UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper
Statistics and Econometrics
17-04
ISSN 2387-0303
Abril 2017

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

CLUSTERING BIG DATA BY EXTREME KURTOSIS PROJECTIONS

Daniel Peña[a], Javier Prieto[b] and Carolina Rendón[c]

Abstract

Clustering Big Data is an important problem because large samples of many variables are usually heterogeneous and include mixtures of several populations. It often happens that only some of a large set of variables are useful for clustering and working with all of them would be very inefficient and may make more difficult the identification of the clusters. Thus, searching for spaces of lower dimension that include all the relevant information about the clusters seems a sensible way to proceed in these situations. Peña and Prieto (2001) showed that the extreme kurtosis directions of projected data are optimal when the data has been generated by mixtures of two normal distributions. We generalize this result for any number of mixtures and show that the extreme kurtosis directions of the projected data are linear combinations of the optimal discriminant directions if we knew the centers of the components of the mixture. In order to separate the groups we want directions that split the data into two groups, each corresponding to different components of the mixture. We prove that these directions can be found from extreme kurtosis projections. This result suggests a new procedure to deal with many groups, working in a binary decision way and deciding at each step if the data should be split into two groups or we should stop. The decision is based on comparing a single distribution with a mixture of two distribution. The performance of the algorithm is analyzed through a simulation study.

Keywords: High dimension; Projection Pursuit; Mixture models.

[a] Peña, Daniel, Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain. e-mail: dpena@est-econ.uc3m.es.
[b] Prieto, Javier, Department of Statistics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain. e-mail: fjp@est-econ.uc3m.es.
[c] Rendón, Carolina, Department of Statistics, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganés (Madrid), Spain. e-mail: jrendon@est-econ.uc3m.es.

Clustering Big Data by Extreme Kurtosis Projections

Daniel Peña, Javier Prieto and Carolina Rendón

April 27, 2017

Abstract

Clustering Big Data is an important problem because large samples of many variables are usually heterogeneous and include mixtures of several populations. It often happens that only some of a large set of variables are useful for clustering and working with all of them would be very inefficient and may make more difficult the identification of the clusters. Thus, searching for spaces of lower dimension that include all the relevant information about the clusters seems a sensible way to proceed in these situations. Peña and Prieto (2001) showed that the extreme kurtosis directions of projected data are optimal when the data has been generated by mixtures of two normal distributions. We generalize this result for any number of mixtures and show that the extreme kurtosis directions of the projected data are linear combinations of the optimal discriminant directions if we knew the centers of the components of the mixture. In order to separate the groups we want directions that split the data into two groups, each corresponding to different components of the mixture. We prove that these directions can be found from extreme kurtosis projections. This result suggests a new procedure to deal with many groups, working in a binary decision way and deciding at each step if the data should be split into two groups or we should stop. The decision is based on comparing a single distribution with a mixture of two distribution. The performance of the algorithm is analyzed through a simulation study.

Key words: High dimension; Projection Pursuit; Mixture models.

1 Introduction

The classification of observations is a basic problem that occurs in many disciplines. The increasing availability of large sets of data with many variables and observations, which are expected to originate from mixtures of different populations, requires cluster procedures able to work in this large data set. Many useful procedures are available for clustering. Partitioning algorithms such as K-Means, see MacQueen (1967), PAM or K-Medoids, see Kaufman and Rousseeuw (1990), and MCLUST, Banfield and Raftery (1993) are very popular for small data sets but all of them have limitations with large data sets with many variables and observations. An alternative, as in the CLARA algorithm to apply PAM in larger data sets (Kaufman and Rousseeuw, (1990)), apply the procedure to several samples from the data and select the best solution, but the problem of many variables is not addressed. Hierarchical methods, see for instance Everit (1993) are also useful but they need to be adapted for Big Data.

In this paper we focus on two problems that high dimensionality presents in clustering. First, the presence of irrelevant attributes, because they negatively affect proximity measures. Second, the dimensionality curse, that is a lack of data separation in high dimensional space. In order to solve this problem, two main approaches have been used. The first one is variable selection and the second one is dimension reduction. Variable selection can be made by using some penalty function, such as the Lasso method. For instance in model-based clustering we can maximize the likelihood of the mixture of normals adding some penalty function in order to introduce variable selection (see Pan and Shen (2007) and Wang and Zhu (2008)). Also, we can select variables as a model selection problem, as proposed by Raftery and Dean (2006) and generalized by Maugis et al. (2009). Other variable selection approaches are due to Steinley and Brusco (2008), who introduce measures of the ability of each variable to detect a fixed number of clusters, and to Fraiman et al. (2008), who propose a method to detect the noninformative variables in clustering. Witten and Tibshirani (2010) developed a cluster algorithm that can be applied to obtain sparse versions of K-means and hyerarchical clustering. Some comparison of these methods and other related references can be found in Galimberti et al. (2017) and Bouveyron and Brunet (2014) present a review of model-based clustering for high-dimensional data.

The second approach is dimensionality reduction methods, where we try to identify some relevant subspace which includes the relevant information for clustering. Several articles have proposed building this subspace using principal components. However, Chang (1983) showed that the components with large eigenvalues may not be useful to separate the groups, see also Peña et al. (2010). A more general approach to space selection is projection pursuit, Friedman and Tukey (1974), where "interesting" projections of multidimensional data are analyzed in order to show the cluster structure. Peña and Prieto (2001) showed that projections onto directions with extreme kurtosis of the projected data can be optimal to reveal the cluster structure, and described a procedure to identify clusters in multivariate data using information obtained from the univariate projections of the sample data on the directions that minimize and maximize the kurtosis coefficient of the projected data. The clustering algorithm proposed by these authors is based on the analysis of a set of $2p$ orthogonal directions for a p-dimensional random variable, such that each direction minimizes or maximizes the kurtosis coefficient. The criteria used to identify the clusters is based on the sample spacings or first-order gaps between the ordered statistics of the projections. This method works well when the data dimension is low and when the number of groups present in the sample is small, but may fail when the data dimension increases.

In this article we propose three modifications of the algorithm proposed by Peña and Prieto (2001). First, in addition to the directions of extreme kurtosis we add random directions computed by the modified Stahel Donoho procedure proposed by Peña and Prieto (2007). Second, instead of using the gap statistics to find groups in the univariate projections we fit a mixture of two normals and test using the BIC criterion for the presence of two or more distributions. Third, the algorithm works in a binary-decision way and decide at each step if the data should be split into two groups or we should stop.

The paper is structured as follows: Section 2 reviews the use of kurtosis coefficient for clustering and proves that if the data has been generated by a mixture of normal distributions with the same covariance matrices the extreme directions of the kurtosis coefficients span the space generated by the differences between pairs of means that are the optimal directions for discrimination. In Section 3 we prove first that there exits directions that projects all the observations into two groups and that these directions can be found by the extreme directions of the kurtosis coefficient. In Section 4, a cluster identification algorithm for high-dimensional data

with several clusters which is based on the previous results is presented. Section 5 presents some examples and computational results, and the proposed algorithm is compared with the clustering algorithm proposed by Peña and Prieto (2001a). We finish with some remarks and conclusions in Section 6.

2 Extreme Projected Kurtosis as optimal directions for discrimination

In symmetrical univariate models, the kurtosis is a measure of the peakedness of the probability distribution of a real-valued random variable. Large values indicate heavy tails or outliers whereas small values indicate bimodality in the data (Darlington, 1970). For the multivariate case, Mardia (1970) proposed a scalar value for the kurtosis coefficient as the second moment of the Mahalanobis distances. A kurtosis matrix was introduced by Cardoso (1989) and Móri et al. (1993) for a random vector X is $K = \mathbb{E}(Z^T Z Z Z^T)$, where $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$ is the covariance matrix and $Z = \Sigma^{-1/2}(X - \mathbb{E}[X])$ denotes the corresponding standardized vector.
Both in the univariate case and in the multivariate case, some proposals have been described in the literature related to the use of kurtosis for outlier detection and as a measure of heterogeneity, see Schwager and Margolin (1982), Peña and Prieto (2001a), Peña and Prieto (2001b), Peña et al. (2010) and Jobson (2012).
We are interested in studying the behavior of the kurtosis coefficient when we have a p-dimensional variable corresponding to a mixture of k normal distributions with the same covariance matrix. Let X be the p-dimensional random variable such that $k \leq p+1$

$$X \sim \alpha_1 N(\mu_1, \Sigma) + \cdots + \alpha_k N(\mu_k, \Sigma), \quad X \in \mathbb{R}^p,$$

where $\mathbb{E}(X) \equiv \bar{\mu} = \sum_{i=1}^k \alpha_i \mu_i$ and $\sum_{i=1}^k \alpha_i = 1$. We assume that $\alpha_i \neq 0$ for all i, as otherwise we could study an equivalent mixture having less than k components. In what follows we will also assume that the following condition holds:

A1. The vectors $\{\mu_i - \mu_k\}_{i=1}^{k-1}$ are linearly independent.

Consider now an arbitrary direction d, with $\|d\| = 1$, and the univariate projection $z = d^T X$, with distribution

$$z \sim \alpha_1 N(m_1, s) + \cdots + \alpha_k N(m_k, s), \quad z \in \mathbb{R} \tag{1}$$

where $s = d^T \Sigma d$, $m_i = d^T \mu_i$ and $\mathbb{E}(z) = \sum_{i=1}^k \alpha_i m_i$.
Our interest is to study those directions d that have information of interest for the detection of heterogeneity in the data X, by revealing this information in the univariate projections z. In particular, we are interested in considering the directions obtained as extreme points for the kurtosis coefficient.
Therefore, our function of interest is then the coefficient of univariate kurtosis defined as

$$\kappa_z = \frac{m_z(4)}{m_z(2)^2} \tag{2}$$

where $m_z(k) = \mathbb{E}[(z - \mathbb{E}(z))^k]$. Considering the univariate projection of the data given by (1), it holds that

$$m_z(2) = \mathbb{E}[(z - \mathbb{E}(z))^2] = s^2 + v_2,$$

where

$$v_2 \equiv \sum_{i=1}^k \alpha_i (m_i - \mathbb{E}(z))^2,$$

is the variance of the projected means. Also,

$$m_z(4) = \mathbb{E}[(z - \mathbb{E}(z))^4] = 3s^4 + 6s^2 v_2 + v_4,$$

where

$$v_4 \equiv \sum_{i=1}^k \alpha_i (m_i - \mathbb{E}(z))^4,$$

is the kurtosis of the projected means.
Then the kurtosis coefficient can be written as,

$$\kappa_z(d) = \frac{3s^4 + 6s^2 v_2 + v_4}{(s^2 + v_2)^2}$$

Note that in the function $\kappa_z(d)$ the arguments v_2 and v_4 are not completely independent but one is not a function of the other. We can write

$$v_4 = \sum_{i=1}^{k} \alpha_i (m_i^4 - 4\mathbb{E}(z) m_i^3 + 3\mathbb{E}(z)^4) + 6\mathbb{E}(z)^2 v_2,$$

that shows that the kurtosis of the projected means depends on the variance of the projected means, but also of the asymmetry on the distribution of the projected means.

Theorem 1 *The stationary points of the problem*

$$\begin{array}{ll} \min_d & \kappa_z(d) \\ s.t. & d^T d = 1 \end{array} \tag{3}$$

satisfy $d \in span\{\mu_i - \mu_k\}$.

Proof To solve this problem we have to study the Lagrangian and the derivatives of the $\kappa_z(d)$ function are

$$\frac{\partial \kappa_z(v_2, v_4)}{\partial v_2} = \frac{-2(3s^2 v_2 + v_4)}{(s^2 + v_2)^3} \equiv A,$$

and

$$\frac{\partial \kappa_z(v_2, v_4)}{\partial v_4} = \frac{1}{(s^2 + v_2)^2} \equiv B$$

The derivatives satisfy

$$\nabla_d \mathcal{L}(d, \lambda) = A \nabla_d v_2 + B \nabla_d v_4 - 2\lambda d,$$

where

$$\nabla_d v_2 = 2 \sum_{i=1}^{k} \alpha_i (m_i - \mathbb{E}(z))(\mu_i - \mathbb{E}(X))$$

$$\nabla_d v_4 = 4 \sum_{i=1}^{k} \alpha_i (m_i - \mathbb{E}(z))^3 (\mu_i - \mathbb{E}(X))$$

Thus, the stationary points satisfy

$$d = \sum_{i=1}^{k} c_i (\mu_i - \bar{\mu}), \tag{4}$$

for $c_i = \frac{1}{\lambda} \alpha_i (m_i - \mathbb{E}(z))(A + 2B(m_i - \mathbb{E}(z))^2)$.
As a consequence, any stationary point d is a linear combination of the vectors $\{\mu_i - \bar{\mu}\}$. Note that this is also valid if $\lambda = 0$.
Finally, as $\mu_i - \bar{\mu} = \mu_i - \mu_k - \sum_{j=1}^{k-1} \alpha_j (\mu_j - \mu_k)$, it holds that $d = \sum_{i=1}^{k-1} \bar{c}_i (\mu_i - \mu_k)$, for $\bar{c}_i = c_i - \alpha_i \sum_{j=1}^{k} c_j$, the desired result. □

From Theorem 1, it holds that there exists an optimal direction d in the subspace generated by $\{\mu_1 - \mu_k, \ldots, \mu_{k-1} - \mu_k\}$.
We have shown in this section that the extreme directions of the kurtosis generates the same space as the optimal directions for discrimination for a mixture of normal distributions with the same covariance matrix. In the next section we analyze the use of this directions to find clusters

3 Interesting Projection Directions

We are interested in the study of directions that would allow the detection of the different groups present in the mixture from the study of the univariate projections of the observations. Suppose that we can find directions where the projected data appear in two separated groups, each one corresponding to a subset of the components in the mixture.. Then a iterative binary separation would be possible and we may have a powerful procedure for many groups. These directions would satisfy

$$\begin{aligned} d^T\mu_i &= V > 0, \quad i \in I_1 \\ d^T\mu_i &= 0, \qquad\quad i \in I_2, \end{aligned} \tag{5}$$

where $d^T d = 1$ for some value V, where I_1 and I_2 denote a partition of the labels $\{1, \ldots, k\}$. These directions would help to separate the groups associated with I_1 from the groups associated with I_2, as long as these groups are sufficiently separated, that is, whenever V is large enough. The value V can be written in terms of the vectors μ_i, and it is a property of the geometry of these centers.
The following result proves that the directions given in (5) exist, and that there is a unique such direction in the subspace spanned by $\{\mu_i - \mu_k\}$.

Lemma 1 *Under condition* A1*, the directions d defined in (5) always exist and are unique on span$\{\mu_i - \mu_k\}_{i=1}^{k-1}$ for any partition (I_1, I_2).*

Proof We consider directions d defined as a linear combination of the vectors $\{\mu_i - \mu_k\}$,

$$d = \sum_{i=1}^{k-1} \gamma_i(\mu_i - \mu_k) = M\gamma, \tag{6}$$

for $M \in \mathbb{R}^{p\times(k-1)}$, a full-rank matrix with columns corresponding to the vectors $\mu_i - \mu_k$, and $\gamma \in \mathbb{R}^{k-1}$.
Assume that $k \in I_2$ (otherwise exchange I_1 with I_2 and d with $-d$); then $d^T\mu_i = d^T(\mu_i - \mu_k)$.
As V is arbitrary, we can write the conditions in (5) as a system of equations of the form $N\gamma = 0$, where

$$N_{ij} = \begin{cases} (\mu_i - \mu_k)^T(\mu_j - \mu_k) & \text{if } i \in I_2,\ j = 1, \ldots, k-1, \\ (\mu_i - \mu_l)^T(\mu_j - \mu_k) & \text{if } i \in I_1 \backslash \{l\},\ j = 1, \ldots, k-1, \end{cases} \tag{7}$$

for some $l \in I_1$. Note that $N \in \mathbb{R}^{(k-2)\times(k-1)}$ and under assumption *A1*, it has full row rank, $k-2$.
From the property that the span of N^T and the null space of N are orthogonal complements of $\mathbb{R}^{k-1}$, and as $\dim(\text{span}(N^T)) = k-2$, it holds that $k-1 = \dim(\text{span}(N^T)) + \dim(\text{null}(N))$, implying $\dim(\text{null}(N)) = 1$. Thus, there exist two vectors satisfying $d^T d = 1$ with $N\gamma = 0$, and one of these vectors is such that $(\mu_l - \mu_k)^T d > 0$, completing the proof. □

Now we will show the relationship between these interesting projection directions and the extreme points of the kurtosis coefficient of the projected data. We will consider first some criteria which include the kurtosis coefficient and study the behavior of the family of optimality criteria for the direction d, which can be written as rational functions of the values $m_i \equiv d^T\mu_i$.
We will consider a representation for these criteria as

$$\tau_z(d) \equiv \frac{t_n(m) + q_n(m)}{t_d(m) + q_d(m)}, \tag{8}$$

where t_n and t_d are polynomials of degree r in the values m_i, for a positive integer $r > 1$, and q_n and q_d are polynomials of degree g at most, with $g < r$. Note that this representation is not unique, just like the polynomials t and q and be selected in many different ways.
Not any criteria $\tau_z(d)$ will provide directions related to the optimization of (5), but we will show that the desired relationship will hold if the following condition holds, at least for one of the possible representations of $\tau_z(d)$ according to (8):

A2. Whenever $m_i = 0$ for $i \in I_2$ and $m_i = V$ for $i \in I_1$, defined in (5), the criterion τ_z defined in (8) satisfies

$$\frac{\partial t_n(m)}{\partial m_i} t_d(m) - \frac{\partial t_d(m)}{\partial m_i} t_n(m) = 0, \tag{9}$$

for all $i = 1, \ldots, k$.

Note that the kurtosis coefficient given in (2) can be represented according to (8), and we will be particularly interested in the following representation:

$$\kappa_z(d) = \frac{3s^4 + 6(s^2 + \mathbb{E}(z)^2) \sum_{i=1}^k \alpha_i m_i^2 + \sum_{i=1}^k \alpha_i m_i^4 - 6s^2\mathbb{E}(z)^2 - 3\mathbb{E}(z)^4 - 4\mathbb{E}(z) \sum_{i=1}^k \alpha_i m_i^3}{(s^2 + \sum_{i=1}^k \alpha_i m_i^2 - \mathbb{E}(z)^2)^2} \tag{10}$$

where $\mathbb{E}(z) = \sum_i \alpha_i m_i$. In particular, for κ_z we have $r = 4$, $g = 2$ and

$$\begin{aligned}
t_n &= 6\mathbb{E}(z)^2 \sum_i \alpha_i m_i^2 + \sum_i \alpha_i m_i^4 - 3\mathbb{E}(z)^4 - 4\mathbb{E}(z) \sum_i \alpha_i m_i^3 \\
q_n &= 3s^4 + 6s^2 \sum_i \alpha_i m_i^2 - 6s^2\mathbb{E}(z)^2 \\
t_d &= \mathbb{E}(z)^4 + \left(\sum_i \alpha_i m_i^2 \right)^2 - 2\mathbb{E}(z)^2 \sum_i \alpha_i m_i^2 \\
q_d &= s^4 + 2s^2 \sum_i \alpha_i m_i^2 - 2s^2\mathbb{E}(z)^2
\end{aligned}$$

The following lemma shows that the kurtosis coefficient satisfies condition *A2* for this particular representation.

Lemma 2 *For $\tau_z(d)$ defined in (10) and d such that $m_i = 0$ for $i \in I_2$ and $m_i = V$ for $i \in I_1$, condition* A2 *holds.*

Proof For τ_z we have that

$$\begin{aligned}
t_n(m) &= 6\mathbb{E}(z)^2 \sum_i \alpha_i m_i^2 + \sum_i \alpha_i m_i^4 - 3\mathbb{E}(z)^4 - 4\mathbb{E}(z) \sum_i \alpha_i m_i^3 \\
t_d(m) &= \mathbb{E}(z)^4 + \left(\sum_i \alpha_i m_i^2 \right)^2 - 2\mathbb{E}(z)^2 \sum_i \alpha_i m_i^2 \\
\frac{\partial t_n(m)}{\partial m_j} &= 12\alpha_j \mathbb{E}(z) \sum_i \alpha_i m_i^2 + 12\mathbb{E}(z)^2 \alpha_j m_j + 4\alpha_j m_j^3 - 12\alpha_j \mathbb{E}(z)^3 \\
&\quad - 4\alpha_j \sum_i \alpha_i m_i^3 - 12\mathbb{E}(z)\alpha_j m_j^2 \\
\frac{\partial t_d(m)}{\partial m_j} &= 4\alpha_j \mathbb{E}(z)^3 + 4\alpha_j m_j \sum_i \alpha_i m_i^2 - 4\alpha_j \mathbb{E}(z) \sum_i \alpha_i m_i^2 - 4\mathbb{E}(z)^2 \alpha_j m_j
\end{aligned}$$

Under the conditions of the Lemma, replacing $m_i = 0$ for $i \in I_2$ and $m_i = V$ for $i \in I_1$, and letting

$\tilde{\alpha} = \sum_{i \in I_1} \alpha_i$, we have

$$\begin{aligned}
t_n &= \tilde{\alpha}(1-\tilde{\alpha})\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right) V^4 \\
t_d &= \tilde{\alpha}^2(1-\tilde{\alpha})^2 V^4 \\
\frac{\partial t_n}{\partial m_j} &= 4\alpha_j(1-\tilde{\alpha})\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right) V^3, \quad j \in I_1 \\
\frac{\partial t_d}{\partial m_j} &= 4\alpha_j\tilde{\alpha}(1-\tilde{\alpha})^2 V^3, \qquad\qquad j \in I_1 \\
\frac{\partial t_n}{\partial m_j} &= -4\alpha_j\tilde{\alpha}\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right) V^3, \qquad j \in I_2 \\
\frac{\partial t_d}{\partial m_j} &= -4\alpha_j\tilde{\alpha}^2(1-\tilde{\alpha}) V^3, \qquad\qquad j \in I_2
\end{aligned}$$

Replacing these results, we obtain for $i \in I_1$

$$\frac{\partial t_n}{\partial m_i} t_d - \frac{\partial t_d}{\partial m_i} t_n = 4V^7\left(\alpha_j\tilde{\alpha}^2(1-\tilde{\alpha})^3\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right) - \alpha_j\tilde{\alpha}^2(1-\tilde{\alpha})^3\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right)\right) = 0,$$

and for $i \in I_2$,

$$\frac{\partial t_n}{\partial m_i} t_d - \frac{\partial t_d}{\partial m_i} t_n = 4V^7\left(-\alpha_j\tilde{\alpha}^3(1-\tilde{\alpha})^2\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right) + \alpha_j\tilde{\alpha}^3(1-\tilde{\alpha})^2\left(1 - 3\tilde{\alpha} + 3\tilde{\alpha}^2\right)\right) = 0.$$

□

Now we will relate the directions defined in (5) and the extreme points of the optimization problem

$$\begin{aligned}
\min_d \quad & \tau_z(d) \\
\text{s.t.} \quad & d^T d = 1
\end{aligned} \tag{11}$$

From Theorem 1, we only need to consider directions in the subspace generated by $\{\mu_i - \mu_k\}_{i=1}^{k-1}$. To simplify the formal justifications, we first reparametrize the direction d as

$$d = \sum_{i=1}^{k-1} \theta_i e_i, \tag{12}$$

where $\sum_i \theta_i^2 = 1$, for a set of $k-1$ orthonormal vectors spanning $\{\mu_1 - \mu_k, \ldots, \mu_{k-1} - \mu_k\}$. Using this parametrization to remove the norm constraint, we have

$$d = \left(\sqrt{1 - \sum_{i=2}^{k-1} \theta_i^2}\right) e_1 + \sum_{i=2}^{k-1} \theta_i e_i. \tag{13}$$

We assume e_1 is the unique vector satisfying (5) for some partition (I_1, I_2), while all other vectors are arbitrary, but form an orthonormal basis of the subspace.
Based on this characterization, the problem of interest can be written as

$$\min_\theta \quad \kappa_z\left(\left(\sqrt{1 - \textstyle\sum_{i=2}^{k-1} \theta_i^2}\right) e_1 + \textstyle\sum_{i=2}^{k-1} \theta_i e_i\right), \tag{14}$$

and if we use $m_i = d^T\mu_i = \sqrt{1 - \sum_{i=2}^{k-1} \theta_i^2}\, e_1^T\mu_i + \sum_{i=2}^{k-1} \theta_i e_i^T\mu_i$, we can also write the problem as

$$\min_\theta \quad \kappa_z\left(m_1(\theta), \ldots, m_k(\theta)\right) \tag{15}$$

An interesting result is introduced in the following theorem, providing an asymptotic relationship for data with arbitrarily large separation between the groups. To establish this relationship, we need an additional condition to control the separation of the groups along orthogonal directions to e_1,

A3. There exists a constant L such that

$$|e_i^T \mu_j| \leq L V^{r-g}, \tag{16}$$

for all $i = 2, \ldots, k-1$ and $j = 2, \ldots, k$.

Theorem 2 *If conditions* A1, A2 *and* A3 *hold and* e_1 *satisfies (5), the gradient of the objective function of problem (15) satisfies*

$$\lim_{V \to \infty} \left| \frac{\partial \tau_z(e_1)}{\partial \theta_i} \right| = 0, \qquad i = 1, \ldots, k. \tag{17}$$

Proof The partial derivatives of the objective function of (15) are given by

$$\begin{aligned} \frac{\partial \tau_z(d)}{\partial \theta_i} &= \sum_{j=1}^{k} \frac{\partial \tau_z}{\partial m_j} \frac{\partial m_j}{\partial \theta_i} \\ &= \sum_{j=1}^{k} \frac{\partial \tau_z}{\partial m_j} \left(\frac{-\theta_i}{\sqrt{1 - \sum_{l=2}^{k-1} \theta_l^2}} e_1^T \mu_j + e_i^T \mu_j \right) \end{aligned}$$

From the definition of τ_z in (8) it follows that

$$\frac{\partial \tau_z}{\partial m_j} = \frac{1}{(t_d + q_d)^2} \left(\left(\frac{\partial t_n}{\partial m_j} + \frac{\partial q_n}{\partial m_j} \right) (t_d + q_d) - (t_n + q_n) \left(\frac{\partial t_d}{\partial m_j} + \frac{\partial q_d}{\partial m_j} \right) \right)$$

If we consider the case when $d = e_1$ ($\theta_i = 0,\ i = 2, \ldots, k-1$) and use Condition *A2*, we have

$$\frac{\partial \tau_z(e_1)}{\partial \theta_i} = \sum_{j=1}^{k} \frac{e_i^T \mu_j}{(t_d + q_d)^2} \left(\frac{\partial t_n}{\partial m_j} q_d + \frac{\partial q_n}{\partial m_j} (t_d + q_d) - q_n \frac{\partial t_d}{\partial m_j} - (t_n + q_n) \frac{\partial q_d}{\partial m_j} \right)$$

Dividing by V^{2r} both numerator and denominator, we have

$$\frac{\partial \tau_z(e_1)}{\partial \theta_i} = \sum_{j=1}^{k} \frac{e_i^T \mu_j / V^{r+1-g}}{(t_d + q_d)^2 / V^{2r}} \left(\frac{\partial t_n}{\partial m_j} \frac{q_d}{V^{r+g-1}} + \frac{\partial q_n}{\partial m_j} \frac{t_d + q_d}{V^{r+g-1}} - \frac{q_n}{V^{r+g-1}} \frac{\partial t_d}{\partial m_j} - \frac{t_n + q_n}{V^{r+g-1}} \frac{\partial q_d}{\partial m_j} \right)$$

Taking limits when $V \to \infty$ and noting that $t_d + q_d$ is a polynomial of degree r in V, that $\frac{\partial t_n}{\partial m_j} q_d$, $\frac{\partial q_n}{\partial m_j}(t_d + q_d)$, $\frac{\partial t_d}{\partial m_j} q_n$ and $\frac{\partial q_d}{\partial m_j}(t_n + q_n)$ are polynomials of degree $r + g - 1$ in V. If we use Condition *A3*, we obtain $e_i^T \mu_j / V^{r+1-g} \to 0$, which is the desired result. □

4 The proposed cluster algorithm

The previous result suggest an iterative procedure to find the possible clusters, as follows : (1)The data are projected on the directions of maximum and minimum kurtosis; (2) A criterion is applied to decide if the projected points can be divided into two groups along these directions; (3) Assuming that the data are divided into two groups, consider each of the groups as new samples and apply to each of then steps (1) and (2); (4) The procedure is repeated until no more groups are identified.
These ideas led to the following algorithm

1. The algorithm starts by standardizing the sample data, $Z = \Sigma^{-1}(X - \mu)$.
2. With standardized data, compute the directions d_{max} and d_{min} that maximizes and minimizes the kurtosis coefficient $\kappa(d)$ of the projected data $\{d^T Z\}$, respectively.
3. For each one of the directions, d_{max} and d_{min}, compute the univariate projections of the standardized observations, $p_{max} = d_{max}^T Z$ and $p_{min} = d_{min}^T Z$.

4. For each of the projections, p_{max} and p_{min}, we analyze if we have the mixture of two distributions according to BIC criteria. The BIC values for $G = 1$ and $G = 2$ are obtained. Where G is the number of mixtures present in the sample. If the BIC value for the mixture of two distributions is greater than the BIC value for one distribution, then this projection is considered to continue the procedure. If both projections have have a greater BIC value for the mixture of two distributions, then the BIC values of each projection are compared. The projection with greater BIC are considered to continue the procedure.

   In the case where none projection has a greater BIC value for the mixture of two distributions, then no groups are obtained and the procedure is finalized.

5. With the original data (non-normalized data) of the two groups obtained. Repeat steps 1 to 4 for each group until that no more groups are identified in the sample.

5 Monte Carlo Results

For the simulation examples we will consider two cases where we have samples formed by mixtures of normal distributions with the same covariance matrix.
In the first case we will analyze a sample formed by a mixture of three populations as follows: we will generate the three populations and use some criteria to analyze the success for the clustering procedure. The results obtained from 100 repetitions of the model will be presented in a table with the percentage representing the number of cases in which the clustering coincides with the original data. We will present and compare the results obtained with the Peña and Prieto Clustering Algorithm and with the Clustering Algorithm with Multivariate Mixtures.
In the second case, we will consider a particular example for a sample formed by the mixture of five populations. In this example we will generate the populations and we will show the results that we will be obtained step by step in the application of the Clustering Algorithm with Multivariate Mixtures.

5.1 Three Populations

We will generate populations as follows: populations 1 and 2 are generated on the first coordinate axis. The populations are separated by a distance $dst1$ and the mean of population 1 is at a distance $dst1/2$ from the origin and the mean of population 2 is located at the same distance $dst1/2$ from the origin but in sense contrary to population 1. The population 3 is at a distance $dst2$ from the origin with an inclination angle. The angles that are used in the simulations are 30°, 60° and 90°. Figure 1 shows an example of data generated with this set-up.

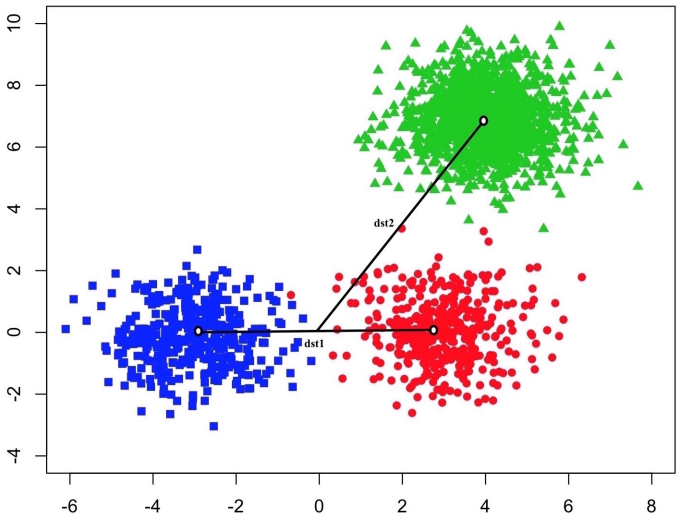

Figure 1: Original Data Three Populations

The parameters in the simulations are given in Table 1.

Parameter	
n	Number of total observations
p	Dimension of the data
r	Cosine of the angle in which the population 3 is located
α_1	Percentage of data in population 1
α_2	Percentage of data in population 2
$\alpha_3 = 1 - (\alpha_1 + \alpha_2)$	Percentage of data in population 3
$dst1$: $6\sqrt{p}/\sqrt{2}$	Distance between the means of populations 1 and 2
$dst2$: $8\sqrt{p}/\sqrt{2}$	Distance from the origin to the mean of the population 3

Table 1: Simulation Parameters for Three Populations

Our interest is to study the existence of clusters in the data using the kurtosis coefficient when the parameters α_1, α_2 and α_3 change. The cases that we will consider in the simulations are in Table 2.

Case	α_1	α_2	α_3
050590	0.05	0.05	0.90
101080	0.10	0.10	0.80
151570	0.15	0.15	0.70
201070	0.20	0.10	0.70
202060	0.20	0.20	0.60
301060	0.30	0.10	0.60
302050	0.30	0.20	0.50
401050	0.40	0.10	0.50
402040	0.40	0.20	0.40
303040	0.30	0.30	0.40

Table 2: Cases to Study

In order to compare the algorithms, we need criteria of success for the clustering procedure. In the case of three groups the clusters detection is done in two stages. The first stage consists in the separation of the first two groups and in the second stage the missing group is detected. Therefore, the following criteria of success in the clustering during the two stages are established:

First stage. If two groups are obtained in the application of the algorithm the first time, we compare each group with the three original populations and analyze the coincidences. If one of the groups obtained belongs to at least 80% of the initial populations and to the other group at least 80% of another population, we consider that the clustering is successful during this stage.

Second stage. The algorithm is applied to the two groups obtained in the first stage. We consider the clustering is successful during this second stage if we have: one of the groups is divided into two subgroups. Each subgroup must match one of the initial populations, at least 80%. To the other group, which is not divided into sugbrupos, must belong to at least 80% of the population that does not belong to the previous subgroups.
The results are presented in a table with the percentage representing the number of cases in which the clustering coincides with the original data during each stage. The table is divided as follows: in the rows are the proportions $n/p = 10, 20, 50, 100$ for each p. The columns are divided into five: the first and second columns indicate p and the corresponding proportions, in the third and fourth columns the results of success obtained in the first and second stages respectively are presented. These columns are divided into three columns corresponding to the angle at which the third population is located, which may be 30°, 60° y 90°. The fifth column shows the percentage of times in which the first and second stage succeed in the clustering at the same time. The results were obtained from 100 repetitions of the model.
In Table 3 we have the results obtained applying the algorithm proposed by Peña and Prieto. In Table 4 the results applying the clustering algorithm with multivariate mixtures, see Section 4.

		Average Success Rate								
		First Stage			***Second Stage***			***Procedure***		
p	Angle / n/p	30°	60°	90°	30°	60°	90°	30°	60°	90°
10	**20**	0.87	0.90	0.89	0.17	0.49	0.40	**0.15**	**0.45**	**0.36**
	50	0.97	0.96	0.97	0.35	0.59	0.63	**0.33**	**0.56**	**0.61**
	100	0.99	0.98	0.99	0.41	0.71	0.71	**0.40**	**0.69**	**0.70**
	250	1	0.99	1	0.40	0.79	0.79	**0.40**	**0.78**	**0.79**
	Mean	**0.96**	**0.96**	**0.96**	**0.33**	**0.65**	**0.63**	**0.32**	**0.62**	**0.61**
20	**20**	0.78	0.84	0.85	0.16	0.28	0.28	**0.16**	**0.25**	**0.26**
	50	0.93	0.91	0.91	0.39	0.44	0.47	**0.35**	**0.41**	**0.41**
	100	0.96	0.95	0.95	0.47	0.52	0.54	**0.45**	**0.49**	**0.50**
	250	0.99	0.97	0.97	0.48	0.60	0.60	**0.48**	**0.58**	**0.58**
	Mean	**0.92**	**0.92**	**0.02**	**0.38**	**0.46**	**0.47**	**0.36**	**0.43**	**0.44**
50	**20**	0.34	0.39	0.43	0.11	0.13	0.12	**0.10**	**0.12**	**0.11**
	50	0.81	0.81	0.81	0.18	0.18	0.16	**0.14**	**0.17**	**0.15**
	100	0.90	0.90	0.91	0.26	0.30	0.27	**0.20**	**0.25**	**0.23**
	250	0.95	0.95	0.95	0.39	0.41	0.41	**0.35**	**0.37**	**0.37**
	Mean	**0.75**	**0.76**	**0.78**	**0.23**	**0.25**	**0.24**	**0.20**	**0.22**	**0.22**

Table 3: Average Success Rate with Peña and Prieto Clustering Algorithm

		Average Success Rate								
		First Stage			***Second Stage***			***Procedure***		
p	Angle / n/p	30°	60°	90°	30°	60°	90°	30°	60°	90°
10	**20**	1	1	1	0.84	0.87	0.85	**0.84**	**0.87**	**0.85**
	50	1	1	1	1	1	1	**1**	**1**	**1**
	100	1	1	1	1	1	1	**1**	**1**	**1**
	250	1	1	1	1	1	1	**1**	**1**	**1**
	Mean	**1**	**1**	**1**	**0.96**	**0.97**	**0.96**	**0.96**	**0.97**	**0.96**
20	**20**	1	1	1	0.75	0.85	0.85	**0.75**	**0.85**	**0.85**
	50	1	1	1	0.97	0.99	1	**0.97**	**0.99**	**1**
	100	1	1	1	1	1	1	**1**	**1**	**1**
	250	1	1	1	1	1	1	**1**	**1**	**1**
	Mean	**1**	**1**	**1**	**0.93**	**0.96**	**0.96**	**0.93**	**0.96**	**0.96**
50	**20**	0.87	0.91	0.92	0.41	0.42	0.43	**0.39**	**0.41**	**0.43**
	50	0.99	1	1	0.60	0.62	0.63	**0.60**	**0.62**	**0.63**
	100	1	1	1	0.94	0.93	0.92	**0.94**	**0.93**	**0.92**
	250	1	1	1	0.97	0.96	0.96	**0.97**	**0.96**	**0.96**
	Mean	**0.96**	**0.98**	**0.98**	**0.73**	**0.73**	**0.74**	**0.73**	**0.73**	**0.74**

Table 4: Average Success Rate with Clustering Algorithm with Multivariate Mixtures

In Table 3 we can see that the success in clustering is significantly better in the first stage than in the second.This could indicate that the algorithm clearly identifies two groups, but does not identify the third group correctly. We can also see that as the dimension of the data increases, success in clustering decreases significantly.

From the results of Table 4 we can conclude that, for each p, success in clustering increases as the value of n/p increases. On the other hand, we can also see that success in clustering decreases as the value of p increases. This could be because the parameter estimation increases as the dimension increases. For example, in the

case of $p = 50$ and $n/p = 20$, n = 1000 would be few data for estimation of approximately 30 parameters. From the results obtained in Tables 3 and 4, we can conclude that our clustering algorithm is more efficient than the clustering algorithm proposed by Peña and Prieto (2001a) when the data dimension and the clusters present in the sample are high.
We will show below, with a detailed example, that the proposed method is also efficient when we have more groups in the sample.

5.2 Five Populations. An example

We now present an example of a sample formed by a mixture of five populations with normal distribution and with the same covariance matrix. The populations are generated as follows: populations 1 and 2 are generated on the first coordinate axis. The populations are separated by a distance $dst1$ as follows: the average of the population 1 is at a distance $dst1/2$ from the origin and the average of the population 2 is located at the same distance $dst1/2$ from the origin but in sense contrary to population 1. The population 3 is generated on the second coordinate axis, is at a distance $dst2$ from the origin. The population 4 is at a distance $dst3$ from the origin with an inclination angle of 60°. The population 5 is at the same distance $dst3$ from the origin, but with an inclination angle of 120°.
The parameters in the simulations are given in Table 5 .

Parameter	
$n = 2000$	Number of total observations
$p = 10$	Dimension of the data
$\alpha_1 = 0.20$	Percentage of data in population 1
$\alpha_2 = 0.25$	Percentage of data in population 2
$\alpha_3 = 0.15$	Percentage of data in population 3
$\alpha_4 = 0.15$	Percentage of data in population 4
$\alpha_5 = 0.25$	Percentage of data in population 5
$dst1 = 6\sqrt{p}/\sqrt{2}$	Distance between the means of populations 1 and 2
$dst2 = 8\sqrt{p}/\sqrt{2}$	Distance from the origin to the mean of the population 3
$dst3 = 10\sqrt{p}/\sqrt{2}$	Distance from the origin to the means of the populations 4 and 5

Table 5: Simulation Parameters for Five Populations

The first population has 400 data, the second population 500, the third population 300, the fourth population 300 and the fifth population 500 data. In the figure 2 we plot the first two principal components and we can see that the populations are mixed.

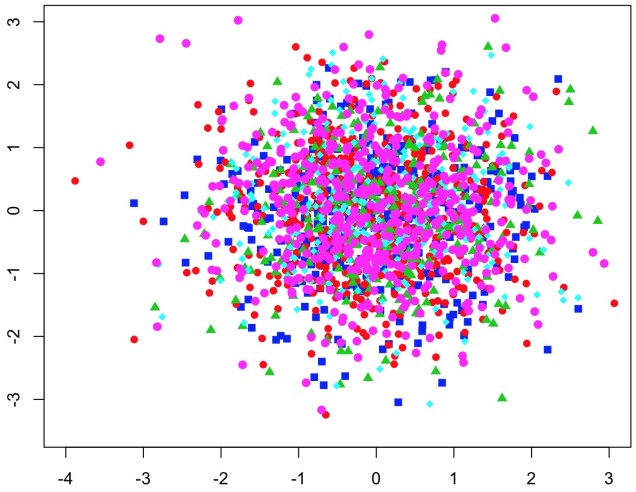

Figure 2: First Two Principal Components for Five Populations Case

The algorithm splits first the sample are separated into two groups. The Group 1 contains 900 data and is made up of populations 1 and 2. The Group 2 contains 1100 data and is made up of populations 3, 4 and 5. In the figure 3 we plot the projection on the direction of minimum kurtosis and we can see the separation of the sample into two groups.

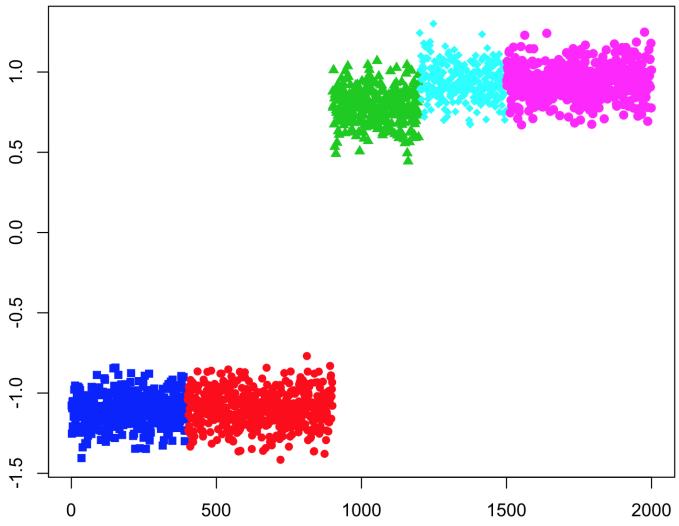

Figure 3: First Clustering for Five Populations

Applying the procedure to the Group 1, we obtain two subgroups. The first subgroup contains 400 data and it is composed by the population 1 and the second subgroup contains 500 data and it is composed by the population 2. Applying the procedure again to each subgroup, no further groups are obtained. In the figure 4 we plot the projection on the direction of minimum kurtosis and we can see the separation of the Group 1 into two subgroups.

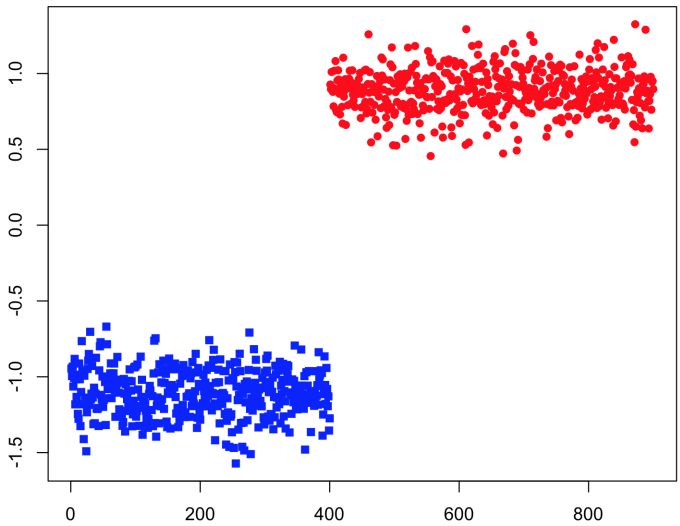

Figure 4: Second Clustering for Five Populations

We now apply the procedure to the Group 2 and we obtain two subgroups. The Subgroup 1 contains 600 data and is made up of populations 3 and 4. The Subgroup 2 contains 500 data and is composed of the population 5, see figure 5.

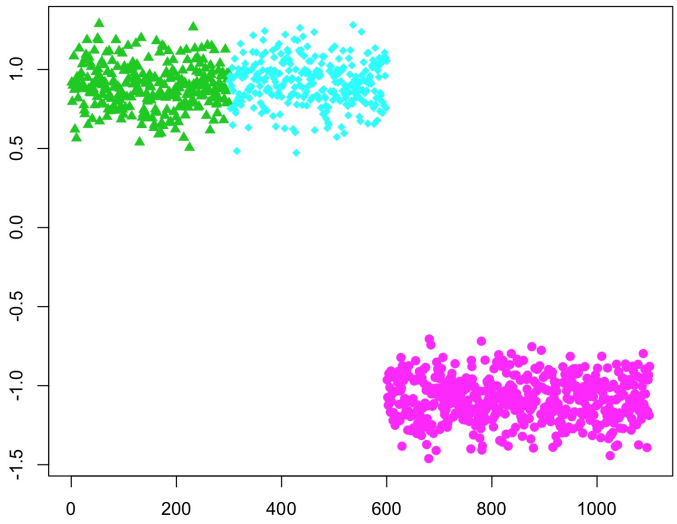

Figure 5: Third Clustering for Five Populations

Applying the procedure to the Subgroup 1, we obtain two subgroups. The first subgroup contains 300 data and it is composed by the population 3 and the second subgroup contains 300 and it is composed by the population 4, see figure 6. Applying the procedure to the Subgroup 2, no further groups are obtained.

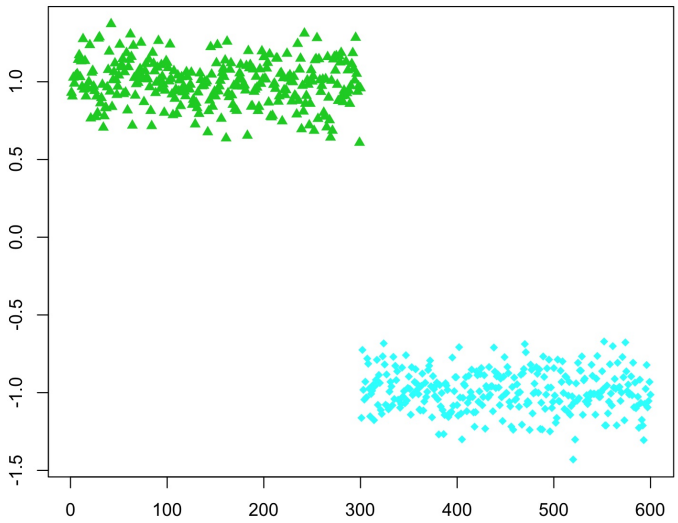

Figure 6: Fourth Clustering for Five Populations

Finally, we apply the procedure to each of the subgroups obtained from Subgroup 1 and and no further groups are obtained.
From Figures 4, 5 and 6 we can conclude that the procedure has efficiently identified in this example the existence of the five groups.

6 Conclusions

In this paper we have presented an iterative binary clustering algorithm based on directions that project the observations onto two separate groups. We have shown that these directions can be found by the extreme directions of kurtosis. Then we have proposed an algorithm where in each one of the projections of the data on the directions of maximum and minimum kurtosis we check for a mixture of two distributions using the BIC criterion. Finally, by some simulation examples, we shown that the algorithm with a mixture of normals is more efficient than the algorithm proposed by Peña and Prieto (2001a) when the data dimension and the conglomerates present in the sample are high.

References

[1] Banfield, J.D. and Raftery, A. E. (1993). Model-based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49, 803-821.

[2] Bouveyron, C., and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics Data Analysis*, 71, 52-78.

[3] Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, 25-71. Springer Berlin Heidelberg.

[4] Cardoso, J. F. (1989). Source separation using higher order moments. *Acoustics, Speech, and Signal Processing. ICASSP-89*, 2109-2112.

[5] Darlington, R. B. (1970). Is kurtosis really "Peakedness"?. *The American Statistician*, 24 (2), 19-22.

[6] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39 (1), 1-38.

[7] Everitt, B.S. (1993), Cluster Analysis, Oxford University Press.

[8] Fraley, C. and Raftery, A. (1999). MCLUST: Software for model-based cluster and discriminant analysis. *Technical Report* 342, Dept. Statistics, Univ. of Washington.

[9] Friedman, J. H. and Tukey, J. W. (1974).A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, 23(9), 881-890.

[10] Fraiman, R., Justel, A., & Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, 103(483), 1294-1303.

[11] Galimberti, G., Manisi, A., & Soffritti, G. (2017). Modelling the role of variables in model-based cluster analysis. *Statistics and Computing*, 1-25.

[12] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, p. 417-441, and 498-520.

[13] Huber, P. J. (1985). Projection Pursuit. *The Annals of Statistics*, 13, 435-525.

[14] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4), 411-430.

[15] Hyvärinen, A., Karhunen, J. and Oja E. M. (2001). Independent Component Analysis. *New York: John Wiley.*

[16] Jain, A. K., Dubes, R. C. (1988). Algorithms for Clustering Data. *Prentice Hall.*

[17] Jobson, J. (2012). Applied multivariate data analysis: volume II: Categorical and Multivariate Methods. *Springer Science and Business Media.*

[18] Jones, M. C. and Sibson, R. (1987). What is projection pursuit?. *Journal of the Royal Statistical Society.* Series A (General), 1-37.

[19] Kaufman, L. and Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis. *Wiley, New York.*

[20] Liu, J., Zhang, J., Palumbo, M. and Lawrence C. (2003): Bayesian clustering with variable and transformation selections. *Bayesian Statistics*, 7, eds, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: University Press, 249 - 75.

[21] Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519-530.

[22] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, 281-297.

[23] Maugis, C., Celeux, G. and Martin-Magniette, M.L. (2009). Variable selection in model-based clustering: a general variable role modeling. *Computational Statistics and Data Analysis*, 53(11), 3872- 3882.

[24] Móri, T. F., Rohatgi, V. K. and Székely, G. J. (1993). On multivariate skewness and kurtosis. *Theory of Probability and its Applications*, 38, 547-551.

[25] Pan, W., and Shen, X. (2007). Penalized model-based clustering with application to variable selection.*Journal of Machine Learning Research*, 8(May), 1145-1164.

[26] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 (11), 559-572.

[27] Pearson, K. (1905). Das Fehlergesetz und Seine Verallgemeinerungen Durch Fechner und Pearson. A Rejoinder. *Biometrika*, 4, 169-212.

[28] Peña, D. and F. J. Prieto (2001a). Cluster identication using projections. *Journal of the American Statistical Association*, 96, 1433 - 1445.

[29] Peña, D. and F. J. Prieto (2001b). Robust covariance matrix estimation and Multivariate outlier detection. *Technometrics*, 43, 286-310.

[30] Peña, D., and Prieto, F. J. (2007). Combining random and specific directions for outlier detection and robust estimation in high-dimensional multivariate data. *Journal of Computational and Graphical Statistics*, 16(1), 228-254.

[31] Peña, D., Prieto, F. J. and Viladomat, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *Journal of Multivariate Analysis*, 101, 1995-2007.

[32] Raftery, A. E., and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473), 168-178.

[33] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6 (2), 461-464.

[34] Schwager, S. J. and Margolin, B. H. (1982). Detection of multivariate normal outliers. *The Annals of Statistics*, 10(3), 943-954.

[35] Steinley, D., and Brusco, M. J. (2008). A new variable weighting and selection procedure for K-means cluster analysis. *Multivariate Behavioral Research*, 43(1), 77-108.

[36] Wang, S., and Zhu, J. (2008), "Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. *Biometrics*, 64, 440-448.

[37] Witten, D. M., and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 713-726.