Contributions to the problem of cluster analysis

Doctoral Thesis

Author: Júlia Viladomat Comerma

Advisors: Daniel Peña Sánchez de Rivera

Francisco Javier Prieto Fernández

PhD in Mathematical Engineering

Department of Statistics

Universidad Carlos III de Madrid

Getafe, May 2009

To my mother and brothers.

Acknowledgements

This thesis would not have come to the form and shape it is today without the help and support of several people who I would like to thank and to whom I am happily in debt.

First of all, I would like to thank my advisors, Dr. Daniel Peña Snchez de Rivera and Dr. Francisco Javier Prieto Fernndez for supporting, guiding, and working very closely with me for the last five years. I am very grateful for their patience, encouragement, criticism, persistence, and friendly relationship with me. They will always be very good friends to me.

I am also very grateful to Dr. Ruben Zamar, with who I worked very closely during a semester and two summers I spent at University of British Columbia in Vancouver. A chapter of this thesis is the result of this collaboration, and I am thankful for his hospitality and the opportunity he gave me to visit the department.

I would like to mention the entire faculty and staff of the department of Statistics for always being friendly, helpful, and appreciative of our work. I would also like to show my gratitude to the Spanish Ministry of Education and Science and to the education and culture section of the Community of Madrid for financial support. In particular, I would like to thank Professors Juan Romo and Daniel Peña for allowing me to participate in the research projects MEC 2002-02054, MEC 2005-03424, MEC 2009-00035, and CAM 2006-03563, CAM 2007-04081, MEC 2009-00176, respectively.

Special thanks to my wonderful friends, to those who have known me since we were kids and to those I met later, for the good times we have had together, for their warmth, closeness and good advice. They will always be in my heart.

My thesis dedication and my most wholehearted thanks go to my family, and specially to my mom, Montse. Without her unconditional love, patience, care and guidance, I would have never come this far. I cannot thank her enough for all she has taught me, and for supporting me in all the decisions I have made in my life.

Resumen

Dada una muestra aleatoria generada por una mezcla de distribuciones, el objetivo del análisis de conglomerados es partir la muestra en grupos homogéneos en relación a las poblaciones que los han generado.

Algoritmos como KMEANS y MCLUST resuelven el problema de conglomerados en el espacio original. Un enfoque alternativo es reducir primero la dimensión de los datos proyectando la muestra en un espacio de dimensión menor, e identificar los grupos en este subespacio. De esta forma, la maldición de la dimensión puede evitarse, pero hay que asegurarse de que los datos proyectados preservan la estructura de conglomerados de la muestra original. En este contexto, los métodos de búsqueda de proyecciones tienen como objetivo encontrar direcciones, o subespacios de baja dimensión, que muestren las vistas más interesantes de los datos (Friedman and Tukey, 1974; Friedman, 1987). Reducir la dimensión de la muestra es efectivo ya que no toda la información de los datos está ligada a la estructura de grupos de la muestra. Con la reducción se pretende eliminar la información no relevante, y quedarse con un espacio de dimensión menor donde el problema de conglomerados sea más fácil de resolver. Para ello hace falta un procedimiento que mantenga la información clave de los grupos.

En este contexto, Peña and Prieto (2001) demuestran que las direcciones que minimizan y maximizan la kurtosis tienen propiedades óptimas para visualizar los grupos, y proponen un algoritmo de conglomerados que proyecta los datos en ambos tipos de direcciones y asigna las observaciones a los grupos en consonancia con los huecos encontrados en éstas.

En el capítulo 1 de la tesis el concepto de kurtosis se revisa en detalle. El coeficiente de kurtosis univariante y las distintas interpretaciones que se le han dado en la literatura son analizadas. También estudiamos de que maneras puede definirse la kurtosis en una muestra multivariante y exploramos sus propiedades para detectar grupos.

En el Capítulo 2 estudiamos las propiedades de una matriz de kurtosis y proponemos

un subconjunto de sus vectores propios como direcciones interesantes para revelar la posible estructura de grupos de los datos. Esta idea es una extensión al caso multivariante del algoritmo propuesto en Peña and Prieto (2001). La ventaja de usar los vectores propios de una matriz para especificar el subespacio de interés radica en que no es necesario usar un algoritmo de optimización para encontrarlo, como ocurre en Peña and Prieto (2001). Por otra parte, ante una mezcla de distribuciones elípticas con matrices de covarianzas proporcionales, demostramos que un subconjunto de vectores propios de la matriz coincide con el subespacio lineal discriminante de Fisher. Los vectores propios de la matriz de kurtosis estimada son estimadores consistentes de este subespacio, y su calculo es fácil de implementar y computacionalmente eficiente. La matriz, por tanto, proporciona una forma de reducir la dimensión de los datos en vistas a resolver el problema de conglomerados en un subespacio de dimensión menor.

Siguiendo la discusión en el Capítulo 2, en el capítulo 3 estudiamos matrices alternativas de kurtosis basadas en modificaciones locales de los datos, con la intención de mejorar los resultados obtenidos con los vectores propios de la matriz de kurtosis estudiada en el Capítulo 2. Mediante la sustitución de las observaciones de la muestra por la media de sus vecinos, las matrices de covarianzas de las componentes de la mezcla de distribuciones se contraen, dando un rol predominante a la variabilidad entre grupos en la descomposición de la matriz de kurtosis. En particular, se demuestra que las propiedades de separación de los vectores propios de la nueva matriz de kurtosis son mejores en el sentido que la modificación de las observaciones propuesta produce medias estandarizadas más alejadas entre sí que las de las observaciones originales.

El Capítulo 4 propone algunas ideas en relación a la identificación de grupos no lineales en un espacio de baja dimensión, proyectando en direcciones aleatorias solamente las observaciones contenidas en un entorno local definido a partir de la dirección. Estas direcciones pueden ser entendidas como direcciones recortadas, y permiten detectar formas específicas que los algoritmos de conglomerados tradicionales con buenos resultados en baja dimensión no detectan con facilidad. El algoritmo sugerido está pensado para usarse una vez la dimensión del espacio de los datos ha sido reducida.

Finalmente, en el Capítulo 5 proponemos un algoritmo de conglomerados no paramétrico basado en medianas locales. Cada observación es sustituida por su mediana local, moviéndose de esta manera hacia los picos y lejos de los valles de la distribución. Este proceso es repetido iterativamente hasta que cada observación converge a un punto fijo. El resultado es un partición de la muestra basado en donde convergen las secuencias de medianas locales. El algoritmo determina el número de grupos y la partición de las

observaciones dada la proporción de vecinos. Una versión rápida del algoritmo, donde solamente se trata un subconjunto de las observaciones, también se proporciona. En el caso univariante, se demuestra la convergencia de cada observación al punto fijo más próximo, así como la existencia y unicidad de un punto fijo en un entorno de cada moda de la distribución.

Contents

List of Tables

List of Figures

Introduction and summary

Given a multivariate sample drawn from a mixture of k distributions, cluster analysis attempts to partition the sample into homogeneous groups according to the populations that generate them.

The KMEANS algorithm proposed in Hartigan and Wong (1979) starts with an initial partition of the sample and iteratively reassigns the observations to clusters according to an homogeneity criterion. The criterion that is generally used is the sum of squares within groups, which can be written as

$$SSW = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)^2, \tag{1}$$

where x_{ig} is the observation i in group g and $\bar{x}_g$ is the mean of group g. The algorithm iterates until the criterion is minimized. Since minimizing (1) is equivalent to minimizing the euclidean distances of the observations to the mean of the group they belong, the k-means algorithm tends to find spherical clusters.

The algorithm MCLUST (Banfield and Raftery, 1993; Dasgupta and Raftery, 1998; Fraley and Raftery, 1999) assumes the sample has been generated from a mixture of G distributions, usually assumed to be normal, and estimates the parameters of each population of the mixture together with the probability of membership for each observation of the sample, which is the so-called probability *a posteriori*

$$\pi_{ig} = \frac{\pi_g f_g(x_i)}{\sum_{g=1}^{G} \pi_g f_g(x_i)}, \tag{2}$$

where f_g is the density function of population g, and π_g is the *a priori* probability of membership to the group g. The observation x_i will be assigned, thus, to the cluster g that maximizes (2). In order to compute (2) we need to estimate the parameters of the mixture, which is done via the logarithm of the correspondent likelihood function, which again will depend on (2). The EM algorithm is used to jointly estimate both. The estimation is repeated for different assumptions on the number of components in the

mixture and covariance matrices of the components, and the BIC criteria is used to choose the assumption more likely to be true. The performance of MCLUST is better than the performance of k-means, and in general works well for low dimensional spaces. However, when the dimension of the space is large, the computational time may become prohibitive; MCLUST estimates several covariance matrices, and thus requires a large sample if the dimension of the data is large.

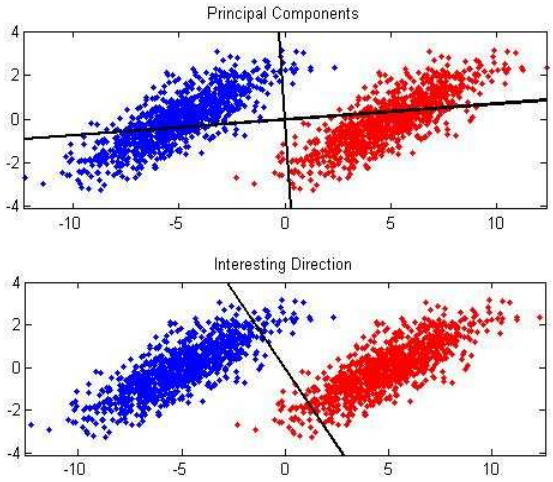
Note that algorithms such as KMEANS and MCLUST perform cluster analysis in the original space. An alternative approach to the problem may be to first reduce the dimension of the sample by projecting the data onto a lower dimensional subspace and identifying the clusters there. The curse of dimensionality can thus be avoided, but care needs to be taken to make sure that the projected data preserve the cluster structure of the original sample. In this context, projection pursuit aims to find the directions, or subspaces of low dimension, that show the most interesting views of the data, see Friedman and Tukey (1974); Friedman (1987).

Reducing the dimension of the sample is effective because not all the information in the dataset is relevant for clustering. We aim to remove the non-relevant, random information and look in a lower dimensional space where the cluster problem is significantly easier to solve. For that, we need a procedure that maintains the key information about the clusters and, since in general the cluster structure is not found in all variables, the selection of the variables to consider must be done carefully.

The dimension reduction approach for clustering is analyzed in Liu et al. (2003), where the data is projected onto the first principal components, and a Bayesian model for a mixture of normal distributions is adjusted in the resulting subspace. However, as we illustrate in Figure 1 with the help of a mixture of two normal populations, using principal components to reduce the dimension is not always appropriate. If we project the data onto one of the two principal components, the groups will overlap. The interesting direction in this case is the one perpendicular to the main axis of the elliptically shaped components of the mixture, which is Fisher's discriminant direction. The principal components fail to detect the clusters because they are the eigenvectors of the covariance matrix of the whole mixture, and not of the components of the mixture.

Independent Components Analysis (ICA) is a relatively new technique whose purpose is to find the independent latent factors that generate the observed multivariate sample, see Hyvärinen et al. (2001). ICA is a step forward from Principal Components Analysis (PCA), as the data are first standardized to be uncorrelated (PCA) and then rotated so that independent factors can be found. Huber (1985) emphasized that interesting

Figure 1: Principal Components and Fisher's discriminant direction.



projections are those that produce non-normal distributions and therefore non-normality is one of the criteria used to find the factors. However, non-normality is a general condition, and it is important to specify how to measure it. One of the ICA algorithms, for example, searches for the factors that maximize the absolute value of the univariate kurtosis coefficient. The idea of maximizing the kurtosis has also been used in cluster analysis, see Jones and Sibson (1987). In addition to that, Peña and Prieto (2001) showed that the directions that minimize the kurtosis can be as useful as, if not more than, the ones that maximize it, and present a cluster algorithm that projects the data in both the directions that minimize and maximize the kurtosis coefficient, and then assign the observations to groups according to the clusters found in the directions.

This thesis presents several approaches for the identification of clusters in the data, that are elaborations of several basic ideas: the use of the kurtosis coefficient to select subspaces of interest, the iterative application of local aggregation steps to improve the cluster structure of the original data, and a combination of ideas from local analysis of the data and kurtosis information to improve the detection of nonlinear structures in the data.

The contributions of this thesis are organized in chapters as follows.

In Chapter 1 the concept of kurtosis is carefully reviewed. The univariate kurtosis coefficient is studied and the different interpretations given to it in the literature are revised. Different attempts to measure what is understood as kurtosis in a multivariate sample are also analyzed in the chapter. Finally, we summarize the use that has been given to kurtosis as a tool to perform cluster analysis.

In Chapter 2 we study the properties of a kurtosis matrix and propose a subset of its eigenvectors as interesting directions to reveal the possible cluster structure of a data set. It is an extension to the multivariate case of the kurtosis-based algorithm in Peña and Prieto (2001), where instead of looking at directions, we look at low-dimensional subspaces. Note that the eigenvectors of the matrix provide the subspace where to project without the need to use an optimization algorithm, as in Peña and Prieto (2001). In addition to that, we prove that the subspace has optimal properties for clustering. In particular, under a mixture of elliptical distributions with proportional scatter matrices, it is shown that a subset of the eigenvectors of the fourth-order moment matrix corresponds to Fisher's linear discriminant subspace. The eigenvectors of the estimated kurtosis matrix are consistent estimators of this subspace and its calculation is easy to implement and computationally efficient, which is specially favourable when the ratio n/p is large. The matrix, thus, provides a way of reducing the dimension of the space of the data in order to perform cluster analysis in a subspace of lower dimension.

Following the discussion in Chapter 2, Chapter 3 studies alternative kurtosis matrices based on local modifications of the data, with the intention of improving the performance of the eigenvectors of the kurtosis matrix studied in Chapter 2. By substituting each observation of the sample with the mean of its neighbours, the covariance matrices of the components of a mixture of distributions will shrink, giving a more predominant role to the variability between clusters in the decomposition of the kurtosis matrix. Specifically, we prove that the separation properties of the eigenvectors of the new kurtosis matrix are better in the sense that the proposed modification of the observations produces standardized means that are further from each other than those of the original observations, and thus the clusters will appear more separated.

Chapter 4 draws some ideas on how to identify non-linearly shaped clusters in a low dimensional space by projecting onto several random directions only those observations contained in a local neighbourhood defined from the directions. These directions can be understood as trimmed projections, and allow to identify specific shapes that traditional clusters methods with good performance in low dimensional spaces fail to detect. The suggested cluster algorithm is intended to be used once the dimension of a high dimensional data set has been reduced.

A non-parametric cluster algorithm based on local medians is proposed in Chapter 5. Each observation is substituted by its local median and this new observation moves towards the peaks and away from the valleys of the distribution. The process is repeated until each observation converges to a fixpoint. We obtain a partition of the sample based

on where the sequences of local medians have converged. The algorithm determines the number of clusters and the partition of the observations given a value of α, the proportion of neighbours. A fast version of the algorithm, where only a subset of observations from the sample are treated, is also given. Furthermore, and for a univariate random variable, we prove the convergence of each point to the closest fixpoint, and the existence and uniqueness of a fixpoint on the neighbourhood of each mode.

Finally, we outline our contributions and give directions for future work in a concluding chapter.

Chapter 1

A review of kurtosis

In this chapter the concept of kurtosis is carefully reviewed. The univariate kurtosis coefficient is studied and the different interpretations given to it in the literature are revised. Different attempts to measure what is understood as kurtosis in a multivariate sample are also analyzed in the chapter. Finally, we review the use that has been given to kurtosis as a tool to perform cluster analysis.

1.1 The univariate kurtosis

The word kurtosis comes from the Greek word *kyrtos* or *kurtos* which means bulging, "a curved shape sticking out from the surface of something". The way the kurtosis distribution characterizes the shape of the distribution is a controversial matter that has been discussed extensively in the literature. In this section we review this discussion and the different interpretations that have been given to what the kurtosis exactly measures.

Let X be a random variable with mean μ and standard deviation σ. The classical univariate kurtosis coefficient was defined by Pearson (1905) as

$$\frac{\mu_4}{\sigma^4}$$

where $\mu_4 = E(X-\mu)^4$ is the fourth-order central moment of X.

Given a univariate random sample $x_1, \ldots, x_n$ drawn from the random variable X, the sample univariate kurtosis coefficient is

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum(x_i - \bar{x})^2\right]^2},$$

where $\bar{x}$ and s are the mean and standard deviation of the sample.

It is easy to see that the kurtosis coefficient reaches its minimum value at one. In effect, if we denote $a_i = x_i - \bar{x}$, k can be expressed as

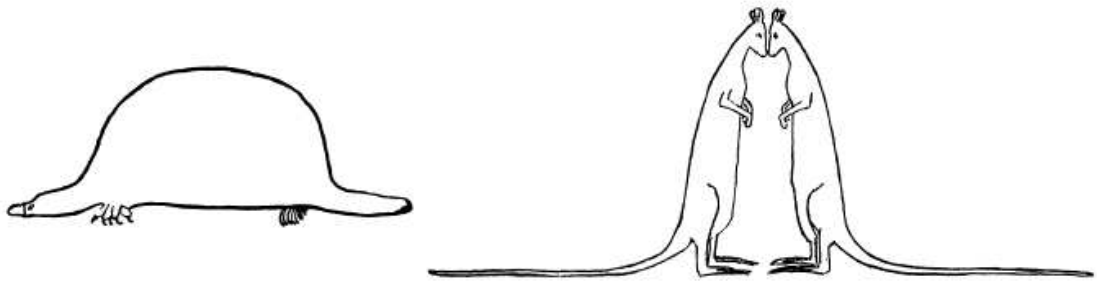
$$k = \frac{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i^4 + a_j^4)}{\sum_{i=1}^{n}\sum_{j=1}^{n} a_i^2 a_j^2},$$

and, since $(a_i^2 - a_j^2)^2 = a_i^4 + a_j^4 - 2a_i^2 a_j^2 \geq 0$, the numerator is always larger than the denominator and therefore $k \geq 1$. The larger the difference between a_i^2 and a_j^2, for two pairs of observations, the higher the value of the kurtosis, and thus k can be seen as a measure of variability of the observations with respect to their mean, as we will see later.

1.1.1 Traditional interpretation of the kurtosis coefficient

In the past, in most elementary statistical books, kurtosis has been used to define whether a unimodal distribution is platykurtic or leptokurtic. "Platy" means flat in Greek and characterizes the distribution as being the opposite of a peaked distribution, which is what leptokurtic means. Specifically, if $k > 3$ the distribution was classified as leptokurtic, and if $k < 3$ the distribution was platykurtic, where 3 is the value of the kurtosis for a normal distribution and therefore the peakedness is defined as relative to that distribution. As a matter of fact, sometimes the kurtosis coefficient is defined as $k' = k - 3$ to standardize it to the normal distribution. In Figure 1.1 we annex an amusing mnemonic provided by "Student" (1927).

* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3," while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the



normal curve of error and leptokurtic longer "tails." I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping," though, perhaps, with equal reason they should be hares!

Figure 1.1: Mnemonic for platykurtic and leptokurtic distributions

However, because of the averaging nature of moments, the kurtosis relationship to shape is a little more complicated than that. Chissom (1970) claimed that more evidence

than the sole value of the kurtosis coefficient should be considered to label a distribution as leptokurtic or platykurtic. By progressively modifying the shape of a (discrete) distribution, he illustrates that a peaked distribution can have a negative kurtosis value ($k' < 0$), and concludes that in order to have a positive kurtosis the distribution must not only be peaked, but contain a good number of cases in the tails, acknowledging the importance of the tails when measuring kurtosis.

1.1.2 Kurtosis as a measure of bimodality

The kurtosis is unaffected by changes in the mean and variance of the sample and therefore can be expressed as a function of the z scores,

$$k = \frac{1}{n}\sum_{i=1}^{n} z_i^4$$

where $z_i = s^{-1}(x_i - \bar{x})$. If we calculate the variance of the squared scores we obtain

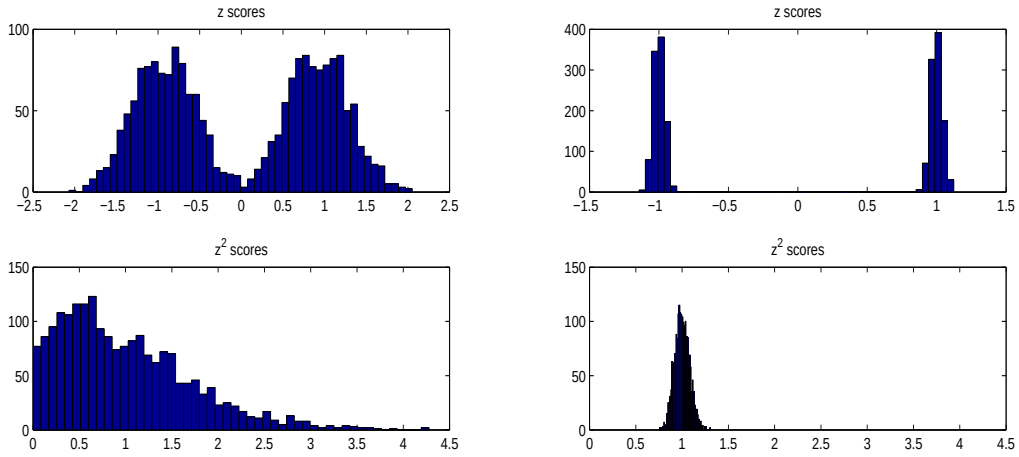
$$s_{z^2} = \frac{1}{n}\sum_{i=1}^{n}(z_i^2 - \bar{x}_{z^2})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^4 - 1 = k - 1, \tag{1.1}$$

using $\bar{x}_{z^2} = 1$, and the kurtosis can be interpreted as the variance of these distances to their mean. Consequently, if all observations of the sample are approximately at the same distance to the mean, the variance of these distances is near zero, and the kurtosis will have a small value. From that, again, since $s_{z^2} \geq 0$, the minimum value for the kurtosis is 1.

More particularly, Darlington (1970) pointed out that k can be understood as a measure of the degree to which the values of z^2 cluster around their mean, of value 1. For the distribution of the z's, since $z = 1$ or $z = -1$ when $z^2 = 1$, the kurtosis can also be interpreted as a measure of the degree to which the z-scores cluster around $+1$ and -1, which is a description of a bimodal distribution. In Figure 1.2 we observe this behaviour in a mixture of two normal distributions. The means are more separated in Figure 1.2(b) than in Figure 1.2(a) and thus the clustering around one is more accentuated in the second case. Darlington illustrates the same idea considering the family of all two-point distributions with densities p and $1-p$ respectively, whose kurtosis value is proven in Darlington (1970) to be

$$k = \frac{1}{p(1-p)} - 3.$$

The minimum value is reached when $p = \frac{1}{2}$, which agrees with the results above regarding bimodality. On the other hand, k approaches infinity when $p \to 1$ or $p \to 0$, i.e. as the



(a) $\mu_1 = 0$, $\mu_2 = 5$.

(b) $\mu_1 = 0$, $\mu_2 = 50$.

Figure 1.2: z- and z^2-scores for a mixture of two normal distributions.

distribution concentrates on one point or the other. Note that the symmetric two point-mass distribution is the only distribution that reaches the minimum kurtosis value of 1.

In the same direction, Hildebrand (1971) considers the symmetric beta distribution family

$$f(x) = \frac{\Gamma(2\alpha)}{\Gamma^2(\alpha)} x^{\alpha-1}(1-x)^{\alpha-1},\ 0 < x < 1,$$

and shows that its kurtosis value is

$$k'_\alpha = \frac{-6}{2\alpha + 3}.$$

If $\alpha = 1$ the distribution is uniform (non-modal) and $k' = -1.2$. For $\alpha < 1$ the distribution is bimodal and $k'_\alpha < -1.2$ approaching -2, the minimum value for k', as $\alpha \to 0$. On the other hand, k' approaches 0 as $\alpha \to \infty$. This example confirms Darlington's statement.

However, when he studies the double gamma distribution family whose density is

$$f(x) = \frac{\beta^\alpha}{2\Gamma(\alpha)} |x|^{\alpha-1} \exp(-\beta|x|),\ -\infty < x < \infty,$$

the value of k' is given by

$$k'_\alpha = \frac{(\alpha+3)(\alpha+2)}{\alpha(\alpha+1)} - 3,$$

regardless the value of the parameter β. For $\alpha = 1$, f is double-exponential and $k' = 3$. If $\alpha < 1$ the distribution is unimodal and $k'_\alpha > 3$ since k'_α is decreasing in α, whereas if $\alpha > 1$ the distribution is bimodal and k'_α ranges from 3 to -2 in the limit. This family, therefore, is inconsistent with Darlington's interpretation since it has values of the kurtosis up to 3 for bimodal distributions.

Moors (1986) claims that bimodal distributions can have large kurtosis and that Darlington's result regarding bimodality should be reexamined. He states that kurtosis measures the dispersion around the values $\mu-\sigma$ and $\mu+\sigma$, instead of the values -1 and $+1$. More explicitly, the kurtosis is an inverse measure of the concentration in these two points. According to Moors, high values of kurtosis arise in two situations; concentration of the probability mass near μ, which corresponds to a peaked distribution, or concentration of the mass in the tails of the distribution.

1.1.3 The influence function for the kurtosis coefficient

Darlington (1970) studied how the kurtosis coefficient changes when new observations are added to a distribution, and calculated the derivative of k with respect to the total change in the size of the distribution, which is proven to be

$$\mathrm{SIF}(z, F, k) = (z^2 - k)^2 - (k^2 - k), \tag{1.2}$$

where F is the distribution function of X, and z is a particular point in the probability distribution of the z-scores. Interestingly, the expression (1.2) is what is now known as the influence function, which was only available in an unpublished thesis at the time of Darlington's paper. The influence function measures what happens to an estimator when the distribution of the data is changed slightly. It was first published by Hampel (1974) and it describes the effect on the estimate of an infinitesimal contamination at a point x of a distribution F. For simplification purposes, (1.2) corresponds to a symmetric influence function for k, in the sense that contamination is considered at the points $-z$ and z. The function is positive if $z^2 < k - (k^2 - k)^{\frac{1}{2}}$ or $z^2 > k + (k^2 - k)^{\frac{1}{2}}$, which implies that both low and high values of z^2 raise the value of k, and intermediate values lower it. If we consider the standard normal distribution, $\mathrm{SIF}(z, \Phi, k)$ is negative for $|z| \in (0.742, 2.334)$ and positive elsewhere, which goes along with Darlington's result of bimodality, since the interval is a neighbourhood of ± 1. And thus, the center can be identified as the range of values $|z| < 0.742$, the flanks are in $.742 < |z| < 2.334$, and the tails correspond to $|z| > 2.334$. Contamination in both the tails and the center of the distribution increases kurtosis. Ruppert (1987) contextualizes Darlington's result within the theory of influence

functions and highlights that in his discussion, Darlington did not pay enough attention to the effect of tail contamination as opposed to center contamination. In effect, if we

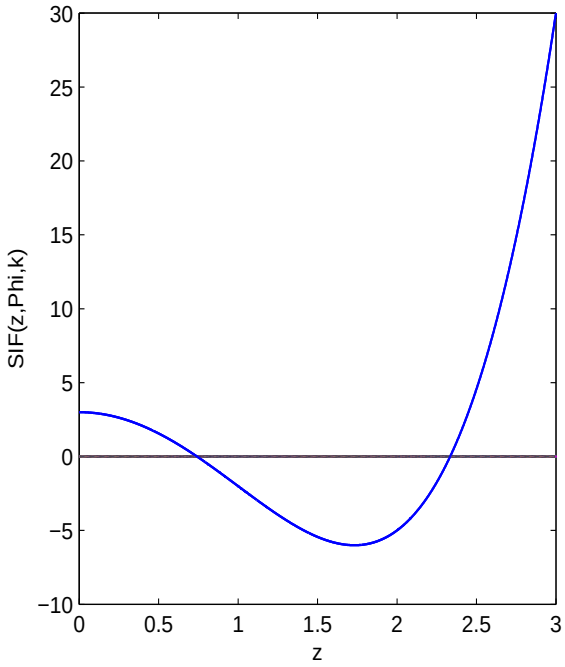


Figure 1.3: Symmetric influence function of the kurtosis coefficient for a normal distribution.

take a look at the symmetric influence function for k in a normal distribution plotted in Figure 1.3, we observe that the function grows fast with z, and so large values of z will raise k considerably. Instead, for values $|z| < 0.742$, the influence function reaches a maximum of only 3 at $z = 0$, and therefore contamination at the center has far less influence than that in the extreme tails. Ruppert states that k is primarily a measure of tail behaviour, and only to a lesser extent of peakedness.

1.1.4 Density crossings to predict the kurtosis value

Dyson (1943) gives a sufficient condition for one distribution to have larger kurtosis than another. Let f_1 and f_2 be standardized to have mean 0 and equal variances, and let μ_{13}, μ_{23}, μ_{14}, μ_{24} be their respective third and fourth moments, a sufficient condition for

$\mu_{14} \leq \mu_{24}$ is that there should exist four abscissae $a_1 < a_2 < a_3 < a_4$ such that

$$\left.\begin{array}{r} -\infty < x < a_1 \\ a_2 < x < a_3 \\ a_4 < x < \infty \end{array}\right\} \Rightarrow f_1 \leq f_2, \qquad \left.\begin{array}{r} a_1 < x < a_2 \\ a_3 < x < a_4 \end{array}\right\} \Rightarrow f_1 \geq f_2$$

and $a_1 + a_2 + a_3 + a_4$ and $\mu_{13} - \mu_{23}$ are not both strictly positive or both strictly negative (in particular that the curves should have equal skewness).

If the conditions hold, the values a_1, a_2, a_3 and a_4 are the points where the densities cross and divide both densities in three parts; tails, shoulders and peak. The first group of conditions identify the tails ($-\infty < x < a_1$, $a_4 < x < \infty$) and the peak ($a_2 < x < a_3$), whereas the second group identifies the flanks ($a_1 < x < a_2$, $a_3 < x < a_4$). Peakedness combined with tailedness or lack of shoulders of one distribution compared to the other imply higher kurtosis. Figure 1.4 illustrates the result for the normal and double-

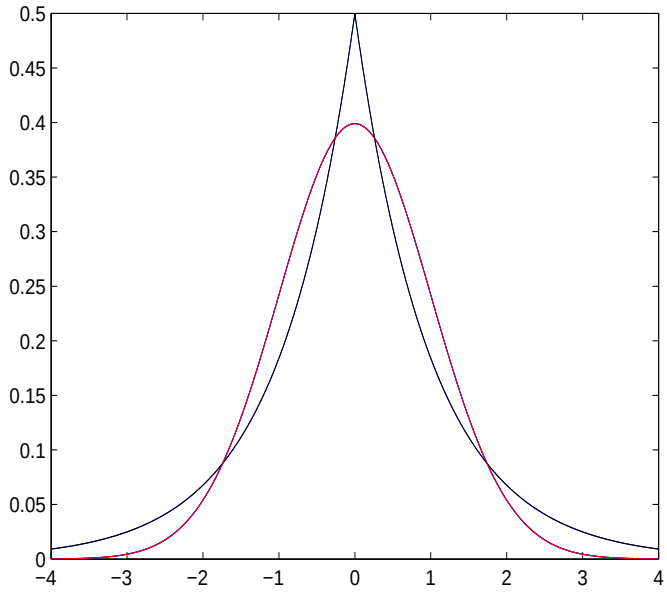


Figure 1.4: Normal and double-exponential distributions satisfying the Dyson's condition.

exponential distributions. The conditions are satisfied for these pair of distributions, while the kurtosis for the normal is smaller. It is emphasized in the paper that although the previous condition is sufficient, it is not necessary. An example of two density curves failing the conditions but with $\mu_{14} \leq \mu_{24}$ is given. In the example, the two curves cross one another four times on each side of the mean. Balanda and MacGillivray (1988) suggest that if the distributions cross more than the required minimum number of times, the

value of k cannot be predicted without more information. According to them, it is the failure to recognize this that causes most of the mistakes and problems in interpreting k.

1.1.5 An ordering-based approach for kurtosis

The previous sections reviewed different uses and interpretations given to the kurtosis coefficient, as well as different attempts to describe those shape characteristics that affect the value of k.

Balanda and MacGillivray (1988) argue that all the interpretations are consistent with the definition of kurtosis as the location- and scale-free movement of probability mass from the shoulders or flanks of a distribution into its center and tails. This definition implies that peakedness and tail weight are best viewed as components of kurtosis, since any movement of mass from the shoulders into the tails must be accompanied by a movement of mass into the center if the scale is to be left unchanged. As it happens with the concepts location, scale, and skewness, the definition is necessarily vague because the movement can be formalized in many ways.

Given that, other definitions of kurtosis, peakedness and tail weight have appeared in the literature. Some of them attempt to measure peakedness or tail weight but they end up measuring both. For example, Horn (1983) proposes an alternative measure of peakedness for symmetric distributions, arguing that the kurtosis coefficient does not exist for all densities. Given the rectangle $R_p(f)$ defined by the lines $x = 0$, $y = 0$, $y = f(0)$ and $x = F^{-1}(p + 0.5)$, for $0 < p < \frac{1}{4}$ and $\mu = 0$, the measure of peakedness is the proportion of area of $R_p(f)$ covered by the density f. Note that the area under the density contained in $R_p(f)$ is always p. This measure ranks in an increasing order of peakedness the normal, t-student with 6 degrees of freedom, Cauchy (from whom kurtosis does not exist) and double-exponential, which seems quite reasonable.

However, as Balanda and MacGillivray (1988) point out, the measure-based approach has received some criticism. For example, van Zwet (1964) claimed that many of the comparisons made with the kurtosis coefficient, and any other measure for that matter, are meaningless. In principle, any two distributions with finite fourth moments could be compared using k, "whereas one feels there are pairs of such distributions that are totally incomparable in this regard". This is due to the fact that a single value for the parameter may correspond to many different density shapes. For example, the normal distribution and the double gamma distribution with $\alpha = \frac{1}{2}(1 + 13^{\frac{1}{2}})$ have both kurtosis $k = 3$, as well as the symmetric Tukey lambda distribution with parameter $\lambda = 5.2$,

and the three distributions correspond to very different distributional shapes; the double gamma is bimodal whereas the symmetric Tukey is considerably more peaked than the normal distribution.

Such reasoning led to the ordering-based approach. Instead of measuring the kurtosis of a given distribution, an order $<<$ is defined in such a way that $F << G$ means, in some sense, that the distribution G has larger kurtosis than F or, in other words, G has more mass in the center and tails than does F. A measure of kurtosis with respect to $<<$ is then restricted to any location- and scale-free nonnegative functional T such that $T(F) << T(G)$ whenever $F << G$; a functional that preserves the ordering. In Balanda and MacGillivray (1988) words, "we believe that a kurtosis measure should not be used without first identifying the ordering underlying it and that a measure should not be used to make comparisons within a family of distributions unless that family is totally ordered by the underlying ordering. It is only in these circumstances that the measure genuinely summarizes a kurtosis property in a meaningful way".

The strongest order that has been considered is the ordering $\leq_S$ introduced by van Zwet (1964) for the class of symmetric distributions: $F \leq_S G$ if and only if $R_{F,G}(x) = G^{-1}(F(x))$ is convex for $x > m_F$, where m_F is the point of symmetry of F. van Zwet (1964) showed that $\cup$-shaped $\leq_S$ uniform $\leq_S$ normal $\leq_S$ logistic $\leq_S$ double-exponential and logistic $\leq_S$ Cauchy, and both the family of double-gamma distributions and the family of symmetric beta distributions are totally ordered by $\leq_S$. The latter allows to make comparisons within these families using k, since it preserves the order. Nevertheless, the examples in Hildebrand (1971) did show that k was inadequate to describe the shape of individual members.

Although two approaches can be taken when studying kurtosis; the measure-based approach and the ordering-based approach, when a new measure of kurtosis is defined, it generally should respect van Zwet's ordering for it to be considered a valid measure of kurtosis.

1.1.6 Kurtosis as a measure of heterogeneity

Despite all the efforts done in the past to provide a good understanding of what kurtosis really measures, the feeling is that the discussion does not bring an unambiguous and final answer to the question. The understanding of kurtosis as the location- and scale-free movement of mass from the shoulders to the tails or peak presented in Balanda and MacGillivray (1988) is difficult to imagine and illustrate. In effect, we cannot take a

distribution, move mass from the tails to the shoulders and at the same time keep the variance as it was; every movement of mass will imply a change on the shape and variance of the distribution, and therefore it will not be a scale-free movement. This limitation makes the interpretation of the coefficient less obvious and straightforward.

We believe that the only practical interpretation or use of the kurtosis coefficient is seen as a measure of heterogeneity. If we define $d_i = (x_i - \bar{x})^2$ as the distances of the observations to the mean, the variance of these distances is a measure of heterogeneity,

$$\tfrac{1}{n}\sum_{i=1}^{n}(d_i - s^2)^2,$$

where the variance of the sample $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} d_i$ is also the mean of the d_i's. In effect, if the d_i's are very different, it may suggest that some observations are very far from the mean and therefore the sample is heterogeneous. On the other hand, if the d_i's are all very similar it might be due to a sample with small variance or a sample generated by two populations of the same size. In order to have a dimensionless measure, a coefficient of homogeneity is defined as

$$H = \frac{\frac{1}{n}\sum_{i=1}^{n}(d_i - s^2)^2}{s^4},$$

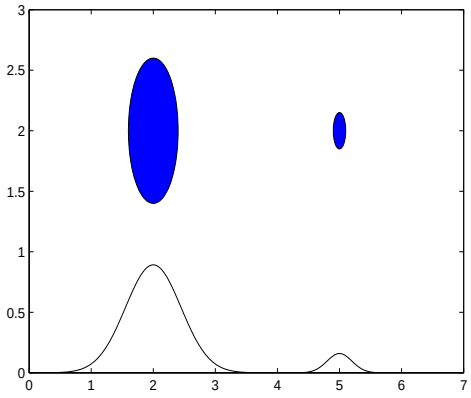
analogous to the coefficient of variation $s/\bar{x}$. Since $\sum_{i=1}^{n}(d_i - s^2)^2 = \sum_{i=1}^{n} d_i^2 + ns^4 - 2s^2\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} d_i^2 - ns^4$, the coefficient H is the variance of the squared scores in (1.1), and thus basically the kurtosis coefficient. Consequently, the univariate kurtosis coefficient can be seen as a measure of heterogeneity. If all observations of the sample are approximately at the same distance to the mean, the variance of these distances is near zero, and the kurtosis will have a small value. This would be the case with two well-separated clusters of the same size and in this case the directions that minimize the kurtosis could reveal the cluster structure.

Heterogeneity arises in several situations. In the following we comment two situations, both giving extreme values of the coefficient of homogeneity/kurtosis.

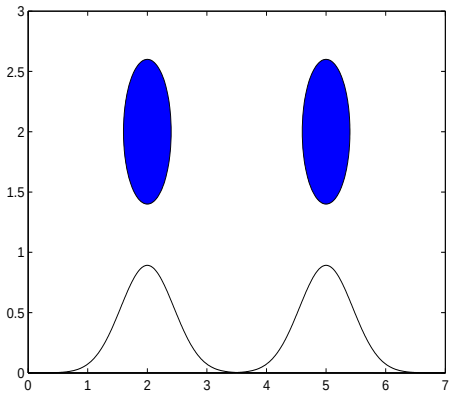
1. In the presence of two clusters of similar size - the mean of the sample will be located in the middle of the two clusters and therefore the distances between the x_i's and the mean will be similar, specially if the clusters are well separated and their variances are small. Then the kurtosis and homogeneity coefficients will have a small value, reaching its minimum in the extreme case of a two point-mass distribution. The same would happen under the presence of three clusters, if the clusters in the extremes have the same size.

2. If we have a sample where most of the observations come from a given distribution, except for some outliers, the mean of the sample will be located near or in the larger cluster, and therefore the distances between the outliers and the mean will be large compared to the other observations, which will make the variance of the distances large, as well as the coefficients of kurtosis and homogeneity.

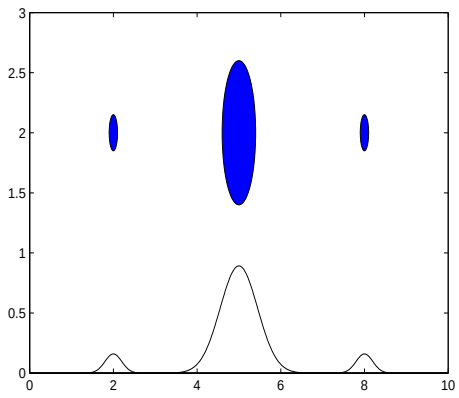
Figure 1.5 illustrates these situations that lead to extreme values of the kurtosis.



(a) A group of outliers - large kurtosis.



(b) Two same-size clusters - small kurtosis.



(c) Two groups of outliers - small kurtosis.

Figure 1.5: The value of the univariate kurtosis coefficient in the presence of clusters.

Therefore, both the directions that minimize the kurtosis coefficient and the ones that maximize it are interesting in the sense that are able to identify structures with more than one cluster. Peña and Prieto (2001) propose a cluster algorithm based on the p directions of minimum and maximum kurtosis. The algorithm starts computing the direction d_i that minimizes k, projects the sample onto the subspace orthogonal to d_i and

searches for a new direction that minimizes k in the subspace. The procedure is repeated iteratively until the whole space is covered, obtaining p directions of minimum kurtosis. Afterwards, the process is conducted again, but this time maximizing k. The algorithm finishes with $2p$ directions that need to be analyzed regarding cluster structure. The second part of the algorithm assigns observations to clusters based on the information found in the projections onto the directions.

In addition to that, they prove that under a mixture of two normal distributions with proportional scatter matrices, either the direction that maximizes or the one that minimizes the kurtosis coefficient is Fisher's linear discriminant function. Let π be the proportion of one of the populations, if $\pi \in (0, (\sqrt{3}-1)/(2\sqrt{3}))$ the Fisher's function is the one that maximizes the kurtosis coefficient, whereas for $\pi \in ((\sqrt{3}-1)/(2\sqrt{3}), 0.5]$ the interesting direction is the one that minimizes it. This result is in agreement with the situations we commented before; if the two clusters are similar in size, with $\pi \in (0.21, 0.5]$, the kurtosis has small value, while if there exists a group of outliers containing at most 20% of the sample, the kurtosis is large.

Heterogeneity can be seen as an extreme case of lack of normality, which explains why some procedures that try to find non-normality use the kurtosis coefficient. For example some of the algorithms used in Independent Component Analysis (ICA, Hyvärinen et al. (2001)) search for those components that maximize the absolute value of the univariate kurtosis coefficient. It is worth mentioning that such algorithms maximize the absolute value of the kurtosis k', which ranges among the values -2 and ∞. But since the range is not symmetric around zero, the maximization of the absolute value would result in prioritizing those directions that maximize the kurtosis as opposed to those that minimize it, and we have already seen in this section and in Section 1.1.2 that minimizing the kurtosis coefficient might also lead to heterogeneity or bimodality and therefore to cases of unequivocal non-normality.

1.2 Kurtosis of multivariate samples

Let $X \in \mathbb{R}^p$ be a multivariate random vector with mean μ and covariance matrix $\Sigma = E[(X-\mu)(X-\mu)^{\mathrm{T}}]$. The p eigenvectors of Σ are found in the space of X. In particular, the eigenvector associated to the largest eigenvalue is the linear combination of the original variables $X_1, \ldots, X_p$ that maximize the variance among all possible linear combinations in $\mathbb{R}^p$, with the value of this variance given by the eigenvalue. The eigenvector associated to the second largest eigenvalue maximizes the variance among all

linear combinations orthogonal to the previous eigenvector, and so on. Geometrically, the eigenvectors represent the axes of the ellipsoid closest to X. The sum of the variances of the p variables coincides with the sum of the variances of the p eigenvectors, since $\text{tr}(\Sigma) = \sum_{i=1}^{p} \sigma_i = \sum_{i=1}^{p} \lambda_i$, where $\sigma_1, \ldots, \sigma_p$ are the variances of the variables and $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of Σ. From that, measures such as the total variation (Seber, 1984) given by $\text{tr}(\Sigma) = \lambda_1 + \ldots + \lambda_p$, the generalized variance (Wilks, 1932) given by $|\Sigma| = \lambda_1 \ldots \lambda_p$, and the effective variance (Peña and Rodríguez, 2003) given by $|\Sigma|^{1/2} = (\lambda_1 \ldots \lambda_p)^{1/2}$ are ways of summarizing in a scalar measure the multivariate variability of the random vector X.

In the multivariate case, ass it happens with the concept of scatter, the concept of kurtosis has to be generalized. In this section we analyze the different attempts that have appeared to define a multivariate kurtosis. Most of these attempts are based on the fourth-order moments and summarize in different ways the information that is found in them.

1.2.1 The Mardia kurtosis and other coefficients

The simplest way to summarize the kurtosis of a multivariate distribution is through a scalar measure. In this section we review some of the multivariate kurtosis coefficients that have been defined in the literature.

As the univariate kurtosis coefficient is the second moment of the squared scores, a natural extension of kurtosis to multivariate random vectors is presented in Mardia (1970) as the second moment of the Mahalanobis distances,

$$\beta_{2,p} = E[(X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)]^2$$

Since $\beta_{2,p}$ can also be expressed as $\beta_{2,p} = \sigma^2_{DM} + \mu^2_{DM}$ and $\mu_{DM} = p$, where $DM = (X-\mu)^{\mathrm{T}}\Sigma^{-1}(X-\mu)$ is the Mahalanobis distance, then $\beta_{2,p} \geq p^2$. Also, if we formulate $\beta_{2,p}$ in terms of the standardized vector $Z = \Sigma^{-1/2}(X-\mu)$,

$$\beta_{2,p} = E[Z^{\mathrm{T}}Z]^2 = \sum_{i=1}^{p} E(Z_i^4) + 2\sum_{i=1}^{p} \sum_{\substack{j=1 \\ i \neq j}}^{p} E(Z_i^2 Z_j^2).$$

Note that $\beta_{2,p}$ depends only on the symmetric fourth-order moments. The coefficient is affine invariant and its sample counterpart is $b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}[(x_i-\bar{x})^{\mathrm{T}}S^{-1}(x_i-\bar{x})]^2$. Mardia (1970) proposes to use $b_{2,p}$ when testing for normality. Under a gaussian distribution $\beta_{2,p} = p(p+2)$, therefore values of $b_{2,p}$ differing significantly from $p(p+2)$ indicate non-normality.

Other coefficients that intend to summarize the kurtosis of a multivariate random vector in a scalar measure are described as follows.

Koziol's kurtosis coefficient. Koziol (1989) defines the following kurtosis measure

$$\tilde{b}_{2,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i^{\mathrm{T}} z_j)^4$$

as the next higher degree analogue to the Mardia's sample measure of skewness $b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i^{\mathrm{T}} z_j)^3$. It can also be written as $\tilde{b}_{2,p} = \sum_{j,k,l,m}^{p}(\frac{1}{n}\sum_{i=1}^{n} z_{ji}z_{ki}z_{li}z_{mi})^2$ and the population counterpart is $\tilde{\beta}_{2,p} = \sum_{j,k,l,m}^{p} E(Z_j Z_k Z_l Z_m)^2$.

The coefficient $\beta_{2,p}$ is the sum of just the symmetric fourth-order moments whereas $\tilde{\beta}_{2,p}$ is the sum of squares for all existing fourth-order moments of Z. As an example, if $p = 2$ then $\beta_{2,p} = \mu_{40} + \mu_{04} + 2\mu_{22}$ and $\tilde{\beta}_{2,p} = \mu_{40}^2 + \mu_{04}^2 + 6\mu_{22}^2 + 4\mu_{31}^2 + 4\mu_{13}^2$.

Oja's kurtosis coefficient. Oja (1983) defines a multivariate kurtosis coefficient by considering the volume of the simplex in a p-dimensional space determined by $p+1$ points as

$$\beta_{2,p}^{*} = \frac{E[\Delta(X_1, \ldots, X_p, \mu)]^4}{[E[\Delta(X_1, \ldots, X_p, \mu)]^2]^2},$$

being $X_1, \ldots, X_p$ independent random vectors distributed as X and Δ the volume of this simplex:

$$\Delta\left(X_1, \ldots, X_{p+1}\right) = \mathrm{abs}\left(\left|\begin{array}{c|ccc} & 1 & \ldots & 1 \\ \frac{1}{p!} & X_{11} & \ldots & X_{p+1,1} \\ & \vdots & & \vdots \\ & X_{1p} & \ldots & X_{p+1,p} \end{array}\right|\right).$$

Malkovich and Afifi's kurtosis coefficient. Malkovich and Afifi (1973) define the multivariate kurtosis as the maximum univariate kurtosis produced by any projection of the p-dimensional distribution onto a direction d; $\beta_2^M = \max_d |\beta_2^d - 3|$, where $\beta_2^d = E\left[\frac{(d^{\mathrm{T}}X - d^{\mathrm{T}}\mu)^4}{d^{\mathrm{T}}\Sigma d}\right]$.

The measures $\beta_{2,p}$, $\beta_{2,p}^{*}$ and β_2^M are invariant under nonsingular affine transformations and reduce to the univariate kurtosis when $p = 1$, which is not the case for $\tilde{\beta}_{2,p}$.

1.2.2 Matrices of kurtosis and cumulants

The mean of the random vector X is a vector of dimension p, the covariance matrix a $p \times p$ matrix that contains the $\frac{p(p+1)}{2}$ distinct second-order moments, and, by analogy, we would

need a cube of dimensions $p \times p \times p$ to contain the third-order central moments and an object in a fourth-dimensional space to contain the fourth-order central moments. Since it is easier to work with matrices, what has been done is to collect in a matrix the η distinct fourth-order central moments, where $\eta = p + \frac{3p(p-1)}{2} + \frac{p(p-1)(p-2)}{2} + \frac{p(p-1)(p-2)(p-3)}{4!}$. In this section we review the different ways of collecting this information in a matrix.

Matrices of kurtosis

The matrix $E[(X-\mu)(X-\mu)^{\mathrm{T}} \otimes (X-\mu)(X-\mu)^{\mathrm{T}}]$ of dimensions $p^2 \times p^2$, where $\otimes$ denotes the Kronecker product, contains the η distinct fourth-order central moments. As it happens with the covariance matrix, the symmetric versions are also included. The univariate kurtosis coefficient k is standardized to be scale-free by dividing it by s^4, and to extend the idea of kurtosis to the multivariate case we want to maintain the invariance property and therefore the corresponding standardized matrix will be

$$M_4 = E\left[\Sigma^{-1/2}(X-\mu)(X-\mu)^{\mathrm{T}}\Sigma^{-1/2} \otimes \Sigma^{-1/2}(X-\mu)(X-\mu)^{\mathrm{T}}\Sigma^{-1/2}\right],$$

which results in

$$M_4 = E\left(ZZ^{\mathrm{T}} \otimes ZZ^{\mathrm{T}}\right).$$

A detailed expression of the matrix is

$$M_4 = E\begin{pmatrix} Z_1^2 \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} [Z_1 \dots Z_p] & \dots & Z_1 Z_p \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} [Z_1 \dots Z_p] \\ \vdots & \ddots & \vdots \\ Z_p Z_1 \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} [Z_1 \dots Z_p] & \dots & Z_p^2 \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} [Z_1 \dots Z_p] \end{pmatrix}. \tag{1.3}$$

Unlike the covariance matrix, which has as dimensions those of the space of X, this matrix has dimensions $p^2 \times p^2$, which complicates its use. For instance, the eigenvectors of this matrix do not belong to the space of the variables. This fact has led to different definitions for a kurtosis matrix of dimensions $p \times p$.

Cardoso (1989) and Móri et al. (1993) define the following kurtosis matrix

$$K = E(Z^{\mathrm{T}} ZZ Z^{\mathrm{T}}). \tag{1.4}$$

The matrix is the sum of the p diagonal blocks of size $p \times p$ of M_4, and contains only $p + \frac{3p(p-1)}{2} + \frac{p(p-1)(p-2)}{2}$ distinct fourth-order moments, since the moments of the kind $Z_i Z_j Z_k Z_l$, with $i \neq j \neq k \neq l$ are not there. Also, observe that the cells contain the sum of p moments, as opposed to (1.3), where a cell corresponded to a single moment:

$$K = I_p * M_4 = E \begin{pmatrix} Z_1^2(Z_1^2 + \ldots + Z_p^2) & \ldots & Z_p Z_1(Z_1^2 + \ldots + Z_p^2) \\ \vdots & \ddots & \vdots \\ Z_1 Z_p(Z_1^2 + \ldots + Z_p^2) & \ldots & Z_p^2(Z_1^2 + \ldots + Z_p^2) \end{pmatrix}.$$

The symbol $*$ denotes the star product defined as follows (McRae, 1974). Let A be a $m \times n$ matrix and B be a $mp \times nq$ matrix, the star product of A and B is a $p \times q$ matrix C defined by

$$C = A * B = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} B_{ij}$$

where a_{ij} is the ijth element of A, and B_{ij} is the ijth block of B, where B is partitioned into blocks of dimension $p \times q$.

The matrix K reduces to the univariate kurtosis coefficient in the univariate case,

$$K = E(ZZZZ) = E(Z^4) = \frac{\mu_4}{\sigma^4}$$

and is positive semidefinite,

$$x^{\mathrm{T}} K x = x^{\mathrm{T}} E(ZZ^{\mathrm{T}} ZZ^{\mathrm{T}}) x = E[(ZZ^{\mathrm{T}} x)^{\mathrm{T}} ZZ^{\mathrm{T}} x] \geq 0, \text{ for } x \in \mathbb{R}^p.$$

The sample counterpart of K is,

$$K_n = \frac{1}{n} \sum_{i=1}^{n} z_i^{\mathrm{T}} z_i z_i z_i^{\mathrm{T}}$$

where $z_i = S^{-\frac{1}{2}}(x_i - \bar{x})$ and $\bar{x}$ and S are the mean and covariance matrix of a random sample $x_1, \ldots, x_n$ of X. The trace of K coincides with the Mardia's kurtosis coefficient,

$$\operatorname{tr} K = \operatorname{tr}[E(Z^{\mathrm{T}} ZZZ^{\mathrm{T}})] = E[Z^{\mathrm{T}} Z \operatorname{tr}(ZZ^{\mathrm{T}})] = E[(Z^{\mathrm{T}} Z)^2] = \beta_{2,p}. \tag{1.5}$$

If X follows an elliptical distribution with density

$$f_X(x) = |V|^{-\frac{1}{2}} h[(x - \mu)^{\mathrm{T}} V^{-1} (x - \ \mu)],$$

for some nonnegative function h, the matrix K is diagonal. In effect, the covariance matrix of the elliptical distribution is $\Sigma = cV$ for some $c \in \mathbb{R}$ and the standardized random vector Z is spherical because its density only depends on z through $z^{\mathrm{T}} z$. The

odd moments are zero $E(Z_i Z_j Z_k^2) = E(Z_i Z_j^3) = 0$, since $f_Z(z)$ is an even function of z_i, and the elements of K are $K_{ij} = E(Z_i Z_j \sum_{k=1}^p Z_k^2)$, where the diagonal elements are given by

$$K_{ii} = E(Z_i^4) + \sum_{\substack{k=1 \\ k \neq i}}^{p} E(Z_i^2 Z_k^2) = E(Z_1^4) + \sum_{\substack{k=1 \\ k \neq 1}}^{p} E(Z_1^2 Z_k^2), \quad 1 \leq i \leq p$$

since the Z_i's are identically distributed, and the off-diagonal elements are zero. Therefore, the matrix K is proportional to the identity.

In particular, if X follows a multivariate normal distribution, the diagonal elements of K are $K_{ii} = p + 2$ since $E(Z_i^4) = 3$ and $E(Z_i^2 Z_j^2) = 1$ and thus $K = (p+2)I$.

Also, if X follows a multivariate t distribution with parameters ν, μ and R, $K = (p+2)(\nu-2)/(\nu-4)I$ since $E(Z_i^4) = 3(\nu-2)/(\nu-4)$ and $E(Z_i^2 Z_k^2) = (\nu-2)/(\nu-4)$. This last result is consistent with the univariate case, where the kurtosis of the Student t distribution is higher than the kurtosis of the normal distribution due to its heavier tails.

Let $\mu_{r_1,\ldots,r_p} = E[\prod_{j=1}^p Z_j^{r_j}]$ be a k-order moment of X, $r_1 + \cdots + r_p = k$, then $\hat{\mu}_{r_1,\ldots,r_p}$ converges to $\mu_{r_1,\ldots,r_p}$ in probability and, since K is a continuous function of the moments, K_n converges to K in probability and the matrix K_n is a consistent estimator of K.

Kollo (2008) defines another kurtosis matrix as

$$B = E\left[(Z^{\mathrm{T}}\mathbf{1})^2 Z Z^{\mathrm{T}}\right], \tag{1.6}$$

The matrix B is the sum of the p^2 blocks of size $p \times p$ of M_4, and therefore contains the η distinct fourth-order moments. This time tzhe cells are sums of p^2 moments:

$$B = \mathbf{1}_{p\times p} * M_4 = E\begin{pmatrix} Z_1^2(Z_1 + \ldots + Z_p)^2 & \ldots & Z_p Z_1(Z_1 + \ldots + Z_p)^2 \\ \vdots & \ddots & \vdots \\ Z_1 Z_p(Z_1 + \ldots + Z_p)^2 & \ldots & Z_p^2(Z_1 + \ldots + Z_p)^2 \end{pmatrix}.$$

The sample counterpart of B is

$$B_n = \frac{1}{n}\sum_{i=1}^{n}(z_i^{\mathrm{T}}\mathbf{1})^2 z_i z_i^{\mathrm{T}},$$

and the trace of B is the sum of all elements of the matrix K

$$\mathrm{tr}(B) = (Z_1^2 + \ldots + Z_p^2)(Z_1 + \ldots + Z_p)^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} K_{ij}.$$

B also reduces to the univariate kurtosis coefficient in the univariate case,

$$B = E(Z^2 Z Z) = E(Z^4) = \frac{\mu_4}{\sigma^4},$$

and is positive semidefinite,

$$x^{\mathrm{T}}Bx = x^{\mathrm{T}}E[(Z^{\mathrm{T}}\mathbf{1})^2 ZZ^{\mathrm{T}}]x = E[\{(Z^{\mathrm{T}}\mathbf{1})Z^{\mathrm{T}}x\}^{\mathrm{T}}(Z^{\mathrm{T}}\mathbf{1})Z^{\mathrm{T}}x] \geq 0, \text{ for } x \in \mathbb{R}^p.$$

Under the assumption of an elliptical distribution for X, B is not diagonal because the term $E(Z_i^2 Z_j^2)$ appears in the off-diagonal elements, but the diagonal elements have the same value as in K. For a multivariate normal distribution, for example, $B_{ij} = 2$ for $i \neq j$ and $B_{ii} = p + 2$ otherwise. For a multivariate t distribution with parameters ν, μ and R, $B_{ij} = 2(\nu - 2)/(\nu - 4)$. The matrix B contains redundant information as it depends for its construction on the vector $\mathbf{1}$. More details on the eigenstructure of B are explained in Section 2.2.1.

Due to the convergence of moments, B_n converges to B in probability and is a consistent estimator for B.

Both matrices K and B can be seen as weighted scatter matrices with weights $Z^{\mathrm{T}}Z$ and $(Z^{\mathrm{T}}\mathbf{1})^2$ respectively. The matrix K in (1.4) has an important invariant property which is not present in B in (1.6).

Let E be an orthogonal matrix whose columns are eigenvectors of K, the new coordinate system $E^{\mathrm{T}}Z$ is invariant under affine transformations of X. In effect, if $Y = AX + b$ with A non-singular, then $K_Y = UKU^{\mathrm{T}}$, where U is some orthogonal matrix. This is true because the standardizations of X and Y are the same up to a rotation, $Z_Y = UZ$, where $Z_Y = \Sigma_Y^{-1/2}(Y - \mu_Y)$. That implies that the eigenvalues of K and K_Y are the same and the eigenvectors are rotated versions of each other (the eigenvectors of K_Y are UE). When applying the same transformation to Z_Y, we obtain the same coordinates $E^{\mathrm{T}}U^{\mathrm{T}}UZ = E^{\mathrm{T}}Z$. The matrix B, however, does not have this desirable property because its weights are not invariant under orthogonal transformations.

Oja et al. (2006) consider a scatter matrix based on fourth-order moments,

$$\tilde{S} = E[Z^{\mathrm{T}}Z(X - \mu)(X - \mu)^{\mathrm{T}}],$$

which is related to K by $\tilde{S} = \Sigma^{1/2}K\Sigma^{1/2}$. If instead of $\tilde{S}$ we consider $\tilde{K} = \tilde{S}\Sigma^{-1}$, then $\tilde{K} = \Sigma^{1/2}K\Sigma^{-1/2}$ and the matrices K and $\tilde{K}$ share the same eigenvalues and trace. Also, if u is an eigenvector of K, $\Sigma^{1/2}u$ is an eigenvector of $\tilde{K}$. Both matrices $\tilde{S}$ and $\tilde{K}$ are positive semidefinite. The matrix $\tilde{K}$ reduces to the univariate kurtosis coefficient in the univariate case,

$$\tilde{K} = E[ZZ(X - \mu)(X - \mu)\sigma^{-2}] = E(Z^4) = \frac{\mu_4}{\sigma^4},$$

which is not the case for $\tilde{S}$. If X follows an elliptical distribution, $\tilde{K} = K$ and $\tilde{S} = K\Sigma$.

In summary, in Chapters 2 and 3, we explore the properties of the matrix K, as we have seen it has an invariance property that the matrix B does not hold. Note also that studying K is equivalent to study $\tilde{K}$.

Matrices of cumulants

Independent Component Analysis (ICA, Hyvärinen et al. (2001)), a methodology whose purpose is to find the independent latent factors that have generated the observed multivariate sample, assumes that the variables X are generated by the independent latent factors S through the following model,

$$X = AS.$$

In order to find $S \in \mathbb{R}^p$ it is necessary to specify the matrix A, or, if we first whiten X with a matrix W, where for example W can be the matrix $\Sigma^{-1/2}$ and it results in standardized variables, the model simplifies to

$$Z = US,$$

where $Z = WX$ and $U = WA$ is an orthogonal matrix since the whitening condition makes $E(ZZ^{\mathrm{T}}) = I$ and the factors are uncorrelated, $E(SS^{\mathrm{T}}) = I$. ICA uses several approaches to specify the orthogonal matrix U. We will focus here on those approaches related to fourth-order moment matrices. More particularly, in Cardoso and Souloumiac (1993) a matrix based on fourth-order cumulants is used to find the latent factors.

The definition of fourth-order cumulants differs from the definition of fourth-order moments in some second-order moments,

$$\begin{aligned}\operatorname{cum}(Z_iZ_jZ_kZ_l) = {} & E(Z_iZ_jZ_kZ_l) - E(Z_iZ_j)E(Z_kZ_l) \\ & -E(Z_iZ_k)E(Z_jZ_l) - E(Z_iZ_l)E(Z_jZ_k)\end{aligned}$$

Two approaches for the determination of U have been reported and are summarize in Cardoso and Souloumiac (1993). In the first approach, the columns of U are the eigenvectors of a $p \times p$ cumulant matrix. They define first a cumulant set denoted by

$$Q_Z = \{\operatorname{cum}(Z_iZ_jZ_kZ_l) \quad | \quad 1 \leq i,\, j,\, k,\, l \leq p\},$$

which contains all p^4 fourth-order cumulants of the vector Z.

A cumulant matrix N_M is a $n \times n$ matrix defined entrywise by

$$n_{ij} = \textstyle\sum_{k=1}^{p} \sum_{l=1}^{p} \operatorname{cum}(Z_iZ_jZ_kZ_l) m_{kl}, \quad 1 \leq i,\, j \leq p$$

where the matrix with entries m_{kl} has to be specified, although the usual choice is $M = b_k b_l^{\mathrm{T}}$, with b_k denoting the $n \times 1$ vector with 1 in the kth position and 0 elsewhere. When this matrix is used, $n_{ij} = \mathrm{cum}(Z_i Z_j Z_k Z_l)$ and therefore M_N contains one cumulant in each cell. This approach uses only a fraction of the fourth-order information; p^2 cumulants out of p^4, and there is no clue a priori to which matrix M should be chosen. An alternative idea is to compute several matrices by randomly selecting k and l, and choose the one whose eigenvalues present the maximum spread, but the information contained in the other cumulant matrices will be still lost. The larger the spread between eigenvalues the higher the possibilities of finding an interesting pattern in X, since it implies that some eigenvector is giving a very spread and therefore informative projection of X.

The problem of which moments/cumulants should be included arises with any fourth-order moment or cumulant matrix of dimension $p \times p$, which is why it does not exist a kurtosis matrix of reference, as it is the covariance matrix for the second-order information, since the natural way of representing this information is not a matrix (it would be an object of four dimensions).

If we choose $M = I$, each cell of M_N is the sum of several cumulants, and it coincides with the choice in K, if we considered cumulants instead of moments.

It is unclear whether the addition process for moments/cumulants in each cell results in a smarter way of arranging the matrix, or in a matrix that contains more information. For instance, the matrix B was redundant with respect to K, but it contained all fourth-order moments, unlike K that uses only p^3 fourth-order moments.

The other approach mentioned in Cardoso and Souloumiac (1993) obtains an estimate of U as the optimizer of some identification criterion which is a function of the whole cumulant set Q_Z, with which "better performance is expected at the expense of solving an optimization problem". Cardoso and Souloumiac (1993) finally propose a technique that combines advantages of both the eigenvalue-based and the criterion-based approach.

1.2.3 Heterogeneity of multivariate samples

Our intention is to explore the properties of multivariate kurtosis measures to perform cluster analysis. We aim to analyze whether heterogeneity is an appropriate interpretation of multivariate kurtosis, and we start by studying the behaviour of a multivariate kurtosis coefficient under a mixture of distributions. In particular, we consider the Mardia (1970)'s kurtosis coefficient, since it is the most widely used and well-known scalar measure of multivariate kurtosis.

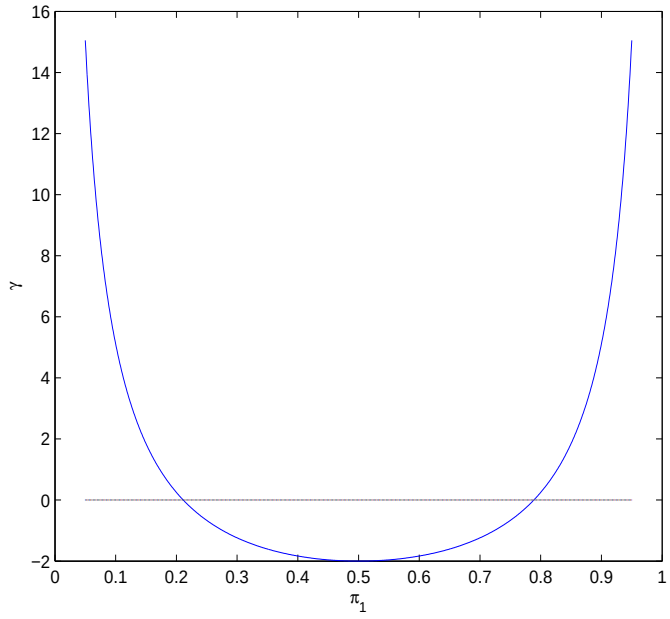


Figure 1.6: The value of γ for different values of π_1.

Let X be distributed as $\pi_1 f_1(X) + \pi_2 f_2(X)$, where f_i is a normal density with mean μ_i and covariance matrix V, and the π_i's are the weights of the mixture. Following expression (1.5), the trace of K is the Mardia's kurtosis coefficient, and thus we can derive the expression for $\beta_{2,p}$ using (2.5),

$$\beta_{2,p} = \operatorname{tr} K = p(p+2) + \beta(\varphi^{\mathrm{T}}\varphi)^2, \tag{1.7}$$

where $\beta = \pi_1\pi_2[1 - 6\pi_1\pi_2]$ and $\varphi = \Sigma^{-1/2}(\mu_2 - \mu_1)$, which can be expressed in terms of the covariance matrix of the components of the mixture as

$$\beta_{2,p} = p(p+2) + \frac{\beta[(\mu_2 - \mu_1)^{\mathrm{T}} V^{-1}(\mu_2 - \mu_1)]^2}{[\pi_1\pi_2(\mu_2 - \mu_1)^{\mathrm{T}} V^{-1}(\mu_2 - \mu_1) + 1]^2}$$

since $\Sigma = V + \pi_1\pi_2(\mu_2 - \mu_1)(\mu_2 - \mu_1)^{\mathrm{T}}$ and from the inverse of the sum property,

$$\Sigma^{-1} = V^{-1} - \frac{\pi_1\pi_2 V^{-1}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^{\mathrm{T}} V^{-1}}{\pi_1\pi_2(\mu_2 - \mu_1)^{\mathrm{T}} V^{-1}(\mu_2 - \mu_1) + 1}.$$

The first term in (1.7) is the value of $\beta_{2,p}$ under a normal distribution, whereas the second term indicates deviations from it. If the means of the two populations are the same then $\beta_{2,p} = p(p+2)$, otherwise we are in the mixture case. In the following we analyze how $\beta_{2,p}$ changes when we move the means away from each other by calculating the value of $\beta_{2,p}$ when the distance between the means tends to infinity,

$$\lim_{\|\mu_2 - \mu_1\| \to \infty} \beta_{2,p} = p(p+2) + \gamma \tag{1.8}$$

where $\gamma = \frac{1-6\pi_1\pi_2}{\pi_1\pi_2}$, and therefore the value of $\beta_{2,p}$ in the limit depends on the proportion of the mixtures: it increases respect to $p(p+2)$ when $1-6\pi_1\pi_2 > 0$ and decreases otherwise. If $1-6\pi_1\pi_2 < 0$, it implies that $\pi_1 \in ((\sqrt{3}-1)/(2\sqrt{3}), 0.5]$ and in the extreme case of $\pi_1 = \pi_2 = 1/2$, $\gamma = -2$ and $\beta_{2,p} = p(p+2) - 2$, which is the maximum distance that can be reached with respect to $p(p+2)$ for negative values of γ. On the other hand, if $1 - 6\pi_1\pi_2 > 0$ then $\pi_1 \in (0, (\sqrt{3}-1)/(2\sqrt{3}))$, Figure 1.6 illustrates how γ depends on values of π_1. In the latter case of $\gamma > 0$, the departure from the normal assumption can be made as large as we want for example by selecting sufficiently small values of π_1, since

$$\lim_{\pi_1 \to 0} \gamma = \infty,$$

Observe that the limit in (1.8) is reached fast, in Figure 1.7 we see that as soon as the means start to separate, $\beta_{2,p}$ reaches its limit value, around 13 in this case for a value of $\pi_1 = 0.1$ and a two-dimensional population. Thus, the Mardia's kurtosis coefficient

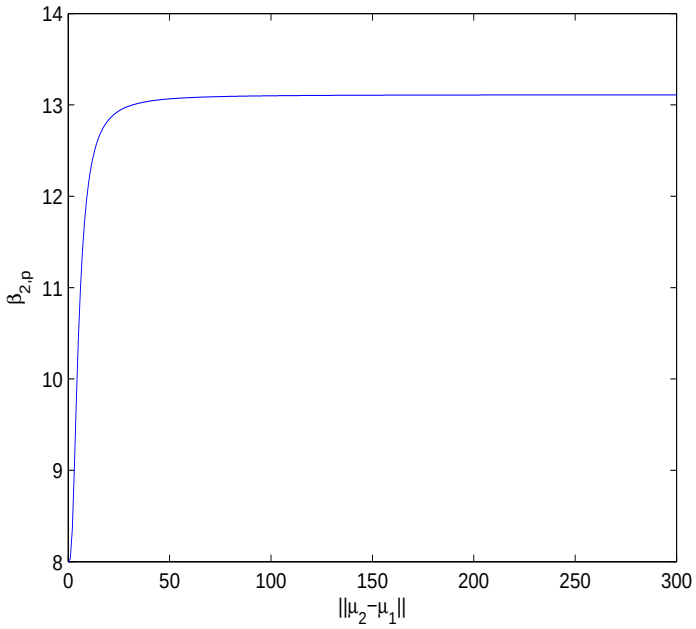


Figure 1.7: The coefficient $\beta_{2,p}$ in function of $\|\mu_2 - \mu_1\|$.

can as well be seen as a measure of heterogeneity. Large values of $\beta_{2,p}$ with respect to $p(p+2)$ may indicate the presence of two different-sized clusters or groups of outliers, whereas small values of the coefficient detect bimodality. Note that this behaviour is a generalization to multivariate samples of the properties of k to detect heterogeneity. The coefficient, thus, may be used to search for optimal subspaces with interesting properties for clustering. In this case, the reduction of the dimension would not be limited to a

direction but to a plane or hyperplane where the identification of the clusters will be easier than in the original space. The subspace might be able to reveal non-linear cluster structures that are not identifiable when projecting onto directions.

However, an optimization algorithm is needed to identify those subspaces with values of minimum and maximum Mardia's kurtosis. An alternative is to explore the properties of the kurtosis matrices introduced in this chapter, and study whether the eigenvectors define any interesting subspace. Using the eigenvectors of a given matrix avoids the need to perform numerical optimization, which can be computationally intensive and its efficacy may depend on the choice of the optimization algorithm to be used. In Chapter 2 we intend to project the multivariate sample onto a subspace generated by some of the eigenvectors of the kurtosis matrix K in (1.4), expecting that this new coordinate system will give us insight on the cluster structure of the data.

Chapter 2

Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure

In this chapter we study the properties of a kurtosis matrix and propose its eigenvectors as interesting directions to reveal the possible cluster structure of a data set. Under a mixture of elliptical distributions with proportional scatter matrices, it is shown that a subset of the eigenvectors of the fourth-order moment matrix corresponds to Fisher's linear discriminant subspace. The eigenvectors of the estimated kurtosis matrix are consistent estimators of this subspace and its calculation is easy to implement and computationally efficient, which is specially favourable when the ratio n/p is large.

2.1 Introduction

Given a multivariate sample in $\mathbb{R}^p$ drawn from a mixture of k populations, cluster analysis attempts to partition the sample into homogeneous groups, according to the populations that generate them.

Projection Pursuit finds subspaces of low dimension that show interesting views of the data according to some criteria, see Friedman and Tukey (1974) and Friedman (1987). Projection Pursuit can be useful in cluster analysis. One may first reduce the dimensionality of the sample by projecting it on a lower dimensional subspace and then finding the clusters there. The curse of dimensionality can thus be avoided, but care needs to be taken to make sure that the projected data preserve the cluster structure of the original

sample. Non-normality is one of the criteria used to find the projections. Huber (1985) emphasized that interesting projections are those that produce non-normal distributions. However, non-normality is a general condition, and we need to specify how to measure it.

The idea of maximizing the kurtosis has also been used in cluster analysis, see Jones and Sibson (1987). Peña and Prieto (2001) showed that for clustering the directions that minimize the kurtosis can be more useful than the ones that maximize it. The reason is that the kurtosis can be seen as the variance of the squared standardized differences between the variable and its mean. Consequently, if all observations of the sample are approximately at the same distance to the mean, the variance of these distances is near zero, and the kurtosis will have a small value. This would be the case with two well-separated clusters of the same size. Therefore, directions that minimize the kurtosis could reveal the cluster structure. The method proposed by Peña and Prieto (2001) (and Projection Pursuit methods in general) needs to perform numerical optimization in order to find the optimal directions. This is computationally intensive and its efficacy may depend on the choice of the optimization algorithm to be used.

An alternative to this approach is to find a matrix whose eigenvectors are related to directions of maximum or minimum kurtosis. In this chapter we study a kurtosis matrix and show that under a mixture of two elliptical distributions with the same scatter matrices, the eigenvector associated to the eigenvalue different from the others coincides with the direction that optimizes the kurtosis coefficient, which is Fisher's linear discriminant function. The kurtosis matrix, thus, has similarities to the nonlinear cluster algorithm in Peña and Prieto (2001). Based on this result, we explore the general case of k groups and we prove that the subspace orthogonal to the eigenspace associated to an eigenvalue with multiplicity $p-k+1$ is Fisher's linear discriminant subspace. Similar results are found in Caussinus and Ruiz-Gazen (1993) and Caussinus and Ruiz-Gazen (1995), where it is shown that Fisher's subspace can be estimated using the k largest eigenvectors of some Generalized Principal Components matrix based on W-estimates of dispersion. Recently, Tyler et al. (2009) prove that a subset of eigenvectors of $V_1^{-1}V_2$ generate Fisher's subspace, being V_1 and V_2 any pair of affine equivariant scatter matrices.

The kurtosis matrix, however, is based on an existent kurtosis-based algorithm which can always be used. The advantage of using the eigenvectors of a kurtosis matrix instead of the univariate kurtosis directions is dependent on the ratio n/p, where n is the sample size and p the dimension. If this ratio is large, the estimation of the kurtosis matrix of dimension p is reliable and therefore the estimation of its eigenvectors becomes accu-

rate and useful. Also, in this case numerical optimization is computationally intensive. However, when n/p is small the estimation of the elements of the matrix has very low precision and we have found that the eigenvalues are not as useful. We will illustrate in which situations is more adequate to use one approach or another. Also, we will show that these eigenvectors are consistent estimators of Fisher's subspace, which ensures their convergence.

This chapter is organized as follows. In Section 2.2 we study the theoretical properties of the eigenvectors of a kurtosis matrix for cluster analysis and present results regarding the convergence of their estimators. In Section 2.3 the behaviour of the eigenvectors to perform cluster analysis is analyzed through a simulation study. We finish with some remarks in Section 2.4.

2.2 The eigenvectors of a kurtosis matrix and its cluster properties

Let X follow a mixture of k elliptical distributions such that, with probability $\pi_i > 0$, X_i has density

$$f_{X_i}(x) = |V_i|^{-1/2} h_i[(x-\mu_i)^{\mathrm{T}} V_i^{-1}(x-\mu_i)], \tag{2.1}$$

for some nonnegative function h_i, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} \pi_i = 1$. We standardize X using its global mean $\mu = \sum_i \pi_i \mu_i$, and covariance matrix $\Sigma = \sum_i \pi_i V_i + \sum_i \pi_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$. The standardized variable $Z = \Sigma^{-1/2}(X - \mu)$ is also a mixture of elliptical distributions Z_i with means and scatter matrices δ_i and W_i, $\delta_i = \Sigma^{-1/2}(\mu_i - \mu)$ and $W_i = \Sigma^{-1/2} V_i \Sigma^{-1/2}$. Using expectation properties the kurtosis matrix K is,

$$K = E(Z^{\mathrm{T}} Z Z Z^{\mathrm{T}}) = \sum_{i=1}^{k} \pi_i E(Z_i^{\mathrm{T}} Z_i Z_i Z_i^{\mathrm{T}}).$$

The fourth-order moment matrix can be expressed as

$$\begin{aligned} E(Z_i^{\mathrm{T}} Z_i Z_i Z_i^{\mathrm{T}}) &= E[(Z_i - \delta_i)^{\mathrm{T}}(Z_i - \delta_i)(Z_i - \delta_i)(Z_i - \delta_i)^{\mathrm{T}}] \\ &\quad + \operatorname{tr} W_i \delta_i \delta_i^{\mathrm{T}} + \delta_i^{\mathrm{T}} \delta_i W_i + 2(\delta_i \delta_i^{\mathrm{T}} W_i + W_i \delta_i \delta_i^{\mathrm{T}}) + \delta_i^{\mathrm{T}} \delta_i \delta_i \delta_i^{\mathrm{T}}, \end{aligned}$$

where we have used that $Z_i = W_i^{1/2} Y + \delta_i$, with Y following a spherical distribution, the intermediate results $E(Z_i Z_i^{\mathrm{T}}) = W_i + \delta_i \delta_i^{\mathrm{T}}$, $E(Y^{\mathrm{T}} W_i Y) = \operatorname{tr} W_i$, $E(\delta_i^{\mathrm{T}} W_i^{1/2} Y W_i^{1/2} Y \delta_i^{\mathrm{T}}) = E(W_i^{1/2} Y Y^{\mathrm{T}} W_i^{1/2} \delta_i \delta_i^{\mathrm{T}})$ and the fact that all odd moments of Y are equal to zero.

The fourth-order central moment matrix of Z_i is

$$\begin{aligned} M_4 &= E[(Z_i-\delta_i)^{\mathrm{T}}(Z_i-\delta_i)(Z_i-\delta_i)(Z_i-\delta_i)^{\mathrm{T}}] \\ &= |W_i|^{-1/2}\int (z-\delta_i)^{\mathrm{T}}(z-\delta_i)(z-\delta_i)(z-\delta_i)^{\mathrm{T}}h_i((z-\delta_i)^{\mathrm{T}}W_i^{-1}(z-\delta_i))dz \\ &= \int y^{\mathrm{T}}W_i y W_i^{1/2} y y^{\mathrm{T}} W_i^{1/2} h_i(y^{\mathrm{T}}y)dy = W_i^{1/2}U\int t^{\mathrm{T}}\Omega t t t^{\mathrm{T}}h_i(t^{\mathrm{T}}t)dtU^{\mathrm{T}}W_i^{1/2} \\ &= W_i^{1/2}U\sum_j \omega_j \int t_j^2 t t^{\mathrm{T}}h_i(t^{\mathrm{T}}t)dtU^{\mathrm{T}}W_i^{1/2} = \sum_j \omega_j \tilde{k}_i W_i + \bar{k}_i W_i^{1/2}U\Omega U^{\mathrm{T}}W_i^{1/2} \\ &= \tilde{k}_i \operatorname{tr} W_i W_i + \bar{k}_i W_i^2, \end{aligned}$$

where we have introduced $y = W_i^{-1/2}(z-\delta_i)$, $t = U^{\mathrm{T}}y$ and $\int t_j^2 t t^{\mathrm{T}} h_i(t^{\mathrm{T}}t)dt = \tilde{k}_i I + \bar{k}_i e_j e_j^{\mathrm{T}}$ for $\tilde{k}_i = \int t_j^2 t_k^2 h_i(t^{\mathrm{T}}t)dt$ where $j \neq k$, and $\bar{k}_i = \int t_j^4 h_i(t^{\mathrm{T}}t)dt - \tilde{k}_i$. Thus, K reduces to

$$\begin{aligned} K &= \sum_{i=1}^{k} \pi_i [\operatorname{tr} W_i(\tilde{k}_i W_i + \delta_i\delta_i^{\mathrm{T}}) + \bar{k}_i W_i^2] \\ &\quad + \sum_{i=1}^{k} \pi_i [2(\delta_i\delta_i^{\mathrm{T}}W_i + W_i\delta_i\delta_i^{\mathrm{T}}) + \delta_i^{\mathrm{T}}\delta_i(W_i + \delta_i\delta_i^{\mathrm{T}})], \end{aligned} \tag{2.2}$$

This explicit expression for the the matrix gives insight on the structure of the problem. Some terms depend on the variability between clusters, the δ_i's, and others on the variability within clusters, the W_i's. We need the eigenstructure of K to capture the cluster structure, which is found in the δ_i's.

2.2.1 Proportional scatter matrices

If the scatter matrices of the groups are proportional, it is seen in Theorem 2.1 that the eigenvectors of K reveal some desirable properties for clustering.

Theorem 2.1. *Suppose X is a mixture of elliptical distributions as stated above with $V_i = V$, for $i = 1, \ldots, k$. The matrix K is*

$$K = \alpha I + \textstyle\sum_{i=1}^{k}\sum_{j=1}^{k}\beta_{ij}\delta_i\delta_j^{\mathrm{T}}, \tag{2.3}$$

with $\alpha = \tilde{k}p + (1-\tilde{k})\sum_{i=1}^{k}\pi_i\delta_i^{\mathrm{T}}\delta_i + \bar{k}$ and where

$$\beta_{ij} = \begin{cases} \gamma\pi_i + (\pi_i + \eta\pi_i^2)\delta_i^{\mathrm{T}}\delta_i & \text{if } i = j \\ \eta\pi_i\pi_j\delta_i^{\mathrm{T}}\delta_j & \text{if } i \neq j \end{cases}$$

with $\gamma = (1-\tilde{k})p - 2\bar{k} + 4$, $\eta = \tilde{k} + \bar{k} - 6$, $\tilde{k} = \sum_{i=1}^{k}\pi_i\tilde{k}_i$ and $\bar{k} = \sum_{i=1}^{k}\pi_i\bar{k}_i$.

We denote by $\Delta = \langle\delta_1, \ldots, \delta_k\rangle$ the subspace spanned by the δ_i's, where $\dim\Delta = q \leq k-1$. If $u \in \Delta^{\perp}$, $Ku = \alpha u$ holds, and α is an eigenvalue of K with multiplicity

$p-q$ associated to the eigenspace $\Delta^{\perp}$. The remaining q eigenvectors of K are found in the Δ subspace. Let $\Phi = \langle \phi_1, \ldots, \phi_k \rangle$ be the subspace spanned by Fisher's directions, $\phi_i = V^{-1}(\mu_i - \mu)$.

Then, the subspaces Φ and Δ_X are the same

$$\Phi = \Delta_X, \tag{2.4}$$

where Δ_X is the Δ-subspace expressed in the space of the original variables, $\Delta_X = \Sigma^{-1/2}\Delta$.

Under the assumption of proportional scatter matrices the best discriminant procedure is linear and Fisher's linear discriminant subspace is optimal in the sense that the relative separation between means is maximized. The theorem states that an identifiable subset of q eigenvectors of the kurtosis matrix K generates the subspace on which the clusters appear more separated. Some details of the theorem are found in the following proof.

Proof of Theorem 2.1. The result in (2.3) is obtained using in expression (2.2) the result $W_i = \Sigma^{-1/2} V \Sigma^{-1/2} = I - \sum_i \pi_i \delta_i \delta_i^{\mathrm{T}}$, where $V = \Sigma - \sum_i \pi_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$.

Denote $\Sigma = V + MPM^{\mathrm{T}}$, with $M = (\mu_1 - \mu, \ldots, \mu_k - \mu)$ and P diagonal with elements $(\pi_1, \ldots, \pi_k)$, then, from the inverse of the sum property, we have $\Sigma^{-1} = V^{-1} - V^{-1}M(M^{\mathrm{T}}V^{-1}M + P^{-1})^{-1}M^{\mathrm{T}}V^{-1}$, and multiplying by M,

$$\Sigma^{-1}M = V^{-1}M\left[I - (M^{\mathrm{T}}V^{-1}M + P^{-1})^{-1}M^{\mathrm{T}}V^{-1}M\right].$$

Therefore, $\Sigma^{-1}M = V^{-1}MT$. And, if we add and subtract P^{-1} appropriately, we can see that $T = [P(M^{\mathrm{T}}V^{-1}M + P^{-1})]^{-1}$ is invertible. Therefore, the columns of $\Sigma^{-1}M$ and $V^{-1}M$ generate the same subspace and thus $\Phi{=}\Delta_X$ and (2.4) is proven. □

Corollary 2.2. *In the particular case of a mixture of normal distributions, the constants are respectively $\tilde{k}_i = 1$ and $\bar{k}_i = 2$ and the eigenvalue associated to $\Delta^{\perp}$ has known value $\alpha = p + 2$. Also, if there are no clusters, from (2.3) we have $K = \alpha I$.*

Mixture of two normal distributions In the particular case of a mixture of two normal distributions, the matrix K simplifies to

$$K = (p+2)I + \beta \varphi^{\mathrm{T}} \varphi \varphi \varphi^{\mathrm{T}}, \tag{2.5}$$

where $\beta = \pi_1\pi_2(1 - 6\pi_1\pi_2)$ and $\varphi = \Sigma^{-1/2}(\mu_2 - \mu_1)$. The vector φ is an eigenvector of K with associated eigenvalue $\lambda = p + 2 + \beta(\varphi^{\mathrm{T}}\varphi)^2$, the rest of the eigenvalues are equal

to $p+2$. Also, $\mathrm{tr}(K) = p(p+2) + \beta(\varphi^{\mathrm{T}}\varphi)^2$ and $\det(K) = (p+2)^p + \beta(p+2)^{p-1}(\varphi^{\mathrm{T}}\varphi)^2$. Note that φ is Fisher's best linear discriminant function in the Z-space. The eigenvalue λ is the largest if $\beta > 0$ and the smallest otherwise. The parameter β is positive if $\pi_1 \in (0, (\sqrt{3}-1)/(2\sqrt{3}))$ and negative if $\pi_1 \in ((\sqrt{3}-1)/(2\sqrt{3}), 0.5]$. Therefore, if we have homogeneous clusters, the eigenvector associated with the smallest eigenvalue will be the one that better separates the clusters, while whenever the two clusters have very different sizes, the largest eigenvalue is the one that identifies the significant eigenvector. These values are the same ones that arise in Corollary 2 in Peña and Prieto (2001), where it is shown that the direction that optimizes the univariate kurtosis coefficient corresponds to Fisher's best linear discriminant function, maximizing it if $\pi_1 \in (0, (\sqrt{3}-1)/(2\sqrt{3}))$ and minimizing it if $\pi_1 \in ((\sqrt{3}-1)/(2\sqrt{3}), 0.5]$. Both approaches give estimations of Fisher's linear discriminant function, and a reasonable question is in which circumstances one procedure is more satisfactory than the other. On one hand, the estimation of eigenvectors may suffer from lack of precision when the sample size is small, but on the other hand a non-linear computationally intensive algorithm is needed to solve the optimization problem of finding the direction of kurtosis. We will address this issue in the next section with the help of some simulations.

Theorem 2.1 is in agreement with Theorem 5.2 in Tyler et al. (2009) and is similar to Proposition 1 in Caussinus and Ruiz-Gazen (1993). In the former the authors present a general method to generate an affine invariant coordinate system to reveal interesting departures from an elliptical distribution by using the eigenvectors of $V_1^{-1}V_2$, the relative scatter matrix. The idea is to first 'standardize' the data with respect to one scatter matrix V_1, and then perform generalized principal components on the 'standardized' data using a different scatter statistic V_2. Calculating the eigenvectors of the kurtosis matrix K is equivalent to choosing $V_1 = \Sigma$, and $V_2 = E[Z^{\mathrm{T}}Z(X-\mu)(X-\mu)^{\mathrm{T}}]$. In this case $V_1^{-1}V_2 = \Sigma^{-1/2}K\Sigma^{1/2}$, and the eigenvalues of $V_1^{-1}V_2$ and K are the same while the eigenvectors are $\Sigma^{-1/2}u$ and u respectively. As a matter of fact, these choices are the ones proposed in Caussinus and Ruiz-Gazen (1993), where more generally they study $V_2 = E[\omega(\beta Z^{\mathrm{T}}Z)(X-\mu)(X-\mu)^{\mathrm{T}}]/E[\omega(\beta Z^{\mathrm{T}}Z)]$, being ω a positive decreasing function and β a positive parameter.

The general case of different scatter matrices, however, is not considered in these references. In particular, the use of just any pair of robust scatter matrices in Tyler et al. (2009) does not guarantee the identification of the clusters, while the kurtosis has already proven to be effective in this situation. Also, the computation of most robust matrices is computationally very expensive. A discussion of the paper is found in Peña and Viladomat (2009).

Comparison with the kurtosis matrix B Under the same assumptions considered when calculating (2.3) plus normality for the components of the mixture, the matrix B in (1.6) is

$$B = pI + 2\mathbf{1}\mathbf{1}^{\mathrm{T}} + \sum_{i=1}^{k}\sum_{j=1}^{k}\gamma_{ij}\delta_i\delta_j^{\mathrm{T}}\mathbf{1}\mathbf{1}^{\mathrm{T}},$$

where

$$\gamma_{ij} = \begin{cases} (\pi_i - 3\pi_i^2)\delta_i^{\mathrm{T}}\delta_i & \text{if } i = j \\ -3\pi_i\pi_j\delta_i^{\mathrm{T}}\delta_j & \text{if } i \neq j \end{cases}$$

Let $\Delta_{\mathbf{1}} = \langle \Delta, \mathbf{1} \rangle$ be the subspace spanned by the $\mathbf{1}$ and the δ_i's and suppose we are in the general case $\mathbf{1} \notin \Delta$ and $\mathbf{1} \not\perp \Delta$. If $u \in \Delta_{\mathbf{1}}^{\perp}$, $Bu = pu$ holds, and p is an eigenvector of B with multiplicity $p-k$ associated to the eigenspace $\Delta_{\mathbf{1}}^{\perp}$. The remaining k eigenvectors are found in the $\Delta_{\mathbf{1}}$ subspace. When using the matrix K, the Δ subspace can be identified by selecting the eigenvectors with eigenvalues different from $p+2$. Instead, if we were to use the matrix B, we could only isolate the $\Delta_{\mathbf{1}}$ subspace, which is a non-informative choice. The procedure thus becomes dependent on the position of the δ_i's with respect to the vector $\mathbf{1}$. This dependency is the reason why the matrix B is not invariant under affine transformations. In the two special cases where $\mathbf{1} \in \Delta$ or $\mathbf{1} \perp \Delta$, the Δ subspace can still be identified using eigenvectors of B. In effect, if $\mathbf{1} \in \Delta$ then we can choose $p-k+1$ orthogonal eigenvectors from $\Delta^{\perp}$ with eigenvalues equal to p. And if $\mathbf{1} \perp \Delta$ then $\mathbf{1}$ is an eigenvector itself with eigenvalue $3p$, which also brings the total number of eigenvectors in $\Delta^{\perp}$ with known eigenvalues to $p-k+1$. The remaining $k-1$ eigenvectors are therefore an orthogonal basis of Δ.

2.2.2 Consistency of the eigenvectors of the estimated matrix K_n

Let $\mu_{r_1,\ldots,r_p} = E(\prod_{j=1}^{p} X_j^{r_j})$ be a k-order moment of X, $r_1 + \cdots + r_p = k$, then $\hat{\mu}_{r_1,\ldots,r_p}$ converges to $\mu_{r_1,\ldots,r_p}$ in probability and, since K is a continuous function of the moments, K_n converges to K in probability and therefore the matrix K_n is a consistent estimator of K. The spectral set of K, denoted Λ, is the set of all eigenvalues of K. The eigenspace of K associated with λ is $V(\lambda) = \{x \in \mathbb{R}^p \mid Kx = \lambda x\}$, whose dimension is the algebraic multiplicity of λ. Since K is symmetric, then $\mathbb{R}^p = \sum_{\lambda\in\Lambda} V(\lambda)$ holds. The eigenprojection of K associated with λ, denoted $P(\lambda)$, is the projection operator onto $V(\lambda)$ with respect to the decomposition of $\mathbb{R}^p$. If v is any subset of the spectral set Λ, then the total eigenprojection for K associated with the eigenvalues in v is defined to be $\sum_{\lambda\in v} P(\lambda)$. The following lemma (Tyler, 1981) states that, for any subset v of

eigenvalues of Λ, we can identify the corresponding subset v_n (because of the relative position of the eigenvalues), and the subspace defined as the sum of subspaces $\sum_{\lambda \in v_n} V_n(\lambda)$ will converge in probability to the subspace $\sum_{\lambda \in v} V(\lambda)$. That is, the subspace generated by eigenvectors of K_n associated to the eigenvalues v_n is a consistent estimator for the subspace generated by eigenvectors of K associated to the corresponding eigenvalues v.

Lemma 2.3. *Let K_n be a $p \times p$ symmetric matrix with eigenvalues $\lambda_1^n \geq \ldots \geq \lambda_p^n$. Let $P_{j,t}^n$ represent the subspace generated by the eigenvectors of K_n associated with $\lambda_j^n, \ldots, \lambda_t^n$ for $t \geq j$. If K_n converges to K in probability, then*

1. *λ_j^n converges to λ_j in probability,*
2. *$P_{j,t}^n$ converges to $P_{j,t}$ in probability, provided $\lambda_{j-1} \neq \lambda_j$ and $\lambda_t \neq \lambda_{t+1}$.*

The distance between two subspaces is measured using $\|P_1 - P_2\|_2$, the matrix spectral norm, and the proof of the lemma can be found in Section VIII-§3.5 of Kato (1980). A corollary of this lemma is that, when the scatter matrices are the same, the subspace orthogonal to the eigenspace associated to an eigenvalue of multiplicity q and value α, is a consistent estimator for Fisher's subspace.

Table 2.1: Factors f used to generate the samples of a mixture of normal populations.

p	k	f
2	2	16
	4	22
	8	30
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

We analyze this convergence through a simulation study. Throughout the thesis, we draw samples from mixtures of distributions as follows. Sets of $100p$ random observations,

with dimensions $p = 2, 4, 8, 15, 30$, are generated from a mixture of k multivariate normal distributions. The number of observations in each population is 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 $N_p(0, fI)$, for a factor f selected to be as small as possible whereas ensuring that the probability of overlapping between groups is roughly equal to 1%, see Table 2.1 for the values of f. The covariance matrices are generated as $S = UDU^{\mathrm{T}}$, using a random orthogonal matrix U and a diagonal matrix D with entries from a uniform distribution on $[10^{-3}, 5\sqrt{p}]$.

In Table 2.2 we consider the case of a mixture of two normal distributions with equal scatter matrices and present the angle between Fisher's discriminant function $V^{-1}(\mu_2 - \mu_1)$ and the eigenvector of K_n associated to the eigenvalue that differs most from the value $p + 2$. Also, we compare the results with the angle between Fisher's direction and the direction of kurtosis that maximizes $|\log(\kappa_d) - \log(3)|$ among the $2p$ considered in Peña and Prieto (2001), where κ_d is the univariate kurtosis coefficient of the direction d.

Table 2.2: Two groups and equal scatter matrices. Angle between Fisher's direction and: 1. the direction (kurt) that maximizes $|\log(\kappa_d) - \log(3)|$ and 2. the eigenvector of K_n (eigK) whose eigenvalue maximizes $|\lambda_i - (p + 2)|$.

p	kurt	eigK	kurt	eigK	kurt	eigK	kurt	eigK
4	16.03	35.39	10.10	21.45	6.91	15.08	3.64	8.01
8	16.03	36.44	12.93	21.74	6.88	18.15	4.36	7.52
15	11.25	42.86	8.96	25.92	14.82	19.61	9.60	10.28
30	24.99	50.30	12.41	26.37	8.32	19.95	4.77	8.70
Average	17.08	41.25	11.10	23.87	9.23	18.20	5.60	8.63
	n=100p		*n=500p*		*n=1000p*		*n=5000p*	

The results for small sample sizes are better for the kurtosis directions due to the limited precision of the eigenvectors and therefore we suggest using the optimization algorithm in these circumstances. However, the angles become more similar as the sample size increases, as expected.

We generate now mixtures of three normal distributions. In this case the subspace of interest is a plane and we want to measure how close Fisher's plane is to the plane generated by the two eigenvectors associated to eigenvalues that differ most from the value $p + 2$. Again, in order to compare the results with the kurtosis directions, we will also consider the plane generated by the two directions that maximize $|\log(\kappa_{d_i}) - \log(3)|$. When comparing directions, the angle between them is a convenient measure.

As a measure of distance between subspaces we will compute the angle between two hyperplanes, which is defined in Section 12.4.3 of Golub and van Loan (1996). Section 16.5 of Peña (2002) provides a geometrical interpretation of the angle. Let F and G be planes in $\mathbb{R}^p$, the angle between F and G is defined as the angle θ^* between u^* and v^*, the vectors that maximize $cos\ \theta = u^{\mathrm{T}}v$, where $u \in F$ and $v \in G$, subject to $\|u\| = \|v\| = 1$. Geometrically, u^* is collinear with the projection of v^* into F and v^* is collinear with the projection of u^* into G. In practice, to obtain θ^* we perform the singular value decomposition of $Q_F^{\mathrm{T}}Q_G$, where the columns of the $p \times 2$ matrices Q_F and Q_G define orthonormal bases for F and G respectively. The smallest singular value is the cosine of θ^*. The angles in Table 2.3 are calculated using this decomposition. This case is

Table 2.3: Three groups and equal scatter matrices. Angle between Fisher's plane and: 1. the plane generated by the directions (kurt) that maximize $|\log(\kappa_d) - \log(3)|$ and 2. the plane generated by the two eigenvectors of K_n (eigK) whose eigenvalues maximize $|\lambda_i - (p+2)|$.

p	kurt	eigK	kurt	eigK	kurt	eigK	kurt	eigK
4	44.90	44.53	37.76	26.75	30.68	19.03	33.47	10.21
8	43.55	51.28	39.69	27.66	31.34	20.47	25.71	12.77
15	51.62	56.05	42.65	35.94	42.10	30.78	35.86	16.54
30	62.79	63.76	45.59	41.80	40.63	33.12	35.94	19.47
Average	50.72	53.91	41.42	33.04	36.19	25.85	32.75	14.75
	n=100p		*n=500p*		*n=1000p*		*n=5000p*	

an example of the benefit of using the matrix K_n. For three groups we know that the the optimal direction is a combination of the directions δ_1 and δ_2, the ones related to the cluster structure, but we cannot identify the directions that would define the best plane. Instead, the eigenvectors do identify the optimal subspace. The angles in both approaches are similar for small samples, but as the sample size increases the distance from the eigenvectors to Fisher's subspace becomes smaller, as expected from the results in Lemma 2.3, while the convergence of the optimization directions is slower.

Another factor in consideration when comparing both approaches is related to the time needed for the kurtosis directions and the eigenvectors to be calculated. We did compute the running times for the p eigenvectors of K_n and the two extreme kurtosis directions. The results were calculated on a PC with Intel 3GHz CPU and are summarized in Table 2.4. Their increase with n is similar for both approaches, slightly faster than linear. This agrees with the fact that the main effort affected by n is the computation of the kurtosis matrix and the evaluation of the kurtosis coefficient, respectively. Regarding

increases in p, the matrix K_n presents a clear advantage, as the time ratios for both algorithms increase from values in the order of 4 for small dimensions to values in the order of 13 to 20 for the largest dimension under consideration ($p = 30$). This growth is associated with the use of Newton's method in the optimization of the kurtosis coefficient, and the need to factorize the corresponding second-derivative matrix in each iteration, as opposed to a single eigenvalue computation for the matrix K_n. In summary, the proposed algorithm seems to be computationally more efficient, particularly for the case of higher-dimensional data.

Table 2.4: Two groups and different scatter matrices. Time ratios in seconds between the two extreme univariate kurtosis directions and the p eigenvectors of K_n to be calculated.

p	kurt/eigK	kurt/eigK	kurt/eigK	kurt/eigK
2	6.56	3.83	4.09	3.63
4	24.50	6.21	5.53	5.22
8	13.91	8.78	7.04	7.12
15	20.42	11.32	10.48	9.68
30	19.08	17.09	13.83	12.75
Average	16.89	9.45	8.19	7.68
	n=100p	*n=1000p*	*n=5000p*	*n=10000p*

2.2.3 Different scatter matrices

In order to study the general case of different scatter matrices in a mixture of elliptical distributions, we start by studying a perturbation of the simpler model, a mixture of two normal distributions with equal scatter matrices. We perturb the covariance matrix of one of the mixtures to see the effect that the relaxation in the condition of equal covariances causes in both the eigenvectors of K and the directions that optimize the kurtosis coefficient.

After standardization and using the same notation as in previous sections, the mixture is characterized as $\pi_1 N(\delta_1, W) + \pi_2 N(\delta_2, W + \Delta W)$, where ΔW is the perturbation added to the model. Consider now the equations that define the solutions for both approaches, an eigenvector of K and the optimum univariate kurtosis direction. For the kurtosis matrix, an eigenvector d is such that $Kd = \lambda d$, which in our case can be formulated as

$$(a_0 - \lambda)d + a_1 \Delta W d + a_2 \Delta W^2 d = -b_1 \delta_1 - b_2 \Delta W \delta_1. \tag{2.6}$$

For the kurtosis direction, the equivalent equation comes from $\nabla \kappa_d = \lambda d$, and reduces to

$$(c_0 - \lambda)d + c_1 \Delta W d = -f_1 \delta_1. \tag{2.7}$$

Details of the derivations are found in Appendix 2.A.

When the scatter matrices are the same, the solution to both approaches is $d = c\delta_1$, for some constant c. Deviations from this solution appear as terms related to the perturbation such as ΔW and ΔW^2, the latter found only in (2.6). Consequently, in addition to ΔW, the eigenvectors of K differ from Fisher's discriminant function also in a quadratic term that does not arise in (2.7). Nevertheless, as we will see in simulation studies, the use of K is helpful when the sample size is not small, as in these cases the nonlinear algorithm for finding the optimal directions is time consuming and the results are similar to the ones obtained using K.

Table 2.5: Two groups and equal scatter matrices. Proportion of variance explained by the clusters, ($\hat{\phi}$), for the optimum direction (d.opt), the eigenvector of K_n associated with the max/min eigenvalue (max/min eigK), the max/min kurtosis direction (max/min kurt), the best eigenvector of K_n (best eigK) and the best kurtosis direction (best kurt).

p	n	d. opt	max/min eigK	max/min kurt	best eigK	best kurt
2	200	0.80	0.77	0.77	0.77	0.79
4	400	0.86	0.79	0.77	0.79	0.83
8	800	0.89	0.79	0.82	0.79	0.84
15	1500	0.93	0.78	0.86	0.78	0.87
30	3000	0.95	0.75	0.87	0.75	0.88
2	1000	0.78	0.78	0.76	0.78	0.78
4	2000	0.84	0.80	0.79	0.81	0.82
8	4000	0.89	0.85	0.85	0.85	0.88
15	7500	0.94	0.87	0.90	0.87	0.92
30	15000	0.96	0.86	0.92	0.86	0.93
2	2000	0.82	0.81	0.79	0.81	0.81
4	4000	0.84	0.82	0.82	0.83	0.83
8	8000	0.88	0.85	0.85	0.85	0.86
15	15000	0.93	0.86	0.89	0.86	0.90
30	30000	0.96	0.87	0.92	0.87	0.93
Average		0.88	0.82	0.84	0.82	0.86

Moreover, this result provides hints on how one might modify the matrix K in order to improve the performance when the scatter matrices are different, which has not been addressed yet in the literature. Further research will we focus in finding a matrix that could manage to reduce the impact of the terms ΔW and ΔW^2.

2.3 Computational results

Table 2.6: Two groups and equal scatter matrices. Percentage (%) of misclassified observations for the optimum direction (d.opt), the eigenvector of K_n associated with the max/min eigenvalue (max/min eigK), the max/min kurtosis direction (max/min kurt), the best eigenvector of K_n (best eigK) and the best kurtosis direction (best kurt).

p	n	d. opt	max/min eigK	max/min kurt	best eigK	best kurt
2	200	2.0	3.9	5.1	3.9	2.7
4	400	0.7	4.9	6.4	3.9	1.5
8	800	0.1	6.1	7.0	5.2	3.4
15	1500	0.0	6.9	6.1	6.2	4.2
30	3000	0.0	8.4	7.6	8.1	5.6
2	1000	2.8	3.7	4.6	3.7	3.2
4	2000	0.7	4.0	5.4	2.3	2.0
8	4000	0.1	2.5	3.7	2.2	0.9
15	7500	0.0	3.4	3.5	2.7	1.8
30	15000	0.0	2.8	3.3	2.6	2.3
2	2000	1.9	2.3	3.5	2.3	2.1
4	4000	0.9	2.0	3.2	1.6	1.5
8	8000	0.1	2.7	3.6	1.7	1.5
15	15000	0.0	3.4	4.2	2.9	2.2
30	30000	0.0	3.0	3.3	2.9	2.5
Average		0.6	4.0	4.7	3.5	2.5

We perform a set of simulations to evaluate the properties of the eigenvectors of K_n for cluster analysis. The measure chosen to assess the performance is the proportion of total projected variance explained by the projected clusters, given by $\phi = d^{\mathrm{T}}Bd/(d^{\mathrm{T}}\Sigma d)$, where $B = \pi_1\pi_2(\mu_2 - \mu_1)(\mu_2 - \mu_1)^{\mathrm{T}}$. The larger the gap between the projected means, the more separated the clusters are. Thus, we are interested in the directions that make ϕ large. If we search for the direction d that maximizes ϕ, it is well-known that Fisher's direction $d = (\pi_1 V_1 + \pi_2 V_2)^{-1}(\mu_2 - \mu_1)$ satisfies the optimality condition $\delta\phi/\delta d = 0$. We will estimate ϕ for the eigenvectors of K_n and for the directions of minimum and maximum kurtosis by generating random samples from a mixture of two p-variate normal populations. In order to have an idea on how close we are to the optimum, we will include the value $\hat{\phi}$ corresponding to Fisher's direction. Also, to estimate ϕ without assuming that we know the parameters of the two distributions, we need a procedure to assign observations to clusters. Once we project the data onto the direction d, we choose the particular assignation maximizing $\hat{\phi}$. In particular, since the cluster problem reduces to

one dimension, we choose n_1 such that $\hat{\phi} = d^{\mathrm{T}}\hat{B}d/[(n-1)d^{\mathrm{T}}Sd]$ is maximized, where $\hat{B} = n_1 n_2/(n_1+n_2)(\bar{x}_2 - \bar{x}_1)(\bar{x}_2 - \bar{x}_1)^{\mathrm{T}}$ and $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{(i)}$. We assume that we know of the existence of just two clusters in the data.

2.3.1 Proportional scatter matrices

We start analyzing the results when the scatter matrices are the same. Table 2.5 presents the measure $\hat{\phi}$ for the optimum direction $V^{-1}(\mu_2 - \mu_1)$, the eigenvector of K_n ('max/min eigK') that maximizes $\hat{\phi}$ among the two eigenvectors corresponding to the maximum and minimum eigenvalue, the univariate kurtosis direction ('max/min kurt') that maximizes $\hat{\phi}$ among the maximum and minimum univariate kurtosis directions, the eigenvector of K_n ('best eigK') that maximizes $\hat{\phi}$ among the p existing eigenvectors and the univariate kurtosis direction ('best kurt') that maximizes $\hat{\phi}$ among the $2p$ directions considered in Peña and Prieto (2001). In Table 2.6 we present the proportion of misclassified observations after assigning them to clusters as stated above. Each value has been replicated 100 times.

Table 2.7: Two groups and equal scatter matrices. Number of times out of 100 where the eigenvalue of K_n corresponding to the eigenvector that maximizes $\hat{\phi}$ does not belong to the 30%-40% largest or smallest eigenvalues.

p	n	30%	40%
2	200	-	-
4	400	9	9
8	800	15	8
15	1500	13	9
30	3000	13	8
2	1000	-	-
4	2000	11	11
8	4000	6	3
15	7500	5	4
30	15000	4	3
2	2000	-	-
4	4000	2	2
8	8000	8	3
15	15000	7	5
30	30000	5	3

When considering only two eigenvectors and two kurtosis directions, the results in the two tables are similar. We observe that the extreme eigenvector of K_n performs

better when the dimension of the space is small (2,4,8), whereas the univariate kurtosis has better results when p is larger. We also observe that the values are very close to the optimum ones, indicating the appropriateness of the two methods. However, when all eigenvectors and kurtosis directions are considered, the results for the eigenvectors are very similar (column 'max/min eigK' and 'best eigK' are practically identical) whereas there is some improvement in the projected kurtosis directions, especially for large p. Note that, for a given p, the eigenvectors improve as n increases, while the kurtosis directions behave more stable in this sense. Also, if we count the number of times that the selected eigenvector in 'best eigK' does not correspond to one of the extreme eigenvalues, we obtain that this number is very small, specially when n is large, in Table 2.7 we summarize these results. Thus we conclude that the maximum/minimum eigenvalue of the kurtosis matrix provides a useful direction for clustering which is very fast to compute. The computation of the matrix K and its eigenvectors is computationally very efficient, while the directions of kurtosis require an optimization algorithm and are computationally more expensive.

Table 2.8: Two groups and different scatter matrices. Proportion of variance explained by the clusters ($\hat{\phi}$) for the optimum direction (d.opt), the best eigenvector of K_n (best eigK) and the best kurtosis direction (best kurt).

p	n	d. opt	best eigK	best kurt
2	200	0.78	0.75	0.77
4	400	0.82	0.74	0.76
8	800	0.87	0.73	0.78
15	1500	0.90	0.76	0.81
30	3000	0.93	0.69	0.80
2	1000	0.78	0.75	0.77
4	2000	0.81	0.76	0.77
8	4000	0.87	0.79	0.80
15	7500	0.90	0.76	0.79
30	15000	0.93	0.75	0.82
2	2000	0.77	0.75	0.76
4	4000	0.82	0.77	0.77
8	8000	0.87	0.77	0.78
15	15000	0.90	0.80	0.81
30	30000	0.93	0.75	0.81
Average		0.86	0.75	0.79

Table 2.9: Two groups and different scatter matrices. Percentage(%) of misclassified observations for the optimum direction (d.opt), the best eigenvector of K_n (best eigK) and the best kurtosis direction (best kurt).

p	n	d. opt	best eigK	best kurt
2	200	3.20	5.30	4.30
4	400	1.10	4.00	3.90
8	800	0.30	5.00	3.30
15	1500	0.10	4.40	5.20
30	3000	0.00	6.10	5.50
2	1000	2.80	4.80	3.50
4	2000	1.30	4.90	4.10
8	4000	0.30	3.80	3.50
15	7500	0.10	3.30	5.10
30	15000	0.00	3.90	5.00
2	2000	3.30	5.00	4.00
4	4000	0.90	4.40	3.70
8	8000	0.30	4.00	3.10
15	15000	0.10	2.60	4.70
30	30000	0.00	3.80	5.40
Average		0.92	4.35	4.29

2.3.2 Different scatter matrices

In the general case of different scatter matrices, the optimum direction for ϕ is $(\pi_1 V_1 + \pi_2 V_2)^{-1}(\mu_2 - \mu_1)$. If we compare in Table 2.8 the columns 'best eigK' and 'best kurt' we observe that the kurtosis directions perform slightly better. However, if we look at the same columns in Table 2.9, the proportion of misclassified observations, the results are very similar. In particular, the eigenvectors perform better when the sample size is large. This behaviour could be due to the lack of precision in the eigenvectors when the sample size is small. As before, we check which eigenvalue is associated to the selected eigenvector; in Table 2.10 we observe that it does not seem to follow a strong pattern in terms of its eigenvalue, even though it looks that most of the times the eigenvalue associated to the chosen eigenvector is one of the extreme ones.

Table 2.10: Two groups and different scatter matrices. Number of times out of 100 where the eigenvalue of K_n corresponding to the eigenvector that maximizes $\hat{\phi}$ does not belong to the 30%-40% largest or smallest eigenvalues.

p	n	30%	40%
2	200	-	-
4	400	17	17
8	800	37	15
15	1500	41	29
30	3000	41	25
2	1000	-	-
4	2000	13	13
8	4000	32	16
15	7500	36	26
30	15000	46	30
2	2000	-	-
4	4000	16	16
8	8000	32	20
15	15000	42	31
30	30000	45	23

2.4 Discussion

In Chapter 3 we study alternative kurtosis matrices based on local modifications of the data, with the intention of improving the performance of the eigenvectors of the kurtosis matrix studied in this chapter. In particular, we explore variations of the kurtosis matrix where the terms in (2.2) that depend on the scatter matrices W_i have less influence on the eigenstructure of the matrix. By substituting each observation of the sample with the mean of its neighbours, the covariance matrices of the components of a mixture of distributions would be expected to shrink, giving a more predominant role to the variability between clusters in the decomposition of the kurtosis matrix.

Appendix 2.A Derivations for the case of different scatters

We have that $\delta_2 = -\frac{\pi_1}{\pi_2}\delta_1$ and, from the decomposition of the covariance matrix in the case of mixture distributions, $I = \pi_1 W + \pi_2 W + \pi_2 \Delta W + \sum_i \pi_i \delta_i \delta_i^{\mathrm{T}}$, and thus $W = \bar{W} - \pi_2 \Delta W$, where $\bar{W} = I - \frac{\pi_1}{\pi_2}\delta_1 \delta_1^{\mathrm{T}}$ corresponds to the equal scatter matrices case. Also $W + \Delta W = \bar{W} + \pi_1 \Delta W$. Replacing $W_1 = W = \bar{W} - \pi_2 \Delta W$ and $W_2 = W + \Delta W =$

$\bar{W} + \pi_1 \Delta W$ in (2.2) we obtain

$$\begin{aligned} K &= \bar{K} + \pi_1\pi_2 \Delta W \operatorname{tr} \Delta W + 2\pi_1\pi_2 \Delta W^2 + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1\delta_1^{\mathrm{T}} \operatorname{tr} \Delta W \\ &\quad + 2\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1\delta_1^{\mathrm{T}}\Delta W + \Delta W \delta_1\delta_1^{\mathrm{T}}) + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}}\delta_1 \Delta W, \end{aligned}$$

where $\bar{K} = (p+2)I + \frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)\delta_1^{\mathrm{T}}\delta_1\delta_1\delta_1^{\mathrm{T}}$. The kurtosis coefficient on a direction is $\kappa_d = 3\sum_i \pi_i (d^{\mathrm{T}} W_i d)^2 + 6\sum_i \pi_i (d^{\mathrm{T}} W_i d)(\delta_i^{\mathrm{T}} d)^2 + \sum_i \pi_i (\delta_i^{\mathrm{T}} d)^4$, and substituting in our case

$$\kappa_d = \bar{\kappa}_d + 3\pi_1\pi_2 (d^{\mathrm{T}} \Delta W d)^2 + 6\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1^{\mathrm{T}} d)^2 d^{\mathrm{T}} \Delta W d,$$

where $\bar{\kappa}_d = 3(d^{\mathrm{T}} d)^2 + \frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)(\delta_1^{\mathrm{T}} d)^4$. The parameters in equations (2.6) and (2.7) derived from these results are $a_0 = p+2$, $a_1 = \pi_1\pi_2 \operatorname{tr} \Delta W + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}}\delta_1$, $a_2 = 2\pi_1\pi_2$, $b_1 = \frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)\delta_1^{\mathrm{T}}\delta_1\delta_1^{\mathrm{T}} d + \frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1^{\mathrm{T}} d \operatorname{tr} \Delta W + 2\delta_1^{\mathrm{T}} \Delta W d)$, $b_2 = 2\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}} d$, $c_0 = 12$, $c_1 = 12\pi_1\pi_2 d^{\mathrm{T}} \Delta W d + 12\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)(\delta_1^{\mathrm{T}} d)^2$ and $f_1 = 4\frac{\pi_1}{\pi_2^3}(1 - 6\pi_1\pi_2)(\delta_1^{\mathrm{T}} d)^3 + 12\frac{\pi_1}{\pi_2}(\pi_1 - \pi_2)\delta_1^{\mathrm{T}} d d^{\mathrm{T}} \Delta W d$.

Chapter 3

Kurtosis matrices based on local modifications of the data

Following the discussion in Chapter 2, this chapter studies alternative kurtosis matrices based on local modifications of the data, with the intention of improving the performance of the eigenvectors of the kurtosis matrix studied in Chapter 2. By substituting each observation of the sample with the mean of its neighbours, the covariance matrices of the components of a mixture of distributions would be expected to shrink, giving a more predominant role to the variability between clusters in the decomposition of the kurtosis matrix. Specifically, we prove that the separation properties of the eigenvectors of the new kurtosis matrix are better in the sense that the proposed modification of the observations produces standardized means that are further from each other than those of the original observations, and thus the clusters will appear to be more clearly separated.

3.1 Using the kurtosis matrix for concentrated data

The kurtosis matrix K that we studied in Chapter 2 can be decomposed as a sum of two matrices, $K = K_W + K_B$, where

$$K_B = \sum_{i=1}^{k} \pi_i \delta_i^{\mathrm{T}} \delta_i \delta_i \delta_i^{\mathrm{T}}$$

$$K_W = \sum_{i=1}^{k} \pi_i [\operatorname{tr} W_i (\tilde{k}_i W_i + \delta_i \delta_i^{\mathrm{T}}) + \bar{k}_i W_i^2 + 2(\delta_i \delta_i^{\mathrm{T}} W_i + W_i \delta_i \delta_i^{\mathrm{T}}) + \delta_i^{\mathrm{T}} \delta_i W_i],$$

which can be understood as a decomposition of the variability. In effect, K_W is function of the covariance matrices W_i and therefore it measures the variability within clusters,

while K_B measures the variability between clusters as it only depends on the cluster means. We want the eigenstructure of K to capture the cluster information, which is found in the δ_i's. If the covariance matrices W_i are "big" enough, the eigenvectors of K will depend mainly on them, hiding the cluster structure, to avoid that we need K_B to have a sufficiently large contribution to K to dominate the effect of K_W.

Suppose we replace each observation from a sample $\{x_i\}$ of X, with sample size n, by the average of the $\lfloor \kappa n \rfloor$ closest observations to x_i (in the euclidean norm), $\tilde{x}_i$.

In population terms the new random variable is defined as

$$\tilde{x}(w) = \frac{1}{\kappa} \int_S y f_X(y) dy, \quad S = \{z : \|z - x(w)\| \leq \epsilon\}, \tag{3.1}$$

where ϵ, the size of the ball, is related to κ through

$$\int_S f_X(y) dy = \kappa. \tag{3.2}$$

Our interest is to study the moments of the new random variable to obtain the expression for the modified matrix $\bar{K}$. In particular, we wish to search for a relationship between the covariance matrices of the original random variable and the modified one, particularly in the case when κ is small.

We start by linking the original and modified observations, where we obtain

$$\tilde{x} = x + \beta \epsilon^2 V^{-1}(x - \mu) + O(\epsilon^4), \tag{3.3}$$

the details are found in Appendix 3.A. This relationship is our starting point to analyze the moments of interest. Our first step is to consider the density associated to the new variable $\tilde{x}$. We have,

$$\tilde{x} - \mu = (I - \beta \epsilon^2 V^{-1})(x - \mu) + O(\epsilon^4).$$

Note first that by taking expectations in (3.3) we have that

$$\tilde{\mu} = E[\tilde{x}] = \mu + (I - \beta \epsilon^2 V^{-1})(E[x] - \mu) + O(\epsilon^4) = \mu + O(\epsilon^4). \tag{3.4}$$

The density for the new variable will be given by

$$\begin{aligned} f_{\tilde{x}}(\tilde{x}) &= |V|^{-1/2} h\left((x-\mu)^{\mathrm{T}} V^{-1}(x-\mu)\right)\left(|(I - \beta\epsilon^2 V^{-1})^{-1}| + O(\epsilon^4)\right) \\ &= |V|^{-1/2} h\left((\tilde{x}-\mu)^{\mathrm{T}}(I - \beta\epsilon^2 V^{-1})^{-1} V^{-1}(I - \beta\epsilon^2 V^{-1})^{-1}(\tilde{x}-\mu) + O(\epsilon^4)\right) \\ &\quad \times \left(|(I - \beta\epsilon^2 V^{-1})^{-1}| + O(\epsilon^4)\right) \\ &= |V - 2\beta\epsilon^2 I|^{-1/2} h\left((\tilde{x}-\mu)^{\mathrm{T}}(V - 2\beta\epsilon^2 I)^{-1}(\tilde{x}-\mu)\right) + O(\epsilon^4). \end{aligned} \tag{3.5}$$

This density function corresponds, up to order ϵ^4, to an elliptical distribution with the same function h as the original observations, mean μ and covariance matrix proportional to $V - 2\beta\epsilon^2 I$.

A consequence of this result is that we can use the moment results for the original observations x, replacing μ with $\mu + O(\epsilon^4)$, and V with $V - 2\beta\epsilon^2 I + O(\epsilon^4)$.

3.2 The model of interest: a mixture of elliptical distributions

Consider now the case where we have k groups of observations, each one generated from an elliptical distribution with density as in (2.1), and weights π_i.

We start by standardizing the observations as $Z = \Sigma^{-1/2}(X - \mu)$, where $\mu = \sum_i \pi_i \mu_i$ and $\Sigma = \sum_i \pi_i c_i V_i + \sum_i \pi_i (\mu_i - \mu)(\mu_i - \mu)^{\mathrm{T}}$ are the mean and covariance matrix of the mixture. The resulting observations can be considered to have been generated from elliptical distributions with new means δ_i, covariance matrices W_i, the same functions h_i and weights π_i. The values of these parameters are given by

$$\delta_i = \Sigma^{-1/2}(\mu_i - \mu), \quad W_i = c_i \Sigma^{-1/2} V_i \Sigma^{-1/2}. \tag{3.6}$$

In the next step, we modify the observations replacing each z with the average $\tilde{z}$ of a percentage of the observations closest to it. Assuming that the groups are sufficiently removed from each other, we obtain new observations defined from (3.3).

The new mixture of observations can be considered (for small values of ϵ) to follow ellipsoidal distributions with parameters $\tilde{\delta}_i$ and $\tilde{W}_i$, and the same functions h_i and weights π_i. From (3.4) and (3.6), the values of the means are given by

$$\tilde{\delta}_i = \delta_i + O(\epsilon^4) = \Sigma^{-1/2}(\mu_i - \mu) + O(\epsilon^4), \tag{3.7}$$

and for the covariance matrices, from (3.5),

$$\tilde{W}_i = W_i - 2\beta_i \epsilon^2 I + O(\epsilon^4) = \Sigma^{-1/2} \left(V_i - 2\beta_i \epsilon^2 \Sigma \right) \Sigma^{-1/2} + O(\epsilon^4). \tag{3.8}$$

As a last step prior to the computation of the new kurtosis matrix, these transformed observations have to be standardized again. This is equivalent to introducing a new linear transformation of the form $\bar{Z} = \tilde{\Sigma}^{-1/2}(\tilde{Z} - \tilde{\delta})$, where $\tilde{\delta}$ denotes the mean of the transformed observations, which from (3.7) satisfies

$$\tilde{\delta} = \sum_i \pi_i \tilde{\delta}_i = O(\epsilon^4),$$

and $\tilde{\Sigma}$ denotes their covariance matrix, which from (3.7) and (3.8) satisfies

$$\begin{aligned}\tilde{\Sigma} &= \sum_i \pi_i \tilde{W}_i + \sum_i \pi_i (\tilde{\delta}_i - \tilde{\delta})(\tilde{\delta}_i - \tilde{\delta})^{\mathrm{T}} \\ &= \sum_i \pi_i (W_i - 2\beta_i \epsilon^2 I) + \sum_i \pi_i \delta_i \delta_i^{\mathrm{T}} + O(\epsilon^4) \\ &= \sum_i \pi_i W_i + \sum_i \pi_i \delta_i \delta_i^{\mathrm{T}} - 2\epsilon^2 \bar{\beta} I + O(\epsilon^4) \\ &= (1 - 2\epsilon^2 \bar{\beta}) I + O(\epsilon^4),\end{aligned}$$

where $\bar{\beta} = \sum_i \pi_i \beta_i$. Note that from this result,

$$\tilde{\Sigma}^{-1/2} = (1 + \epsilon^2 \bar{\beta}) I + O(\epsilon^4).$$

Using these moments, the values of the parameters for the new standardized observations will be given by

$$\bar{\delta}_i = \tilde{\Sigma}^{-1/2}(\tilde{\delta}_i - \tilde{\delta}) = (1 + \bar{\beta}\epsilon^2)\delta_i + O(\epsilon^4), \tag{3.9}$$

$$\begin{aligned}\bar{W}_i &= \tilde{\Sigma}^{-1/2} \tilde{W}_i \tilde{\Sigma}^{-1/2} = (1 + \epsilon^2 \bar{\beta})^2 (W_i - 2\beta_i \epsilon^2 I) + O(\epsilon^4) \\ &= (1 + 2\bar{\beta}\epsilon^2) W_i - 2\beta_i \epsilon^2 I + O(\epsilon^4).\end{aligned} \tag{3.10}$$

3.3 The definition of the kurtosis matrix $\bar{K}$

Given the parameters derived in the preceding section for the different transformed observations, we now analyze their impact on the kurtosis matrix.

Consider first a mixture of variables Z_i with ellipsoidal distributions with parameters δ_i and W_i and weights π_i, and introduce a shift $Y_i = Z_i - \delta_i$. The kurtosis matrix is defined as

$$\begin{aligned}K &= E[Z^{\mathrm{T}} Z Z Z^{\mathrm{T}}] = \sum_i \pi_i E[Z_i^{\mathrm{T}} Z_i Z_i Z_i^{\mathrm{T}}] \\ &= \sum_i \pi_i E[(Y_i + \delta_i)^{\mathrm{T}} (Y_i + \delta_i)(Y_i + \delta_i)(Y_i + \delta_i)^{\mathrm{T}}] = \sum_i \pi_i K_i.\end{aligned}$$

Using the property that all odd moments of Y_i are equal to zero, we have that

$$\begin{aligned}K_i &= E[Y_i^{\mathrm{T}} Y_i Y_i Y_i^{\mathrm{T}}] + E[Y_i^{\mathrm{T}} Y_i] \delta_i \delta_i^{\mathrm{T}} + \delta_i^{\mathrm{T}} \delta_i E[Y_i Y_i^{\mathrm{T}}] \\ &\quad + 2\delta_i \delta_i^{\mathrm{T}} E[Y_i Y_i^{\mathrm{T}}] + 2E[Y_i Y_i^{\mathrm{T}}] \delta_i \delta_i^{\mathrm{T}} + \delta_i^{\mathrm{T}} \delta_i \delta_i \delta_i^{\mathrm{T}}.\end{aligned}$$

If we now analyze each one of the terms, using the results in Appendix 3.C, we have

$$\begin{aligned} E[Y_i^{\mathrm{T}} Y_i Y_i Y_i^{\mathrm{T}}] &= \tilde{k}_i \operatorname{tr}(W_i) W_i + \bar{k}_i W_i^2 \\ E[Y_i^{\mathrm{T}} Y_i] \delta_i \delta_i^{\mathrm{T}} &= k_i \operatorname{tr}(W_i) \delta_i \delta_i^{\mathrm{T}} \\ \delta_i^{\mathrm{T}} \delta_i E[Y_i Y_i^{\mathrm{T}}] &= k_i \delta_i^{\mathrm{T}} \delta_i W_i \\ \delta_i \delta_i^{\mathrm{T}} E[Y_i Y_i^{\mathrm{T}}] &= k_i \delta_i \delta_i^{\mathrm{T}} W_i \\ E[Y_i Y_i^{\mathrm{T}}] \delta_i \delta_i^{\mathrm{T}} &= k_i W_i \delta_i \delta_i^{\mathrm{T}}. \end{aligned}$$

Therefore,

$$\begin{aligned} K &= \sum_i \pi_i [\tilde{k}_i \operatorname{tr}(W_i) W_i + \bar{k}_i W_i^2 + k_i \operatorname{tr}(W_i) \delta_i \delta_i^{\mathrm{T}}] \\ &\quad + \sum_i \pi_i [k_i \delta_i^{\mathrm{T}} \delta_i W_i + 2 k_i \delta_i \delta_i^{\mathrm{T}} W_i + 2 k_i W_i \delta_i \delta_i^{\mathrm{T}} + \delta_i^{\mathrm{T}} \delta_i \delta_i \delta_i^{\mathrm{T}}]. \end{aligned}$$

Now we consider the same matrix for the transformed and standardized observations $\bar{Z}_i$. The main change is that the parameters δ_i and W_i are replaced by $\bar{\delta}_i$ and $\bar{W}_i$, defined in (3.9) and (3.10). We obtain

$$\begin{aligned} \bar{K}_i &= (1 + 4\bar{\beta}\epsilon^2) K_i - 2\beta_i \epsilon^2 (p\tilde{k}_i + 2\bar{k}_i) W_i - 2\beta_i \epsilon^2 [\tilde{k}_i \operatorname{tr}(W_i) + k_i \delta_i^{\mathrm{T}} \delta_i] I \\ &\quad -2\beta_i \epsilon^2 (p+4) k_i \delta_i \delta_i^{\mathrm{T}} + O(\epsilon^4), \end{aligned}$$

and the corresponding matrix $\bar{K}$ is given by

$$\begin{aligned} \bar{K} &= (1 + 4\bar{\beta}\epsilon^2) K - 2\epsilon^2 \sum_i \pi_i \beta_i [(p\tilde{k}_i + 2\bar{k}_i) W_i + (\tilde{k}_i \operatorname{tr}(W_i) + k_i \delta_i^{\mathrm{T}} \delta_i) I] \\ &\quad -2\epsilon^2 \sum_i \pi_i \beta_i (p+4) k_i \delta_i \delta_i^{\mathrm{T}} + O(\epsilon^4). \end{aligned}$$

3.4 Properties of the modified data: separation of the observations

We now consider the impact of the modification of the observations on the separation properties of the directions obtained from the kurtosis matrices. To simplify the analysis we will analyze the case where we only have two different groups.

The quality of the directions will be studied by comparing for the selected projection direction d the value of the criterion

$$\frac{(d^{\mathrm{T}} \delta)^2}{d^{\mathrm{T}} \Sigma d}$$

in both cases. Note that for the standardized observations the denominator is equal to one, and the criterion reduces to the value of the numerator. We should thus compare the values of $(d^{\mathrm{T}}\delta_i)^2$ with those of $(\bar{d}^{\mathrm{T}}\bar{\delta}_i)^2$, where d denotes the eigenvector associated to the largest eigenvalue of K, while $\bar{d}$ denotes the eigenvector associated to the largest eigenvalue of $\bar{K}$. As $\pi_1\delta_1 = -\pi_2\delta_2$, it does not matter which δ_i is considered, as long as it is the same in both cases. Also, since we observed through simulations that most of the times $(\bar{d}^{\mathrm{T}}\bar{\delta}_i)^2 > (d^{\mathrm{T}}\bar{\delta}_i)^2$, it is enough to compare $(d^{\mathrm{T}}\bar{\delta}_i)^2$ with $(d^{\mathrm{T}}\delta_i)^2$ in order to draw conclusions. From (3.9) we have that

$$(d^{\mathrm{T}}\bar{\delta}_i)^2 = (1 + 2\bar{\beta}\epsilon^2)(d^{\mathrm{T}}\delta_i)^2 + O(\epsilon^4) = (d^{\mathrm{T}}\delta_i)^2 + 2\bar{\beta}\epsilon^2(d^{\mathrm{T}}\delta_i)^2 + O(\epsilon^4).$$

Therefore, $(d^{\mathrm{T}}\bar{\delta}_i)^2 > (d^{\mathrm{T}}\delta_i)^2$ for small enough values of ϵ, which implies that the proposed modification of the observations produces standardized means that are further from each other than those of the original observations, and thus the clusters will appear more separated.

Appendix 3.A Linking the original and modified observations

Consider $\tilde{x}$ in (3.1), it can be written as

$$\begin{aligned}\tilde{x} &= \frac{1}{\kappa}x\int_S f_X(y)dy + \frac{1}{\kappa}\int_S f_X(y)(y-x)dy \\ &= x + \frac{1}{\kappa}\int_S f_X(y)(y-x)dy,\end{aligned}$$

and introducing the Taylor series expansion for $f_X(y)$ around x,

$$\begin{aligned}\tilde{x} &= x + \frac{f_X(x)}{\kappa}\int_S (y-x)dy \\ &\quad + \frac{2|V|^{-1/2}}{\kappa}h'\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)\int_S (x-\mu)^{\mathrm{T}}V^{-1}(y-x)(y-x)dy \\ &\quad + \frac{2|V|^{-1/2}}{\kappa}h''\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)\int_S ((x-\mu)^{\mathrm{T}}V^{-1}(y-x))^2(y-x)dy \\ &\quad + \frac{|V|^{-1/2}}{2\kappa}h'\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)\int_S (y-x)^{\mathrm{T}}V^{-1}(y-x)(y-x)dy \\ &\quad + \frac{1}{\kappa}\int_S |V|^{-1/2}O(\|y-x\|^4)dy \\ &= x + \frac{2}{\kappa}|V|^{-1/2}h'\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)\int_S (x-\mu)^{\mathrm{T}}V^{-1}(y-x)(y-x)dy + \frac{1}{\kappa}O(\epsilon^4 v_p(\epsilon)) \\ &= x + \frac{h'\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)}{h\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)}\frac{2}{v_p(\epsilon)}\int_S (x-\mu)^{\mathrm{T}}V^{-1}(y-x)(y-x)dy + O(\epsilon^4), \qquad (3.11)\end{aligned}$$

where we have used symmetry to cancel the third-order terms, together with $\int_S (x - \mu)^{\mathrm{T}} V^{-1}(y - x)(y - x)dy = O(\epsilon^2 v_p(\epsilon))$ and the result in (3.12) to replace the terms depending on κ.

Consider now the remaining integral in the preceding expression,

$$\int_S (x - \mu)^{\mathrm{T}} V^{-1}(y - x)(y - x)dy = \int_{\bar{S}} (x - \mu)^{\mathrm{T}} V^{-1} \bar{x}\bar{x} d\bar{x} = \dots$$

for $\bar{x} = y - x$ and $\bar{S} = \{z : \|z\| \leq \epsilon\}$. Let $\bar{x} = Uy$, where U is an orthogonal matrix having its first column equal to $V^{-1}(x - \mu)/\|V^{-1}(x - \mu)\|$, we have that

$$\begin{aligned} \dots &= \|V^{-1}(x - \mu)\| \int_{\bar{S}} y_1 U y dy = \|V^{-1}(x - \mu)\| \sum_i \int_{\bar{S}} y_1 y_i u_i dy \\ &= \|V^{-1}(x - \mu)\| u_1 \int_{\bar{S}} y_1^2 dy = V^{-1}(x - \mu) \int_{\bar{S}} y_1^2 dy = \dots \end{aligned}$$

where we have used symmetry to cancel the terms $\int_{\bar{S}} y_1 y_i dy$ with $i \neq 1$ from the sum. We can write

$$\dots = \frac{1}{p} V^{-1}(x - \mu) \int_{\bar{S}} y^{\mathrm{T}} y dy = \dots$$

again from symmetry, as $\int_{\bar{S}} y_1^2 dy = \int_{\bar{S}} y_i^2 dy = (1/p) \int_{\bar{S}} y^{\mathrm{T}} y dy$. Letting $y^{\mathrm{T}} y = z^2$,

$$\begin{aligned} \dots &= \frac{1}{p} V^{-1}(x - \mu) \int_0^\epsilon z^2 v_p'(z) dz \\ &= \frac{1}{p} V^{-1}(x - \mu) K_p p \int_0^\epsilon z^2 z^{p-1} dz = V^{-1}(x - \mu) K_p \frac{\epsilon^{p+2}}{p + 2} \\ &= V^{-1}(x - \mu) \frac{1}{p + 2} v_p(\epsilon) \epsilon^2. \end{aligned}$$

where we have used $v_p(\epsilon) = \epsilon^p \pi^{p/2}/\Gamma(p/2 + 1) = K_p \epsilon^p$. Replacing the result for the integral in (3.11), we obtain

$$\tilde{x} = x + \beta \epsilon^2 V^{-1}(x - \mu) + O(\epsilon^4).$$

where $\beta = -\frac{2}{p+2} \frac{h'\left((x-\mu)^{\mathrm{T}} V^{-1}(x-\mu)\right)}{h((x-\mu)^{\mathrm{T}} V^{-1}(x-\mu))} > 0$, since we assume that $h(z) > 0$ and $h'(z) < 0$ for all $z > 0$.

Appendix 3.B Neighbourhood size

We relate the value of κ and ϵ. Using Taylor series expansions for $f_X(y)$ around x in (3.2),

$$\begin{aligned}\kappa &= \int_S f_X(y)dy = f_X(x)\int_S dy \\ &\quad + 2|V|^{-1/2}h'\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)\int_S (x-\mu)^{\mathrm{T}}V^{-1}(y-x)dy \\ &\quad + |V|^{-1/2}\int_S O(\|y-x\|^2)dy \\ &= f_X(x)v_p(\epsilon) + O(\epsilon^2 v_p(\epsilon)),\end{aligned}$$

where $v_p(\epsilon)$ denotes the volume of S (a hypersphere in dimension p with radius equal to ϵ), $\int_S dy = v_p(\epsilon)$, and we have used $\int_S (y-x)dy = 0$ (from symmetry) to cancel the second term in the expansion.

From this result we have that

$$\frac{\kappa}{v_p(\epsilon)} - f_X(x) = O(\epsilon^2),$$

and for $f_X(x) > 0$ we also have

$$\frac{v_p(\epsilon)}{\kappa} - \frac{1}{f_X(x)} = \frac{f_X(x) - \kappa/v_p(\epsilon)}{f_X(x)\kappa/v_p(\epsilon)} = O(\epsilon^2),$$

and thus

$$f_X(x)\frac{v_p(\epsilon)}{\kappa} = 1 + O(\epsilon^2). \tag{3.12}$$

Appendix 3.C Moments of an elliptical distribution

Consider a random variable X following an elliptical distribution with density as in (2.1). Note that the covariance matrix of X is given by

$$\begin{aligned}E[(X-\mu)(X-\mu)^{\mathrm{T}}] &= |V|^{-1/2}\int (x-\mu)(x-\mu)^{\mathrm{T}}h\left((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu)\right)dx \\ &= |V|^{-1/2}V^{1/2}\int yy^{\mathrm{T}}h(y^{\mathrm{T}}y)|V|^{1/2}dyV^{1/2} \\ &= V^{1/2}\int yy^{\mathrm{T}}h(y^{\mathrm{T}}y)dyV^{1/2} = kV,\end{aligned}$$

where we have used the change of variable $y = V^{-1/2}(x-\mu)$, and also symmetry to obtain

$$\int yy^{\mathrm{T}}h(y^{\mathrm{T}}y)dy = kI,$$

for $k = \int y_i^2 h(y^{\mathrm{T}}y)dy = (1/p)\int y^{\mathrm{T}}yh(y^{\mathrm{T}}y)dy$, with $k > 0$.

Also, its fourth-order central moments $M_4 = E[(X-\mu)^{\mathrm{T}}(X-\mu)(X-\mu)(X-\mu)^{\mathrm{T}}]$ are

$$\begin{aligned} M_4 &= |V|^{-1/2}\int (x-\mu)^{\mathrm{T}}(x-\mu)(x-\mu)(x-\mu)^{\mathrm{T}}h((x-\mu)^{\mathrm{T}}V^{-1}(x-\mu))dx \\ &= \int y^{\mathrm{T}}VyV^{1/2}yy^{\mathrm{T}}V^{1/2}h(y^{\mathrm{T}}y)dy = V^{1/2}U\int z^{\mathrm{T}}\Omega zzz^{\mathrm{T}}h(z^{\mathrm{T}}z)dzU^{\mathrm{T}}V^{1/2} \\ &= V^{1/2}U\sum_i \omega_i \int z_i^2 zz^{\mathrm{T}}h(z^{\mathrm{T}}z)dzU^{\mathrm{T}}V^{1/2} = \sum_i \omega_i \tilde{k}V + \bar{k}V^{1/2}U\Omega U^{\mathrm{T}}V^{1/2} \\ &= \tilde{k}\,\mathrm{tr}(V)V + \bar{k}V^2, \end{aligned}$$

where we have introduced $y = V^{-1/2}(x-\mu)$, $z = U^{\mathrm{T}}y$ and

$$\int z_i^2 zz^{\mathrm{T}}h(z^{\mathrm{T}}z)dz = \tilde{k}I + \bar{k}e_i e_i^{\mathrm{T}},$$

for $\tilde{k} = \int z_i^2 z_j^2 h(z^{\mathrm{T}}z)dz$ where $i \neq j$, and $\bar{k} = \int z_i^4 h(z^{\mathrm{T}}z)dz - \tilde{k}$.

Chapter 4

Cluster analysis using trimmed projections

This chapter describes some ideas to help identify non-linearly shaped clusters in a low dimensional space. The procedure projects onto several affine subspaces those observations that are closest to the subspaces. In our proposal the affine subspaces are one-dimensional (straight lines) and are defined from observations in the data. The projections are then examined to determine the possible existence of clusters. This procedure can be interpreted as the computation of trimmed projections, and allows the identification of specific shapes that traditional clusters methods with good performance in low dimensional spaces may fail to detect. The suggested cluster algorithm is intended to be used once the dimension of a high dimensional data set has been reduced.

4.1 Identifying the local structure of the data

In previous chapters we have seen techniques to reduce the dimension of the space previous to clustering. This chapter presents a new method that searches for clusters in a space of low dimension, by detecting the areas of low or no density in the sample.

The method attempts to identify the presence of empty spaces in the data and use them as evidence of the existence of clusters. The algorithm we present generalizes to multivariate samples the idea proposed in Peña and Prieto (2001), where a large distance or gap between two consecutive observations was an indication of the end of one cluster and the beginning of the next, and therefore an indication of heterogeneity. In a space of dimension larger than one we cannot use the concept of ordering, and the way we

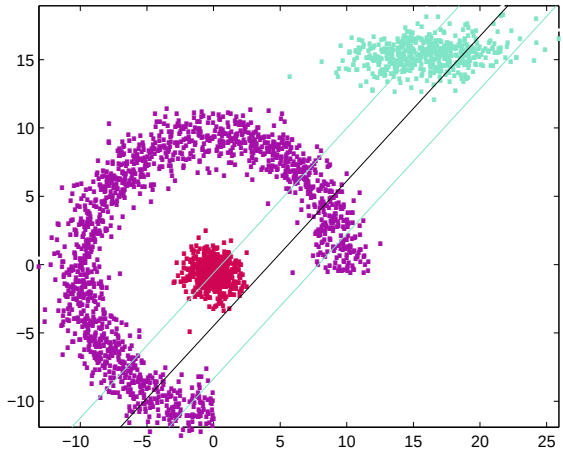


Figure 4.1: Clusters non linearly separable.

identify the gaps must be modified.

As an alternative, we explore the space considering only those observations contained in a sequence of bands. We select several random directions from observations in the data and project only the α-nearest observations onto the one-dimensional affine subspace defined by the direction and the observations. The parameter α, that indicates the proportion of observations projected onto any given line, has to be chosen. The resulting values can be understood as trimmed projections, and if we find a gap in the projections onto at least one of the lines, we may conclude that the sample is heterogeneous. By doing that, we identify the local structure of the sample closest to each line every time we select one. After defining enough affine subspaces, and computing the corresponding projections, we should be able to reconstruct the structure of the whole data.

If we choose to project the whole sample onto a given subspace, we may not be able to see the different clusters unless they are linearly separable. The linearity may not be present globally, but it can still be present locally, and projecting a subset of the sample might be enough to reveal part of the cluster structure. For example, in Figure 4.1 there is no direction able to discriminate the three clusters. But if we consider the line in black and project onto it the observations within the cyan lines, the spherical cluster would be set apart from the other clusters. Cluster methods such as KMEANS or MCLUST have problems dealing with this type of structures because they tend to identify elliptically shaped groups, while our strategy would seem better suited to detect them.

We want each observation to be projected onto at least one line, so that we can classify

the whole sample and capture its structure. Therefore, we need to draw a sufficient number of lines. Each line is defined by linking two chosen random observations from the sample.

The gaps in each set of projections are identified using a procedure proposed in Peña and Prieto (2001), where it is assumed that a lack of clusters in the data implies that the sample has been generated from a unimodal multivariate distribution, and therefore any projection of this sample will also be unimodal. The sample spacings of the projected observations $z_i = d^{\mathrm{T}} x_i$ onto the direction d are used to detect patterns that may indicate the presence of clusters. Thus, a subset of observations can be split into two clusters when a sufficiently large gap is found. The gaps or spacings of the sample are defined as the differences between two consecutive order statistics

$$w_i = z_{(i+1)} - z_{(i)}.$$

It is known that when the sample comes from a uniform distribution, the expected value for the gaps is $E(w_i) = 1/(n+1)$, which does not depend on i and so all gaps are expected to be equal. 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. More details on the properties of the gaps are found in Peña and Prieto (2001). We analyze the observations and identify the gaps assuming they follow a normal distribution function. If an inverse transformation is applied to the gaps using the normal distribution function, the resulting distribution for the modified gaps should be uniform in $[0, 1]$, where the distances between consecutive observations are expected to be of the same length. If any of these distances is significantly larger than the others, we conclude that the unimodal assumption does not hold and instead the data have been generated from a mixture of distributions, where the gaps indicate the different clusters.

4.2 Assigning labels to observations

Once we have looked for gaps in each trimmed projection, we need an algorithm that combines all this information and assigns a label to each observation, according to the group they belong.

The basic idea of this phase of the algorithm is to assign different labels to observations found in different clusters in any of the trimmed projections. The process is done iteratively: we analyze a first line d_1 and study the observations projected onto it, giving them appropriate labels. We then proceed analyzing the next line d_i and treat the

correspondent observations, and continue until all lines are treated. The fact that not all the observations are projected onto a given line adds extra complications that need to be commented.

For a given line d_i, two kinds of situations can arise regarding the observations that have been projected onto d_i;

- the observation is already labelled: it means that it has been projected onto previous line/s.
- the observation is not yet labelled: it is the first time that we treat it.

These situations may arise for any line we consider, except for the first one, when all observations are still unlabelled.

For the first projection where gaps were detected, we identify the clusters from the values of the gaps, defining as many clusters as the number of significantly large gaps plus one, and label the projected observations according to the groups they belong. Note that after completing this step, only the observations projected onto the first line may have been assigned to a group, the rest remain unclassified at this stage.

For the subsequent projections with gaps, we identify the partition from the values of the gaps and treat the observations as follows:

- If the observation is already labelled, we might need to assign a new label in case other observations with the same label are found in other group/s for this projection. In this case, we proceed to partition the sample according to the groups found for the current line.
- If the observation is not labelled, two situations may arise.
  - If the observation is found in a group containing only non-labelled observations, we assign a new label to the whole group.
  - On the contrary, if the observation appears in a group with other labelled observations, again two things might happen.
    - The group is homogeneous in the sense that only one label is involved. We assign to the non-labelled observations the label of the group. If in fact they did not belong to this group, another direction will eventually partition the group.

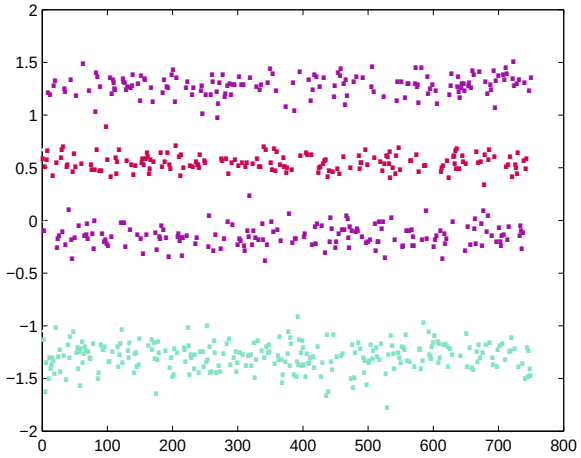


Figure 4.2: Trimmed projection with gaps.

  * Or, the group is heterogeneous, and different labels are present. We do not treat the non-labelled observations and wait until they are projected for an upcoming line. There is not enough evidence to decide to which group they should be assigned.

We summarize the different cases in the following scheme:

- labelled observations: assign a new label to observations having the same label as other observations found in another group of the current direction.
- non-labelled observations:
  - non-labelled group: new label to the whole group.
  - labelled group:
    * homogeneous group: give the label of the group to the non-labelled observations.
    * heterogeneous group: do nothing.

We are aware of the complicated nature of this procedure, although it is the one that provides better results in the simulations. One of the pitfalls associated to the way the observations are assigned to clusters it is the danger of ending up with a partition of the sample into too many clusters. There are several ways of regrouping clusters, but we realize the merging has to be done in a way that the non-linear structure detected

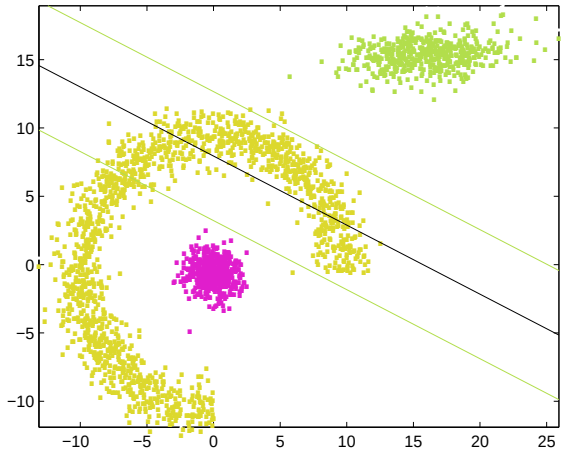


Figure 4.3: Non informative direction: no gaps.

with the help of the trimmed projections is not lost. We have tried different techniques, and although the results obtained are reasonable, they are still subject to improvement. We intend to work on the improvement of this stage of the algorithm in the future by studying non-linear merging strategies.

An example of a projection with gaps is found in Figure 4.1. Figure 4.2 shows the observations that were projected onto this direction, coloured according to the cluster they belong. The direction is represented in the vertical axis, while the horizontal axis is non informative. The number of gaps found in this direction is three, and they separate observations from the three original groups. Thus, this was an informative direction that helped classify 30% of the sample, because that is the proportion of the sample projected onto the direction. On the other hand, in Figure 4.3 a non informative direction is represented, where no gaps were found.

4.3 The GAPS algorithm

Let $x_1, \ldots, x_n$ be a sample drawn from X. The following steps define the GAPS algorithm.

1. Choose α, the proportion of observations to project. Let $m = [\alpha n]$ be the number of observations to project.

   Let G be a vector that assigns a label to each observation from the sample, and set $g = 1$, the number of groups.

Choose w_0, a cutoff that decides whether a distance between consecutive observations is large enough to be considered a gap or not.

2. Repeat the following for $i = 1 \dots n_d$, where n_d is the number of random directions to be drawn.

   (a) Choose two random observations x_{i_1}, x_{i_2} from the sample and define the direction $d_i = \frac{x_{i_1} - x_{i_2}}{\|x_{i_1} - x_{i_2}\|}$ that links them, for $i_1, i_2 = 1, \dots, n$, $i_1 \neq i_2$.

   (b) Find $x_{(1)}, \dots, x_{(m)}$, the m-nearest observations to the line defined by d_i, and project them onto d_i: $u_j = x_{(j)}^{\mathrm{T}} d_i d_i^{\mathrm{T}}$.

   (c) Let $z_j = (u_j - \bar{u})/s$ be the standardization of u_j, where $\bar{u}$ and s are the mean and standard deviation of $u_1, \dots, u_m$.

   (d) Sort out and transform $z_1, \dots, z_m$ using the standard normal distribution function: $\bar{z}_j = \Phi(z_{(j)})$. Store in the jth component of a vector named *pos* the position of $z_{(j)}$ before sorting out, for $j = 1, \dots, m$.

   (e) Let $w_j = \bar{z}_{j+1} - \bar{z}_j$ be the distances between consecutive values.

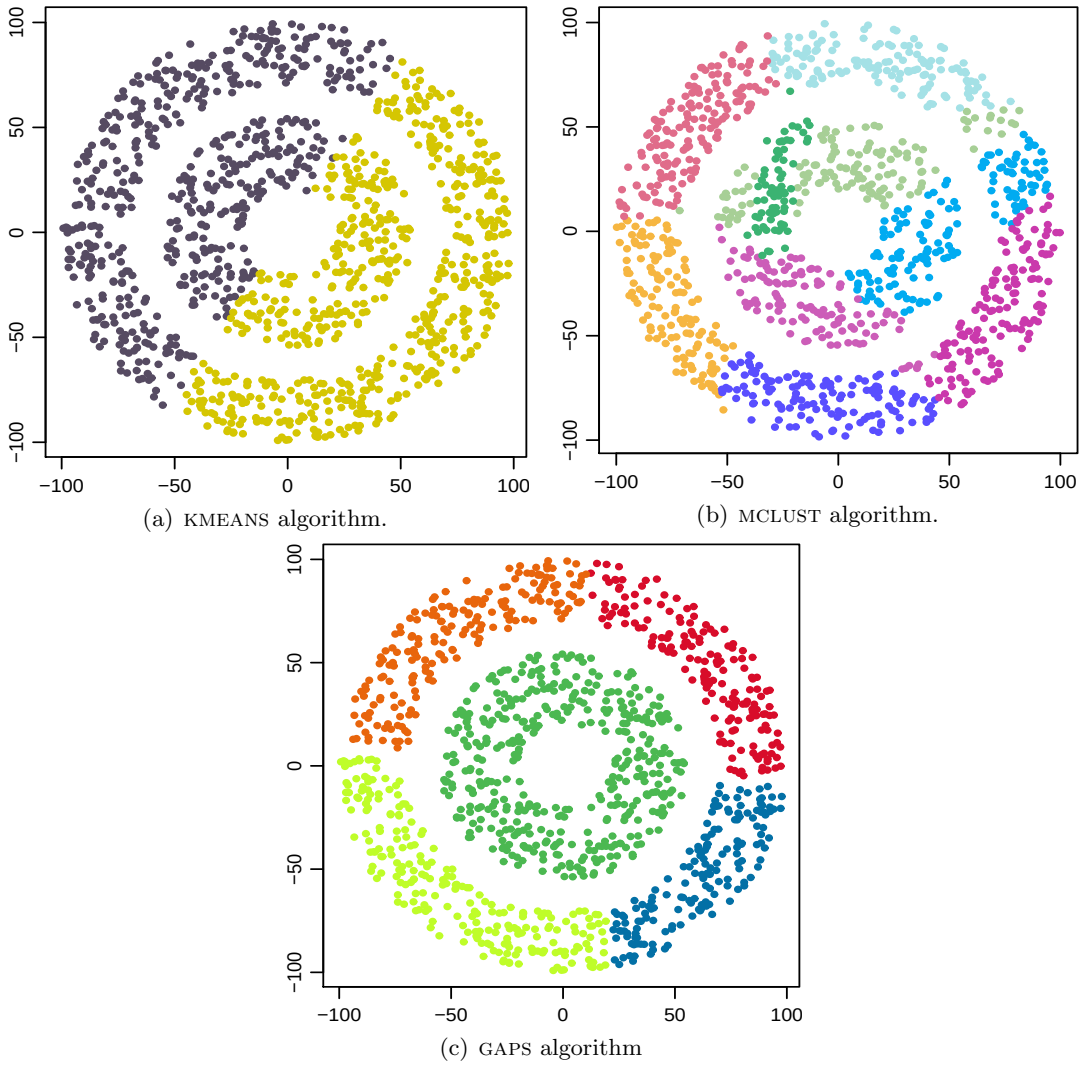
   (f) Let $J = \{1 \leq j \leq m - 1 : w_j > w_0\}$ and $J = J \cup \{0, m\}$. If $|J| > 2$ we found at least a gap in d_i:

   - If d_i is the first projection with gaps:
     i. For $k \in 1 : |J| - 1$ repeat: set $G_{(pos(t))} = g$ for $t = J_{(k)} + 1 : J_{(k+1)}$ and $g = g + 1$.
   - Otherwise:
     i. Let $N = \{G_{(pos(1))}, \dots, G_{(pos(J_{(2)}))}\}$.
     ii. For $k \in 2 : |J| - 1$ repeat:
       - For $s \in 1 : |N|$, if $N_s \neq 0$ repeat: if $G_{(pos(t))} = N_s$ then set $G_{(pos(t))} = g$, for any $t = J_{(k)} + 1 : J_{(k+1)}$ and $g = g + 1$.
       - Update $N = N \bigcup_{t=J_{(k)}+1}^{J_{(k+1)}} \{G_{(pos(t))}\}$.
     iii. For $k \in 1 : |J| - 1$ repeat: if $G_{(pos(t))} = 0$ for all $t = J_{(k)} + 1 : J_{(k+1)}$, set $G_{(pos(t))} = g$ for $t = J_{(k)} + 1 : J_{(k+1)}$ and $g = g + 1$.

3. The vector G returns, for each observation of the sample, a label indicating one of the corresponding $g - 1$ clusters.

4.4 Implementation details and examples

To illustrate the contribution of our algorithm, we use the example shown in Figure 4.4, where two uniformly distributed rings of different sizes are simulated. One ring is located inside the other, and thus they cannot be linearly separated by any direction. The two clusters are not spherically or elliptically shaped, which makes it a challenging example, specially for algorithms such as KMEANS or MCLUST. The sample size of the example is $n = 1200$, partitioned in clusters of sizes 400 and 800 respectively.



(a) KMEANS algorithm.

(b) MCLUST algorithm.

(c) GAPS algorithm

Figure 4.4: An example of two rings on a two dimensional space, and the results obtained with the algorithms KMEANS, MCLUST and GAPS.

Figure 4.4 presents the results for the algorithms KMEANS, MCLUST and GAPS. The KMEANS algorithm needs the number of clusters to be specified by the user. We perform

KMEANS iteratively by increasing the number of groups and select the choice that maximizes the following standardized difference between consecutive sum of squares within groups

$$(n-g)\frac{[SSW_{g-1} - SSW_g]}{SSW_g},$$

where SSW_g is the sum of squares within groups defined in (1), for a partition of the sample in g clusters. In Figure 4.4(a) we observe that KMEANS overlooks both clusters, providing a pretty bad solution. Approximately half of the observations are misclassified, since half of each ring is classified to another cluster. Table 4.1 shows how the observations have been classified, with 396 observations of the larger ring being classified to another cluster, as well as 205 observations of the smaller ring. What happens is that KMEANS partitions the sample using linear discrimination, and ignores the shape of the clusters.

Table 4.1: Results obtained with the KMEANS algorithm for the ring example in Figure 4.4.

	Cluster 1	Cluster 2	
Ring 1	395	405	800
Ring 2	198	202	400
	593	607	1200

The algorithm MCLUST, which uses the BIC criteria to choose the number of clusters, also fails in the identification of the two rings. The algorithm is suited to find elliptical shapes for the clusters. For the smaller ring, it expects the observations to be denser in the center than in the extremes, and therefore does not identify it as a cluster and combines observations from the two rings, as the blue- and green-coloured observations in Figure 4.4(b) show. In Table 4.2 we can see the number of observations from each ring assigned to each cluster, observe that clusters 6 to 9 are formed with observations from both rings, indicating the confusion between clusters mentioned.

Table 4.2: Results obtained with the MCLUST algorithm for the ring example in Figure 4.4.

	Clus 1	Clus 2	Clus 3	Clus 4	Clus 5	Clus 6	Clus 7	Clus 8	Clus 9	
Ring 1	160	155	142	137	127	6	17	55	1	800
Ring 2	0	0	0	0	0	134	124	79	63	400
	160	155	142	137	127	140	141	134	64	1200

Our algorithm performs significantly better than the preceding two algorithms, the results can be seen in Figure 4.4(c). The GAPS algorithm is able to completely discrim-

inate the inner ring as one cluster. On the other hand, the larger ring is partitioned in four clusters. Nevertheless, we think that as long as the two clusters are identified, the fact that one cluster is partitioned in several pieces is a minor problem. But this example shows that, as we mentioned above, the method is in need of a non-linear merging strategy, and that is what we intend to study in the future. Observe that this strategy should be able to merge clusters that are contiguous to each other. The algorithm is applied with a parameter $\alpha = 0.4$, and thus, the 40% of the sample is projected onto each random direction. In this case, projecting almost half of the sample is enough to detect the clusters. Table 4.3 shows the number of observations from each ring assigned to each cluster.

Table 4.3: Results obtained with the GAPS algorithm for the ring example in Figure 4.4.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	
Ring 1	231	208	199	162	0	800
Ring 2	0	0	0	0	400	400
	231	208	199	162	400	1200

4.5 Discussion

Further research will be conducted to reevaluate the group assignments. At present, the algorithm partitions the sample into too many clusters and a merging strategy is needed to be applied after the GAPS algorithm. This strategy must be able to merge non-linear clusters, as this is the target of our algorithm. If we look at the example in Figure 4.4, we observe that adjacent clusters need to be merged, that is, clusters that leave no gap between them. This could be done by computing a measure that summarized the distance of the neighbours to observations belonging to the extremes of different clusters. Given an observation in the border of the two clusters, the distance of this observation to neighbours within its cluster does not differ much from the distance of the same observation to neighbours belonging to the other cluster, since they touch each other. Therefore, we could consider a measure of this kind in order to merge adjacent clusters. Other procedures that recombine observations could be applied, as for example the one proposed in Peña et al. (2003).

Chapter 5

Nearest-neighbours median cluster algorithm [§]

In Chapter 2 we presented a method to reduce the dimension of the space in order to perform cluster analysis in a subspace of lower dimension. Chapter 4 described a way of identifying non-linearly shaped clusters in this low dimensional space based on projection pursuit ideas. In this chapter we propose a non-parametric cluster algorithm based on local medians that may also be applied after the dimension has been reduced and can as well be used to detect non-linear clusters. The detection of the clusters is carried on in the original space, and not based on projections, as the previous methods. Each observation is substituted by its local median and this new observation moves towards the peaks and away from the valleys of the distribution. The process is repeated until each observation converges to a fixpoint. We obtain a partition of the sample based on where the sequences of local medians have converged. The algorithm determines the number of clusters and the partition of the observations given a value of α, the proportion of neighbours. A fast version of the algorithm, where only a subset of the observations from the sample are treated, is also proposed. Furthermore, and for a univariate random variable, we prove the convergence of each point to the closest fixpoint, and the existence and uniqueness of a fixpoint on the neighbourhood of each mode.

[§]This chapter is a joint work with Professor Ruben Zamar from University of British Columbia.

5.1 Introduction

Given a multivariate sample in $\mathbb{R}^p$ drawn from a mixture of g populations, cluster analysis attempts to partition the sample into homogeneous groups according to the populations that generate them. Numerous cluster algorithms, such as k-means (Hartigan and Wong, 1979) or its robust version PAM (Kaufman and Rousseeuw, 1990), need the number of clusters to be specified by the user. Choosing the number of groups is one of the most difficult problems in cluster analysis and several approaches have been considered. One way of dealing with it is obtaining partitions of the data for different values of g and choosing the one that optimizes a certain measure of the strength of the clusters (Tibshirani et al., 2001). For instance, MCLUST algorithm (Fraley and Raftery, 1999) uses the BIC criteria to choose the number of components in a mixture of elliptical distributions. A second strategy that can be considered is to first partition the data into many small clusters, and merge the clusters on a second stage (Frigui and Krishnapuram, 1999). Other approaches consider extracting one cluster at a time (Zhung et al., 1996) or using methods that detect modes or bumps (Cheng and Hall, 1998).

Recently, a new strategy for the estimation of g has appeared. The purpose is to iteratively move the data points towards the centers of the clusters and use the number of convergence points as the number of clusters. In this sense, gravitational clustering (Wright, 1977; Kundu, 1999; Sato, 2000; Wang and Rau, 2001) assumes the data points are particles of unit mass with zero velocity which move towards clusters centers due to gravitational forces. Furthermore, mean-shift clustering (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 1999, 2000, 2001, 2002) uses kernel functions in density estimation to move data points towards denser areas.

In this chapter we also present an algorithm that moves the observations towards their cluster centers, but using the nearest neighbour approach (Mardia et al., 1979). In particular we benefit from the properties of the nearest neighbour median. Let X be a p-variate random vector with density function f, the α-nearest neighbour median at $x \in \mathbb{R}^p$ is the median of the distribution of X conditioned on $X \in B_x$, where B_x is a ball around x such that $P(X \in B_x) = \alpha$. If the local median at x is equal to x, x is a fixpoint. Otherwise, the local median has the property of moving towards the peaks and away from the valleys of f because it is located at the denser region of B_x. We can iterate the process by calculating the local median at the local median of x, and so on. The sequence will converge to a neighbourhood of a mode. If we repeat the above for all points in $\mathbb{R}^p$, we obtain a partition of them based on where the sequences of local medians have converged (fixpoints). The properties of the local median suggest a cluster

algorithm. Starting by calculating the local median at each observation of a p-variate sample of size n, we will obtain n sequences of local medians. The sequences will converge to $k < n$ different fixpoints, which returns a partition of the sample in g clusters. We call it ATTRACTORS algorithm.

A similar algorithm was presented in Wang et al. (2007). Attractors algorithm is a modified version of it where some improvements have been made. For each observation, both algorithms calculate its local median until convergence and therefore neighbours need to be identified at each iteration. While in Wang et al. (2007) the observations are updated on the value of its local medians, ATTRACTORS does not update them and therefore the neighbours are always observations from the sample. This difference makes possible to deduce some theoretical results for the ATTRACTORS algorithm, which was not possible with its previous version due to the mathematical complexity of updating the observations after each iteration. In particular, we prove, for the univariate case, the convergence of each point to the closest fixpoint and the existence and uniqueness of a fixpoint on the neighbourhood of each mode. Details are found in Section 5.4.

Furthermore, and from a computational point of view, ATTRACTORS algorithm allows for some improvement of the efficiency based on not considering all the observations of the sample, which permits a considerable saving on computational time. Section 5.2.2 addresses this issue.

The chapter is organized as follows. Section 5.2 describes the algorithm in detail and introduces the fast modified version. In Section 5.3 we study its behaviour through real and simulated examples. The theoretical results of the method are given in Section 5.4 and we conclude with some final remarks in Section 5.5.

5.2 Nearest neighbours and cluster analysis

Let X be a p-variate random vector with density function f and support S. The α-nearest neighbours median at $x \in \mathbb{R}^p$ is $g_\alpha(x) = (m_1, \ldots, m_p)^{\mathrm{T}}$, where m_j is the median of the marginal distribution Y_j and $Y = X \mid X \in B_x$ is the distribution of X conditioned on $X \in B_x$, with B_x being a ball around x such that $P(X \in B_x) = \alpha$. Several definitions of the multivariate median can be found in the literature. We use the coordinate-wise median for computational reasons.

Definition 5.1. *A fixpoint of g_α is any $x \in S$ such that $g_\alpha(x) = x$.*

If x is a fixpoint, the local median of f at x is x. If not, the local median has the

property of moving towards the peaks and away from the valleys of f because it is located at the denser region of B_x. We can iterate the process by calculating the local median at the local median of x, and so on. In effect, suppose that we iterate

$$x_{k+1} = g_\alpha(x_k),$$

for any starting value $x_0 \in \mathbb{R}^p$, the sequence $\{x_k\}$ will converge to a fixpoint of g_α. This process returns a partition of $\mathbb{R}^p$ based on where the sequences of local medians have converged (fixpoints).

Based on these results we suggest an algorithm for clustering.

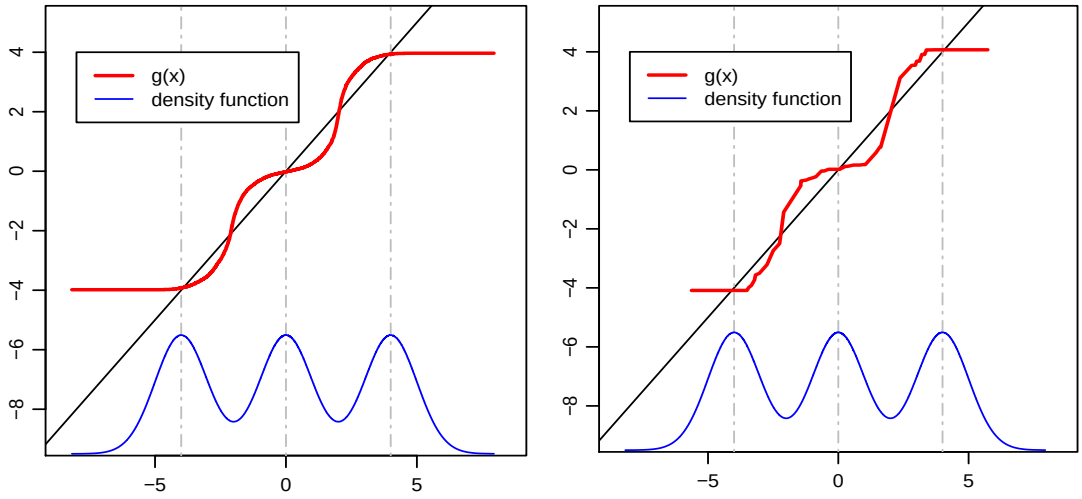
Let $x_1, \ldots, x_n$ be a sample from the random vector X. Let $m = [\alpha n]$ be the number of neighbours, given α. For each element of the sample, the algorithm iterates as

$$x^i_{k+1} = \hat{g}_\alpha(x^i_k),$$

starting at $x^i_0 = x_i$, and where $\hat{g}_\alpha(x^i_k) = (\hat{m}_1, \ldots, \hat{m}_p)^{\mathrm{T}}$ is the m-nearest neighbour median at x^i_k, where $\hat{m}_j$ is the median of the jth component of $x_{(1)}, \ldots, x_{(m)}$, the m observations from the sample that minimize the euclidean distances $\|x^i_k - x_l\|$, $l = 1 \ldots n$. The algorithm stops when $x^i_{k+1} = x^i_k$ for $i = 1, \ldots, n$. This phase of the algorithm finalizes with a division of the sample into as many clusters as fixpoints.

A representation of the local medians when X is distributed as a univariate mixture of three normal populations with means $\mu_1 = -4$, $\mu_2 = 0$ and $\mu_3 = 4$ and variances equal to 1 is found in Table 5.1. In Figure 5.1(a) we illustrate the density function f and the local median function g_α of f for $\alpha = \frac{1}{3}$. In Figure 5.1(b) we represent the estimated $\hat{g}_\alpha$ evaluated at a random sample of size 100 drawn from X. The black line corresponds to $g_\alpha(x) = x$, therefore every x in this line is a fixpoint. The function g_α has five fixpoints, three of them (attractors) correspond to the three modes. Since the populations are symmetric around the mode, the fixpoints coincide with the modes ($x^*_1 = \mu_1$, $x^*_2 = \mu_2$ and $x^*_3 = \mu_3$). The other two fixpoints correspond to the two valleys of the distribution (these fixpoints attract no x's and thus they are not of interest because they do not reveal any population). Observe that all points in $(-\infty, v_1)$ converge to μ_1, the points in (v_1, v_2) converge to μ_2 and the ones in (v_2, ∞) converge to μ_3, where v_1 and v_2 are the two valleys. In effect, if you try to delineate the sequence of local medians for a given x, you can see that they terminate in one of the three modes (proven in Theorem 5.3). In fact, the points in the extremes have already converged after the first iteration ($g_\alpha((-\infty, \mu_1)) = \mu_1$ and $g_\alpha((\mu_3, \infty)) = \mu_3$).

In practice, α is a parameter to be chosen by the user when invoking the algorithm, and consequently it is the user's choice to decide which α is appropriate for his purposes.



(a) g_α and f (b) $\hat{g}_\alpha$ for $n = 100$ and f

Figure 5.1: Function g_α, $\hat{g}_\alpha$ and density function f for a mixture of three normal distributions with means $\mu_1 = -4$, $\mu_2 = 0$ and $\mu_3 = 4$.

If, for example, there exists prior information about the size of the clusters expected to be found in the sample, α should be set accordingly. In theory, choosing α sufficiently small guarantees the identification of all clusters represented by a mode, see Theorem 5.8 for univariate samples. However, due to finite samples, small values of α could result on sequences of local medians stoping before reaching a fixpoint. In effect, it can happen that even when x is not a fixpoint, the same number of neighbours is found in each side of (each component of) x. The presence of these spurious fixpoints is more likely to occur when m is small. On the other hand, if α is too large, not all fixpoints will be detected, and consequently not all the populations that generate the sample are identified. Consequently, the choice of α is a tradeoff. Being aware of that, we recommend small values of α and introduce a second phase for the algorithm to get rid of the spurious fixpoints that attract very little observations and are a consequence of the sampling inaccuracies. In this phase, all fixpoints attracting less than $[\frac{\alpha}{3}n]$ are eliminated, and the observations converging to them are assigned to the closest fixpoint. In addition to that, we consider a last step where we merge two clusters if their means are close enough in terms of the Mahalanobis distance. The algorithm is described in the next section.

5.2.1 The ATTRACTORS algorithm

Let $x_1, \ldots, x_n$ be a sample drawn from X. The following steps constitute the ATTRACTORS algorithm.

1. Choose α, the proportion of neighbours. Let $m = [\alpha n]$ be the number of neighbours.

2. Repeat the following for each observation x_i, $i = 1 \dots n$.

   (a) Let $x_0^i = x_i$ and $k = 0$.

      i. Calculate the local median at x_k^i, $x_{k+1}^i = \hat{g}_\alpha(x_k^i)$.

      ii. If $x_k^i \neq x_{k+1}^i$ set $k = k + 1$ and return to i. Otherwise $\phi(x_i) = x_k^i$ is the fixpoint where the sequence $\{x_k^i\}$ converges.

3. Let $x_1^*, \dots, x_g^*$ be the elements of $\bigcup_{i=1}^n \{\phi(x_i)\}$. For each $t = 1 \dots g$, define the group $G_t = \{x_i \mid \phi(x_i) = x_t^*\}$ as the set of observations attracted by the fixpoint x_t^*.

4. Discard from being a fixpoint any x_j^* such that $|G_j| < G_{low}$, where $G_{low} = [\frac{\alpha}{3} n]$ and $j = 1 \dots g$. Update g, the number of fixpoints. Reassign the elements of G_j to the cluster G_t, where t is such that the Mahalanobis distance $\mathrm{MD}(\bar{x}_{G_j}, \bar{x}_{G_t}, S_{G_t})$ is minimum, for $t = 1 \dots g$. Substitute x_t^* for the weighted mean of x_j^* and x_t^*.

5. Sort out the groups by descending number of observations and repeat the following for all $j = 1 \dots g - 1$.

   (a) For $t = j + 1 \dots g$, merge the groups G_j and G_t if the Mahalanobis distance $\mathrm{MD}(\bar{x}_{G_j}, \bar{x}_{G_t}, S_{G_j}) < \chi^2_{0.9}$. Update x_j^* on the weighted mean of x_j^* and x_t^*.

5.2.2 Improvement of the computational efficiency

The algorithm chooses the neighbours and calculates the local median several times for each observation until it converges. If n is large, the process can be time consuming. We propose a modified version of the algorithm where we only consider the convergence of a subset of the n observations. The treated observations are chosen randomly. The key problem is to decide the number of observations to be treated, n_{sub}. Let x^* be a fixpoint attracting a proportion $p > 0$ of the points. Thus, the probability of an observation of the sample to converge to x^* is p. Moreover, $(1 - p)^n$ is the probability of finding n consecutive observations not converging to the fixpoint. If n tends to ∞, $(1 - p)^n$ tends to 0, and there exist N such that for any $n > N$ the probability $(1 - p)^n$ is very small. Therefore, if after treating N consecutive observations none of them have converged to x^*, we assume x^* does not exist. We set $\mathrm{prob} = (1 - p)^N$ to be very small and thus $N = \log(\mathrm{prob}) / \log(1 - p)$, where p is chosen to be the maximum size for a fixpoint to be considered fixpoint, in the sense that we do not mind not to detect fixpoints attracting less than a proportion p of points. The procedure starts treating observations

and marking to which fixpoint they converge using a counter to keep track of the number of observations treated. Every time an observation converges to a fixpoint none of the previous observations have converged (a new fixpoint appears), we set the counter to zero. If we find N consecutive observations converging to "old" fixpoints, that is, if the counter reaches the value N, we stop. Each non-treated observation will be assigned to the cluster defined by the closest fixpoint.

Depending on the values of p, prob and the sample size n, we may encounter n_{sub} being larger than n. In this case, all observations of the sample are treated and we experience no improvement of the efficiency. Nevertheless, this will happen for small sample sizes which does not cause the algorithm to be inefficient. Instead, if n is large, $n - n_{sub}$ will also be large and the efficiency will improve significantly since only n_{sub} observations are treated, as opposed to n.

Fast-ATTRACTORS algorithm

Let $x_1, \ldots, x_n$ be a sample drawn from X. The following steps constitute the fast version of the ATTRACTORS algorithm.

1. Choose α, the proportion of neighbours. Set prob to a small value and choose p to be the maximum size for a cluster. Set $N = \log(\text{prob})/\log(1-p)$, $m = [\alpha n]$ to be the number of neighbours, counter $= 0$ and $i = 1$. Order the n observations randomly.

2. While counter $< N$ and $i \leq n$ repeat the following:

   (a) Let $x_0^i = x_i$ and $k = 0$.

      i. Calculate the local median at x_k^i, $x_{k+1}^i = \hat{g}_\alpha(x_k^i)$.

      ii. If $x_k^i \neq x_{k+1}^i$ set $k = k + 1$ and return to i. Otherwise $\phi(x_i) = x_k^i$ is the fixpoint where the sequence $\{x_k^i\}$ converges.

      iii. If $\phi(x_i) \in \Phi$ then counter $=$ counter $+ 1$. Otherwise set counter $= 0$ and $\Phi = \Phi \bigcup \{\phi(x_i)\}$. Set $i = i + 1$.

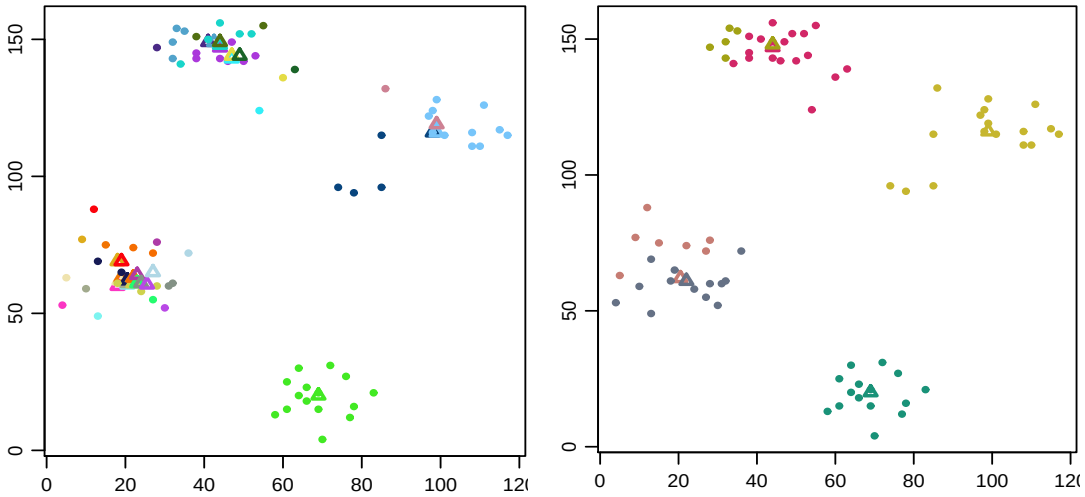
3. Set $n_{sub} = i - 1$ and let $x_1^*, \ldots, x_g^*$ be the elements of Φ. For each $j = 1 \ldots g$, define the group $G_j = \{x_i \mid \phi(x_i) = x_j^*\}$ as the set of observations attracted by the fixpoint x_j^*, where $i = 1 \ldots n_{sub}$.

4. For each $i = n_{sub} + 1 \ldots n$, assign x_i to the cluster G_j, where j is such that the euclidean distance $\|x_i - \bar{x}_{G_j}\|$ is minimized, for $j = 1 \ldots g$.

5. Apply steps 4 and 5 of the previous version of the algorithm in Section 5.2.1.

The values p and prob can be changed using prior information if is available.

5.3 Examples and simulation results



(a) 1st iteration: 27 different local medians.

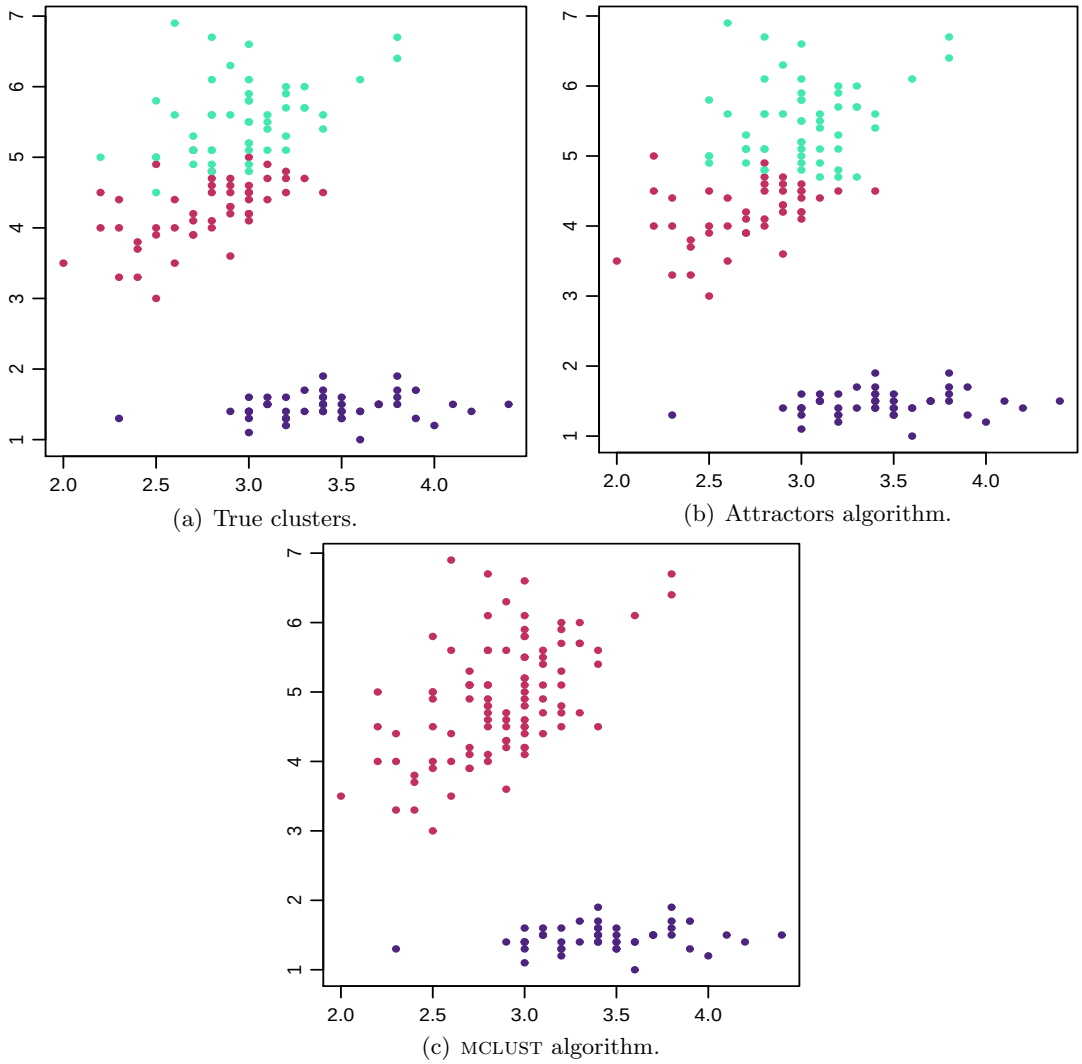
(b) 2nd iteration: 6 different local medians.

(c) 3rd iteration: 4 fixpoints.

Figure 5.2: Ruspini data and the local medians (triangles) after each iteration when invoking the ATTRACTORS algorithm with $\alpha = 0.2$.

We start by illustrating the behaviour of the algorithm on some well-known examples from the literature such as those of Ruspini (1970) and Fisher (1936). The Ruspini data set is a two-dimensional example consisting of 75 observations divided into four well-separated clusters. In Figure 5.2 we represent the observations and the sequences of local

medians when the algorithm is invoked with $\alpha = 0.2$. We start calculating the local median at each observation of the sample. In Figure 5.2(a) we plot with the same colour the observations that share local median, and represent the local median with the triangle sign of the same colour. We obtain 27 different local medians after this first iteration. We calculate now the local medians at the 27 points and obtain six new different local medians, which are plotted in Figure 5.2(b). After three iterations all sequences have already converged to four different fixpoints, as shows Figure 5.2(c). The four groups have been found correctly and in this case steps 4 and 5 of the algorithm were not needed.



(a) True clusters.

(b) Attractors algorithm.

(c) MCLUST algorithm.

Figure 5.3: Iris data set on the two-dimensional space of the variables sepal-width and petal-length and results for the MCLUST and ATTRACTORS algorithm ($\alpha = 0.3$).

The Iris dataset described in Fisher (1936) consists in 50 flowers from each of the species *Iris setosa*, *Iris versicolor* and *Iris virginica*. The four variables are the length

and the width of the sepal and petal respectively. One specie is linearly separable from the other two, while the latter tend to overlap and are hard to distinguish. Figure 5.3(a) shows the dataset in the two-dimensional space of the sepal-width and the petal-length variables. The results obtained with the algorithm MCLUST are shown in Figure 5.3(c). MCLUST assumes that the sample comes from a mixture of elliptical populations and estimates the parameters for several options on the number of clusters, selecting the one that optimizes the BIC criteria. This approach is called model-based clustering. The algorithm does not perform well with the iris dataset and confuses two of the clusters giving as a result two clusters with 50 and 100 observations each, instead of three. When using the ATTRACTORS algorithm instead, the observations are clustered as shown in Figure 5.3(b), where 12 observations of the two overlapped groups were classified incorrectly.

In order to asses the performance of the fast-ATTRACTORS algorithm, we use the simulated data set shown in Figure 5.4. The clusters are reasonably separated but four of them have unusual shapes. The ATTRACTORS algorithm is able to identify the five different clusters with no classification error, as it is also achieved in Wang et al. (2007). Fast-ATTRACTORS algorithm returns a proportion of misclassified observations equal to 0.014, which is a total of 14 observations wrongly classified (see Figure 5.4). However, the algorithm treats only $n_{sub} = 160$ observations, more than 6 times less than treating the whole sample of $n = 1000$ observations, and which reduces the computational time significantly. The values chosen for the algorithm were $p = 0.1$ and prob $= 0.001$, and thus $N = 66$, the number of consecutive observations with non new fixpoints that needed to appear as a condition to stop.

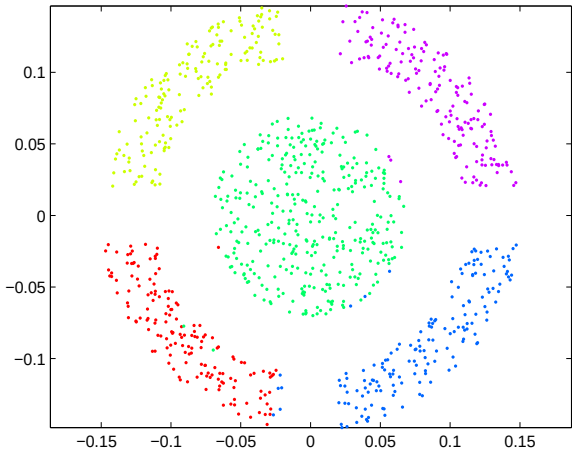


Figure 5.4: Partition of the data set using the fast-ATTRACTORS algorithm with $\alpha = 0.1$.

We also study the properties of the algorithm through a computational experiment on randomly generated samples. We start generating samples from mixtures of g multivariate normal populations with distinct scatter matrices, as stated previously in Chapter 2. The results can be seen in Table 5.1. The measure to assess the performance of the algorithms is the proportion of misclassified observations. Attractors is invoked with $\alpha = 0.05$ to assure all clusters are detected for the different values of g. MCLUST is designed to estimate mixtures of elliptical populations and therefore performs very well in this situation, but nevertheless the results for the ATTRACTORS algorithm are comparable to those of MCLUST, performing as good except for the case of fifteen dimensions and two groups, where two out of every ten observations are misclassified.

Table 5.1: Proportion of misclassified observations for the algorithms ATTRACTORS ($\alpha = 0.05$), MCLUST and kurtosis under a mixture of g normal distributions.

p	g	Attractors	Kurtosis	Mclust
4	2	0.003	0.080	0.014
	4	0.011	0.091	0.041
	8	0.023	0.111	0.027
8	2	0.040	0.146	0.011
	4	0.013	0.115	0.037
	8	0.024	0.082	0.061
15	2	0.224	0.299	0.003
	4	0.099	0.332	0.024
	8	0.031	0.084	0.057
Average		0.052	0.149	0.031

If we consider mixtures of non-normal populations, where for example each cluster follows marginal univariate t-students with two degrees of freedom so that the shape of the clusters resembles a star, the results change substantially. MCLUST algorithm fails clearly on detecting the clusters specially when the dimension is large, while ATTRACTORS behaves very well in all situations (see Table 5.2). MCLUST assumes the data comes from a normal mixture and estimates the parameters of the components of the mixtures, therefore is troubled when dealing with non-elliptical mixtures. Attractors is not a model-based algorithm, it does not assume any model underneath the data, and consequently does not depend strongly on the shape of the clusters. Table 5.3 shows the percentage of times that the number of clusters that ATTRACTORS and MCLUST return coincides with g, for both sets of simulations, the mixture of normal distributions and the mixture of non-normal distributions.

In addition to the results obtained, it is worth mentioning that MCLUST is computa-

Table 5.2: Proportion of misclassified observations for the algorithms ATTRACTORS ($\alpha = 0.05$), MCLUST and Kurtosis under a mixture of g non-normal distributions (marginal t-students).

p	g	Attractors	Kurtosis	Mclust
4	2	0.014	0.219	0.279
	4	0.016	0.202	0.199
	8	0.023	0.172	0.125
8	2	0.011	0.280	0.369
	4	0.013	0.258	0.282
	8	0.021	0.236	0.201
15	2	0.094	0.350	0.473
	4	0.020	0.303	0.339
	8	0.020	0.300	0.227
Average		0.026	0.258	0.277

Table 5.3: Percentage of times (%) that the number of clusters that ATTRACTORS ($\alpha = 0.05$) and MCLUST return coincides with g, for a mixture of normal distributions and a mixture of non-normal distributions (marginal t-students).

		Normals		T-students	
p	g	Attractors	Mclust	Attractors	Mclust
4	2	96	96	95	0
	4	85	77	87	0
	8	47	62	61	2
8	2	82	96	96	0
	4	77	66	81	1
	8	39	29	51	0
15	2	30	99	69	2
	4	40	64	83	1
	8	28	32	51	0
Average		58.22	69	74.89	0.67

tionally more intensive than ATTRACTORS, even when not using the fast version of the algorithm.

5.4 Univariate nearest-neighbours median study

The results in this section only apply to the univariate case. The extension to the multivariate case has proved to be highly non-trivial and may require a significant amount of

original work. Nonetheless, the general idea behind them is still valid for the multivariate case.

Let X be a random variable with distribution function F and density function f with convex support S.

The local median g_α of f at $x \in \mathbb{R}$ is the median of the interval of weight $\alpha \in [0,1]$, centered at x:

$$F(g_\alpha(x)) - F(x - d_x) = \frac{\alpha}{2} \tag{5.1}$$

where d_x is such that

$$F(x + d_x) - F(x - d_x) = \alpha. \tag{5.2}$$

Substituting (5.2) in (5.1), g_α can also be written as

$$g_\alpha(x) = F^{-1}\left[\frac{F(x + d_x) + F(x - d_x)}{2}\right]$$

where d_x satisfies (5.2).

Following Definition 5.1, if x is a fixpoint, the local median of f at x is x, the center of the interval. In Theorem 5.2 we prove that any density with convex support has at least one point with these properties.

For $\alpha = 1$, the local median of f is just the median of the distribution f, for any $x \in \mathbb{R}$. The median, therefore, is the unique fixpoint of g_α. We will not consider this case because is not of interest for our purpose.

Theorem 5.2. *Let f be a density with convex support S, for $0 < \alpha < 1$, the function g_α has at least one fixpoint.*

Proof of Theorem 5.2. From (5.1) we have

$$F(g_\alpha(x)) = \frac{\alpha}{2} + F(x - d_x) \geq \frac{\alpha}{2}$$

Similarly, from (5.1) and (5.2)

$$F(g_\alpha(x)) = F(x + d_x) - \frac{\alpha}{2} \leq 1 - \frac{\alpha}{2}$$

Thus, g_α is bounded by

$$F^{-1}\left(\frac{\alpha}{2}\right) \leq g_\alpha(x) \leq F^{-1}\left(1 - \frac{\alpha}{2}\right). \tag{5.3}$$

Therefore,

$$g_\alpha(x) > x, \text{ for } x < F^{-1}\left(\frac{\alpha}{2}\right)$$
$$\text{and } g_\alpha(x) < x, \text{ for } x > F^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Since F and F^{-1} are continuous, g_α is continuous and therefore there exists an $x^* \in (F^{-1}(\frac{\alpha}{2}), F^{-1}(1-\frac{\alpha}{2}))$ such that $g_\alpha(x^*) = x^*$. □

The following theorem states that any $x \in \mathbb{R}$ will eventually converge to a fixpoint if we substitute x by its local median $g_\alpha(x)$, $g_\alpha(x)$ again by its local median $g_\alpha(g_\alpha(x))$ and so on, repeating the process until convergence.

Theorem 5.3. *Let f be a density with convex support S. Suppose that we iterate*

$$x_{k+1} = g_\alpha(x_k),$$

then, for any starting value $x_0 \in \mathbb{R}$, and for $0 < \alpha < 1$, the sequence $\{x_k\}$ converges to a fixpoint of g_α. In particular, if $x_0 < g_\alpha(x_0)$, $\{x_k\}$ converges to the smallest fixpoint greater than x_0. If $x_0 > g_\alpha(x_0)$, $\{x_k\}$ converges to the greatest fixpoint smaller than x_0.

Theorem 5.3 also gives results on where the sequence $\{x_k\}$ converges. If x_0 is located at a part of f with positive slope, $\{x_k\}$ converges to the first fixpoint on the right of x_0, and viceversa, which implies that the sequence escalates the density function towards the local mode.

Proof of Theorem 5.3. In order to prove that g_α is non-decreasing we want to show that $g_\alpha(x) \geq g_\alpha(y)$ if $x > y$. Due to the monotonicity of F^{-1}, it is sufficient to prove that $F(x + d_x) \geq F(y + d_y)$ and $F(x - d_x) \geq F(y - d_y)$. Again, due to the monotonicity of F, it is enough to show

$$\begin{aligned} x + d_x &\geq y + d_y \\ x - d_x &\geq y - d_y. \end{aligned} \tag{5.4}$$

Let us suppose the contrary, $x + d_x < y + d_y$, then $d_x < d_y$ and so $x - d_x < y - d_y$. Therefore

$$\alpha = F(x + d_x) - F(x - d_x) < F(y + d_y) - F(y - d_y) = \alpha, \tag{5.5}$$

which is a contradiction. The proof for the second part of (5.4) is analogous. The inequality in (5.5) is strict because it can only be equal if both $F(x + d_x) = F(y + d_y)$ and $F(x - d_x) = F(y - d_y)$, which can happen if the four points are not in S, and that is only possible for the excluded case $\alpha = 1$.

Consider first $x_0 < g_\alpha(x_0) = x_1$, then, since g_α is non-decreasing, $g_\alpha(x_0) \leq g_\alpha(x_1)$. Thus,

$$x_0 < g_\alpha(x_0) = x_1 \leq g_\alpha(x_1) = x_2 \leq \ldots \leq g_\alpha(x_{k-1}) = x_k \leq \ldots$$

since the sequence $\{x_k\}$ is non-decreasing and bounded (see (5.3)), there exists x^* such that $\lim_{k\to\infty} x_k = x^*$. Moreover, x^* is a fixpoint:

$$x^* = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g_\alpha(x_k) = g_\alpha(\lim_{k\to\infty} x_k) = g_\alpha(x^*)$$

Also, for $x \in (x_k, x_{k+1})$, $g_\alpha(x) \geq g_\alpha(x_k) = x_{k+1} > x$, which means that there are no fixpoints in (x_k, x_{k+1}). Therefore the fixpoint x^* is the smallest fixpoint greater than x_0.

Analogously, if $x_0 > g_\alpha(x_0)$, $\{x_k\}$ converges to the greatest fixpoint smaller than x_0.

If $x_0 = g_\alpha(x_0)$, x_0 is already a fixpoint. □

The next theorem claims that, if the distribution is unimodal, the corresponding local median function g_α has only one fixpoint, regardless the value of α.

Theorem 5.4. *Let f with convex support S be a strictly unimodal density, then, for $0 < \alpha < 1$, the function g_α of f has a unique fixpoint.*

Proof of Theorem 5.4. In Theorem 5.2 we proved the existence of at least one fixpoint, for any f. In this proof we deal with its uniqueness for f unimodal.

Suppose there exist two fixpoints $x_1, x_2 \in \mathbb{R}$ such that $x_1 < x_2$. Assume that, without loss of generality, $f(x_1) < f(x_2)$. Otherwise consider the random variable $Y = -X$ with density function $f_Y(x) = f(-x)$ instead.

Let d_1 and d_2 be such that $F(x_1 + d_1) - F(x_1) = F(x_1) - F(x_1 - d_1) = F(x_2 + d_2) - F(x_2) = F(x_2) - F(x_2 - d_2) = \frac{\alpha}{2}$.

Note that $x_1 + d_1 < x_2 + d_2$, otherwise $(x_2, x_2 + d_2) \subset (x_1, x_1 + d_1)$ and, since the integrals of $f(x)$ on these intervals are $\frac{\alpha}{2}$, it is a contradiction because S is a convex support.

When f is a unimodal density

$$f(x) > \min\{f(a), f(b)\}, \text{ for any } a < x < b. \tag{5.6}$$

The following results hold too,

$$f(x) < f(x_1), \text{ for any } x < x_1 \tag{5.7}$$

$$f(x) > f(x_1), \text{ for any } x \in (x_1, x_2) \tag{5.8}$$

the expression (5.8) is due to (5.6).

Observe that

$$f(x_1 + d_1) < f(x_1). \tag{5.9}$$

Indeed, since

$$\alpha/2 = \int_{x_1-d_1}^{x_1} f(x)dx < d_1 f(x_1),$$

because of (5.7), and

$$\alpha/2 = \int_{x_1}^{x_1+d_1} f(x)dx > d_1 \min\{f(x_1), f(x_1+d_1)\},$$

using (5.6), and we obtain that $\min\{f(x_1), f(x_1+d_1)\} < f(x_1)$ which leads to (5.9).

This result implies that $x_2 < x_1 + d_1$, otherwise $x_1 < x_1 + d_1 < x_2$, and we know that $f(x_2) > f(x_1) > f(x_1 + d_1)$, which contradicts (5.6).

Therefore, we established the following order

$$x_1 < x_2 < x_1 + d_1 < x_2 + d_2.$$

We will see now that $d_1 > d_2$. In effect,

$$\alpha/2 = \int_{x_1-d_1}^{x_1} f(x)dx = \int_{x_2-d_2}^{x_2} f(x)dx,$$

and the values of $f(x)$ in the second integral are larger than in the first, because the expressions (5.7) and (5.8) hold, so the interval of integration should be shorter. Thus, the interval (x_1^+, x_2^+), where $x_1^+ = x_1 + d_1$ and $x_2^+ = x_2 + d_2$, is shorter than (x_1, x_2) because $x_2^+ - x_1^+ = (x_2 - x_1) - (d_1 - d_2) < x_2 - x_1$.

Finally,

$$\begin{aligned} F(x_2) - F(x_1) &> (x_2 - x_1) f(x_1) > (x_2^+ - x_1^+) f(x_1) \\ &> (x_2^+ - x_1^+) \max_{x \in (x_1^+, x_2^+)} f(x) > F(x_2^+) - F(x_1^+). \end{aligned}$$

The first inequality is due to (5.8), the second due to $d_1 > d_2$, and the third inequality is because $f(x_1) > \max_{x \in (x_1^+, x_2^+)} f(x)$, which is true since (5.9) and the fact that f is strictly decreasing after x_1^+ because the mode of f is in (x_1, x_1^+).

This result leads to a contradiction because $F(x_2) - F(x_1) = F(x_1^+) - F(x_1) - (F(x_1^+) - F(x_2)) = \alpha/2 - (F(x_1^+) - F(x_2)) = F(x_2^+) - F(x_2) - (F(x_1^+) - F(x_2)) = F(x_2^+) - F(x_1^+)$, therefore $x_1 = x_2$ and we have shown that it is not possible to have two distinct fixpoints x_1 and x_2. Therefore, for any unimodal distribution, the function g_α has one and only one fixpoint (the existence was already proved in Theorem 5.2). □

The following Corollaries refer to where the fixpoint is located. In particular, the smaller α the closer the fixpoint to the mode.

Corollary 5.5. *Let x_m be the mode of f, then $|F(x^*) - F(x_m)| \leq \frac{\alpha}{2}$, where x^* is the fixpoint.*

Proof of Corollary 5.5. Since x^* is a fixpoint, d_{x^*} is such that

$$F(x^* + d_{x^*}) - F(x^*) = F(x^*) - F(x^* - d_{x^*}) = \frac{\alpha}{2} \tag{5.10}$$

Then, x_m must be in $(x^* - d_{x^*}, x^* + d_{x^*})$, otherwise the density is strictly monotonous and the two integrals in (5.10) can not be equal.

Therefore, $|F(x^*) - F(x_m)| \leq \frac{\alpha}{2}$. □

Corollary 5.6. *If $\alpha \to 0$ then $x^* \to x_m$.*

Proof of Corollary 5.6. From the previous proof,

$$|x^*_\alpha - x_m| < d_{x^*} = F\left(x^*_\alpha + \frac{\alpha}{2}\right) - F\left(x^*_\alpha\right).$$

Since F is continuous, $d_{x^*} \to 0$ as $\alpha \to 0$. Therefore $|x^*_\alpha - x_m| \to 0$ as well. □

In the following theorem we show that, for small enough values of α, there exist a unique fixpoint in the neighbourhood of a mode of a distribution f.

Definition 5.7. *x_m is a (δ_1, δ_2)-mode if it is a mode and f is strictly unimodal in the interval $\left[F^{-1}(y_m - \delta_1), F^{-1}(y_m + \delta_2)\right]$, where $y_m = F(x_m)$ and $\delta_1, \delta_2 > 0$.*

Theorem 5.8. *Let x_m be a (δ_1, δ_2)-mode, then, for any $\alpha \leq \min(\delta_1, \delta_2)$, there exists a fixpoint $x^* \in \left(F^{-1}(y_m - \frac{\alpha}{2}), F^{-1}(y_m + \frac{\alpha}{2})\right)$ and it is the only fixpoint in the interval $\left[F^{-1}(y_m - \delta_1 + \frac{\alpha}{2}), F^{-1}(y_m + \delta_2 - \frac{\alpha}{2})\right]$.*

Proof of Theorem 5.8. In order to prove the existence of a fixpoint in the interval $\left(F^{-1}(y_m - \frac{\alpha}{2}), F^{-1}(y_m + \frac{\alpha}{2})\right)$, we define

$$\delta_x^- = x - F^{-1}\left(F(x) - \frac{\alpha}{2}\right)$$
$$\delta_x^+ = F^{-1}\left(F(x) + \frac{\alpha}{2}\right) - x$$

which implies that $F(x + \delta_x^+) - F(x) = F(x) - F(x - \delta_x^-) = \frac{\alpha}{2}$. If x is a fixpoint, then $\delta_x^- = \delta_x^+$.

Let $x_l = F^{-1}(y_m - \frac{\alpha}{2})$ be on the left of the mode, then $\delta_{x_l}^- > \delta_{x_l}^+$ because f increases in $\left(F^{-1}(y_m - \delta_1), x_m\right)$. Let also $x_r = F^{-1}(y_m + \frac{\alpha}{2})$, on the right of the mode, then

$\delta_{x_r}^- < \delta_{x_r}^+$ because f decreases in $\big(x_m, F^{-1}(y_m+\delta_2)\big)$. Note that $\delta_{x_l}^-$ and $\delta_{x_r}^+$ are contained in $\big[F^{-1}(y_m-\delta_1), F^{-1}(y_m+\delta_2)\big]$, so that proper monotonicity is in place. Therefore, since $\delta_x^+ - \delta_x^-$ is a continuous function of x, because F and F^{-1} are continuous in $\big[F^{-1}(y_m-\frac{\alpha}{2}), F^{-1}(y_m+\frac{\alpha}{2})\big]$, there exist an $x^* \in \big(F^{-1}(y_m-\frac{\alpha}{2}), F^{-1}(y_m+\frac{\alpha}{2})\big)$ such that $\delta_{x^*}^+ = \delta_{x^*}^-$, which implies that x^* is a fixpoint.

Regarding the uniqueness of the fixpoint in $\big[F^{-1}(y_m-\delta_1+\frac{\alpha}{2}), F^{-1}(y_m+\delta_2-\frac{\alpha}{2})\big]$, we refer to the proof of Theorem 5.4. However, we should mention a number of things that changed now. Since we are proving the uniqueness of the fixpoint on a finite interval, we start assuming that x_1 and x_2 are two different fixpoints in the interval $\big[F^{-1}(y_m-\delta_1+\frac{\alpha}{2}), F^{-1}(y_m+\delta_2-\frac{\alpha}{2})\big]$. Also, inequalities in (5.6) and (5.7) should be restricted to the interval of interest, so that $f(x) > \min\{f(a), f(b)\}$, for any $F^{-1}(y_m-\delta_1) \leq a < x < b \leq F^{-1}(y_m+\delta_2)$, and $f(x) < f(x_1)$, for any $x \in \big[F^{-1}(y_m-\delta_1), x_1\big)$. The rest of the proof is exactly the same. We can conclude now that the unique fixpoint in $[F^{-1}(y_m-\delta_1+\frac{\alpha}{2}), F^{-1}(y_m+\delta_2-\frac{\alpha}{2})]$ is located in $\big(F^{-1}(y_m-\frac{\alpha}{2}), F^{-1}(y_m+\frac{\alpha}{2})\big)$. □

Corollary 5.9. *Following Theorems 5.3 and 5.8, for any starting value* $x_0 \in \big[F^{-1}(y_m-\delta_1+\frac{\alpha}{2}), F^{-1}(y_m+\delta_2-\frac{\alpha}{2})\big]$, *the sequence* $\{x_k\}$ *converges to* x^*.

Theorem 5.8 is the main result of the chapter. It states that, if f is strictly unimodal in an interval of weight $\delta_1+\delta_2$, for any $\alpha \leq \min(\delta_1, \delta_2)$ the identification of the population that induces the mode is guaranteed. Any $x_0 \in \big[F^{-1}(y_m-\delta_1+\frac{\alpha}{2}), F^{-1}(y_m+\delta_2-\frac{\alpha}{2})\big]$ will be attracted by a fixpoint x^*, which points out the existence of a mode in its proximity. Therefore, any population in f characterized by a (δ_i, δ_j)-mode such that $\alpha \leq \min(\delta_i, \delta_j)$ will be revealed. Theorem 5.8, thus, provides tools to use the algorithm.

It is worth mentioning that the restriction $\alpha \leq \min(\delta_1, \delta_2)$ in Theorem 5.8 is a sufficient condition but not always necessary. In practice, good performance can be achieved with values of α significantly exceeding the bound. As a matter of fact, the weight of each component of the mixture in Figure 5.1 is 1/3 and, since the three densities are symmetric, the corresponding δ_1 and δ_2 are all equal to 1/6. The restriction $\alpha \leq \min(\delta_1, \delta_2)$ does not hold since $\alpha = 2\delta_1$, but the three fixpoints representing the three populations were still identified.

5.4.1 Examples of some univariate distributions

In the following be give some results on the location and domain of attraction of the fixpoints for some particular choices of the distribution f.

- If X follows a uniform distribution in the interval $[a, b]$, all points in the interval $\left[F^{-1}\left(\frac{\alpha}{2}\right), F^{-1}\left(1-\frac{\alpha}{2}\right)\right]$ are fixpoints. Moreover, any $\{x_k\}$ starting at $x_0 < F^{-1}\left(\frac{\alpha}{2}\right)$ or $x_0 > F^{-1}\left(1-\frac{\alpha}{2}\right)$ converge to $F^{-1}\left(\frac{\alpha}{2}\right)$ and $F^{-1}\left(1-\frac{\alpha}{2}\right)$ respectively.

- If X follows a normal distribution with parameters μ and σ, $x^* = \mu$ is the unique fixpoint and attracts any $x_0 \in S$.

- If X follows an exponential distribution with density function $f(x) = \lambda e^{-\lambda x}$, $x^* = F^{-1}\left(\frac{\alpha}{2}\right) = -\frac{1}{\lambda}\ln(1-\frac{\alpha}{2})$ is the unique fixpoint of g_α of f since $g_\alpha(x) = F^{-1}\left(\frac{\alpha}{2}\right)$ for $x \leq \frac{F^{-1}(\alpha)}{2}$ and f is strictly unimodal. Therefore, for any starting value $x_0 \in S$, the sequence $\{x_k\}$ converges to x^*.

5.5 Discussion

Further research will be focused on extending to the multivariate case the theoretical results we have proved for the univariate case. The proof of the results for the multivariate case is highly non-trivial and may require a significant amount of original work, although we do believe they hold.

Conclusions and further research

In this final chapter we review the results of this thesis and mention possible future directions of research.

In Chapter 2 we identify a subset of the eigenvectors of a kurtosis matrix as a subspace with optimal properties for clustering in the sense that it coincides with Fisher's linear discriminant subspace, which maximizes the standardized distance between the cluster centers. We also provide an explicit formula for the kurtosis matrix under a mixture of elliptical distributions. The eigenvectors are identified by looking for the eigenvalues whose values are different from the value $p + 2$, and thus we are able to identify the subspace without knowing the cluster centers in advance. We also prove that the eigenvectors of the sample kurtosis matrix are consistent estimators of this subspace. The method is easy to implement and computationally efficient, providing specially favourable results when the ratio n/p is large. This matrix, therefore, provides a way of reducing the dimension of the space of the data in order to perform cluster analysis in a subspace of lower dimension. Future research will be focused on modifying the kurtosis matrix to improve the performance of the eigenvectors, specially when the scatter matrices are different, a case that has not been addressed yet in the literature.

Following the discussion in Chapter 2, in Chapter 3 we present alternative kurtosis matrices based on local modifications of the data, with the intention of improving the performance of the eigenvectors of the kurtosis matrix studied in Chapter 2. By substituting each observation of the sample for the mean of its neighbours, the covariance matrices of the components of a mixture of distributions will shrink, giving a more predominant role to the variability between clusters in the decomposition of the kurtosis matrix. Specifically, we prove that the separation properties of the eigenvectors of the new kurtosis matrix are improved, in the sense that the proposed modification of the observations produces standardized means that are further from each other than those of the original observations, and thus the clusters will appear more separated.

We also propose in Chapter 4 a procedure to identify non-linearly shaped clusters in a low dimensional space by projecting the sample onto straight lines. A trimmed projection is computed, such that only a subset of observations are projected onto it, the ones closest to the line. This idea allows the identification of clusters that would overlap if we projected the whole sample. Further research will be conducted to reevaluate the rules used to perform group assignments in the algorithm. At the present, the algorithm partitions the sample into too many clusters and a merging strategy needs to be applied after the GAPS algorithm. The required strategy should be able to merge efficiently non-linear clusters.

We present in Chapter 5 a new non-parametric cluster algorithm based on local medians. Each observation is substituted by its local median and this new observation moves towards the peaks and away from the valleys of the distribution. The process is repeated until each observation converges to a fixpoint. We obtain a partition of the sample based on where the sequences of local medians have converged. The algorithm determines the number of clusters and the partition of the observations given a value of α, the proportion of neighbours. A fast version of the algorithm, where only a subset of observations from the sample are treated, is also given. Furthermore, and for a univariate random variable, we prove the convergence of each point to the closest fixpoint, and the existence and uniqueness of a fixpoint on the neighbourhood of each mode. In the future, we will focus on extending to the multivariate case the theoretical results we have proved for the univariate case. The proof of the results for the multivariate case is highly non-trivial and may require a significant amount of original work.

Bibliography

Balanda, K. P. and H. L. MacGillivray (1988). Kurtosis: A critical review. *Am. Statist. 42*, 111–9.

Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–21.

Cardoso, J. F. (1989). Source separation using higher order moments. *Proc. ICASSP 4*, 2109–12.

Cardoso, J. F. and A. Souloumiac (1993). Blind beamforming for non-gaussian signals. *IEE Proceedings-F 140*, 362–70.

Caussinus, H. and A. Ruiz-Gazen (1993). Projection pursuit and generalized principal component analysis. In S. Morgenthaler, E. Ronchetti, and W. Stahel (Eds.), *New directions in statistical data analysis and robustness*, pp. 35–46. Basel: Birkhuser Verlag.

Caussinus, H. and A. Ruiz-Gazen (1995). Metrics for finding typical structures by means of principal component analysis. In Y. Escoufier and C. Hayashi (Eds.), *Data science and its applications*, pp. 177–92. Tokyo: Academy Press.

Cheng, M. Y. and P. Hall (1998). Calibrating the excess mass and dip tests of modality. *J. R. Statist. Soc. B 60*, 579–89.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell. 17*, 790–9.

Chissom, B. S. (1970). Interpretation of the kurtosis statistic. *Am. Statist. 24-4*, 19–22.

Comaniciu, D. and P. Meer (1999). Mean shift analysis and applications. *Proceedings of the Seventh International Conference on Computer Vision*, 1197–203.

Comaniciu, D. and P. Meer (2000). Real-time tracking of non-rigid objects using mean shift. *IEEE Conference on Computer Vision and Pattern Recognition 2*, 142–9.

Comaniciu, D. and P. Meer (2001). The variable bandwidth mean shift and data-driven scale selection. *Proceedings of the Eighth International Conference on Computer Vision 1*, 438–45.

Comaniciu, D. and P. Meer (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell. 24*, 603–19.

Darlington, R. B. (1970). Is kurtosis really "Peakedness"? *Am. Statist. 24-2*, 19–22.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *J. Am. Statist. Assoc. 93*, 294–302.

Dyson, F. J. (1943). A note on kurtosis. *J. R. Statist. Soc. B. 106*, 360–1.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenic. 7*, 179–88.

Fraley, C. and A. E. Raftery (1999). Mclust: Software for model-based cluster analysis. *J. Classification 16*, 297–306.

Friedman, J. H. (1987). Exploratory projection pursuit. *J. Am. Statist. Assoc. 82*, 249–66.

Friedman, J. H. and J. W. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers C-23*, 881–9.

Frigui, H. and R. Krishnapuram (1999). A robust competitive clustering algorithm with applications in computer vision. *IEEE Trans. Pattern Anal. Mach. Intell. 21*, 450–65.

Fukunaga, K. and L. D. Hostetler (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory 21*, 32–40.

Golub, G. H. and C. F. van Loan (1996). *Matrix computations*. The Johns Hopkins University Press.

Gordon, A. D. (1999). *Classification*. Chapman and Hall.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc. 69*, 383–93.

Hartigan, J. A. and M. A. Wong (1979). A k-means clustering algorithm. *J. R. Statist. Soc. C*.

Hildebrand, D. K. (1971). Kurtosis measures bimodality? *Am. Statist. 25*, 42–3.

Horn, P. S. (1983). A measure for peakedness. *Am. Statist. 37*, 55–6.

Huber, P. J. (1985). Projection pursuit. *Ann. Statist. 13*, 435–75.

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

Jones, M. C. and R. Sibson (1987). What is projection pursuit? *J. R. Statist. Soc. 150*, 1–37.

Kato, T. (1980). *Perturbation theory for linear operators*. Berlin: Springer.

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

Kollo, T. (2008). Multivariate skewness and kurtosis measures with an application in ICA. *J. Multivariate Anal. 99*, 2328–38.

Koziol, J. A. (1989). A note on measures of multivariate kurtosis. *Biom. J. 31*, 619–24.

Krzanowski, W. J. and F. H. C. Marriot (1994). *Multivariate Analysis Part I: Distributions, Ordination and Inference*. London: Edward Arnold.

Kundu, S. (1999). Gravitational clustering: A new approach based on the spatial distribution of the points. *Pattern Recognition 32*, 1149–60.

Liu, J. S., J. L. Zhang, M. J. Palumbo, and C. E. Lawrence (2003). Bayesian clustering with variable and transformation selections. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 7*, pp. 249–75. Oxford: University Press.

Malkovich, J. F. and A. A. Afifi (1973). On tests for multivariate normality. *J. Am. Statist. Assoc. 68*, 176–9.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika 57*, 519–30.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. New York: Academic Press.

McRae, E. C. (1974). Matrix derivatives with an application to an adaptive linear decision problem. *Ann. Statist. 2*, 337–46.

Moors, J. J. A. (1986). The meaning of kurtosis: Darlington reexamined. *Am. Statist. 40*, 283–4.

Móri, T. F., V. K. Rohatgi, and G. J. Székely (1993). On multivariate skewness and kurtosis. *Theory Probab. Appl. 38*, 547–51.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett. 1*, 327–32.

Oja, H., S. Sirkiä, and J. Eriksson (2006). Scatter matrices and independent component analysis. *Au. J. Statist. 35*, 175–89.

Pearson, K. (1905). Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A rejoinder. *Biometrika 4*, 169–212.

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill.

Peña, D. and F. J. Prieto (2001). Cluster identification using projections. *J. Am. Statist. Assoc. 96*, 1433–45.

Peña, D. and J. Rodríguez (2003). Descriptive measures of multivariate scatter and linear dependence. *J. Multivariate Anal. 58*, 361–74.

Peña, D., J. Rodríguez, and G. C. Tiao (2003). Identifying mixtures of regression equations by the SAR procedure. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 7*. Oxford: University Press.

Peña, D. and J. Viladomat (2009). Discussion of "Invariant co-ordinate selection" by D. E. Tyler, F. Critchley, L. Dümbgen and H. Oja. *J. R. Statist. Soc. B. 71-3*.

Petersen, K. B. and M. S. Petersen (2008). *The Matrix Cookbook*. Version 20080216.

Ruppert, D. (1987). What is kurtosis? An influence function approach. *Am. Statist. 41*, 1–5.

Ruspini, E. H. (1970). Numerical methods for fuzzy clustering. *Information Science 2*, 319–50.

Sato, Y. (2000). An autonomous clustering technique. In A. L. H. Kiers, J. P. Rasson, P. J. E. Groenen, and M. Schader (Eds.), *Data Analysis, Classification, and Related Methods*. Berlin: Springer.

Schwager, S. J. (1985). Multivariate skewness and kurtosis. In N. L. Johnson and S. Kotz (Eds.), *Encyclopedia of statistical sciences*. New York: John Wiley.

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

"Student" (1927). Errors of routine analysis. *Biometrika 19*, 151–164.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B 63*, 411–23.

Tyler, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist. 9*, 725–36.

Tyler, D. E., F. Critchley, L. Dümbgen, and H. Oja (2009). Invariate co-ordinate selection (with discussion). *J. R. Statist. Soc. B. 71-3*, 1–27.

van Zwet, W. R. (1964). Convex transformations of random variables. In *Mathematics Centre Tract 7*. Amsterdam: Mathematisch Centrum.

Wang, J. H. and J. D. Rau (2001). VQ-agglomeration: A novel approach to clustering. *IEE Proc. Vis. Image Signal Process 148*, 36–44.

Wang, X., W. Qiu, and R. Zamar (2007). CLUES: A non-parametric clustering method based on local shrinking. *Comput. Statist. Data Anal. 52*, 286–98.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika 24*, 471–94.

Wright, W. E. (1977). Gravitational clustering. *Pattern Recognition 9*, 151–66.

Zhung, X., Y. Huang, K. Palaniappan, and J. S. Lee (1996). Gaussian mixture modelling, decomposition and applications. *IEEE Trans. Signal Process 5*, 1293–302.