_j - m_j)\sigma_j^{-1})\alpha_i\alpha_j^{-1} \leq .05$, where $m_j = \mathbf{d}_j'\boldsymbol{\mu}_j$ and $\sigma_j^2 = \mathbf{d}_j'\mathbf{V}_j\mathbf{d}_j$.

Consider the projections of the observed data onto a direction $\mathbf{d}$. This direction will be interesting if the projected observations show the presence of at least two clusters, indicating that the data comes from two or more distributions. Thus, on this direction, the data shall look as a sample of univariate data from a mixture of unimodal distributions. Consider the scalar random variable $z = \mathbf{X}'\mathbf{d}$, with distribution function $(1-\alpha)G_1 + \alpha G_2$ having finite moments. Let us call $m_i = \int z dG_i = \mathbf{d}'\boldsymbol{\mu}_i$ and $m_i(k) = \int (z - m_i)^k dG_i$, and in particular $m_i(2) = \mathbf{d}'\mathbf{V}_i\mathbf{d}$ for $i = 1, 2$. It is easy to see that these two distributions can be linearly separated with high power if the ratio

$$w = \frac{(m_2 - m_1)^2}{\left(m_1^{\frac{1}{2}}(2) + m_2^{\frac{1}{2}}(2)\right)^2} \tag{1}$$

is large. To prove this result we let $c_1 = m_1 + m_1^{1/2}(2)/\sqrt{\varepsilon}$ and from Chebychev inequality we have that

$$P(z \leq c_1 | z \sim G_1) \geq P(|z - m_1| \leq c_1 - m_1 | z \sim G_1) \geq 1 - \varepsilon.$$

In the same way, taking $c_2 = m_2 - m_2^{1/2}(2)/\sqrt{\varepsilon}$ we have that $P(z \geq c_2 | z \sim G_2) \geq 1 - \varepsilon$. The condition $c_1 = c_2$ then implies $w = \varepsilon^{-2}$ and the power will be large if w is large.

In particular, if (1) is maximized, the corresponding extreme directions would satisfy

$$\mathbf{d} = \varepsilon^{-1}\left((\mathbf{d}'\mathbf{V}_1\mathbf{d})^{-\frac{1}{2}}\mathbf{V}_1 + (d'\mathbf{V}_2\mathbf{d})^{-\frac{1}{2}}\mathbf{V}_2\right)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \tag{2}$$

To compute these directions we would need to make use of the parameters of the two distributions, that are, in general, unknown. We are interested in deriving equivalent criteria that provide directions that can be computed without any knowledge of the individual distributions. We consider criteria defined by a measure of the distance between the two projected distributions of the form

$$D(f_1, f_2) = \frac{(\mathbf{d}'(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1))^2}{\lambda_1\mathbf{d}'\mathbf{V}_1\mathbf{d} + \lambda_2\mathbf{d}'\mathbf{V}_2\mathbf{d}}.$$

For this criterion we would have the extreme direction

$$\mathbf{d} = (\lambda_1\mathbf{V}_1 + \lambda_2\mathbf{V}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \tag{3}$$

that, as shown in Anderson and Bahadur (1962), has the form required for any admissible linear classification rule for multivariate normal populations with different covariance matrices. The following result indicates that, under certain conditions, the directions with extreme kurtosis coefficient would fit the preceding rule, for specific values of λ_1 and λ_2.

Theorem 1. Consider a p-dimensional random variable $\mathbf{X}$ distributed as $(1-\alpha)f_1(\mathbf{X}) + \alpha f_2(\mathbf{X})$, with $\alpha \in (0, 1)$. We assume that $\mathbf{X}$ has finite moments up to order 4 for any α, and we denote by $\boldsymbol{\mu}_i$, $\mathbf{V}_i$ the vector of means and the covariance matrix under f_i, $i = 1, 2$. Let $\mathbf{d}$ be a unit vector on $\mathbb{R}^p$ and let $z = \mathbf{d}'\mathbf{X}$, $m_i = \mathbf{d}'\boldsymbol{\mu}_i$. The directions that maximize or minimize the kurtosis coefficient of z are of the form

$$\mathbf{V}_m\mathbf{d} = \lambda_3(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \lambda_4((1-\alpha)\boldsymbol{\phi}_1 + \alpha\boldsymbol{\phi}_2) + \lambda_5(\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1),$$

where $\mathbf{V}_m = \lambda_1\mathbf{V}_1 + \lambda_2\mathbf{V}_2$, λ_i are scalars, $\boldsymbol{\phi}_i = 4\int_{\mathbb{R}^p}(z - m_i)^3 \times (\mathbf{X} - \boldsymbol{\mu}_i)f_i(\mathbf{X})\,d\mathbf{X}$ and $\boldsymbol{\tau}_i = 3\int_{\mathbb{R}^p}(z - m_i)^2(\mathbf{X} - \boldsymbol{\mu}_i)f_i(\mathbf{X})\,d\mathbf{X}$.

Proof. If we introduce the notation

$$\begin{aligned}
\Delta &= m_2 - m_1, \\
\sigma_m^2 &= (1-\alpha)m_1(2) + \alpha m_2(2), \\
\tilde{\sigma}_m^2 &= \alpha m_1(2) + (1-\alpha)m_2(2), \\
r^2 &= \Delta^2/\sigma_m^2,
\end{aligned}$$

the kurtosis coefficient for the projected data can be written as

$$\begin{aligned}
\gamma_z(\mathbf{d}) = ((1-\alpha)m_1(4) + \alpha m_2(4) + \alpha(1-\alpha) \\
\times \Delta(4m_2(3) - 4m_1(3) + 6\Delta\tilde{\sigma}_m^2 \\
+ \Delta^3(\alpha^3 + (1-\alpha)^3)))/(\sigma_m^2 + \alpha(1-\alpha)\Delta^2)^2,
\end{aligned} \tag{4}$$

where $m_i(k) = E_{f_i}(z - m_i)^k$. The details of the derivation are given in Appendix A. Any solution of the problem

$$\begin{aligned}
\max_d \quad & \gamma_z(\mathbf{d}) \\
\text{s.t.} \quad & \mathbf{d}'\mathbf{d} = 1
\end{aligned}$$

must satisfy $\nabla\gamma_z(\mathbf{d}) = 0$, where $\nabla\gamma_z(\mathbf{d})$ is the gradient of $\gamma_z(\mathbf{d})$ and $\mathbf{d}'\mathbf{d} = 1$. We have used that γ_z is homogeneous in $\mathbf{d}$ to simplify the first-order condition. The same condition is necessary for a solution of the corresponding minimization problem. From (4), this condition can be written as

$$\begin{aligned}
(\lambda_1\mathbf{V}_1 + \lambda_2\mathbf{V}_2)\mathbf{d} = \lambda_3(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \lambda_4((1-\alpha)\boldsymbol{\phi}_1 \\
+ \alpha\boldsymbol{\phi}_2) + \lambda_5(\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1),
\end{aligned} \tag{5}$$

where the scalars λ_i, dependent on $\mathbf{d}$, are given by

$$\begin{aligned}
\lambda_1 &= (1-\alpha)\big(\gamma_z + \alpha r^2((1-\alpha)\gamma_z - 3\alpha)\big),\\
\lambda_2 &= \alpha\big(\gamma_z + (1-\alpha)r^2(\alpha\gamma_z - 3(1-\alpha))\big),\\
\lambda_3 &= \alpha(1-\alpha)\sigma_m\big((m_2(3) - m_1(3))/\sigma_m^3\\
&\quad + r\big(3\tilde{\sigma}_m^2/\sigma_m^2 - \gamma_z\big) + r^3(\alpha^3 + (1-\alpha)^3\\
&\quad - \alpha(1-\alpha)\gamma_z)\big),\\
\lambda_4 &= 1/(4\sigma_m^2),\\
\lambda_5 &= \alpha(1-\alpha)r/\sigma_m. \qquad (6)
\end{aligned}$$

See Appendix A for its derivation.

To gain some additional insight on the behavior of the kurtosis coefficient, consider the expression given in (4). If Δ grows without bound (and the moments remain bounded), then

$$\gamma_z \to \frac{\alpha^3 + (1-\alpha)^3}{\alpha(1-\alpha)}.$$

In the limit, if $\alpha = .5$, then the kurtosis coefficient of the observed data will be equal to one, the minimum possible value. On the other hand, if $\alpha \to 0$, then the kurtosis coefficient will increase without bound. Thus, when the data projected onto a given direction is split into two groups of very different size, we expect that the kurtosis coefficient will be large. On the other hand, if the groups are of similar size, then the kurtosis coefficient will be small. Therefore, it would seem reasonable to look for interesting directions among those with maximum and minimum kurtosis coefficient, and not just the maximizers of the coefficient.

From the discussion in the preceding paragraphs, a direction satisfying (5), although closely related to the acceptable directions defined by (3), is not equivalent to them. To ensure that a direction maximizing or minimizing the kurtosis coefficient is acceptable, we would need that both ϕ_i and τ_i should be proportional to $\mathbf{V}_i\mathbf{d}$. Next we show that this will be true for a mixture of elliptical distributions.

Corollary 1. Consider a p-dimensional random variable $\mathbf{X}$ distributed as $(1-\alpha)f_1(\mathbf{X}) + \alpha f_2(\mathbf{X})$, with $\alpha \in (0,1)$ and f_i, $i = 1, 2$, is an elliptical distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{V}_i$. Let $\mathbf{d}$ be a unit vector on $\mathbb{R}^p$ and $z = \mathbf{d}'\mathbf{X}$. The directions that maximize or minimize the kurtosis coefficient of z are of the form

$$(\bar{\lambda}_1\mathbf{V}_1 + \bar{\lambda}_2\mathbf{V}_2)\,\mathbf{d} = \bar{\lambda}_3(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \qquad (7)$$

Proof. From Theorem 1, these directions will satisfy (5). The values of ϕ_i and τ_i are the gradients of the central moments $m_i(k)$ for $k = 3, 4$. We first show that these values can be obtained (in the continuous case) from integrals of the form

$$\int\cdots\int(\mathbf{d}'\mathbf{Y})^k\mathbf{Y}f(\mathbf{Y})\,d\mathbf{Y},$$

for $k = 2, 3$, where $\mathbf{Y}$ is a vector random variable with zero-mean in $\mathbb{R}^p$. If the characteristic function of the vector random variable $\mathbf{Y}$ is denoted by

$$\varphi(\mathbf{t}) = \int\cdots\int\exp(i\mathbf{t}'\mathbf{Y})f(\mathbf{Y})\,d\mathbf{Y},$$

for $\mathbf{t} \in \mathbb{R}^p$, the characteristic function of its univariate projections onto the direction $\mathbf{d}$ will be given by $\varphi(t\,\mathbf{d})$, where $t \in \mathbb{R}$ and $\mathbf{d} \in \mathbb{R}^p$. It is straightforward to show that

$$\phi = 4\,\frac{d^3\Psi(t,\mathbf{d})}{i^3dt^3}\bigg|_{t=0}, \qquad \tau = 3\,\frac{d^2\Psi(t,\mathbf{d})}{i^2dt^2}\bigg|_{t=0},$$

where

$$\Psi(t,\mathbf{d}) = \frac{1}{it}\nabla\varphi(t\,\mathbf{d}),$$

and $\nabla\varphi(t\,\mathbf{d})$ is the gradient of φ with respect to its argument. The characteristic function of a member $\mathbf{Y}$ of the family of elliptical symmetric distributions with zero-mean and covariance matrix $\mathbf{V}$ is (see for instance Muirhead, 1982)

$$\varphi(\mathbf{t}) = g(-\tfrac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}).$$

Letting $\mathbf{Y}_i = \mathbf{X}_i - \boldsymbol{\mu}_i$ and $z_i = \mathbf{d}'\mathbf{Y}_i$, the univariate random variables z_i would have characteristic functions

$$\varphi_i(t\,\mathbf{d}) = g_i(-\tfrac{1}{2}t^2\,\mathbf{d}'\mathbf{V}_i\,\mathbf{d}).$$

It is easy to verify that $\Psi(t\,\mathbf{d}) = g'(u)it\mathbf{V}\,\mathbf{d}$, where $u = -\frac{1}{2}t^2\,\mathbf{d}'\mathbf{V}\,\mathbf{d}$, and

$$\begin{aligned}
m_i(3) &= 0,\\
\tau_i &= 0,\\
\phi_i &= 12g_i''(0)\,\mathbf{d}'\mathbf{V}_i\,\mathbf{d}\,\mathbf{V}_i\,\mathbf{d}.
\end{aligned}$$

From (5) it follows that the direction that maximizes (or minimizes) the kurtosis coefficient has the form indicated in (7), where

$$\begin{aligned}
\bar{\lambda}_1 &= \lambda_1 - 3(1-\alpha)g_1''(0)m_1(2)/\sigma_m^2,\\
\bar{\lambda}_2 &= \lambda_2 - 3\alpha g_2''(0)m_2(2)/\sigma_m^2,\\
\bar{\lambda}_3 &= \alpha(1-\alpha)r\sigma_m\big(\big(3\tilde{\sigma}_m^2/\sigma_m^2 - \gamma_z\big)\\
&\qquad + r^2(\alpha^3 + (1-\alpha)^3 - \alpha(1-\alpha)\gamma_z)\big),
\end{aligned}$$

and λ_1, λ_2 are given in (6).

If the distributions are multivariate normal with the same covariance matrix, then we can be more precise in our characterization of the directions that maximize (or minimize) the kurtosis coefficient.

Corollary 2. Consider a p-dimensional random variable $\mathbf{X}$ distributed as $(1-\alpha)f_1(\mathbf{X}) + \alpha f_2(\mathbf{X})$, with $\alpha \in (0,1)$ and $f_i, i = 1, 2$ is a normal distribution with mean $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{V}_i = \mathbf{V}$, the same for both distributions. Let $\mathbf{d}$ be a unit vector on $\mathbb{R}^p$ and $z = \mathbf{d}'\mathbf{X}$. If $\mathbf{d}$ satisfies

$$\mathbf{V}\,\mathbf{d} = \bar{\lambda}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \qquad (8)$$

for some scalar $\bar{\lambda}$, then it maximizes or minimizes the kurtosis coefficient of z. Furthermore, these directions minimize the kurtosis coefficient if $|\alpha - 1/2| < 1/\sqrt{12}$, and maximize it otherwise.

Proof. The normal mixture under consideration is a particular case of Corollary 1. In this case $g_i(x) = \exp(x)$, $g_i''(0) = 1$, $m_1(2) = m_2(2) = \sigma_m^2 = \tilde{\sigma}_m^2$ and as a consequence (7) holds with the following expression

$$\tilde{\lambda}_1 \mathbf{V}\mathbf{d} = \tilde{\lambda}_2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \tag{9}$$

where the values of the parameters are

$$\tilde{\lambda}_1 = (\gamma_z - 3)(1 + \alpha(1-\alpha)r^2)$$
$$\tilde{\lambda}_2 = r\alpha(1-\alpha)\sigma_m\big(3 - \gamma_z + r^2(\alpha^3 + (1-\alpha)^3 - \alpha(1-\alpha)\gamma_z)\big).$$

Also, from (4), for this case we have that

$$\gamma_z = 3 + r^4 \frac{\alpha(1-\alpha)(1-6\alpha+6\alpha^2)}{(1+\alpha(1-\alpha)r^2)^2}. \tag{10}$$

Replacing this value in $\tilde{\lambda}_1$ we obtain

$$\tilde{\lambda}_1 = r^4 \frac{\alpha(1-\alpha)(1-6\alpha+6\alpha^2)}{1+\alpha(1-\alpha)r^2}$$
$$\tilde{\lambda}_2 = r\alpha(1-\alpha)\sigma_m\big(3 - \gamma_z + r^2(\alpha^3 + (1-\alpha)^3 - \alpha(1-\alpha)\gamma_z)\big).$$

From (9), a direction that maximizes or minimizes the kurtosis coefficient must satisfy that either (i) $\tilde{\lambda}_1 \neq 0$ and $\mathbf{d} = \bar{\lambda}\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ for $\bar{\lambda} = \tilde{\lambda}_2/\tilde{\lambda}_1$, and we obtain the Fisher linear discriminant function, or (ii) $\tilde{\lambda}_1 = \tilde{\lambda}_2 = 0$, implying $r = 0$, that is, the direction is orthogonal to $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. From (10) we have that if $\mathbf{d}$ is such that $r = 0$, then $\gamma_z = 3$, and if $\mathbf{d} = \bar{\lambda}\mathbf{V}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, then $r^2 = 1$ and

$$\gamma_z = 3 + \frac{\alpha(1-\alpha)(1-6\alpha+6\alpha^2)}{(1+\alpha(1-\alpha))^2}.$$

This function of α is smaller than 3 whenever $|\alpha - 1/2| < 1/\sqrt{12}$, and larger than 3 if $|\alpha - 1/2| > 1/\sqrt{12}$.

This corollary generalizes the result by Peña and Prieto (2000) which showed that if the distributions f_i are multivariate normal with the same covariance matrix $\mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}$ and $\alpha = .5$, the direction that minimizes the kurtosis coefficient corresponds to the Fisher best linear discriminant function.

We conclude that in the normal case there exists a close link between the directions obtained by maximizing or minimizing the kurtosis coefficient and the optimal linear discriminant rule. Also, in other cases where the optimal rule is not in general linear, as is the case for symmetric elliptical distributions with different means and covariance matrices, the directions obtained from the maximization of the kurtosis coefficient have the same structure as the admissible linear rules. Thus maximizing and minimizing the kurtosis coefficient of the projections seems to provide a sensible way to obtain directions that have good properties in these situations.

3. THE CLUSTER IDENTIFICATION PROCEDURE

If the projections were computed for only one direction, then some clusters might mask the presence of others. For example, the projection direction might significantly separate one cluster, but force others to be projected onto each other, effectively masking them. To avoid this situation, we propose to analyze a full set of $2p$ orthogonal directions, such that each direction minimizes or maximizes the kurtosis coefficient on a subspace "orthogonal" to all preceding directions. Once these directions have been computed, the observations are projected onto them, and the resulting $2p$ sets of univariate observations are analyzed to determine the existence of clusters of observations.

The criteria used to identify the clusters rely on the analysis of the sample spacings or first-order gaps between the ordered statistics of the projections. If the univariate observations come from a unimodal distribution, then the gaps should exhibit a very specific pattern, with large gaps near the extremes of the distribution and small gaps near the center. This pattern would be altered by the presence of clusters. For example, if two clusters are present, it should be possible to observe a group of large gaps separating the clusters, towards the center of the observations. Whenever these kinds of unusual patterns are detected, the observations are classified into groups by finding anomalously large gaps, and assigning the observations on different sides of these gaps to different groups. We now develop and formalize these ideas.

3.1 The Computation of the Projection Directions

Assume that we are given a sample of size n from a p-dimensional random variable $\mathbf{x}_i$, $i = 1, \ldots, n$. The projection directions $\mathbf{d}_k$ are obtained through the following steps. Start with $k = 1$, let $\mathbf{y}_i^{(1)} = \mathbf{x}_i$ and define

$$\bar{\mathbf{y}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i^{(k)},$$
$$\mathbf{S}_k = \frac{1}{(n-1)}\sum_{i=1}^{n}\big(\mathbf{y}_i^{(k)} - \bar{\mathbf{y}}^{(k)}\big)\big(\mathbf{y}_i^{(k)} - \bar{\mathbf{y}}^{(k)}\big)',$$

1. Find a direction $\mathbf{d}_k$ that solves the problem

$$\max k(\mathbf{d}_k) = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{d}_k' \mathbf{y}_i^{(k)} - \mathbf{d}_k'\bar{\mathbf{y}}_k\big)^4 \quad \text{s.t.} \quad \mathbf{d}_k'\mathbf{S}_k\mathbf{d}_k = 1, \tag{11}$$

   that is, a direction that maximizes the kurtosis coefficient of the projected data.
2. Project the observations onto a subspace that is $\mathbf{S}_k$-orthogonal to the directions $\mathbf{d}_1, \ldots, \mathbf{d}_k$. If $k < p$, define

$$\mathbf{y}_i^{(k+1)} = \left(\mathbf{I} - \frac{1}{\mathbf{d}_k'\mathbf{S}_k\,\mathbf{d}_k}\mathbf{d}_k\,\mathbf{d}_k'\mathbf{S}_k\right)\mathbf{y}_i^{(k)},$$

   let $k = k+1$ and compute a new direction by repeating step 1. Otherwise, stop.
3. Compute another set of p directions $\mathbf{d}_{p+1}, \ldots, \mathbf{d}_{2p}$ by repeating steps 1 and 2, except that now the objective function in (11) is minimized instead of maximized.

Several aspects of this procedure may need further clarification.

Remark 1. The optimization problem (11) normalizes the projection direction by requiring that the projected variance along the direction is equal to one. The motivation for this condition is twofold: it simplifies the objective function and its derivatives, as the problem is now reduced to optimizing the fourth central moment, and it preserves the affine invariance of the procedure. Preserving affine invariance would imply computing equivalent directions for observations that have been modified through an affine transformation. This seems a reasonable property for a cluster detection procedure, as the relative positions of these observations are not modified by the transformation, and as a consequence, the same clusters should be present for both the sets of data.

Remark 2. The sets of p directions that are obtained from either the minimization or the maximization of the kurtosis coefficient are defined to be $\mathbf{S}_k$-orthogonal to each other (rather than just orthogonal). This choice is again made to ensure that the algorithm is affine equivariant.

Remark 3. The computation of the projection directions as solutions of the minimization and maximization problems (11) represents the main computational effort incurred in the algorithm. Two efficient procedures can be used: (a) applying a modified version of Newton's method, or (b) solving directly the first-order optimality conditions for problem (11). As the computational efficiency of the procedure is one of its most important requirements, we briefly describe our implementation of both approaches.

1. The computational results shown later in this article have been obtained by applying a modified Newton method to (11) and the corresponding minimization problem. Taking derivatives in (11), the first-order optimality conditions for these problems are

$$\nabla k(\mathbf{d}) - 2\lambda S_k \mathbf{d} = 0,$$
$$\mathbf{d}'\mathbf{S}_k \mathbf{d} - 1 = 0.$$

Newton's method computes search directions for the variables $\mathbf{d}$ and constraint multiplier λ at the current estimates $(\mathbf{d}_l, \lambda_l)$ from the solution of a linear approximation for these conditions around the current iterate. The resulting linear system has the form

$$\begin{pmatrix} \mathbf{H}_l & 2S_k \mathbf{d}_l \\ 2\mathbf{d}_l'\mathbf{S}_k & 0 \end{pmatrix} \begin{pmatrix} \Delta \mathbf{d}_l \\ -\Delta\lambda_l \end{pmatrix} \begin{pmatrix} -\nabla k(\mathbf{d}_l) + 2\lambda_l \mathbf{S}_k \mathbf{d}_l \\ 1 - \mathbf{d}_l'\mathbf{S}_k \mathbf{d}_l \end{pmatrix},$$

where $\Delta \mathbf{d}_l$ and $\Delta\lambda_l$ denote the directions of movement for the variables and the multiplier respectively, and $\mathbf{H}_l$ is an approximation to $\nabla^2 L(\mathbf{d}_l, \lambda_l) \equiv \nabla^2 k(\mathbf{d}_l) - 2\lambda_l \mathbf{S}_k$, the Hessian of the Lagrangian function at the current iterate. To ensure convergence to a local optimizer, the variables are updated by taking a step along the search directions $\Delta \mathbf{d}_l$ and $\Delta\lambda_l$ that ensures that the value of an augmented Lagrangian merit function

$$k(\mathbf{d}_l) - \lambda_l\left(\mathbf{d}_l'\mathbf{S}_k \mathbf{d}_l - 1\right) + \frac{\rho}{2}\left(\mathbf{d}_l'\mathbf{S}_k \mathbf{d}_l - 1\right)^2,$$

decreases sufficiently in each iteration, for the minimization case. To ensure that the search directions are descent directions for this merit function, and a decreasing step can be taken, the matrix $\mathbf{H}_l$ is computed to be positive definite in the subspace of interest from a modified Cholesky decomposition of the reduced Hessian matrix $\mathbf{Z}_l'\nabla^2 L_l \mathbf{Z}_l$, where $\mathbf{Z}_l$ denotes a basis for the null-space of $\mathbf{S}_k \mathbf{d}_l$, see Gill, Murray, and Wright (1981) for additional details. It also may be necessary to adjust the penalty parameter ρ; in each iteration, if the directional derivative of the merit function is not sufficiently negative (again, for the minimization case), the penalty parameter is increased to ensure sufficient local descent. This method requires a very small number of iterations for convergence to a local solution, and we have found it to perform much better than other suggestions in the literature, such as the gradient and conjugate gradient procedures mentioned in Jones and Sibson (1987). In fact, even if the cost per iteration is higher, the total cost is much lower as the number of iterations is greatly reduced, and the procedure is more robust.

2. The second approach mentioned above is slightly less efficient, particularly when the sample space dimension p increases, although running times are quite reasonable for moderate sample space dimensions. It computes $\mathbf{d}_k$ by solving the system of nonlinear equations

$$4\sum_{i=1}^{n} (\mathbf{d}_k' \mathbf{y}_i^{(k)})^3 \mathbf{y}_i^{(k)} - 2\lambda\, \mathbf{d}_k = 0,$$
$$\mathbf{d}'\mathbf{d} = 1. \qquad (12)$$

These equations assume that the data have been standardized in advance, a reasonable first step given the affine equivariance of the procedure. From (12),

$$\sum_{i=1}^{n} (\mathbf{d}_k' \mathbf{y}_i^{(k)})^2 \mathbf{y}_i^{(k)} \mathbf{y}_i^{(k)\prime}\, \mathbf{d}_k = \frac{1}{2}\lambda\, \mathbf{d}_k,$$

implies that the optimal $\mathbf{d}$ is the unit eigenvector associated with the largest eigenvalue (the eigenvalue provides the corresponding value for the objective function) of the matrix

$$\mathbf{M}(\mathbf{d}) \equiv \sum_{i=1}^{n} \left(\mathbf{d}' \mathbf{y}_i^{(k)}\right)^2 \mathbf{y}_i^{(k)} \mathbf{y}_i^{(k)\prime},$$

that is, of a weighted covariance matrix for the sample, with positive weights (depending on $\boldsymbol{d}$). The procedure starts with an initial estimate for $\mathbf{d}_k$, $\mathbf{d}_0$, computes the weights based on this estimate and obtains the next estimate $\mathbf{d}_{l+1}$ as the eigenvector associated with the largest eigenvalue of the matrix $\mathbf{M}(\mathbf{d}_l)$. Computing the largest eigenvector is reasonably inexpensive for problems of moderate size (dimensions up to a few hundreds, for example), and the procedure converges at a linear rate (slower than Newton's method) to a local solution.

3. It is important to notice that the values computed from any of the two procedures are just local solutions, and perhaps not the global optimizers. From our computational experiments, as shown in a latter section, this

does not seem to be a significant drawback, as the computed values provide directions that are adequate for the study of the separation of the observations into clusters. Also, we have conducted other experiments showing that the proportion of times in which the global optimizer is obtained increases significantly with both the sample size and the dimension of the sample space.

3.2 The Analysis of the Univariate Projections

The procedure presented in this article assumes that a lack of clusters in the data implies that the data have been generated from a common unimodal multivariate distribution $F_p(\mathbf{X})$. As the procedure is based on projections, we must also assume that F is such that the distribution of the univariate random variable obtained from any projection $z = \mathbf{d}'\mathbf{X}$ is also unimodal. It is shown in Appendix B that this property holds for the class of multivariate unimodal distributions with a density that is a nonincreasing function of the distance to the mode, that is, $\nabla f(\mathbf{m}) = 0$ and if $(\mathbf{x}_1 - \mathbf{m})'\mathbf{M}(\mathbf{x}_1 - \mathbf{m}) \leq (\mathbf{x}_2 - \mathbf{m})'\mathbf{M}(\mathbf{x}_2 - \mathbf{m})$ for some definite positive matrix $\mathbf{M}$, then $f(\mathbf{x}_1) \geq f(\mathbf{x}_2)$. This condition is verified for instance by any elliptical distribution.

Once the univariate projections are computed for each one of the $2p$ projection directions, the problem is reduced to finding clusters in unidimensional samples, where these clusters are defined by regions of high-probability density. When the dimension of the data p is small, a promising procedure would be to estimate a univariate nonparametric density function for each projection and then define the number of clusters by the regions of high density. However, as the number of projections to examine grows with p, if p is large then it would be convenient to have an automatic criterion to define the clusters. Also, we have found that the allocation of the extreme points in each cluster depends very much on the choice of window parameter and there being no clear guide to choose it, we present in this article the results from an alternative approach that seems more useful in practice.

The procedure we propose uses the sampling spacing of the projected points to detect patterns that may indicate the presence of clusters. We consider that a set of observations can be split into two clusters when we find a sufficiently large first-order gap in the sample. Let $z_{ki} = \mathbf{x}_i'\mathbf{d}_k$ for $k = 1, \ldots, 2p$, and let $z_{k(i)}$ be the order statistics of this univariate sample. The first-order gaps or spacings of the sample, w_{ki}, are defined as the successive differences between two consecutive order statistics

$$w_{ki} = z_{k(i+1)} - z_{k(i)}, \quad i = 1, \ldots, n-1.$$

Properties of spacings or gaps can be found in Pyke (1965) and Read (1988). These statistics have been used for building goodness-of-fit tests (see for instance Lockhart, O'Reilly, and Stephens 1986) and for extreme values analysis (see Kochar and Korwar 1996), but they do not seem to have been used for finding clusters. As the expected value of the gap w_i is the difference between the expected values of two consecutive order statistics, it will be in general a function of i and the distribution of the observations. In fact, it is well known that when the data is a random sample from a distribution $F(x)$ with continuous density $f(x)$, the expected value of the ith sample gap is given by

$$E(w_i) = \binom{n}{i} \int_{-\infty}^{\infty} F(x)^i (1 - F(x))^{n-i}\, dx. \tag{13}$$

For instance, if f is an uniform distribution, then $E(w_i) = 1/(n+1)$ and all the gaps are expected to be equal, whereas if f is exponential then $E(w_i) = 1/(n-i)$ and the gaps are expected to increase in the tail of the distribution. In general, for a unimodal symmetric distribution, it is proved in Appendix C that the largest gaps in the sample are expected to appear at the extremes, w_1 and w_{n-1}, whereas the smallest ones should be those corresponding to the center of the distribution. Therefore, if the projection of the data onto $\mathbf{d}_k$ produces a unimodal distribution then we would expect the plot of w_{ki} with respect to k to decrease until a minimum is reached (at the mode of the distribution) and then to increase again. The presence of a bimodal distribution in the projection would be shown by a new decreasing of the gaps after some point. To further illustrate this behavior, consider a sample obtained from the projection of a mixture of three normal multivariate populations; this projection is composed of 200 observations, 50 of these observations have been generated from a univariate $N(-6, 1)$ distribution, another 50 are from a $N(6, 1)$ distribution, and the remaining 100 have been generated from a $N(0, 1)$. Figure 3.1(a) shows the histogram for this sample. Figure 3.1(b) presents the values of the gaps for these observations. Note how the largest gaps appear around observations 50 and 150, and these local maxima correctly split the sample into the three groups.

The procedure will identify clusters by looking at the gaps w_{ki} and determining if there are values that exceed a certain threshold. A sufficiently large value in these gaps would provide indication of the presence of groups in the data. As the distribution of the projections is, in general, not known in advance, we suggest defining these thresholds from a heuristic procedure. A gap will be considered to be significant if it has a very low probability of appearing in that position under a univariate normal distribution. As we see in our computational results, we found that this choice is sufficiently robust to cover a variety of practical situations, in addition to being simple to implement.

Before testing for a significant value in the gaps, we first standardize the projected data and transform these observations using the inverse of the standard univariate normal distribution function $\mathbf{\Phi}$. In this manner, if the projected data would follow a normal distribution, then the transformed data would be uniformly distributed. We can then use the fact that for uniform data, the spacings are identically distributed with distribution function $F(w) = 1 - (1 - w)^n$ and mean $1/(n+1)$, see Pyke (1965).

The resulting algorithm to identify significant gaps has been implemented as follows:

1. For each one of the directions $\mathbf{d}_k$, $k = 1, \ldots, 2p$, compute the univariate projections of the original observations $u_{ki} = \mathbf{x}_i'\mathbf{d}_k$.
2. Standardize these observations, $z_{ki} = (u_{ki} - m_k)/s_k$, where $m_k = \sum_i u_{ki}/n$ and $s_k = \sum_i (u_{ki} - m_k)^2/(n-1)$.

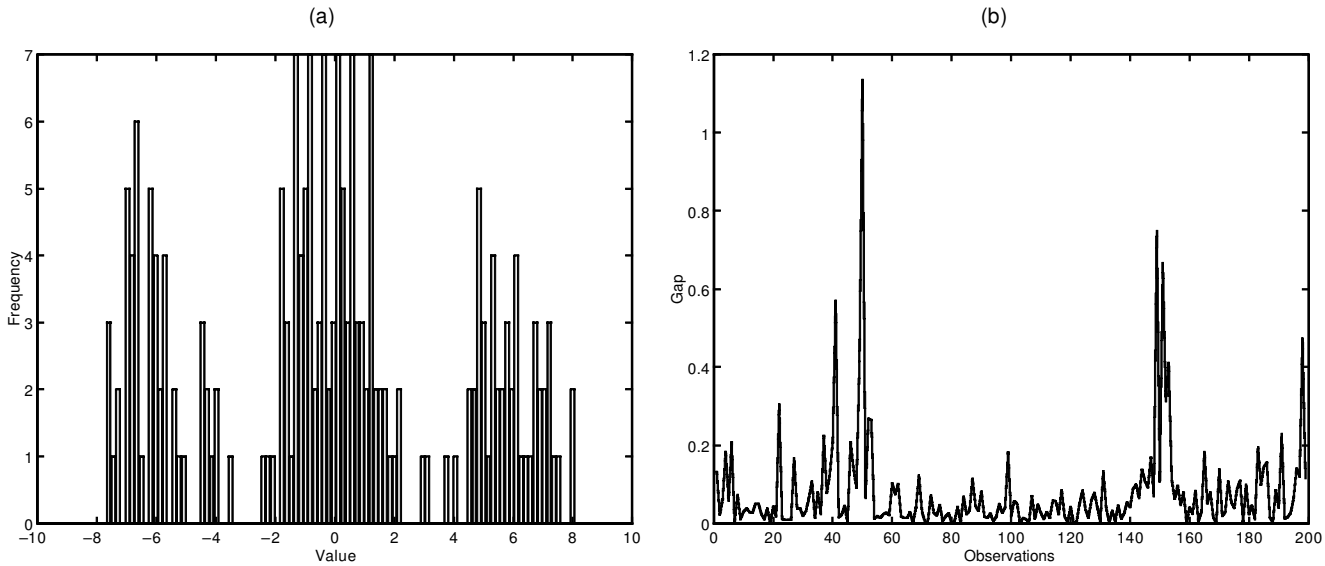

Figure 1. (a) Histogram for a Set of 200 Observations From Three Normal Univariate Distributions. (b) Gaps for the Set of 200 observations.

3. Sort out the projections z_{ki} for each value of k, to obtain the order statistics $z_{k(i)}$ and then transform using the inverse of the standard normal distribution function $\bar{z}_{ki} = \mathbf{\Phi}^{-1}(z_{k(i)})$.
4. Compute the gaps between consecutive values, $w_{ki} = \bar{z}_{k,i+1} - \bar{z}_{ki}$.
5. Search for the presence of significant gaps in w_{ki}. These large gaps will be indications of the presence of more than one cluster. In particular, we introduce a threshold $\kappa = \nu(c)$, where $\nu(c) = 1 - (1-c)^{1/n}$ denotes the cth percentile of the distribution of the spacings, define $i_{0k} = 0$ and
$$r = \inf_j \{n > j > i_{0k} : w_{kj} > \kappa\}.$$
If $r < \infty$, the presence of several possible clusters has been detected. Otherwise, go to the next projection direction.
6. Label all observations l with $\bar{z}_{kl} \le \bar{z}_{kr}$ as belonging to clusters different to those having $\bar{z}_{kl} > \bar{z}_{kr}$. Let $i_{0k} = r$ and repeat the procedure.

Some remarks on the procedure are in order. The preceding steps make use of a parameter c to compute the value $\kappa = \nu(c)$, that is used in step 5 to decide if more than one cluster is present. From our simulation experiments, we have defined $\log(1-c) = \log 0.1 - 10\log p/3$, and consequently $\kappa = 1 - 0.1^{1/n}/p^{10/(3n)}$, as this value works well on a wide range of values of the sample size n and sample dimension p. The dependence on p is a consequence of the repeated comparisons carried out for each of the $2p$ directions computed by the algorithm.

Also note that the directions $\mathbf{d}_k$ are a function of the data. As a consequence, it is not obvious that the result obtained in Appendix C applies here. However, according to Appendix B, the projections onto any direction of a continuous unimodal multivariate random variable will produce a univariate unimodal distribution. We have checked by Monte Carlo simulation that the projections of a multivariate elliptical distribution onto the directions that maximize or minimize the kurtosis coefficient have this property.

3.3 The Analysis of the Mahalanobis Distances

After completing the analysis of the gaps, the algorithm carries out a final step to assign observations within the clusters identified in the data. As the labeling algorithm, as described above, tends to find suspected outliers, but the projection directions are dependent on the data, it is reasonable to check if these observations are really outliers or just a product of the choice of directions. We thus test in this last step if they can be assigned to one of the existing clusters, and if some of the smaller clusters can be incorporated into one of the larger ones.

This readjustment procedure is based on standard multivariate tests using the Mahalanobis distance, see Barnett and Lewis (1978), and the procedure proposed by Peña and Tiao (2001) to check for data heterogeneity. It takes the following steps:

1. Determine the number of clusters identified in the data, k, and sort out these clusters by a descending number of observations (cluster 1 is the largest and cluster k is the smallest). Assume that the observations have been labeled so that observations $i_{l-1}+1$ to i_l are assigned to cluster l ($i_0 = 0$ and $i_k = n$).
2. For each cluster $l = 1, \ldots, k$ carry out the following steps:
   (a) Compute the mean $\mathbf{m}_l$ and covariance matrix $\mathbf{S}_l$ of the observations assigned to cluster l, if the number of observations in the cluster is at least $p+1$. Otherwise, end.
   (b) Compute the Mahalanobis distances for all observations not assigned to cluster l,
$$\delta_j = (\mathbf{x}_j - \mathbf{m}_l)'\mathbf{S}_l^{-1}(\mathbf{x}_j - \mathbf{m}_l), \quad j \le i_{l-1}, \quad j > i_l.$$

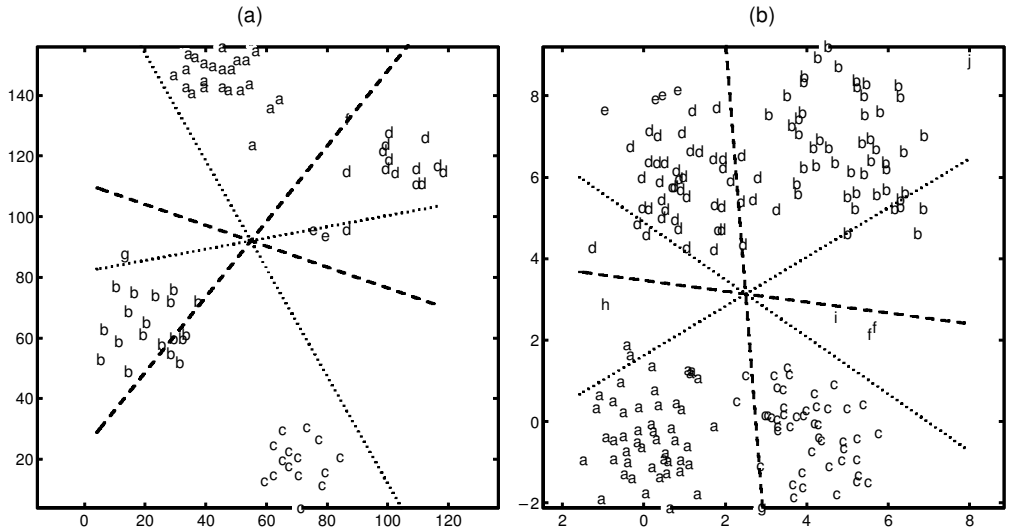

Figure 2. Plots Indicating the Original Observations, Their Assignment to Different Clusters, and the Projection Directions Used by the Algorithm for: (a) the Ruspini Example, and (b) the Maronna Example.

(c) Assign to cluster l all observations satisfying $\delta_j \leq \chi^2_{p,0.99}$.

(d) If no observations were assigned in the preceding step, increase l by one and repeat the procedure for the new cluster. Otherwise, relabel the observations as in step 1, and repeat this procedure for the same l.

4. COMPUTATIONAL RESULTS

We start by illustrating the behavior of the algorithm on some well-known examples from the literature, those of Ruspini (1970) and Maronna and Jacovkis (1974). Both cases correspond to two-dimensional data grouped into four clusters. Figure 2 shows the clusters detected by the algorithm for both the test problems, after two iterations of the procedure. Each plot represents the observations, labeled with a letter according to the cluster they have been assigned to. Also, the $2p = 4$ projection directions are represented in each plot. Note that the algorithm is able to identify every cluster present in all cases. It also tends to separate some observations from the clusters, observations that might be considered as outliers for the corresponding cluster.

The properties of the algorithm have been studied through a computational experiment on randomly generated samples. Sets of $20p$ random observations in dimensions $p = 4, 8, 15, 30$ have been generated from a mixture of k multivariate normal distributions. The number of observations from each distribution has been determined randomly, but ensuring that each cluster contains a minimum of $p+1$ observations. The means for each normal distribution are chosen as values from a multivariate normal distribution $N(0, f\,\mathbf{I})$, for a factor f (see Table 1) selected to be as small as possible whereas ensuring that the probability of overlapping between groups is roughly equal to 1%. The covariance matrices are generated as $\mathbf{S} = \mathbf{UDU}'$, using a random orthogonal matrix $\mathbf{U}$ and a diagonal matrix $\mathbf{D}$ with entries generated from a uniform distribution on $[10^{-3}, 5\sqrt{p}]$.

Table 2 gives the average percentage of the observations that have been labeled incorrectly, obtained from 100 replications for each value. When comparing the labels generated by the algorithm with the original labels, the following procedure has been used to determine if a generated label is incorrect: (i) we find those clusters in the original data having most observations in each of the clusters generated by the algorithm; (ii) we associate each cluster in the output data with the corresponding cluster from the original data, according to the preceding criterion, except when several clusters would be associated with the same original cluster; in this case only the largest cluster from the output data is associated with that original cluster; (iii) an observation is considered to be incorrectly labeled if it belongs to an output cluster associated with the wrong original cluster for that observation; (iv) as the data generating mechanism allows for some overlapping between clusters with small probability, the previous rule is only applied if for a given cluster in the output data the number of observations with a wrong label is larger than 5% of the size of that output cluster.

Table 1. Factors f Used to Generate the Samples for the Simulation Experiment

p	k	f
4	2	14
	4	20
	8	28
8	2	12
	4	18
	8	26
15	2	10
	4	16
	8	24
30	2	8
	4	14
	8	22

Table 2. Percentages of Mislabeled Observations for the Suggested Procedure, the k-means and Mclust Algorithms, and the Jones and Sibson Procedure (normal observations)

p	k	Kurtosis	k means	Mclust	J&S
4	2	.06	.36	.03	.19
	4	.09	.06	.07	.29
	8	.11	.01	.40	.30
8	2	.09	.40	.07	.25
	4	.10	.07	.15	.47
	8	.08	.01	.32	.24
15	2	.15	.53	.09	.30
	4	.32	.20	.25	.58
	8	.09	.04	.47	.27
30	2	.27	.65	.32	.33
	4	.60	.33	.61	.61
	8	.66	.28	.81	.74
Average		.22	.25	.30	.38

Table 4. Percentages of Mislabeled Observations for the Suggested Procedure, the k-means and Mclust Algorithms, and the Jones and Sibson Procedure (different overlaps between clusters)

	Kurtosis	k means	Mclust	J&S
Normal				
1% overlap	.09	.15	.17	.29
8% overlap	.15	.17	.22	.36
Uniform				
1% overlap	.05	.19	.12	.23
8% overlap	.07	.19	.13	.27
Student-t				
1% overlap	.14	.16	.19	.32
8% overlap	.19	.21	.23	.37

To provide better understanding of the behavior of the procedure, the resulting data sets have been analyzed using both the proposed method ("Kurtosis") and the k-means (see Hartigan and Wong, 1979) and Mclust (see Fraley and Raftery, 1999) algorithms as implemented in S-plus version 4.5. The rule used to decide the number of clusters in the k-means procedure has been the one proposed by Hartigan (1975, pp. 90–91). For the Mclust algorithm, it has been run with the option "VVV" (general parameters for the distributions). As an additional test on the choice of projection directions, we have implemented a procedure [column (Jones and Sibson) (J&S) in Table 2] that generates p directions using the Jones and Sibson (1987) projection pursuit criterion, although keeping all other steps from the proposed procedure. The Matlab codes that implement the Kurtosis algorithm, as described in this article, and the Jones and Sibson implementation are available for download at http://halweb.uc3m.es/fjp/download.html

As some of the steps in the procedure are based on distribution dependent heuristics, such as the determination of the cutoff for the gaps, we have also tested if these results would hold under different distributions in the data. The preceding experiment was repeated for the same data sets as above, with the difference that the observations in each group were generated from a multivariate uniform distribution and a multivariate Student-t distribution with p degrees of freedom. The corresponding results are shown in Table 3.

From the results in Tables 2 and 3, the proposed procedure behaves quite well, given the data used for the comparison. The number of mislabeled observations increases with the number of clusters for Mclust, whereas it decreases in general for k means. For kurtosis and J&S there is not a clear pattern because although in general the errors increase with the number of clusters and the dimension of the space, this is not always the case (see Tables 2, 3, and 5). It is important to note that, owing to the proximity between randomly generated groups, the generating process produces many cases where it might be reasonable to conclude that the number of clusters is lower than the value of k (this would help to explain the high rate of failure for all algorithms). The criterion based on the minimization and maximization of the kurtosis coefficient behaves better than the k means algorithm, particularly when the number of clusters present in the data is small. This seems to be mostly owing to the difficulty of deciding the number of clusters present in the data, and this difficulty is more marked when the actual number of clusters is small. On the other hand, the proposed method has a performance similar to that of Mclust, although it tends to do better when the number of clusters is large. Although for both algorithms there are cases in which the proposed algorithm does worse, it is important to note that it does better on the average than both of them,

Table 3. Percentages of Mislabeled Observations for the Suggested Procedure, the k-means and Mclust Algorithms, and the Jones and Sibson Procedure (uniform and student-t observations)

		Uniform				Student-t			
p	k	Kurtosis	k means	Mclust	J&S	Kurtosis	k means	Mclust	J&S
4	2	.05	.41	.01	.23	.10	.39	.04	.20
	4	.04	.13	.02	.21	.13	.15	.12	.28
	8	.07	.01	.41	.17	.16	.24	.41	.36
8	2	.02	.48	.02	.25	.09	.36	.11	.29
	4	.06	.12	.06	.43	.22	.11	.17	.44
	8	.05	.00	.18	.10	.13	.20	.32	.34
15	2	.08	.53	.01	.26	.16	.42	.10	.27
	4	.12	.12	.12	.53	.36	.16	.25	.57
	8	.06	.00	.36	.14	.16	.13	.51	.37
30	2	.21	.57	.09	.27	.28	.50	.30	.30
	4	.28	.18	.39	.60	.57	.14	.62	.62
	8	.07	.00	.65	.51	.70	.16	.80	.77
Average		.09	.21	.19	.31	.25	.25	.31	.40

Table 5. Percentages of Mislabeled Observations for the Suggested Procedure, the k-means and Mclust Algorithms, and the Jones and Sibson Procedure. Normal observations with outliers

p	k	Kurtosis	k means	Mclust	J&S
4	2	.06	.19	.08	.17
	4	.08	.06	.08	.23
	8	.11	.07	.41	.29
8	2	.05	.13	.11	.13
	4	.09	.05	.15	.43
	8	.09	.05	.40	.23
15	2	.05	.19	.12	.10
	4	.12	.10	.23	.53
	8	.13	.07	.51	.34
30	2	.03	.29	.11	.06
	4	.10	.21	.58	.44
	8	.55	.22	.77	.77
Average		.12	.14	.30	.31

and also that there are only 4 cases out of 36 where it does worse than both of them. It should also be pointed out that its computational requirements are significantly lower. Regarding the Jones and Sibson criterion, the proposed use of the directions minimizing and maximizing the kurtosis comes out as far more efficient in all these cases.

We have also analyzed the impact of increasing the overlapping of the clusters on the success rates. The values of the factors f used to determine the distances between the centers of the clusters have been reduced by 20% (equivalent to an average overlap of 8% for the normal case) and the simulation experiments have been repeated for the smallest cases (dimensions 4 and 8). The values in Table 4 indicate the average percentage of mislabeled observations both for the original and the larger overlap in these cases. The results show the expected increase in the error rates corresponding to the higher overlap between clusters, and broadly the same remarks apply to this case.

A final simulation study has been conducted to determine the behavior of the methods in the presence of outliers. For this study, the data have been generated as indicated above for the normal case, but 10% of the data are now outliers. For each cluster in the data, 10% of its observations have been generated as a group of outliers at a distance $4\chi^2_{p,0.99}$ in a group along a random direction, and a single outlier along another random direction. The observations have been placed slightly further away to avoid overlapping; the values of f in Table 1 have now been increased by two. Table 5 presents the numbers of misclassified observations in this case.

The results are very similar to those in Table 2, in the sense that the proposed procedure does better than k-means for small numbers of clusters, and better than Mclust when many clusters are present. It also does better than both procedures on the average. Again, the Jones and Sibson criterion behaves very poorly in these simulations. Nevertheless, the improvement in the k-means procedure is significant. It seems to be owing to its better performance as the number of clusters increases, and the fact that most of the outliers have been introduced as clusters. Its performance is not so good for the small number of isolated outliers.

APPENDIX A: PROOF OF THEOREM 1

To derive (4), note that $E(z)=(1-\alpha)m_1+\alpha m_2$ and $E(z^2)=(1-\alpha)m_1(2)+\alpha m_2(2)+(1-\alpha)m_1^2+\alpha {m_2}^2$; therefore $m_z(2)=E(z^2)-(E(z))^2=\sigma_m^2+\alpha(1-\alpha)\Delta^2$, where $\sigma_m^2=(1-\alpha)m_1(2)+\alpha m_2(2)$ and $\Delta=m_2-m_1$. The fourth moment is given by

$$m_z(4)=(1-\alpha)E_{f_1}\left[(z-m_1-\alpha\Delta)^4\right]+\alpha E_{f_2}\left[(z-m_2+(1-\alpha)\Delta)^4\right],$$

and the first integral is equal to $m_1(4)-4\alpha\Delta m_1(3)+6\alpha^2\Delta^2 m_1(2)+\alpha^4\Delta^4$, whereas the second is $m_2(4)+4(1-\alpha)\Delta m_2(3)+6(1-\alpha)^2\Delta^2 m_2(2)+(1-\alpha)^4\Delta^4$. Using these two results, we obtain that

$$\begin{aligned} m_z(4)=&(1-\alpha)m_1(4)+\alpha m_2(4)+4\alpha(1-\alpha)\Delta(m_2(3)\\ &-m_1(3))+6\alpha(1-\alpha)\Delta^2\tilde{\sigma}_m^2+\alpha(1-\alpha)\Delta^4(\alpha^3+(1-\alpha)^3). \end{aligned}$$

Consider now (6). From (4) we can write $\gamma_z(\mathbf{d})=N(\mathbf{d})/D(\mathbf{d})^2$, where $N(\mathbf{d})=m_z(4)$ and $D(\mathbf{d})=\sigma_m^2+\alpha(1-\alpha)\Delta^2$. Note that $D\neq 0$ unless both projected distributions are degenerate and have the same mean; we ignore this trivial case. We have

$$\begin{aligned} \nabla N=&(1-\alpha)\boldsymbol{\phi}_1+\alpha\boldsymbol{\phi}_2+4\alpha(1-\alpha)\Delta(\boldsymbol{\tau}_2-\boldsymbol{\tau}_1)\\ &+12\alpha(1-\alpha)\Delta^2(\alpha\mathbf{V}_1+(1-\alpha)\mathbf{V}_2)\mathbf{d}\\ &+4\alpha(1-\alpha)\big(m_2(3)-m_1(3)+3\Delta\tilde{\sigma}_m^2\\ &+(\alpha^3+(1-\alpha)^3)\Delta^3\big)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1),\\ \nabla D=&2((1-\alpha)\mathbf{V}_1+\alpha\mathbf{V}_2)\mathbf{d}+2\alpha(1-\alpha)\Delta(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1), \end{aligned}$$

and from the optimality condition $\nabla\gamma_z(\mathbf{d})=0$, for the optimal direction $\mathbf{d}$ we must have

$$\nabla N(\mathbf{d})=2\gamma_z(\mathbf{d})D(\mathbf{d})\nabla D(\mathbf{d}).$$

Replacing the expressions for the derivatives, this condition is equivalent to

$$\begin{aligned} &4(1-\alpha)(D\gamma_z-3\alpha^2\Delta^2)\mathbf{V}_1\mathbf{d}+4\alpha(D\gamma_z-3(1-\alpha)^2\Delta^2)\mathbf{V}_2\mathbf{d}\\ &\quad=(1-\alpha)\boldsymbol{\phi}_1+\alpha\boldsymbol{\phi}_2+4\alpha(1-\alpha)\\ &\qquad\times\big(\Delta(\boldsymbol{\tau}_2-\boldsymbol{\tau}_1)+\big(m_2(3)-m_1(3)\\ &\qquad+3\Delta\tilde{\sigma}_m^2+(\alpha^3+(1-\alpha)^3)\Delta^3-D\Delta\gamma_z\big)(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\big), \end{aligned}$$

and the result in (6) follows after substituting the value of D, dividing both sides by $4\sigma_m^2$ and regrouping terms.

APPENDIX B: PROJECTIONS OF UNIMODAL DENSITIES

Assume a random variable $\mathbf{X}$ with continuous unimodal density $f_X(\mathbf{x})$ with mode at $\mathbf{m}$. We show that its projections onto any direction $\mathbf{d}$, $\mathbf{d}'\mathbf{X}$, are also unimodal, provided that f_X is a nonincreasing function of the distance to the mode, that is, whenever $(\mathbf{x}_1-\mathbf{m})'\mathbf{M}(\mathbf{x}_1-\mathbf{m})\leq(\mathbf{x}_2-\mathbf{m})'\mathbf{M}(\mathbf{x}_2-\mathbf{m})$ for some positive definite matrix $\mathbf{M}$, then $f_X(\mathbf{x}_1)\geq f_X(\mathbf{x}_2)$.

To simplify the derivation and without loss of generality we work with a random variable $\mathbf{Y}$ satisfying the preceding properties for $\mathbf{m}=0$ and $\mathbf{M}=\mathbf{I}$. Note that the projections of $\mathbf{X}$ would be unimodal if and only if the projections of $\mathbf{Y}=\mathbf{M}^{1/2}(\mathbf{X}-\mathbf{m})$ are unimodal. This statement follows immediately from $\mathbf{d}'\mathbf{X}=\mathbf{d}'\mathbf{m}+\mathbf{d}'\mathbf{M}^{-1/2}\mathbf{Y}$, implying the equivalence of the two densities, except for a constant.

From our assumption we have $f_Y(\mathbf{y}_1)\geq f_Y(\mathbf{y}_2)$ whenever $\|\mathbf{y}_1\|\leq\|\mathbf{y}_2\|$; note that this property implies that $f_Y(\mathbf{y})=\varphi(\|\mathbf{y}\|)$, that is, the density is constant on each hypersphere with center as the origin.

As a consequence, for any projection direction $\mathbf{d}$, the density function of the projected random variable, $z = \mathbf{d}'\mathbf{Y}$, will be given by

$$f_z(z)\,dz = \int_{z\le \mathbf{d}'\mathbf{y}\le z+dz} f_Y(\mathbf{y})\,d\mathbf{y} = \int_{z\le w_1\le z+dz} f_Y(\mathbf{U}'\mathbf{w})\,d\mathbf{w},$$

where we have introduced the change of variables $\mathbf{w} = \mathbf{U}\mathbf{y}$ for an orthogonal matrix $\mathbf{U}$ such that $\mathbf{d} = \mathbf{U}'\mathbf{e}_1$, where $\mathbf{e}_1$ denotes the first unit vector, and $\mathbf{d}'\mathbf{y} = \mathbf{e}_1'\mathbf{U}\mathbf{y} = \mathbf{e}_1'\mathbf{w} = w_1$. Also note that $f_Y(\mathbf{U}'\mathbf{w}) = \varphi(\|\mathbf{w}\|) = f_Y(\mathbf{w})$, and as a consequence the density of z will be given by

$$f_z(z) = \int_D f_Y(z, w_2, \ldots, w_p)\,dw_2 \ldots dw_p,$$

where the integration domain D is the set of all possible values of $w_2, \ldots, w_p$. As for any fixed values of $w_2, \ldots, w_p$, we have $f_Y(z_1, w_2, \ldots, w_p) \ge f_Y(z_2, w_2, \ldots, w_p)$ for any $|z_1| \le |z_2|$, it follows that

$$\begin{aligned} f_z(z_1) &= \int_D f_Y(z_1, w_2 \ldots w_p)\,dw_2 \ldots dw_p \\ &\ge \int_D f_Y(z_2, w_2, \ldots, w_p)\,dw_2 \ldots dw_p \\ &= f_z(z_2), \end{aligned}$$

for any $|z_1| \le |z_2|$, proving the unimodality of f_z.

APPENDIX C: PROPERTIES OF THE GAPS FOR SYMMETRIC DISTRIBUTIONS

We now justify the statement that for a unimodal symmetric distribution the largest gaps in the sample are expected to appear at the extremes. Under the symmetry assumption, and using (13) for the expected value of the gap, we would need to prove that for $i > n/2$,

$$E(w_{i+1}) - E(w_i) = \frac{n+1}{i+1}\binom{n}{i}\int_{-\infty}^{\infty} F(x)^i(1-F(x))^{n-i-1} \times \left(F(x) - \frac{i+1}{n+1}\right)dx \ge 0,$$

Letting $g(x) \equiv F(x)^i(1-F(x))^{n-i-1}\big(F(x) - (i+1)/(n+1)\big)$ this is equivalent to proving that

$$\int_{-\infty}^{\infty} g(x)\,dx \ge 0. \tag{C.1}$$

To show that this inequality holds, we use the following property of the Beta function: for any i,

$$\frac{1}{n+1} = \binom{n}{i}\int_{-\infty}^{\infty} F(x)^i(1-F(x))^{n-i} f(x)\,dx.$$

Taking the difference between the integrals for $i+1$ and i, it follows that

$$\begin{aligned} 0 &= \frac{n+1}{i+1}\binom{n}{i}\int_{-\infty}^{\infty} g(x)f(x)\,dx \\ &\Leftrightarrow \quad \int_{-\infty}^{\infty} g(x)f(x)\,dx = 0. \end{aligned} \tag{C.2}$$

This integral is very similar to the one in (C.1), except for the presence of $f(x)$. To relate the values of both integrals, the integration interval $(-\infty, \infty)$ will be divided into several zones. Let $a = F^{-1}((i+1)/(n+1))$, implying that $F(x) - (i+1)/(n+1) \le 0$ and $g(x) \le 0$ for all $x \le a$. As we have assumed the distribution to be symmetric and unimodal, and without loss of generality, we may suppose the mode to be at zero, the density will satisfy $f(x) \ge f(a)$ for any $x \in [-a, a]$, and $f(x) \le f(a)$ for $x \in (-\infty, -a]$ and $x \in [a, \infty)$. As a consequence,

$$\int_{-a}^{a} g(x)\frac{f(x)}{f(a)}\,dx \le \int_{-a}^{a} g(x)\,dx. \tag{C.3}$$

To find similar bounds for the integral in the intervals $(-\infty, -a]$ and $[a, \infty)$ we introduce the change of variables $y = -x$ and use the symmetry of the distribution to obtain the equivalent representation

$$\int_{-\infty}^{-a} g(x)\frac{f(x)}{f(a)}\,dx = -\int_{a}^{\infty} F(x)^{n-i-1}(1-F(x))^i \times \left(F(x) - 1 + \frac{i+1}{n+1}\right)\frac{f(x)}{f(a)}\,dx.$$

From this equation it will hold that

$$\int_{-\infty}^{\infty} g(x)\frac{f(x)}{f(a)}\,dx = \int_{-a}^{a} g(x)\frac{f(x)}{f(a)}\,dx + \int_{a}^{\infty} h(x)\frac{f(x)}{f(a)}\,dx, \tag{C.4}$$

where

$$\begin{aligned} h(x) &\equiv g(x) - F(x)^{n-i-1}(1-F(x))^i\left(F(x) - 1 + \frac{i+1}{n+1}\right) \\ &= F(x)^i(1-F(x))^{n-i-1}\left(F(x) - \frac{i+1}{n+1}\right. \\ &\quad \left. - \left(\frac{1-F(x)}{F(x)}\right)^{2i+1-n}\left(F(x) - 1 + \frac{i+1}{n+1}\right)\right). \end{aligned}$$

If $i > n/2$, it holds that $h(a) < 0$, then the function has a zero at $b \in [a, \infty)$, and this zero is unique in the interval. As f is decreasing on $[a, \infty)$, $h(x) \le 0$ for $a \le x \le b$ and $h(x) \ge 0$ for $x \ge b$, it must follow that

$$\begin{aligned} \int_a^b h(x)\,dx &\ge \int_a^b h(x)\frac{f(x)}{f(b)}\,dx, \\ \int_b^{\infty} h(x)\,dx &\ge \int_b^{\infty} h(x)\frac{f(x)}{f(b)}\,dx \\ \Rightarrow \int_a^{\infty} h(x)\,dx &\ge \int_a^{\infty} h(x)\frac{f(x)}{f(b)}\,dx. \end{aligned}$$

This inequality together with (C.4), (C.3), and (C.2) yield

$$\int_{-\infty}^{\infty} g(x)\,dx \ge \int_{-\infty}^{\infty} g(x)\frac{f(x)}{f(a)}\,dx = 0,$$

and this bound implies (C.1) and the monotonicity of the expected gaps.

[Received July 1999. Revised December 2000.]

REFERENCES

Anderson, T. W., and Bahadur, R. R. (1962), "Classification Into Two Multivariate Normal Distributions With Different Covariance Matrices," *Annals of Mathematical Statistics,* 33, 420–431.

Balanda, K. P., and MacGillivray, H. L. (1988), "Kurtosis: A Critical Review," *The American Statistician,* 42, 111–119.

Banfield, J. D., and Raftery, A. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics,* 49, 803–821.

Barnett, V., and Lewis, T. (1978) *Outliers in Statistical Data*, New York: Wiley.

Bensmail, H., and Celeux, G. (1997), "Inference in Model-Based Cluster Analysis," *Statistics and Computing,* 7, 1–10.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika,* 65, 31–38.

Celeux, G., Hurn, M., and Robert, C. P. (2000), "Computational and Inferencial Difficulties With Mixture Posterior Distributions," *Journal of the American Statistical Association,* 95, 957–970.

Cook, D., Buja, A., Cabrera, J., and Hurley, C. (1995), "Grand Tour and Projection Pursuit," *Journal of Computational and Graphical Statistics,* 4, 155–172.

Dasgupta, A., and Raftery, A. E. (1998), "Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering," *Journal of the American Statistical Association,* 93, 294–302.

Fraley, C., and Raftery, A. E. (1999), "MCLUST: Software for Model-Based Cluster Analysis," *Journal of Classification,* 16, 297–306.

Friedman, H. P., and Rubin, J. (1967), "On some Invariant Criteria for Grouping Data," *Journal of the American Statistical Association,* 62, 1159–1178.

Friedman, J. H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association,* 82, 249–266.

Friedman, J. H., and Tukey, J. W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers,* C-23, 881–889.

Gill, P. E., Murray, W., and Wright, M. H. (1981), *Practical Optimization*, New York: Academic Press.

Gordon, A. D. (1994), "Identifying Genuine Clusters in a Classification," *Computational Statistics and Data Analysis,* 18, 561–581.

Hardy, A. (1996), "On the Number of Clusters," *Computational Statistics and Data Analysis,* 23, 83–96.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

Hartigan, J. A., and Wong, M. A. (1979), "A k-means Clustering Algorithm," *Applied Statistics,* 28, 100–108.

Huber, P. J. (1985), "Projection Pursuit," *The Annals of Statistics,* 13, 435–475.

Jones, M. C., and Sibson, R. (1987), "What Is Projection Pursuit?," *Journal of the Royal Statistical Society*, Series A, 150, 1–18.

Justel, A., and Peña, D. (1996), "Gibbs Sampling Will Fail in Outlier Problems With Strong Masking," *Journal of Computational and Graphical Statistics,* 5, 176–189a.

Kochar, S. C., and Korwar, R. (1996), "Stochastic Orders for Spacings of Heterogeneous Exponential Random Variables," *Journal of Multivariate Analysis,* 57, 69–83.

Lavine, M., and West, M. (1992), "A Bayesian Method for Classification and Discrimination," *Canadian Journal of Statistics,* 20, 451–461.

Lockhart, R. A., O'Reilly, F. J., and Stephens, M. A. (1986), "Tests of Fit Based on Normalized Spacings," *Journal of the Royal Statistical Society,* Ser. B, *Methodological,* 48, 344–352.

Maronna, R., and Jacovkis, P. M. (1974), "Multivariate Clustering Procedures with Variable Metrics," *Biometrics,* 30, 499–505.

Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Theory*, New York: Wiley.

Nason, G. (1995), "Three-Dimensional Projection Pursuit," *Applied Statistics,* 44, 411–430.

Peña, D., and Prieto, F. J. (2000), "The Kurtosis Coefficient and the Linear Discriminant Function," *Statistics and Probability Letters,* 49, 257–261.

——— (2001), "Robust Covariance Matrix Estimation and Multivariate Outlier Detection," *Technometrics,* 43, 3, 286–310.

Peña, D., and Tiao, G. C. (2001), "The SAR Procedure: A Diagnostic Analysis of Heterogeneous Data," (manuscript).

Posse, C. (1995), "Tools for Two-Dimensional Exploratory Projection Pursuit," *Journal of Computational and Graphical Statistics,* 4, 83–100.

Pyke, R. (1965), "Spacings" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, *Methodological,* 27, 395–449.

Read, C. B. (1988), "Spacings," in *Encyclopedia of Statistical Sciences,* (Vol. 8), 566–569.

Ruspini, E. H. (1970), "Numerical Methods for Fuzzy Clustering," *Information Science,* 2, 319–350.

Seber, G. A. F. (1984), *Multivariate Observations*, New York: Wiley.

Stephens, M. (2000) "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society,* Ser. B, 62, 795–809.

Switzer, P. (1985), Comments on "Projection Pursuit," by P. J. Huber, *The Annals of Statistics*, 13, 515–517.