UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 13-07
Statistics and Econometrics Series (06)
March 2013

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

# RECOMBINING PARTITIONS VIA UNIMODALITY TESTS

**Adolfo Álvarez[(1)], Daniel Peña[(2)]**

## Abstract

In this article we propose a recombination procedure for previously split data. It is based on the study of modes in the density of the data, since departing from unimodality can be a sign of the presence of clusters. We develop an algorithm that integrates a splitting process inherited from the SAR algorithm (Peña et al., 2004) with unimodality tests such as the dip test proposed by Hartigan and Hartigan (1985), and finally, we use a network configuration to visualize the results. We show that this can be a useful tool to detect heterogeneity in the data, but limited to univariate data because of the nature of the dip test. In a second stage we discuss the use of multivariate mode detection tests to avoid dimensionality reduction techniques such as projecting multivariate data into one dimension. The results of the application of multivariate unimodality tests show that is possible to detect the cluster structure of the data, although more research can be oriented to estimate the proper fine-tuning of some parameters of the test for a given dataset or distribution.

**Keywords:** Cluster analysis, unimodality, dip test.

(1) Álvarez, Adolfo, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, e-mail: aaapinto@est-econ.uc3m.es.
(2) Peña, Daniel, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, e-mail: dpena@est-econ.uc3m.es.

Work partially supported by Spanish Ministry of Science and Innovation, research projects SEJ2007-64500 and ECO2012-38442.

## 1. Introduction

A useful non parametric strategy to merge partitions is to check whether the data can be assumed as unimodal or not. Unimodality detection is a natural way to identify the presence of clusters, understanding each of them as a mode surrounded by a density and separated enough from other modes, if they exist. Unimodality is well defined for univariate data sets but these techniques can be extended to multivariate analysis by: a) projecting the data into one dimension and then evaluating the unimodality, or b) choosing one of the possible unimodality definitions and techniques for multivariate data.

A distribution $F$ is defined as "unimodal", if $F$ is convex for $x < m$ and concave for $x > m$, where $m$ is the mode. Under this definition is clear that the normal, the student or the chi-squared distributions are unimodal, but also the uniform(a,b) distribution is considered unimodal under this definition, given that $m$ can be any value in $[a, b]$.

Hartigan and Hartigan (1985), introduce the "dip test" to detect the presence of one or multiple modes into the data. Given the empirical distribution, the dip statistic computes the maximum difference between that distribution and an unimodal distribution function in the following way:

Let $x_1, x_2, ..., x_n$be a set of univariate data coming from a density function $f(x)$, and $F_n(x)$ be the sample empirical distribution function. Let $H(x)$ be the closest unimodal c.d.f. respect to the empirical distribution, then the DIP statistic is given by:

$$DIP = \sup_x |F_n(x) - H(x)| \tag{1}$$

Although Bickel and Fan (1996) show that the non-parametric maximum likelihood estimate of the closest unimodal cdf, given the mode location $m_0$, is the greatest convex minorant of $F_n$ on $(-\infty, m_0]$ and the least concave majorant on $[m_0, -\infty)$ (Tantrum et al., 2003), the authors of the test propose the use of an uniform distribution to obtain a critical value to compare the statistic. They claim that the dip is asymptotically larger for the uniform than for any unimodal distribution with exponentially decreasing tails, so this choice implies being very conservative in the assumption of the underlying distribution of the data.

Cluster methods like M-clust (Fraley and Raftery, 1998), model the underlying distribution of the data by a mixture of normal distributions. The

parameters are estimated by the EM algorithm, while the Bayesian Information Criteria (BIC) is used to decide the number of groups, by estimating the number of components of the mixture which maximize the likelihood, penalized by the number of estimated parameters.

The problem with this kind of estimation arises when the true data is not a mixture of normals, and other distribution can fit better, or when the concept of "cluster" is not equivalent with the number of components of the mixture. For example, when a cluster is defined by finding gaps in the density, a mixture of normals can be not appropriate to define the number of groups.

The dip test has been used by Tantrum et al. (2003) as a tool to identify whether a mixture of normal distributions overestimates the real number of clusters in a sample. They propose an algorithm for pruning the cluster tree generated by the mixture model chosen by the Model Based Clustering. It then progressively merges clusters that seems to be unimodal by using the dip test. A similar approach to Tantrum et al. (2003) is proposed by Ahmed and Walther (2012) who project multivariate data on its principal curves and then apply Silverman's multimodality test (Silverman, 1981) to the resulting univariate sample. Other methods specifically designed to merge Gaussian components are reviewed in Hennig (2010).

## 2. Recombining with the dip test

The procedure is as follows, given a data sample $x_1, x_2, ...x_n$ of $n$ i.i.d. observations coming from an unknown distribution function, we apply the discriminator function to classify the observations into $k \leq n$ partitions. We split the sample in the same way as the SAR algorithm (Peña et al., 2004), where $x_l$ is defined as the discriminator of $x_i$ if the latter observation appears as most discrepant (using the heterogeneity measures) with respect to the rest of the data set when the discriminator is deleted from the sample. The underlying idea is the following: If two observations are identical, they must have the same discriminator, thus, if they are close enough to each other, they should still have the same discriminator.

Formally, $x_l$ in the multivariate case, assuming normality, is equivalent to:

$$x_l = arg \max_j \ (x_i - \bar{x}_{(ij)})' \hat{V}^{-1}_{(ij)} (x_i - \bar{x}_{(ij)}) \tag{2}$$

which is the Mahalanobis distance between the element $x_j$ and the rest of the sample, when the $i_{th}$ and $j_{th}$ elements are removed.

In the univariate case, Peña et al. (2004) shows that the discriminator are always the extreme points, while in the multivariate case, Rodriguez (2002) generalize this result demostrating that the discriminators belong to the convex hull of the sample. Therefore, Rodriguez (2002) proofs that discriminators are invariant to scale and positions transformation, because they are a monotonic function of the Mahalanobis Distance. Using these two properties, the observation $x_l$ will be the discriminator of $x_i$ if and only if:

$$x_l(x_i) = \underset{x_j \in \, Convex \; Hull}{arg\ max} \frac{\left(\frac{1}{n} + x_i' x_j\right)^2}{\left(\frac{n}{n-1} - x_j' x_j\right)} \tag{3}$$

Which is an efficient definition in terms of computational time, so it will be used in the algorithms included in this research.

To illustrate the discriminator function in the multivariate case, we present the widely known Old Faithful data set from Azzalini and Bowman (1990), considering the waiting time between eruptions and the duration of them from the geyser "Old Faithful" in Yellowstone Park, Wyoming, USA. This data set form two groups as shown in the Figure 1.

Applying the discriminator function, each data point is assigned to one discriminator following Equation (3) as showed in Figure 2, where is possible to see that the use of the discriminator function split the data into groups, assigning each point to one of the discriminators (observations 19, 58, 76, 149, 158, 161, 197, and 265) and this measure will be used in the SAR to perform the cluster analysis as we will see in the next section.

This splitting process is iteratively repeated until the resultant groups are all of sizes smaller than a minimum size. Following the guidelines of Peña et al. (2004), the minimum size is set as $n_0 = p + log(n - p)$ where as usual $p$ is the number of variables and $n$ is the sample size. As a result of the splitting process, we get a set of basic groups, all of them of relatively small size and internally homogeneous.

Given the structure of the basic groups, it is usual that the number of groups is bigger than the actual number of clusters in the data, so a recombination process is needed. We propose the use of the dip test to contrast if two basic groups conform an unimodal sample or not. The idea behind it is

that if two basic groups are part of the same original clusters, they should share the same mode.

One of the limitations of the original implementation of the dip test is that is only applicable to univariate samples, so when the dimension of the problem is greater than one, we need to project the data into one dimension before performing the test. For each pair of basic groups, the procedure tests if they are unimodal (and they should be merged), or not. To do so, the natural election for the projection is the Fisher's linear discriminator direction, since it maximize the separation of the groups to be tested. In this case two groups should be merged if even in the projection which separate

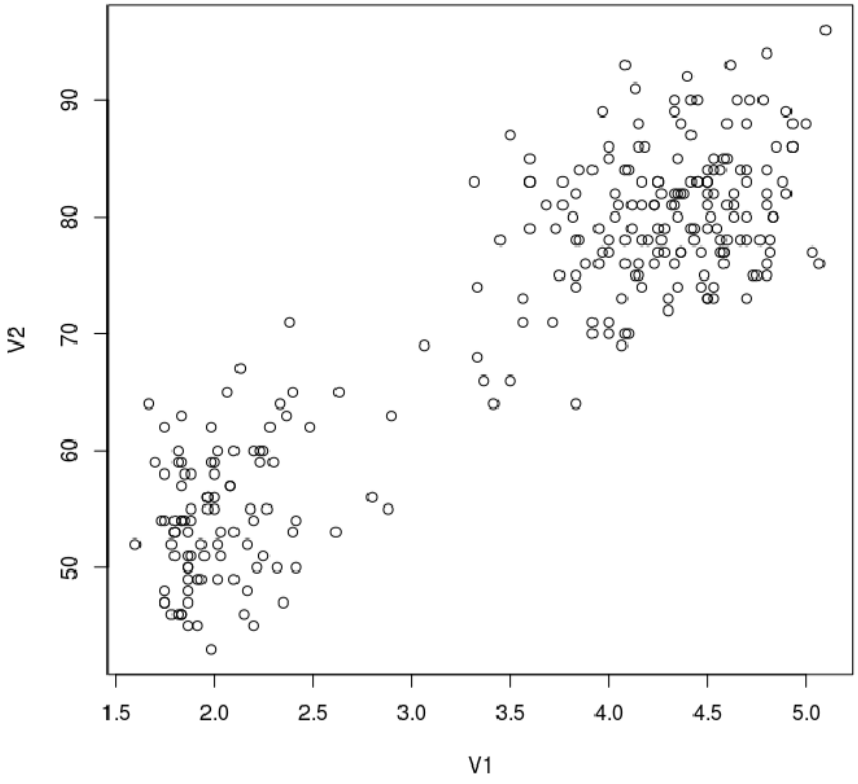


Figure 1: The Old Faithful data set

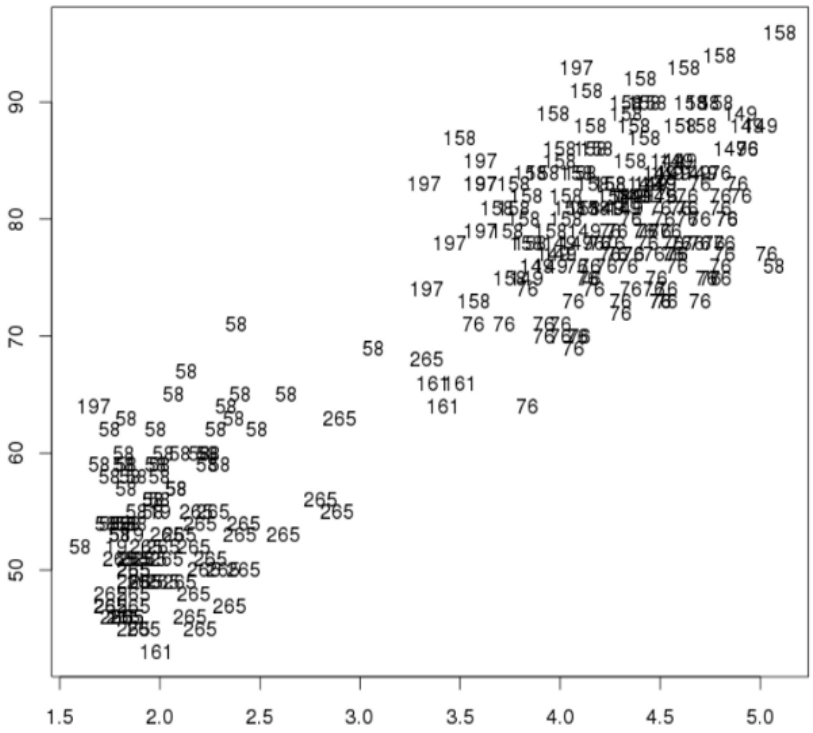


Figure 2: Discriminator function relationships, the number of the discriminator point is plotted

them the most, they still show one mode (See Section 4 for a discussion about the choose of a good direction for the projection).

The output of the test is the value of the dip statistic and the associated p-value calculated with the simulation performed by Maechler (2013), who corrected the original code proposed by Hartigan (1985). The quantiles were obtained using 1000001 samples for each sample size $n$, and a summary of they are shown in Table 1.

Given that all possible combinations of basic groups have been tested via the dip statistic, we propose the use of a graphical tool to identify if the groups should be merged or not. To do so, we plot all groups as nodes in a network, and when for a pair of groups the null hypothesis of unimodality is not rejected (i.e. the groups can be merged) the two nodes will be connected by a line. Varying the minimum level of significance $\alpha$ over the set $[0; 1]$ is possible to see the evolution of the grouping process, although the usual $\alpha = 0.1, 0.05$ and $0.01$ should unveil the structure of the data set.

After combined, the remaining observations which were not previously assigned to the basic sets, can be incorporated to the resulting sets by using the criteria of smaller Mahalanobis distance.

Table 1: Table of quantiles from a large simulation for Hartigan's dip test

| | 0.01 | 0.05 | 0.1 | 0.5 | 0.9 | 0.95 | 0.99 | 0.995 | 0.999 |
|---|---|---|---|---|---|---|---|---|---|
| 4 | 0.1250 | 0.1250 | 0.1250 | 0.1250 | 0.1874 | 0.2073 | 0.2318 | 0.2373 | 0.2444 |
| 5 | 0.1000 | 0.1000 | 0.1000 | 0.1216 | 0.1768 | 0.1864 | 0.1965 | 0.1982 | 0.1996 |
| 6 | 0.0833 | 0.0833 | 0.0833 | 0.1231 | 0.1591 | 0.1648 | 0.1919 | 0.2021 | 0.2195 |
| 7 | 0.0714 | 0.0726 | 0.0817 | 0.1178 | 0.1442 | 0.1599 | 0.1841 | 0.1910 | 0.2023 |
| 8 | 0.0625 | 0.0739 | 0.0820 | 0.1110 | 0.1418 | 0.1540 | 0.1730 | 0.1790 | 0.1945 |
| 9 | 0.0613 | 0.0733 | 0.0804 | 0.1042 | 0.1364 | 0.1466 | 0.1642 | 0.1728 | 0.1887 |
| 10 | 0.0610 | 0.0718 | 0.0780 | 0.0978 | 0.1305 | 0.1396 | 0.1597 | 0.1672 | 0.1806 |
| 15 | 0.0546 | 0.0610 | 0.0643 | 0.0836 | 0.1101 | 0.1188 | 0.1360 | 0.1425 | 0.1555 |
| 20 | 0.0474 | 0.0527 | 0.0568 | 0.0733 | 0.0972 | 0.1051 | 0.1206 | 0.1266 | 0.1386 |
| 30 | 0.0396 | 0.0444 | 0.0474 | 0.0615 | 0.0815 | 0.0882 | 0.1015 | 0.1065 | 0.1172 |
| 50 | 0.0314 | 0.0353 | 0.0377 | 0.0489 | 0.0649 | 0.0703 | 0.0812 | 0.0853 | 0.0941 |
| 100 | 0.0228 | 0.0257 | 0.0274 | 0.0355 | 0.0472 | 0.0511 | 0.0590 | 0.0620 | 0.0684 |
| 200 | 0.0165 | 0.0185 | 0.0198 | 0.0256 | 0.0340 | 0.0368 | 0.0427 | 0.0450 | 0.0497 |
| 500 | 0.0106 | 0.0119 | 0.0127 | 0.0165 | 0.0219 | 0.0237 | 0.0275 | 0.0289 | 0.0320 |
| 1000 | 0.0076 | 0.0085 | 0.0091 | 0.0117 | 0.0156 | 0.0169 | 0.0196 | 0.0206 | 0.0229 |
| 2000 | 0.0054 | 0.0061 | 0.0065 | 0.0084 | 0.0111 | 0.0120 | 0.0140 | 0.0147 | 0.0163 |
| 5000 | 0.0034 | 0.0039 | 0.0041 | 0.0053 | 0.0071 | 0.0077 | 0.0089 | 0.0093 | 0.0103 |
| 10000 | 0.0024 | 0.0027 | 0.0029 | 0.0038 | 0.0050 | 0.0054 | 0.0063 | 0.0066 | 0.0073 |
| 20000 | 0.0017 | 0.0019 | 0.0021 | 0.0027 | 0.0035 | 0.0038 | 0.0045 | 0.0047 | 0.0052 |
| 40000 | 0.0012 | 0.0014 | 0.0015 | 0.0019 | 0.0025 | 0.0027 | 0.0032 | 0.0033 | 0.0037 |
| 72000 | 0.0009 | 0.0010 | 0.0011 | 0.0014 | 0.0019 | 0.0020 | 0.0024 | 0.0025 | 0.0028 |

## 3. Results

To illustrate the behaviour of the procedure, remember the Old Faithful geyser data set from Figure 1, where we can clearly observe two well differentiated groups. If we apply the splitting step we obtain 12 sets and some isolated observations. These basic groups are shown in the Figure 3

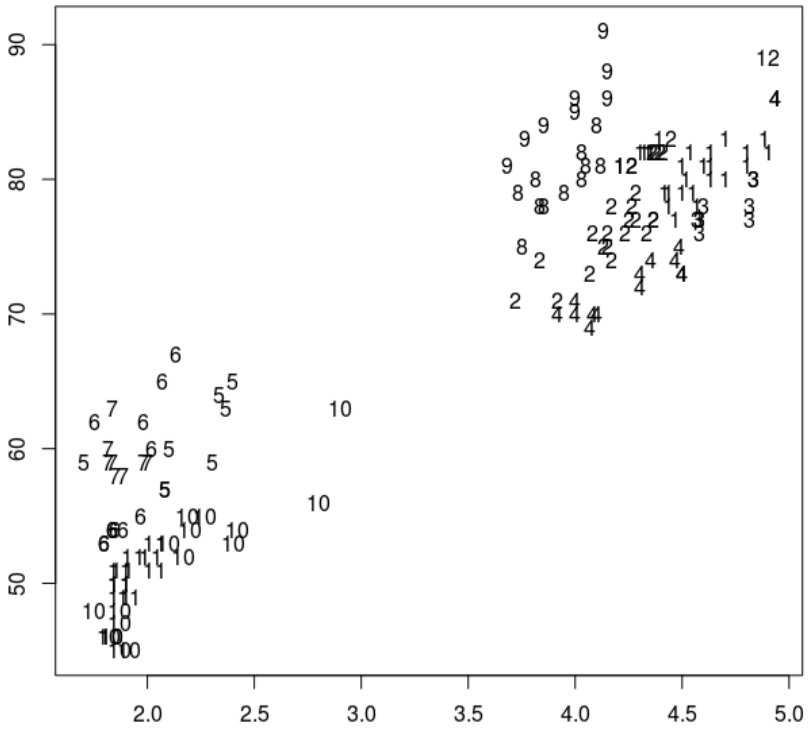


Figure 3: Basic groups from the Old Faithful data set

The following step is to calculate the dip statistic and the correspondent p-value for each of the $\frac{12 \times (12-1)}{2} = 66$ possible pair of groups. These results are given in Table 2

Table 2: Pairwise dip testing of the 12 basic groups obtained from the Old Faithful data set

| Group1 | Group2 | dip | p-value | Group1 | Group2 | dip | p-value |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 0.0413 | 0.9239 | 4 | 8 | 0.1335 | 0.0002 |
| 1 | 3 | 0.0518 | 0.8153 | 4 | 9 | 0.1214 | 0.0040 |
| 1 | 4 | 0.0912 | 0.0169 | 4 | 10 | 0.1403 | 0.0000 |
| 1 | 5 | 0.1091 | 0.0055 | 4 | 11 | 0.1735 | 0.0000 |
| 1 | 6 | 0.1631 | 0.0000 | 4 | 12 | 0.1192 | 0.0050 |
| 1 | 7 | 0.1332 | 0.0001 | 5 | 6 | 0.0763 | 0.4316 |
| 1 | 8 | 0.1002 | 0.0098 | 5 | 7 | 0.0817 | 0.4889 |
| 1 | 9 | 0.0980 | 0.0231 | 5 | 8 | 0.1461 | 0.0005 |
| 1 | 10 | 0.1434 | 0.0000 | 5 | 9 | 0.1683 | 0.0001 |
| 1 | 11 | 0.1522 | 0.0000 | 5 | 10 | 0.0627 | 0.6105 |
| 1 | 12 | 0.1032 | 0.0121 | 5 | 11 | 0.0825 | 0.3744 |
| 2 | 3 | 0.0722 | 0.2995 | 5 | 12 | 0.1887 | 0.0000 |
| 2 | 4 | 0.0541 | 0.6550 | 6 | 7 | 0.0754 | 0.4519 |
| 2 | 5 | 0.1133 | 0.0049 | 6 | 8 | 0.1916 | 0.0000 |
| 2 | 6 | 0.1682 | 0.0000 | 6 | 9 | 0.1545 | 0.0001 |
| 2 | 7 | 0.1328 | 0.0003 | 6 | 10 | 0.0856 | 0.0787 |
| 2 | 8 | 0.0774 | 0.1687 | 6 | 11 | 0.0808 | 0.2675 |
| 2 | 9 | 0.1078 | 0.0099 | 6 | 12 | 0.1617 | 0.0000 |
| 2 | 10 | 0.1299 | 0.0000 | 7 | 8 | 0.1900 | 0.0000 |
| 2 | 11 | 0.1602 | 0.0000 | 7 | 9 | 0.1954 | 0.0000 |
| 2 | 12 | 0.1017 | 0.0212 | 7 | 10 | 0.0907 | 0.0822 |
| 3 | 4 | 0.0647 | 0.5864 | 7 | 11 | 0.1383 | 0.0024 |
| 3 | 5 | 0.1847 | 0.0000 | 7 | 12 | 0.2023 | 0.0000 |
| 3 | 6 | 0.1931 | 0.0000 | 8 | 9 | 0.1009 | 0.0906 |
| 3 | 7 | 0.2207 | 0.0000 | 8 | 10 | 0.1331 | 0.0001 |
| 3 | 8 | 0.1585 | 0.0000 | 8 | 11 | 0.2121 | 0.0000 |
| 3 | 9 | 0.1742 | 0.0000 | 8 | 12 | 0.0987 | 0.1077 |
| 3 | 10 | 0.1399 | 0.0001 | 9 | 10 | 0.1082 | 0.0123 |
| 3 | 11 | 0.2140 | 0.0000 | 9 | 11 | 0.1692 | 0.0000 |
| 3 | 12 | 0.1717 | 0.0000 | 9 | 12 | 0.1526 | 0.0009 |
| 4 | 5 | 0.1249 | 0.0024 | 10 | 11 | 0.0627 | 0.5464 |
| 4 | 6 | 0.1775 | 0.0000 | 10 | 12 | 0.1249 | 0.0013 |
| 4 | 7 | 0.1397 | 0.0003 | 11 | 12 | 0.1801 | 0.0000 |

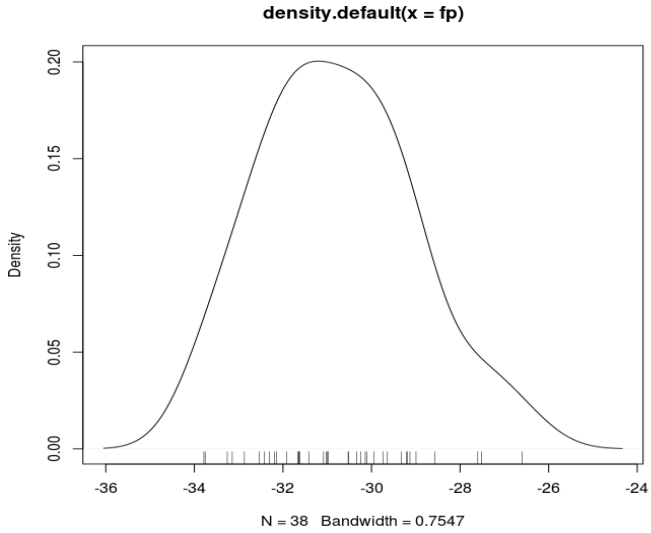


Figure 4: Density function of univariate projection of basic sets 1 and 2

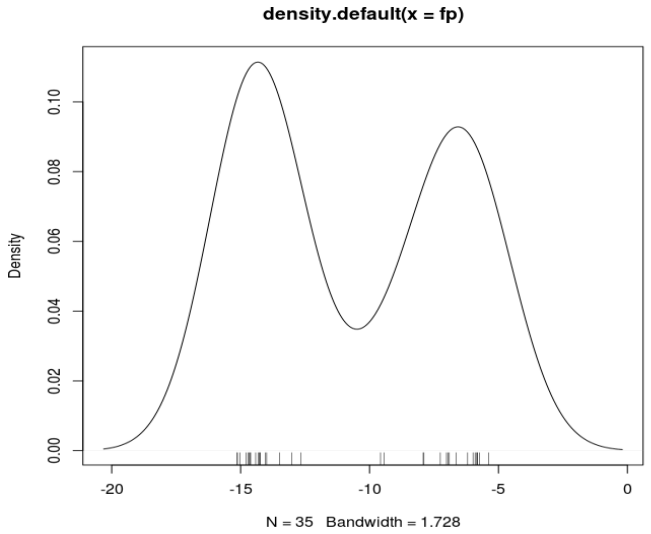


Figure 5: Density function of univariate projection of basic sets 2 and 10

If we take a look into two basic sets which belongs to the same cluster, for example, sets 1 and 2, the density plot of their projection into the Fisher direction does not show a bimodal evidence (See Figure 4 ), and the p-value from Table 2 is 0.9239. In the case of basic groups 2 and 10, the associated p-value is equal to 0, and the corresponding density plot clearly shows two modes. (See Figure 5)

Graphically, the interaction between all basic sets is shown in the Figure 6, where we observed two clearly differentiated groups, one formed by groups 5,6,7,10 and 11; and other by the remaining basic sets, corresponding with the original configuration of the data.

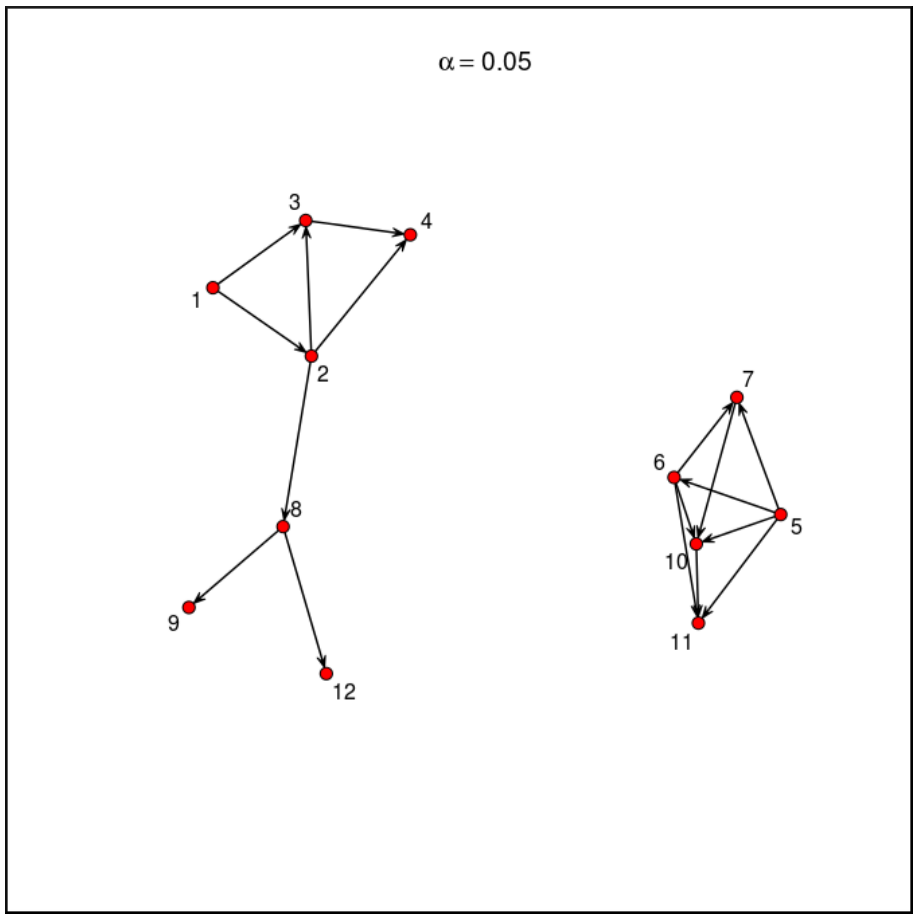


Figure 6: Dip test network for the Old Faithful data set, $\alpha = 0.05$

This graphical tool as an exploratory approach, allows also to see different strengths within the groups. For example, the group formed by sets 5-6-7-10-11 seems to be more internally connected than the group composed by basic sets 1-2-3-4-8-9-10, which can be separated into a "strong group" of sets "1-2-3-4" and other formed by 8-9-12.

As a second example, we consider a case when the data set is not linearly separable. In Figure 7 we show the simulation of two half-moons, each of them consisting of 250 data points in two dimensions. After the splitting procedure, we find 19 basic groups (See Figure 8), while the graphical results of the dip test are shown in the Figure 9.

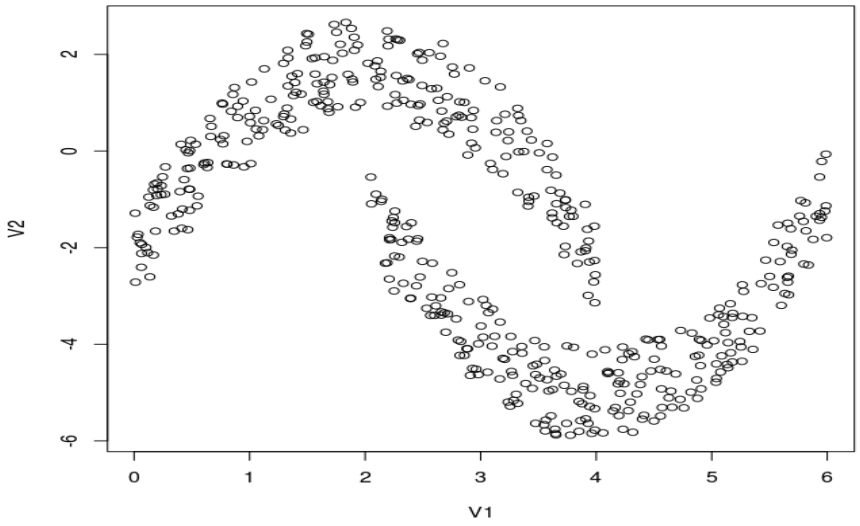


Figure 7: The two half moons data set

When $\alpha = 0.1$, the procedure detects 5 combination of groups and other 4 single groups, not detecting yet the structure of two groups in the data. A similar situation occurs when $\alpha$ is decreased to 0.05, but when we consider $\alpha = 0.01$ the two clusters appears, one in the upper half of the network formed by groups 1-3-4-8-9-13-16-17-19 and other in the lower half composed of groups 2-5-6-7-10-11-12-14-15, plus an isolated group (18) unveiling the more complex structure of the data. Notice that these two clusters are connected by the group 19, reflecting a problem in the partition process, because that

group incorrectly includes observations from the two half moons (See Figure 8).

In complex data sets, the partition process need to be adapted, for example including a cleaning process, therefore, smaller values of $\alpha$ are needed in order to properly reflect the original clusters of the data. This procedure should be intended as exploratory, where is possible to observe the evolution of the combining process while varying the confidence levels.

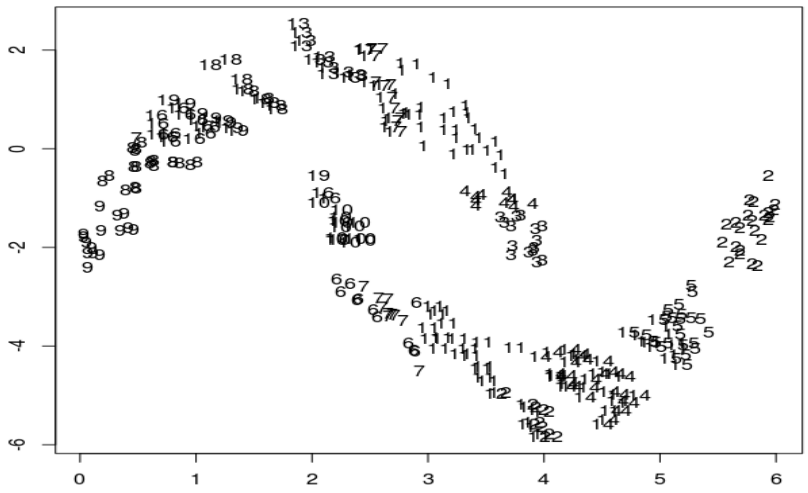


Figure 8: Basic groups of the two half moons data set

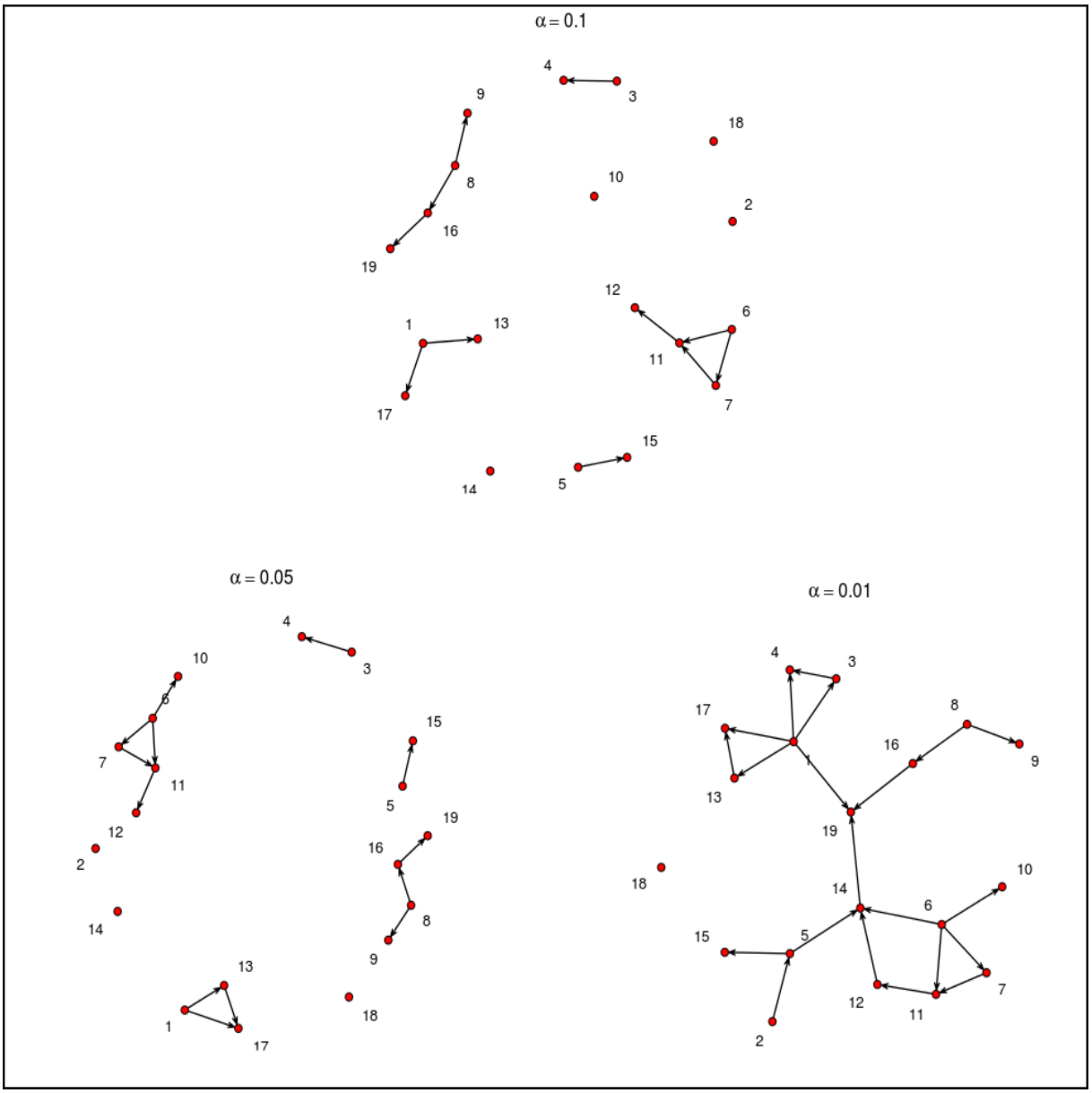


Figure 9: Dip network for the two half moons data set for $\alpha = 0.1, 0.05$ and $0.01$

## 4. Discussion

In order to recombine multivariate subpartitions based on unimodality tests there are two main approaches: First, to keep the dimension of the

problem and to look for an appropriate multivariate modality detection, or second, to use a simpler univariate modality test but to choose a good projection direction reducing the dimensionality of the data. We will briefly discuss this two alternatives and justify the election we made for this research.

*4.1. Multivariate modality tests*

Besides the dip test, Hartigan tried to extend its work to a multivariate framework. On his publications we can find three proposals in that direction, the tests "span" (Hartigan, 1988), "RUNT" (Hartigan and Mohanty, 1992), and "MAP" (Rozál and Hartigan, 1994). All of them are based on a hierarchy of similarities: starting from the $n$ classes corresponding to the $n$ initial points of the sample, and finishing with all data points in one class, the distance between two classes A,B is defined as the smallest distance between an observation from the class A and another observation from the class B.

The RUNT test is based on the fact that for a bimodal distribution is expected that the two modes of the distribution are merged in the last step of the hierarchy, while the span test is a generalization of the dip test where the empirical function $F_n(x)$ is the proportion of points $x_i$ such that $x_i \preceq x$. Starting from a random root point $r = x_k$, $x \preceq y$ if x is further away from r in the hierarchy. Finally, the MAP test is based on the Minimal Ascending Path, calculated from a MAP Spanning Tree which is a tree such that the length of the links are non-increasing from any link to a root node.

Departing from hierarchy trees, another more recent researches have been focused in the mode detection problem:

Burman and Polonik (2009) assume the data is coming from an unknown distribution with isolated modes. The idea of the method is first, to find potential mode candidates and second, determine if they represent different modal regions via pairwise statistical tests. A modal region is defined as a set $R_y$ with $y \in R_y$ , and $f(y + \alpha x)$, with $\alpha \in [0, 1]$, decreasing $\forall x \in Ry$, being $y$ a mode of $f$.

The first candidate $W_1$ to be a mode is selected as the observation which have its $k_1$ neighbour closer. Formally, if $\hat{d}_n(x_j)$ is the distance between an observation $x_j, j = 1, ..n$ and its $k_1$ nearest neighbour, then:

$$W_1 = \underset{x_j}{arg\ min}\ \hat{d}_n(x_j) \tag{4}$$

The second candidate is obtained in a similar way but deleting from the sample the previous candidate and its $k_2$ neighbours, and the procedure is

continued until no more candidates are found.

As a second step, the list of candidates is purged, keeping only those observations which does not significantly differ from the mean of its $k_2$ neighbours, using a Hotelling's test and assuming multivariate normality.

Finally, the candidates are pairwise tested to belong to the same modal region, by considering the existence of "antimodes" between them. One of the possible tests the authors propose to compare two candidates $x$ and $y$ is the following statistic:

$$\hat{SB}(\alpha) = p\left[\log \hat{d_n}(x_\alpha) - \max\left\{\log \hat{d_n}(x), \log \hat{d_n}(y)\right\}\right] \tag{5}$$

where $x_\alpha = \alpha x + (1-\alpha)y, 0 \leqslant \alpha \leqslant 1$. The authors propose to reject the null hypothesis when $\hat{SB}(\alpha) \geqslant \sqrt{\frac{2}{k_1}}\Phi^{-1}(0.95)$, being $\Phi$ the c.d.f. of the multivariate normal distribution

Einbeck (2011) develops a technique for multivariate mode detection, although the main objective of their research is focus on a cluster analysis algorithm. The base of the mode detection is the work of Cheng (1995), who defined the "mean shift" as the shift necessary to move a point $x \in \mathbb{R}^p$ towards the local mean around this point.

Let $K$ be a p-variate kernel function (usually Gaussian), and $H = diag(h_1^2, h_2^2, ..., h_p^2)$, with $h_j > 0$ a bandwidth matrix, then:

$$K_H(x) = |H|^{-1/2} K(H^{-1/2}x) \tag{6}$$

the local mean and the mean shift are then defined as:

$$\mu_H(x) = \frac{\sum_{i=1}^n K_H(x_i - x)x_i}{\sum_{i=1}^n K_H(x_i - x)} \tag{7}$$

$$S_H(X) = \mu_H(x) - x = \frac{\sum_{i=1}^n K_H(x_i - x)(x_i - x)}{\sum_{i=1}^n K_H(x_i - x)} \tag{8}$$

For a given distribution function f, and bandwidth H, at a mode $m_H$ of f, $S_H(m_H) = 0$, then $\mu_H(m_H) = m_H$. The authors recall all points satisfying that condition as "Local principal points"

In order to find those local modes, Cheng (1995) proved that the sequence $m_l, l \geq 0$ will converge to a local principal point $m_H$, with $m_0 = x$, and $m_{l+1} = \mu_H(m_l)$., and this mean shift sequence is iterated the for all data observation.

The application to our original problem is now clear, given two partitions we will recombine them if we found only one mode on its merged set, and keep them separate in other case.

Recalling the Old Faithful example, where the basic groups are plotted in Figure 3, we will apply the procedure of Einbeck (2011), since its method is already implemented in an R package (Einbeck and Evers, 2012). However, for further research to build our own implementation of Burman and Polonik (2009) seems to be feasible in order to compare multivariate mode detection methods.

Several parameters need to be fixed in the procedure, including *taumin*, *taumax* and *gridsize*, all of them related with the grid of bandwidths where the search for modes is focused. Default options are $taumax = 0.02$, $taumin = 0.5$, and $gridsize = 25$, although its application to the Old Faithful basic groups does not properly recognise the two clusters under this parameters (Figure 10). Two other parameter combination are shown in Figures 11 and 12, being the last one which correctly identify the clusters.

As is shown in the figures, the procedure is highly sensitive to the parameters, and its interpretation is not as clear as the dip test proposed in the previous sections. At the same time the parameters cannot be dynamically adjusted from an "all connected" to a "none connected" framework in a simple way, hindering its visualization. Nevertheless, for higher dimensions this procedure take advantage since the projection can produce high loss of information.

### *4.2. Directions to project the data*

Given two multivariate candidate groups for recombining, the choice of the Fisher's direction to project the data in the proposed procedure is natural, since it maximizes the separability of the groups, and has been long used in classical methods as discriminant analysis. For our problem, that choice implies the most conservative scenario, because it tests unimodality even in the case where the separation between groups is maximum.

In the context of cluster analysis, the search for interesting directions to project the data and keep the structure of it has been widely used as a way to avoid the dimensionality curse (Friedman and Tukey, 1974; Friedman, 1987). The choice of Fisher's direction is also supported by the literature: Peña and Prieto (2001) proposed the direction that minimize the kurtosis as appropriate for cluster analysis, and later, Peña et al. (2010) proved that given the kurtosis matrix, the subspace orthogonal to the eigenspace associated to an

eigenvalue with multiplicity $p-k+1$ is Fishers linear discriminant subspace. Similar results can be found in Caussinus and Ruiz-Gazen (1994) and Caussinus and Ruiz-Gazen (1995), where the Fishers subspace is obtained from the $k$ largest eigenvectors of a Generalized Principal Components matrix, or Tyler et al. (2009) who proved that it can be generated from eigenvectors of affine equivariant scatter matrices.

The choice of Fisher's direction is optimal when we actually know the two partitions we want to test for recombine. Only under the assumption of no knowledge about the basic groups, one of the alternative directions presented here can be considered, for example in the case of a splitting step,

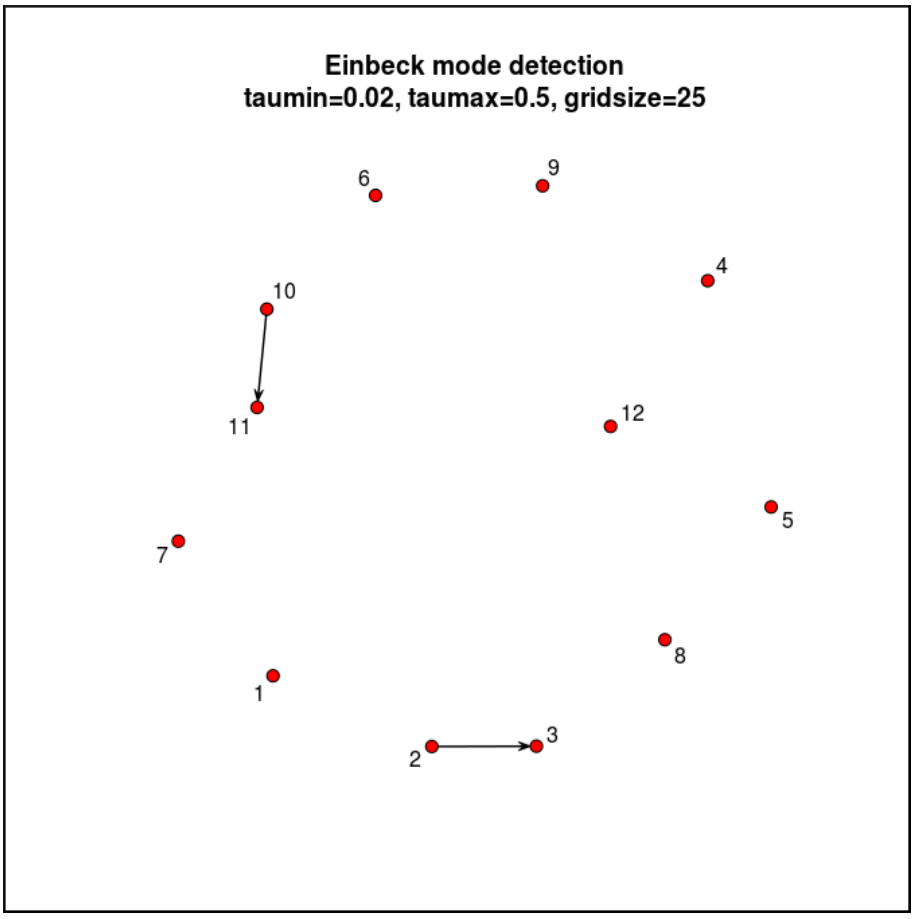


Figure 10: Einbeck mode detection test with default parameters

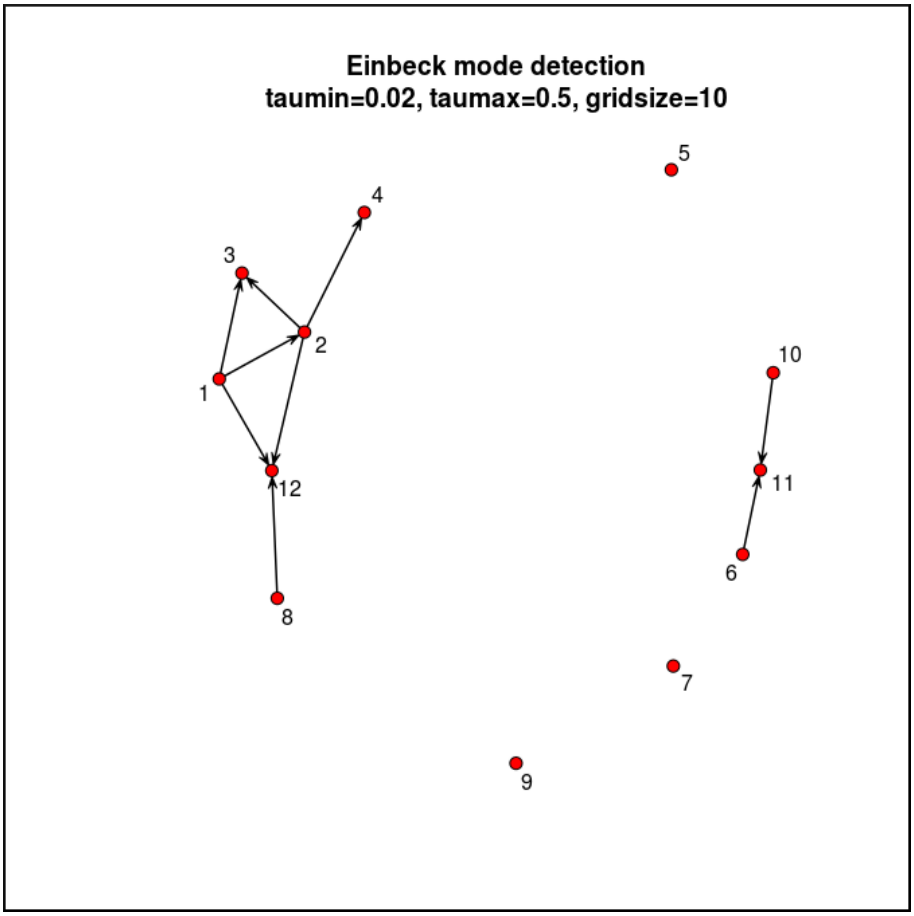


Figure 11: Einbeck mode detection test, gridsize decreased

where we can project the data and split into groups until no bimodality can be detected.

## 5. Conclusions

We have developed a method to split a data set using the discriminator function and recombine the obtained groups to find the final configuration incorporating the dip-statistic to test for unimodality. Also, we presented a graphical tool which allows to see the evolution of the merging procedure, and unveil which groups are more internally connected. The results show

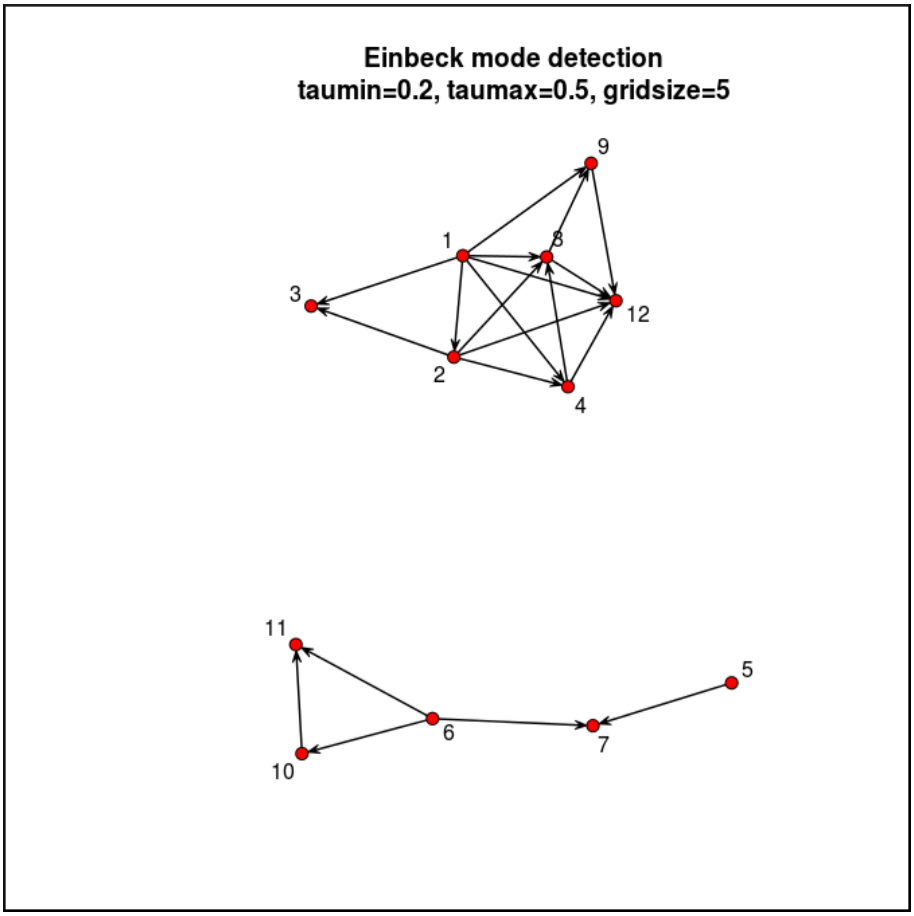


Figure 12: Einbeck mode detection test, gridsize decreased and taumin augmented

that the proposed technique can be a useful tool for exploratory research, since it allows to dynamically vary the level of significance to visualise the merging behaviour of the procedure.

The method have two issues which must be taken into account when applying it, which are also present in other dip-statistic based approaches: the validity and interpretation of p-value, and the chosen projection technique for multivariate data.

The obtained p-values does not hold the assumption of independence of a standard hypothesis test, because the partitions we test for unimodality are obtained from a previous methodology, and they are dependent in the sense

they are disjoint by construction. Therefore, we test the same data several times, because we compare each basic group against all the rest.

Nevertheless, even without the traditional interpretation of p-values, they can be used to show the behaviour of the merging when we modify the minimum level $\alpha$ from 1, where no groups are connected, to 0, where there is a connection between all groups. The most similar groups will merge in values close to 1, and clearly disjoint groups will not merge until values below 0.01.

In the other hand, it is important to notice that some useful information of the structure of the real data can be lost in the reduction of dimensionality. This is specially relevant in complex data sets or high dimension problems, and in this context, multivariate mode detection techniques, as those we reviewed in the discussion section, should be preferred.

## References


Ahmed, M. O. and Walther, G. (2012), "Investigating the Multimodality of Multivariate Data with Principal Curves," *Computational Statistics & Data Analysis*, 56, 4462–4469.

Azzalini, A. and Bowman, A. W. (1990), "A Look at Some Data on the Old Faithful Geyser," *Applied Statistics*, 39, 357–365.

Bickel, P. J. and Fan, J. (1996), "Some Problems on the Estimation of Unimodal Densities," *Statistica Sinica*, 6, 23–45.

Burman, P. and Polonik, W. (2009), "Multivariate Mode Hunting: Data Analytic Tools with Measures of Significance," *Journal of Multivariate Analysis*, 100, 1198–1218.

Caussinus, H. and Ruiz-Gazen, A. (1994), "Projection Pursuit and Generalized Principal Component Analysis," in *New Directions in Statistical Data Analysis and Robustness*, eds. Morgenthaler, S., Ronchetti, E., and Stahel, W., Basel: Birkhuser Verlag, pp. 35–46.

— (1995), "Metrics for Finding Typical Structures by Means of Principal Component Analysis," in *Data Science and its Applications*, eds. Escoufier, Y. and Hayashi, C., Tokyo: Academy Press, pp. 177–192.

Cheng, Y. (1995), "Mean Shift, Mode Seeking, and Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17, 790–799.

Einbeck, J. (2011), "Bandwidth Selection for Mean-Shift Based Unsupervised Learning Techniques: a Unified Approach Via Self-Coverage," *Journal of pattern recognition research*, 6, 175–192.

Einbeck, J. and Evers, L. (2012), "LPCM: Local Principal Curve Methods," *R package version 0.44-6*.

Fraley, C. and Raftery, A. (1998), "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *The Computer Journal*, 41, 578–588.

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

Friedman, J. and Tukey, J. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, c-23, 881 – 890.

Hartigan, J. (1988), "The SPAN Test for Unimodality." in *Classification and Related Methods of Data Analysis*, ed. Book, H. H., Amsterdam: North-Holland Publishing Company, pp. 229 – 236.

Hartigan, J. and Mohanty, S. (1992), "The RUNT Test for Multimodality," *Journal of Classification*, 9, 63–70.

Hartigan, J. J. and Hartigan, P. P. (1985), "The DIP Test of Unimodality," *The Annals of Statistics*, 13, 70–84.

Hartigan, P. (1985), "Algorithm AS 217: Computation of the Dip Statistic to Test for Unimodality," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34, 320–325.

Hennig, C. (2010), "Methods for Merging Gaussian Mixture Components," *Advances in Data Analysis and Classification*, 4, 3–34.

Maechler, M. (2013), "diptest: Hartigan's dip Test Statistic for Unimodality - Corrected Code," *R package version 0.75-5*.

Peña, D. and Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics*, 43, 286–310.

Peña, D., Prieto, F. J., and Viladomat, J. (2010), "Eigenvectors of a Kurtosis Matrix as Interesting Directions to Reveal Cluster Structure," *Journal of Multivariate Analysis*, 101, 1995–2007.

Peña, D., Rodriguez, J., and Tiao, G. (2004), "A General Partition Cluster Algorithm," in *COMPSTAT: Proceedings in Computational Statistics: 16th Symposium held in Prague, Czech Republic, 2004*, Springer, pp. 371–379.

Rodriguez, J. (2002), "Contribuciones al Estudio de la Heterogeneidad y la Dependencia," Ph.D. thesis, Universidad Carlos III de Madrid.

Rozál, G. and Hartigan, J. (1994), "The MAP Test for Multimodality," *Journal of Classification*, 11, 5–36.

Silverman, B. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 43, 97–99.

Tantrum, J., Murua, A., and Stuetzle, W. (2003), "Assessment and Pruning of Hierarchical Model Based Clustering," in *Proceedings of the ninth SIGKDD*, New York, New York, USA: ACM Press, pp. 197–205.

Tyler, D. E., Critchley, F., Dümbgen, L., and Oja, H. (2009), "Invariant Coordinate Selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 549–592.